Data Mining on Folksonomies

Andreas Hotho

Abstract. Social resource sharing systems are central elements of the Web 2.0 and use all the same kind of lightweight knowledge representation, called *folksonomy*. As these systems are easy to use, they attract huge masses of users. Data Mining provides methods to analyze data and to learn models which can be used to support users. The application and adaptation of known data mining algorithms to folksonomies with the goal to support the users of such systems and to extract valuable information with a special focus on the Semantic Web is the main target of this paper.

In this work we give a short introduction into folksonomies with a focus on our own system BibSonomy. Based on the analysis we made on a large folksonomy dataset, we present the application of data mining algorithms on three different tasks, namely spam detection, ranking and recommendation. To bridge the gap between folksonomies and the Semantic Web, we apply association rule mining to extract relations and present a deeper analysis of statistical measures which can be used to extract tag relations. This approach is complemented by presenting two approaches to extract conceptualizations from folksonomies.

1 Introduction

Complementing the Semantic Web effort, a new breed of so-called “Web 2.0” applications recently emerged on the Web. These include user-centric

Andreas Hotho
Knowledge & Data Engineering Group,
University of Kassel,
34121 Kassel, Germany
e-mail: hotho@cs.uni-kassel.de

G. Armano et al. (Eds.): Intelligent Information Access, SCI 301, pp. 57–82.
springerlink.com

publishing and knowledge management platforms like wikis, blogs, and social resource sharing tools. In this paper, we focus on resource sharing systems, which all use the same kind of lightweight knowledge representation, called *folksonomy*.[1]

Social resource sharing systems are web-based systems that allow users to upload all kinds of resources, and to label them with arbitrary words, so-called *tags*. The systems can be distinguished according to what kind of resources are supported. Flickr[2], for instance, allows the sharing of photos, del.icio.us[3] the sharing of bookmarks, CiteULike[4] and Connotea[5] the sharing of bibliographic references, and 43Things[6] even the sharing of goals in private life. Our own system, *BibSonomy*,[7] allows sharing bookmarks and BibTeX entries simultaneously.

According to Fayyad et. al. *Data Mining* is "the nontrivial process of identifying valid, previously unknown, and potentially useful patterns" [12] in a potentially very huge amount of data. Web Mining is the application of data mining techniques on three areas: the content, the structure and the usage of resources in the web [33]. Although Web 2.0 systems – as the name suggests – are still web applications and the analysis of such systems could be subsumed under the term web mining, new challenges for data mining emerge, as new structures and new data can be found in such systems. Therefore, we call the analysis of folksonomies: *Folksonomy Mining*. One example is that structure mining is applied on folksonomies and not – as it is known from web mining – on the web graph as a whole. Given the high number of publications in the short lifetime of folksonomy systems, researchers seem to be very interested in folksonomies and the information and knowledge which can be extracted from them. This can be explained by the tremendous amount of information collected from a very large user basis in a distributed fashion in such systems.

The application of mining techniques on folksonomies bears a large potential. Further, it is in line with the general idea of *Semantic Web Mining* [47]. Two aspects are of central interest: On the one hand, folksonomies form a rich source of data which can be used as a source for full-blown ontologies. This process is known as ontology learning and often utilizes data mining techniques. On the other hand, mining the Semantic Web is a second important application of mining techniques in this area. As folksonomies are considered as weak knowledge representation, analyzing their data can be seen as an implementation of Semantic Web Mining. The goal of this work is therefore to bridge the gap between folksonomies and the Semantic Web and to start to solve this problem with research contributions from various

[1] `http://www.vanderwal.net/folksonomy.html`

[2] `http://www.flickr.com/`

[3] `http://delicious.com`

[4] `http://www.citeulike.org`

[5] `http://www.connotea.org`

[6] `http://www.43things.com`

[7] `http://www.bibsonomy.org`

sides. More precisely, to reach this goal, a better understanding of the hidden and emergent semantics in folksonomies is necessary, as well as methods to extract the hidden information. Data Mining techniques provide methods for solving these issues.

This paper gives an overview of the previously published articles [24, 7, 34, 25, 31, 41, 5, 30] and shows connection between them. There are two lines of research: We analyzed the data we collected in order to get a better understanding of its structure (cf. [7]), and developed algorithms to support the users of folksonomy systems (cf. [25, 34, 31]). Second, we developed our own system BibSonomy as a platform for research experiments (cf. [24]). As we own the system, we have full access on e.g. all data, the user interface and so on. This puts us in the situation which researchers usually do not have: We can perform research experiments to test our new methods and push our research results into BibSonomy to show, to evaluate, and to demonstrate the advantages of our methods. One example are the online recommender experiments we are doing for this year's ECML PKDD discovery challenge.[8] Further, we have implemented many of our research results from the last years into the system. One of the first results having found its way into BibSonomy was a lightweight recommender (cf. Sec. 3.3) followed by the FolkRank ranking (cf. Sec. 3.2).

Related Work

Folksonomies and especially data mining on folksonomies are a relatively young research area. Meanwhile, work for specific areas starts to show up. To discuss the related work for all methods mentioned in this paper is beyond the scope of it. More detailed surveys can be found in the respective papers. To start with folksonomies and to learn more about their strengths and weaknesses one may look into [19, 36, 37]. One of the first works defining a model of semantic-social networks for extracting lightweight ontologies from del.icio.us was [38]. Recently, work on more specialized topics such as structure mining on folksonomies – e. g. to visualize trends [11] and our work on patterns [41] in users' tagging behavior – as well as ranking of folksonomy contents [25], analyzing the semiotic dynamics of the tagging vocabulary [6], or Halpin's analysis of the dynamics and semantics [18] have been presented.

Structure of the Paper

The rest of the paper is structured as follows: After a short introduction into the topic of social bookmarking and folksonomies we present a formal model and first properties we found in the graph formed by folksonomies. In Section 3 we present the three applications spam detection, a ranking method for folksonomies and a tag recommender method. Section 4 goes

[8] `http://www.kde.cs.uni-kassel.de/ws/dc09/`

one step towards more semantics in folksonomies and presents approaches to bridge the gap between folksonomies and the Semantic Web.

2 Basics of Folksonomies

In this section we will review the basic principles of folksonomies. We will start with an introduction into folksonomies followed by the description of our own system BibSonomy. A formal definition and first insights into the properties of such systems are the next part of this section. The presented ideas are the basis for the following steps where we first use these insights to provide valuable services like a better ranking or tag recommendation before we investigate the extraction of semantics out of folksonomies.

2.1 Social Bookmarking Systems

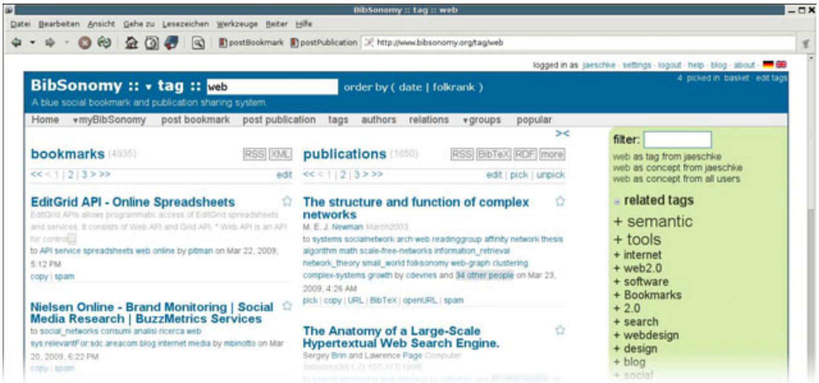
First bookmarking systems were developed at the end of the 90s but without a nice and fast user interface and often with a very weak business model (cf. [22]). Therefore, it was virtually impossible to attract a large number of users – which is necessary to make such systems attractive. One of the first systems which could reach a broad user basis was del.icio.us.[9] It was started in 2003 by Joshua Schachter and is today the best-known social bookmarking system for websites in the world. After he released a first version in 2003, he followed the advice of his users to make it more attractive. In 2004 the system reached a critical mass and the number of users increased dramatically. At the end of 2005 he sold the system to Yahoo. It is still running and the number of users is estimated with more than five million.[10] Similar services followed and provided a comparable service. Some of them focus on different content types, e.g. images, music, videos or places. Others provide an added value in form of additional functionality, e.g. by caching the seen webpage or presenting improved tag clouds for easier browsing. The core structure of all these systems is very similar and is known under the name *folksonomy.*

For research purposes we collected data from del.icio.us. The first time we crawled it was in 2005, where we collected the complete user pages for more than 75000 users. At that time we were able to gather an almost complete snapshot (mostly without spam). The second time we crawled del.icio.us in 2006 and collected data of more than 600000 users. Within our research we used and use these two datasets for our analyses and scientific experiments. Details can be found in the cited works.

In the next section, we will describe our own social bookmark and publication sharing system BibSonomy before we focus on the core structure of social bookmarking systems.

[9] `http://delicious.com/`

[10] `http://blog.delicious.com/blog/2008/11/delicious-is-5.html`


BibSonomy :: tag :: web
Datei Bearbeiten Ansicht Gehe zu Lesezeichen Werkzeuge Beiter Hilfe
postBookmark postPublication
http://www.bibsonomy.org/tag/web
logged in as jaeschke · settings · logout · help · blog · about ·
4 picked in basket · edit tags
BibSonomy :: ▾ tag :: web
order by (date | folkrank)
A blue social bookmark and publication sharing system.
Home ▾myBibSonomy post bookmark post publication tags authors relations ▾groups popular
bookmarks (4935) RSS XML
<< < 1 | 2 | 3 > >> edit
publications (1650) RSS BibTeX RDF more
<< < 1 | 2 | 3 > >> edit | pick | unpick
filter:
related tags

EditGrid API - Online Spreadsheets
EditGrid APIs allows programmatic access of EditGrid spreadsheets and services. It consists of Web API and Grid API. * Web API is an API for control...
to API service spreadsheets web online by pitman on Mar 22, 2009, 5:12 PM
copy | spam

Nielsen Online - Brand Monitoring | Social Media Research | BuzzMetrics Services
to social_networks consumi analisi ricerca web sys:relevantFor:sdc areacom blog internet media by mbinotto on Mar 20, 2009, 6:22 PM
copy | spam

The structure and function of complex networks
M. E. J. Newman March2003
to systems socialnetwork arch web readinggroup affinity network thesis algorithm math scale-free-networks information_retrieval network_theory small_world folksonomy web-graph clustering complex-systems growth by cdevries and 34 other people on Mar 23, 2009, 4:26 AM
pick | copy | URL | BibTeX | openURL | spam

The Anatomy of a Large-Scale Hypertextual Web Search Engine.
Sergey Brin and Lawrence Page Computer Networks 30(1-7):107-117(1998)

web as tag from jaeschke
web as concept from jaeschke
web as concept from all users

+ semantic
+ tools
+ internet
+ web2.0
+ software
+ Bookmarks
+ 2.0
+ search
+ webdesign
+ design
+ blog
+ social

Fig. 1 BibSonomy displays bookmarks and BibTeX based bibliographic references simultaneously

2.2 BibSonomy: A Social Bookmark and Publication Sharing System

Resource sharing systems like BibSonomy provide an easy way to organize and manage different kinds of resources. What is considered as a resource depends on the type of system. For instance, in del.icio.us, the resources are URLs, in flickr, the resources are pictures, and in BibSonomy they are either URLs or publication entries. As described in the previous section, 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 tags to it. The collection of all his assignments is his *personomy*, the collection of all personomies constitutes the *folksonomy*. As in other systems, the user can explore his personomy, as well as the personomies of the other users, in all dimensions: for a given user one can see all resources he has uploaded, together with the tags he has assigned to them; when clicking on a resource one sees which other users have uploaded this resource and how they tagged it; and when clicking on a tag one sees who assigned it to which resources (see Figure 1). The systems allows for additional functionality. For instance, one can copy a resource from another user, and label it with one's own tags. Overall, these systems provide a very intuitive navigation through the data.

BibSonomy[11] is one of the social resource sharing tools that have acquired large numbers of users within the last 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 the individual user (e.g. organizing ones bookmarks in a browser-independent, persistent fashion) without too much overhead. Additionally, BibSonomy allows to share both bookmarks

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

and publication metadata. It started as a student project at the Knowledge and Data Engineering Group of the University of Kassel[12] in spring 2005. The goal was to implement a system for organizing BibTeX entries (cf. [39]) in a way similar to bookmarks in del.icio.us – which was at that time becoming more and more popular. BibTeX is a popular literature management system for LaTeX, which many researchers use for writing scientific papers. We soon decided to integrate bookmarks as a second type of resource into the system. At the end of 2005, we announced BibSonomy first to some colleagues, later in 2006 to the public. Since then, the number of users has grown steadily. Today, BibSonomy has more than 190000 registered users. We implemented several useful features and redesigned the architecture to ease future developments. Our team and other research groups use BibSonomy or its data for research, and we have implemented our research results into the system, e.g. the FolkRank algorithm and tag recommendation methods – both for the benefit of the users and to directly measure the performance of our methods. A more detailed description of BibSonomy can be found in [24]. In the following subsections, we will give pointers to the most interesting parts of the system.

2.2.1 User Interface

A typical list of posts is depicted in Figure 1 which shows bookmark and publication posts containing the tag *web*. The page is divided into four parts: the header (showing information such as the current page and path, navigation links and a search box), two lists of posts – one for bookmarks and one for publications – each sorted by date in descending order, and a list of tags related to the posts. This scheme holds for all pages that are showing posts; it allows for navigation in all dimensions of the folksonomy. The posts in the lists are sorted by date in descending order, while the tags can be sorted lexicographically or by frequency of usage, depending on the user's choice.

Beside this kind of pages, systems like BibSonomy typically contain summary pages representing the content in form of a cloud. The page with the global tag cloud summarizes in a clear way the content of the system by the used tags. A similar functionality is offered by the author and relation pages which are special pages for BibSonomy. Note that on selected pages posts can be ordered by relevance as calculated by the FolkRank algorithm (cf. Sec. 3.2).

2.2.2 Architecture

The basic building blocks of BibSonomy are an Apache Tomcat[13] servlet container using Java Server Pages[14] and Java Servlet[15] technology and a

[12] `http://www.kde.cs.uni-kassel.de/`

[13] `http://tomcat.apache.org/`

[14] `http://java.sun.com/products/jsp`

[15] `http://java.sun.com/products/servlets`

MySQL[16] database as backend. The project uses the Model View Controller (MVC) programming paradigm to separate the logical handling of data from the presentation. This enables us to produce output in various formats (see Section 2.2.3), since adding a new output format is accomplished by simply implementing a JSP as a view of the model.

The central database schema of BibSonomy is based on four tables: one for bookmark posts, one for publication posts, one for tag assignments (*tas*) and one for *relations* between tags. Two further tables store information regarding *users* and *groups*.

The post tables are connected with the tas table by the key *post_ id*. The schema is not normalized – on the contrary we have added a high amount of redundancy to speed up queries. For example, besides storing group, user name and date in the posts table, we also store it in the tas table to minimize the number of rows touched when selecting rows for the various views. Furthermore, several other tables hold counters (i. e., how many people share one resource, how often a tag is used, etc.). Finally, a large set of indexes (12 in the tas table alone) builds the basis for a fast answering of queries.

Overall, we spent a large amount of work on investigating and optimizing SQL queries and table schemas and tested both with folksonomy data of up to 8000000 posts. At the moment, we need no special caching or physical distribution of the database to get reasonable response times.

2.2.3 Features

The most simplistic but also most laborious way to add posts to BibSonomy is by entering their metadata manually into form fields. To lower the effort to get data into BibSonomy, it supports various ways to import resources from files and web pages (e.g. BibTeX or Endnote[17]) or by so called "scrapers"[18] which allow to automatically extract publication metadata from digital libraries like SpringerLink.[19] Nevertheless – forms are still used to edit posts.

Exporting publication references in BibTeX format is accomplished by preceding the path of a URL showing publication posts with the string `/bib` – this returns all publications shown on the respective page in BibTeX format. For example the page `http://www.bibsonomy.org/bib/search/text+clustering` returns a BibTeX file containing all literature references which contain the words "text" and "clustering" in their fulltext.

More general, every page which shows posts can be represented in several different ways by preceding the path of the URL with a specific string to specify the export format, e.g. `/xml` for bookmarks in XML format or `/publ` for publications in a simple HTML format suited for the integration into a

[16] `http://www.mysql.com/`

[17] `http://www.endnote.com/`

[18] `http://scraper.bibsonomy.org/`

[19] `http://www.springerlink.de/`

web page (for an integration example see `http://www.kde.cs.uni-kassel.de/pub`). For an overview of the available export formats for publications, one can use the `/export` path extension which is also linked on all web pages showing publication posts. The export feature allows to generate publication lists for external websites, e.g. for personal and institute webpages or for project pages.

Experience has shown that an Application Programming Interface (API) is crucial for a Web 2.0 system to gain success. Hence we have implemented a lightweight REST API[20] which can be used and accessed also by less experienced programmers. We use the API for the integration of JabRef.[21] The catalogue of the library of the university of Cologne uses the API to access tagging information for its books.

There are several other valuable features like the publication basket, the duplicate detection mechanism, a tag editor, the mirror of the famous DBLP computer science library,[22] or the integration with other systems. A description of those features can be found in [24]. In our blog,[23] we report regularly on new developments. For research purposes, we release a complete snapshot of BibSonomy's public data on a regular basis.[24]

2.3 Folksonomies

As described in Sec. 2.1, folksonomies are the core structure of social bookmarking systems. The word "folksonomy" is a blend of the words "taxonomy" and "folk", and stands for conceptual structures created by the people.[25] The way an folksonomy is emerging is the same in all these systems and can be described as follows: There is a user who is interested in a certain resource. A folksonomy system provides a way to store this resource and to annotate it. Typically, the annotation process is as simple as possible and driven by keywords called *tags*.[26] The tags can serve several purposes [16], e.g. they describe the content of a resource or the reasons why the resource was saved. The central elements of a folksonomy are depicted in Figure 2. The center is formed by a post which connects a user with tags and a resource. Different users can use different tags to describe the same resource and resources are typically tagged by several users. The emergent structure of users, tags and resources is called *folksonomy.*

[20] `http://www.bibsonomy.org/help/doc/api.html`

[21] `http://bibsonomy.blogspot.com/2009/02/feature-of-week-bibsonomy-plugin-for.html`

[22] `http://www.informatik.uni-trier.de/~ley/db/`

[23] `http://bibsonomy.blogspot.com/`

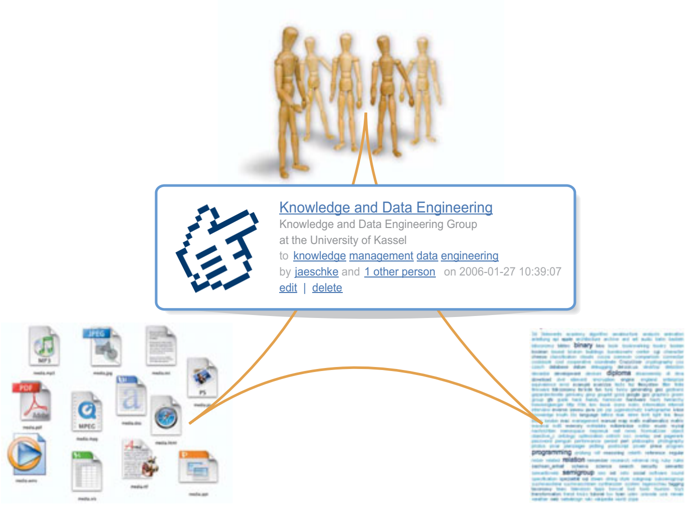
[24] `http://www.bibsonomy.org/help/faq/600_benchmark`

[25] `http://www.vanderwal.net/folksonomy.html`

[26] A more exotic example is the use of geographic coordinates as tags to describe where a photo was taken. In principle this is the same annotation process.

Such systems provide direct benefits which makes them attractive for their users. The process of tagging is very simple and it is very easy for every user to access his own resources. The web based storage of e.g. bookmarks allows to access them at any time from all over the world. Bookmarks are no longer tied to a single computer. Due to the nature of such systems, browsing within the bookmark collection leads to the serendipitous discovery of unknown resources. This property of folksonomies is mostly unexpected by its users, but it makes systems with a folksonomy core so fascinating. Despite the uncoordinated tag assignment of different users in such systems, the emergence of semantics can be observed (details in Sec. 2.5).

Another advantage of folksonomies is the human-contributed annotation which can be seen as a lightweight knowledge representation. Most of the tags describe the content of the annotated resource and as they are assigned by humans which are able to grasp the content, the resulting description is better than automatic solutions used in search engines or categorizations systems. However, the broad range of human contribution is also a major disadvantage of folksonomies. To make the usage of bookmarking system simple, it is allowed to use any arbitrary tag in a totally uncontrolled way (cf. [48]). This results in difficulties, as tags tend to suffer from typical language problems like synonyms, polysemy and singular vs. plural forms. The usage of tags can be driven by a very personal preferences which confuses others and does not contribute to common semantics.



Knowledge and Data Engineering
Knowledge and Data Engineering Group
at the University of Kassel
to knowledge management data engineering
by jaeschke and 1 other person on 2006-01-27 10:39:07
edit | delete

Fig. 2 Visualization of an example bookmark post of a tagging system

2.4 A Formal Model for Folksonomies

A folksonomy describes the users, resources, and tags, and the user-based assignment of tags to resources. We present here a formal definition of folksonomies (cf. [25]), 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., $Y \subseteq U \times T \times R$, whose elements are 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 *is-a relation.*

Definition 2. 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 ith dimension.

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, e.g. in BibSonomy they are either URLs or publication entries.

Definition 3. For convenience we also define the set P of all *posts* as

$$P := \{(u, S, r) \mid u \in U, r \in R, S = \text{tags}(u, r), S \neq \emptyset\}$$

where, for all $u \in U$ and $r \in R$, $\text{tags}(u, r) := \{t \in T \mid (u, t, r) \in Y\}$ denotes all tags the user u assigned to the resource r.

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 [50, 15] as a *triadic context* [35, 46]. 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.

In a typical folksonomy system, every tag assignment is connected with several other properties like date, group or resource type. For sake of simplicity, we disregard these properties for the rest of the work, unless stated otherwise.

2.5 Network Properties of Folksonomies

The new data of folksonomy systems provides a rich resource for data analysis, information retrieval, and knowledge discovery applications. We made a first step towards this end in [7]. The goal is to gain better insights into these systems by analyzing the main network characteristics on two example systems.

To this extent, we investigate the growing network structure of folksonomies over time from different viewpoints, using two datasets from del.icio.us and BibSonomy as examples. First, we investigate the network structure of folksonomies much on the same line as the developments in the research area of complex networks. To that end, we adapt classical network measures like characteristic path length and clustering coefficient for so-called "small world networks" which have been used on a wide variety of graphs in recent years, to the particular tripartite structure of folksonomies. We show that folksonomies do indeed exhibit a small world structure, as we observe in both systems very short path length, (in average around 3) in both systems and a very high clustering coefficient compared to random graphs. This helps to explain to some extent the successful serendipitous browsing of users in such systems. The small world property implies for the users that only few clicks are needed to end up in a new and hopefully interesting topic within a folksonomy. On the other hand, the high clustering coefficient hints a cluster of resources with a similar topic within the direct neighborhood.

Second, beyond the analysis of the whole hypergraph, we also consider specific projections of it by narrowing the scope and focusing on particular features of the structure. We introduced a weighted network of tags where link strengths are based on the frequencies of the tag-tag co-occurrence, and studied the weight distributions and connectivity correlations among nodes in this network. Our analysis and experiments indicate the existence of the emergence of shared semantics in the folksonomy system, implicitly negotiated by users. We find indicators for both hierarchical and social structures in the network of tag-tag co-occurrence.

Our experiments hint that spam – which becomes an increasing nuisance in social resource sharing systems – systematically shows up in the connectivity correlation properties of the weighted tag-tag co-occurrence network. These activities in data from del.icio.us in its early days indicate the need to develop more advanced methods to fight against such misuse of folksonomy systems. We will present our approach to detect spam in Sec. 3.1. A deeper analysis of the emergent semantics in folksonomies appears promising, and results in this direction are presented in Sec. 4.3. A first application to support the user based on the collaborative intelligence hidden in folksonomies is the tag recommender (cf. Sec. 3.3).

3 Applications

3.1 Spam Detection

Web spam detection is a well known challenge for search engines. Spammers add specific information to their web sites that solely serves the purpose to increase the rank of a page in search results, but not its quality or content. They thereby increase the traffic to their web sites – be it for commercial or political interests or to disrupt the service provided. Ranking algorithms need to detect those pages.

Not only search engines struggle with malicious web content. Social bookmarking systems also have become an attractive place for posting web spam. Spammers (mis)use the popularity and the high ranking of social bookmarking systems in search engines for their purposes. All they need is an account; then they can freely post entries which bookmark the target spam web site. In recent months, different spamming techniques have been developed to frequently show up on popular sites, recent post sites or as highly ranked posts on a search for a specific tag. For instance, spammers register several accounts and publish the same post several times. Besides appearing on the "recent post" page, the bookmark may show up on the "popular page", since "many" users have considered the bookmark. Another technique is to add diverse tags to the bookmark or use popular tags.

In order to retain the original benefits of social bookmarking systems, we developed techniques which prevent spammers from publishing in these systems [34]. The problem can be considered as a binary classification task. Based on different features that describe a user and his posts, a model is built from training data to classify unknown examples (on a post or user level) either as "spam" or "non-spam". As we consider "social" systems in which users interact with each other and one incentive to use the system is to see and to be seen, an exclusion of non-spammers from publishing is a severe error which might prevent the user from further participation. Similar to other spam detection settings, this problem needs to be taken into consideration when classifying users.

The adaptation of classification algorithms to this task consists of two major steps. The first one is to select features for describing the users. The second step is the selection of an appropriate classifier for the problem. In [34], we introduce a set of initial features that can be used for spam classificiation. These features are evaluated with well-known classifiers (SVM, Naive Bayes, J48 and logistic regression) against a simple baseline of representing a user by the usage of tags. Combining all features shows promising results exceeding the AUC and F1 measure of the selected baseline. Considering the different feature groups, co-occurrence features show the best ROC curves.

Our results support the claim of [21], that the problem can be solved with classical machine learning techniques – although not perfectly. The difference to web spam classification are the features applied: on the one hand, more information (e. g., IP address, tags) is given, on the other hand spammers reveal their identity by using a similar vocabulary and similar resources. This is why co-occurrence features tackle the problem very well.

Overall, our contribution represents a first step towards the elimination of spam in social bookmarking systems using machine learning approaches. We implemented a framework within BibSonomy following the results of our analysis. The framework automatically flags in average more than 200 new spammers per day. Besides the practical need to eliminate spam, we intend to use this platform to develop and evaluate further social spam detection mechanisms and to tune the performance of the running machine learning approaches. For this we use also results from last year's ECML PKDD discovery challenge[27] organized by us, where one task was the prediction of spam in social bookmarking systems.

3.2 Ranking in Folksonomies

In the past, folksonomies were able to attract a large number of users who created huge amounts of information. But with the growing number of resources stored within each users personomy, it becomes more and more difficult for the user to find and retrieve the saved resources. A first step to searching folksonomy based systems – complementing the browsing interface usually provided as of today – is to employ standard techniques used in information retrieval or, more recently, in web search engines. Since users are used to web search engines, they likely will accept a similar interface for search in folksonomy-based systems. The research question is how to provide suitable ranking mechanisms, similar to those based on the web graph structure, but now exploiting the structure of folksonomies instead. To this end, we proposed in [25] a new algorithm, called *FolkRank*, that takes into account the folksonomy structure for ranking users, tags and resources in folksonomy based systems. Further, the algorithm can be used for a topic-specific ranking.

The general idea of FolkRank is as follows: Given a set of preferred tags, users, and/or resources, a topic specific ranking provides an ordering of the elements of the folksonomy in descending importance with respect to the preferred elements. To that end, FolkRank is a differential approach of a weight-spreading algorithm which compares the resulting rankings with and without preference vector computed on the folksonomy graph. We implemented the weight-spreading ranking scheme on folksonomies in two steps. First, we transform the hypergraph between the sets of users, tags, and resources into an undirected, weighted, tripartite graph. On this graph, we apply a version of PageRank that takes into account the edge weights.

The original formulation of PageRank [3] reflects the idea that a page is important if there are many pages linking to it, and if those pages are important themselves. We employ a similar motivation for our ranking scheme in folksonomies. The basic notion is that a resource which is tagged with important tags by important users becomes important itself. The same holds, symmetrically, for tags and users, thus we have a tripartite graph in which the vertices are mutually reinforcing each other by spreading their weights. It turned out, however, that running an adapted PageRank as is returned results that were largely dominated by the global structure of the folksonomy, yielding the same top elements such as the tags "web" or "blog" on top no matter what the preferences were. Thus, FolkRank circumvents that problem

[27] `http://www.kde.cs.uni-kassel.de/ws/rsdc08`

using a differential approach. It computes a topic-specific ranking in a folksonomy by computing the winners and losers of the mutual reinforcement of resources when a user preference is given, compared to the baseline without a preference vector. More details can be found in [25].

There, the FolkRank ranking scheme has been used to generate personalized rankings of the items in a folksonomy, and to recommend users, tags and resources. Top folksonomy elements which are retrieved by FolkRank tend to fall into a coherent topic area, e.g. "Semantic Web". This leads naturally to the idea of extracting *communities of interest* from the folksonomy, which are represented by their top tags and the most influential persons and resources. This idea found its way into BibSonomy. There is now an option to rank resources not only by date, but also by FolkRank[28]. The shown page displays not only the ranked resource list but also the top ranked similar tags and the most influential users of this topic. These users form some kind of community. By making such an implicit existing communities explicit, interested users can find other users, also interested in the search topic and in this way community members can more easily get to know each other and learn of others' resources.

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 its benefits without bothering untrained users with its rigidity. This could be done by utilizing the strength of the semantic technology within a folksonomy system and using data mining methods to bridge the gap between both worlds. One approach going in this direction is presented in Sec. 3.3 as tag recommenders simplify the posting process and in Sec. 4.3 where different kinds of related tags are extracted which can form a basis for a better organization of tags. We believe that this will become a fruitful research area for the Semantic Web community for the next years.

One application of FolkRank is presented in [26]. There, we analyze the emergence of common semantics by exploring trends in the folksonomy. Since the structure of a folksonomy is symmetric with respect to the dimensions "user", "tag", and "resource", we can apply the same approach to study upcoming users, upcoming tags, and upcoming resources over time. With FolkRank, we compute topic-specific rankings on users, tags, and resources. In a second step, we can then compare these rankings for snapshots of the system at different points in time. We can discover both the absolute rankings (who is in the Top Ten?) and winners and losers (who rose/fell most?). We present a technique for analyzing the evolution of topic-specific trends.

Furthermore, there has been a lively discussion in e.g. the delicious-discuss mailing list about the usefulness of the $\prec$ relation in the folksonomy, which is partially realized as *bundles* in del.icio.us. We will investigate first steps to be

[28] For an example, see `http://www.bibsonomy.org/tag/Semantic_Web?order=folkrank`

able to make use of ontology learning techniques to populate this relation in BibSonomy and augment the underlying semantic structure in the folksonomy in Sec. 4.

3.3 Recommending Tags

Recommenders are a common technique to support users in finding new and interesting items, e.g. movies, books, or other products. In folksonomies, recommenders can be used to recommend similar users, interesting resources or help to find the right tags while posting a new resource. We focus on the third task in this section. One example for a resource recommender can be found in [49]. The literature on tag recommendations in folksonomies is still sparse. The existing approaches usually lie in the area of collaborative filtering and information retrieval. Most recently, the ECML PKDD 2008 Discovery Challenge[29] organized by our research group has addressed the problem of tag recommendations in folksonomies [23]. The provided dataset gives a good basis for the research in this area and the upcoming next challenge[30] shows the need for better recommender approaches and increasing interest of researchers in this area.

To support users in the tagging process and to expose different facets of a resource, most of the systems offered some kind of tag recommendations already at an early stage. Del.icio.us, for instance, had a tag recommender in June 2005 at the latest,[31] and also included resource recommendations.[32] However, no algorithmic details were published. We hypothesize that these recommendations basically provide those tags which were most frequently assigned to the resource.

As of today, nobody has empirically shown the benefits of recommenders in such systems. In [31], we evaluate a tag recommender based on Collaborative Filtering, a graph based recommender using our ranking algorithm FolkRank, and several simple approaches based on tag counts. With this research we start a qualitative comparison of different recommender approaches while simultaneously adopting state of the art techniques to work with the underlying triadic graph. The results presented in [31] built upon results presented at ECML PKDD 2007 [32].

The presented results in [31] show that the graph-based approach of FolkRank is able to provide tag recommendations which are significantly better than those of approaches based on tag counts and even better than those of state-of-the-art recommender systems like Collaborative Filtering.

[29] `http://www.kde.cs.uni-kassel.de/ws/rsdc08/`

[30] `http://www.kde.cs.uni-kassel.de/ws/dc09/`

[31] `http://www.socio-kybernetics.net/saurierduval/archive/2005_06_01_archive.html`

[32] `http://blog.del.icio.us/blog/2005/08/people_who_like.html`

The tradeoff is that the computation of FolkRank recommendations is cost-intensive so that one might prefer less expensive methods to recommend tags in a social bookmarking system. The *most popular tags ρ–mix* approach proposed by us in [31] has proven to be considered as a solution for this problem. It provides results which can almost reach the quality of FolkRank but which are rather cheap to generate. Especially the possibility to use index structures (which databases of social bookmarking services typically provide anyway) makes this approach a good choice for online recommendations. Finally, despite its simplicity and non-personalized aspect, the *most popular tags* achieved reasonable precision and recall on the small datasets (last.fm and BibSonomy) which indicates its adequacy for the cold start problem of young systems.

One result of the ECML PKDD discovery challenge 2008 was the insight that two recommendation tasks can be distinguished. In [31], we focus on the dense part of the folksonomy. We assume that we have information about both the user and the resource and make use of this information to predict the tags the user will use to describe the resource which was already tagged by other users of the system. Contrary to this, most often not all information is available. This means that either the user or the resource or both are new. In this case, one cannot apply the methods described in [31]. We address this issue in [28] where we utilize the content of the webpage which the user will tag in a content based recommender. The underlying methods are known as text classification approaches. J. Illig evaluates the applicability of these methods in general in [27]. In principle, the application of text classification approaches is possible, but the approaches need to be better adapted to the underlying problem. The high number of classes decreases the performance in terms of runtime behavior and accuracy. An interesting next step is the integration of user information in the recommendation process.

4 Towards More Semantics in Folksonomies

As mentioned in Section 2.2, BibSonomy provides the possibility to store tag relations as a kind of conceptualization. One outcome of Sec. 2.5 is the existence of *emergent semantics* [42, 44] in folksonomies. In this section, we present three approaches which will help to understand and to extract the semantics that are implicitly added by the user and hidden in the folksonomy. We will show ways to make it explicit and available for further use. We start with a short comparison of folksonomies and ontologies.

4.1 Folksonomies and Ontologies

Ontologies are a well-known formalism to represent knowledge in a structured way [43] and are the building block of the “Semantic Web” effort. With their well-defined semantics, ontologies offer benefits for a wide spectrum of applications supported by advanced tools from industry and academics. Nevertheless, there are problems to make use of Semantic Web technology in very large application contexts, especially in the web. The web contains huge masses of data but not in any case the data is available in the structured form needed by the Semantic Web, e.g. as ontologies. The knowledge acquisition bottleneck characterize the phenomena that the transformation process from unstructured to structured information is possible but does not scale to the size of the web. The reason is that a certain expertise is needed to create ontologies and to maintain them. This raises the cost of knowledge acquisition and only few people are contributing. Learning ontologies from text [9] is a first way to simplify the acquisition process by utilizing machine learning approaches and linguistic knowledge.

Folksonomies can be seen as a lightweight knowledge representation. Many unexperienced users contribute small pieces of information – unfortunately only in a weakly structured fashion. There is a large amount of information, but it is unstructured and therefore incompatible with semantically rich representations. Both approaches could benefit from each other: While folksonomies need more structure, ontologies need more contributors. Research in this direction has been stimulated in form of the "Bridging the Gap between Semantic Web and Web 2.0" workshop,[33] where the contributions ranged from the use of human contributed information to simplified Web 2.0-like Semantic Web tools.

The emergent semantics in folksonomies can be extracted by using machine learning algorithms or advanced analysis methods. A first approach is presented in the next section. To be able to develop advanced knowledge extraction methods, a better understanding of the kind of underlying semantics is needed. We presented the summary of a first analysis in Sec. 2.5, which supports the existence of semantics in folksonomies. The next steps are a deeper understanding of the type of the relations hidden in folksonomies (cf. Sec. 4.3) and the development of methods to extract them (cf. Sec. 4.4).

4.2 Associations between Tags, Users, and Resources

As folksonomy systems grow larger, the users feel the need for more structure for better organizing their resources. For instance, approaches for tagging tags, or for bundling them, are discussed on the corresponding mailing lists e.g. the delicious-discuss list and are provided by some of the systems. 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 of the already existing tags, and additional tags, ultimately leading towards emergent semantics by converging use of the same vocabulary.

[33] `http://www.kde.cs.uni-kassel.de/ws/eswc2007/`

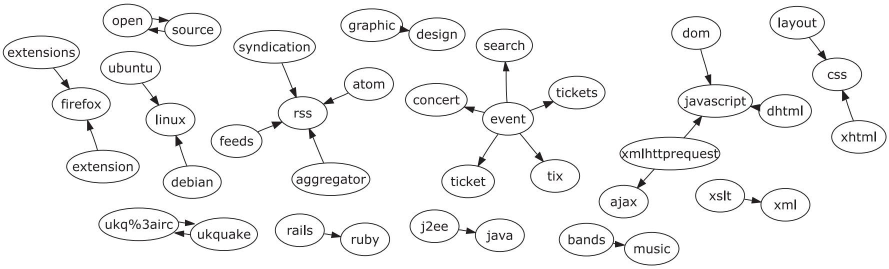


Fig. 3 Example of extracted association rules between tags

In [41], we focus on a certain KDD technique, namely association rules [1]. Since folksonomies provide a three-dimensional dataset (users, tags, and resources) instead of a usual two-dimensional one (items and transactions), we start in [41] with a systematic overview of projecting a folksonomy onto a two-dimensional structure. For one selected projection, we demonstrate here the outcome of association 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. Another example as well as details are described in [41].

To illustrate the outcome of the learning approach, an example from [41] is given in Figure 3. It shows all rules between tags from del.icio.us for a minimum support of 0.05% and a minimum confidence of 50%. In this example, rules of the form $A \rightarrow B$ can be read as "if a user has assigned the tag A to some resources, he often assigned tag B as well". If del.icio.us users are tagging some webpage with *debian*, they are likely to tag it with *linux* as well, and pages about *bands* are probably also tagged with *music*. As discussed in Section 4.1, we are looking for ways to discover subsumption relations which are needed to build ontologies, so that rule mining can be used to learn a taxonomic structure. As an example, consider the case where many resources tagged with *xslt* are also tagged with *xml*. This indicates that *xml* can be considered a supertopic of *xslt* if one wants to automatically populate the $\prec$ relation. Figure 3 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 the chat channel ukq.

We can learn from these examples that it is possible to extract meaningful relations between tags from folksonomy data. To get a better understanding of what was extracted, we have to ground the extracted relations between the tags, users and resources by mapping them to an external knowledge source with a clear semantic meaning or better grounded relationships. We can try to do this for all three dimension, but we focus on tags, as they transport (most of) the semantic information. Such semantic information is captured by large lexical ontologies and thesauri, and we will use both to evaluate the meaning of different similarity measures between tags.

Therefore, in the next section, a fine grained analysis of various techniques we used here and in previous sections, namely, association rule mining and FolkRank ranking, is presented to further contribute to the understanding of the extracted relation by every method.

4.3 Understanding Tag Relatedness in Folksonomies

In this part we focus on the understanding of the specific relationship between tags in folksonomies. As we have seen in Sec. 2.5, the structure of folksonomies differs fundamentally from that of, e.g. natural text or web resources, and poses new challenges for the fields of knowledge discovery and ontology learning. Central to these tasks are the concepts of similarity and relatedness. In the previous section, among others, we introduced the computation of relations between tags by the association rule mining algorithm (based on co-occurrence) which can be easily turned into a tag relatedness measure. In [5], we focus on similarity and relatedness of tags, because they carry most of the semantic information within a folksonomy, and provide thus the link to ontologies and more formal semantics. Additionally, this focus allows for an evaluation with well-established measures of similarity in existing lexical databases.

Budanitsky and Hirst pointed out that similarity can be considered as a special case of relatedness [4]. As both similarity and relatedness are semantic notions, one way of defining them for a folksonomy is to map the tags to a thesaurus or lexicon like Roget's thesaurus[34] or WordNet [13], and to measure the relatedness there by means of well-known metrics. The other option is to define measures of relatedness directly on the network structure of the folksonomy. One important reason for using measures grounded in the folksonomy, instead of mapping tags to a thesaurus, is the observation that the vocabulary of folksonomies includes many community-specific terms which did not make it yet into any lexical resource. Measures of tag relatedness in a folksonomy can be defined in several ways. Most of these definitions use statistical information about different types of *co-occurrence* between tags, resources and users. Other approaches adopt the *distributional hypothesis* [14, 20], which states that words found in similar contexts tend to be semantically similar. This approach also retains the possibility to include "matured" folksonomy vocabulary back into the thesauri or lexicons, which addresses the inherent knowledge acquisition bottleneck problem of these systems. From a linguistic point of view, these two families of measures focus on orthogonal aspects of structural semiotics [10, 8].

[34] `http://www.gutenberg.org/etext/22`

The co-occurrence measures address the so-called syntagmatic relation, where words are considered related if they occur in the same part of text. The contextual measures address the paradigmatic relation (originally called associative relation by Saussure), where words are considered related if they can replace one another without affecting the structure of the sentence.

In most studies, the selected measures of relatedness seem to have been chosen in a rather ad-hoc fashion. We believe that a deeper insight into the semantic properties of relatedness measures is an important prerequisite for the design of ontology learning procedures that are capable of harvesting the emergent semantics of a folksonomy.

In [5], we analyse five measures of tag relatedness: the *co-occurrence count*, three *distributional measures* which use the cosine similarity [40] in the vector spaces spanned by users, tags, and resources, respectively, and *FolkRank* (cf. Sec. 3.2), our graph-based measure. Our analysis is based on data from a large-scale snapshot of the popular social bookmarking system del.icio.us.[35] To provide a semantic grounding of our folksonomy-based measures, we map the tags of del.icio.us to synsets of WordNet and use the semantic relations of WordNet to infer corresponding semantic relations in the folksonomy. In WordNet, we measure the similarity by using both the taxonomic path length and a similarity measure by Jiang and Conrath [29] that has been validated through user studies and applications [4]. The use of taxonomic path lengths, in particular, allows us to inspect the edge composition of paths leading from one tag to the corresponding related tags. This characterization proves to be especially insightful.

As a result, we show that distributional measures, which capture the context of a given tag in terms of resources, users, or other co-occurring tags, establish – in a statistical sense – *paradigmatic* relations between tags in a folksonomy. Strikingly, our analysis shows that the behavior of the most accurate measure of similarity (in terms of semantic distance of the indicated tags) can be matched by a computationally lighter measure (tag context similarity) which only uses co-occurrence with the popular tags of the folksonomy. In general, we show that a semantic characterization of similarity measures computed on a folksonomy is possible and insightful in terms of the type of relations that can be extracted. We show that despite a large degree of variability in the tags indicated by different similarity measures, it is possible to connotate *how* the indicated tags are related to the original one.

Another contribution of [5] addresses the question of emergent semantics: our results indicate clearly that, given an appropriate measure, globally meaningful tag relations can be harvested from an aggregated and uncontrolled folksonomy vocabulary. Specifically, we show that the measures based on tag and resource context are capable of identifying tags belonging to a common semantic concept. Admittedly, in their current status, none of the measures we studied can be seen as *the* way to instant ontology creation. However, we

[35] `http://del.icio.us/`

believe that further analysis of these and other measures, as well as research on how to combine them, will help to close the gap towards the Semantic Web.

Based on the results we have so far, the construction of tag hierarchies is the natural next step. We made a first attempt in [2], where we present results of a learned music style ontology. The data stems from last.fm[36], a music folksonomy system. A more advanced learning approach was applied on our del.icio.us dataset. The idea was to show that learning of ontologies from a large scale folksonomy is possible. In [45], an extended version of the algorithm from [2] is used for learning the ontology. The results are compared with Wordnet[37] and with the categorization scheme of Wikipedia.[38] Several drawbacks of the original algorithm could be solved and led to a better ontology. One central factor was the disambiguation of the word sense of polysemous tags and the calculation of synsets. Both approaches utilize the relatedness measures grounded before. While the synset detection algorithm reduces the number of tags by merging real synonyms as well as spelling variants, the word sense disambiguation component places tags more than once in the generated ontology. An example of a learned ontology is depicted in Fig. 4. As one can see, the tag *language* is placed under the tag *programming*, which hints the meaning of *language* in this case. We see the programming languages *lisp* as a sub-tag of *languages.* No names of natural languages like German are placed as sub-tags of *language* in this part of the graph. A more detailed description of the algorithm and the results can be found in [45].

4.4 Conceptual Structures in Folksonomies

Unlike ontologies, folksonomies do not suffer from the knowledge acquisition bottleneck, as the significant provision of content by many people shows. On the other hand, folksonomies – unlike ontologies [17] – do not explicitly state shared conceptualisations, nor do they force users to use tags consistently. However, the usage of tags of users with similar interests tends to converge to a shared vocabulary as explained in the previous section. Our intention is to discover these shared conceptualisations that are hidden in a folksonomy. To this end, we present in [30] an algorithm, TRIAS, for discovering subsets of folksonomy users who implicitly agree (on subsets of resources) on a common conceptualization.

Our algorithm returns a tri-ordered[39] set of triples, where each triple (A, B, C) consists of a set A of users, a set B of tags, and a set C of resources. These triples – called *tri-concepts* in the sequel – have the property that each user in A has tagged each resource in C with all tags from B, and

[36] `http://www.last.fm/`

[37] `http://wordnet.princeton.edu/`

[38] `http://www.wikipedia.org/`

[39] See [30] for details.

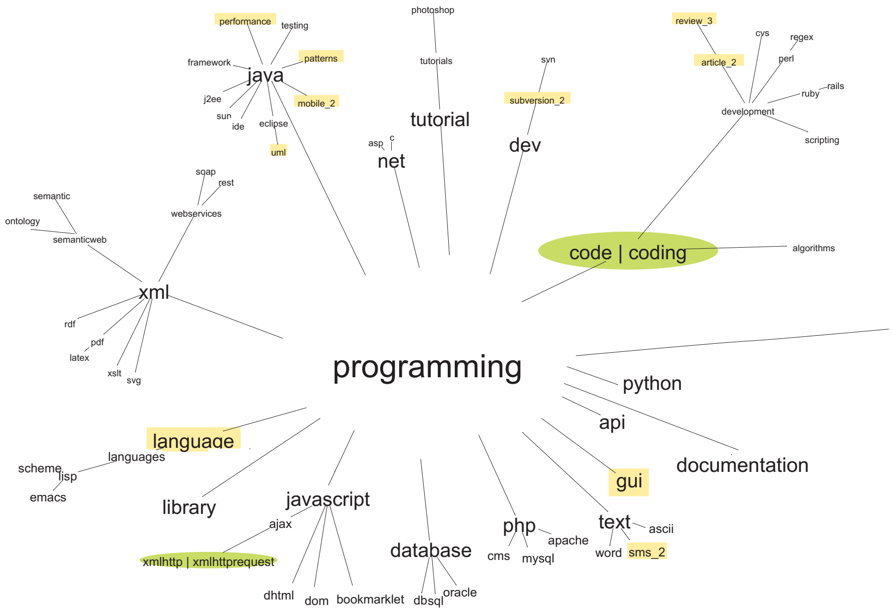


Fig. 4 Fragment of the learned ontology centered around programming. One meaning of language (programming languages) is depicted. (Figure is taken from [45])

that none of these sets can be extended without shrinking one of the other two dimensions. Each retrieved triple indicates thus a set A of users who (implicitly) share a conceptualisation, where the set B of tags is the intension of the concept, and the set C of resources is its extension. We can additionally impose minimum support constraints on each of the three dimensions "users", "tags", and "resources", to retrieve the most significant shared concepts only.

From a data mining perspective, the discovery of shared conceptualizations opens a new research field which may prove interesting also outside the folksonomy domain: "Closed itemset mining in triadic data", which is located on the confluence of the research areas of Association Rule Mining and Formal Concept Analysis.

In contrast to the already presented results of Sec. 4.3 and 4.2, Trias relates elements from different dimensions of the folksonomy. This allows for the simultaneous detection of hidden user groups and their interest expressed by the tags and the tagged resources. Another application could be the extraction of a concept hierarchy to learn ontologies as pointed out in [30].

The next step after discovering shared conceptualisations would be to formalize them in an ontology, and to combine and integrate this approach with the results of Sec. 4.3.

5 Conclusion and Future Work

Data Mining on folksonomies is a new research area attracting a lot of attention in the last years as new types of data with unknown and interesting properties appear. In this paper we presented the analysis of the properties of these new data, the application and adaption of known data mining approaches, and the usage of this data to extract semantic information. The three applications spam detection, ranking and recommendation were introduced and three approaches to extract the hidden semantic information from folksonomies were presented. Our own system BibSonomy was introduced as a platform where researchers manage their publication on a daily basis but also as a research environment to test new methods like ranking and recommendation which already found the way into the system.

In principle, the presented folksonomy mining approaches implement the ideas of Semantic Web Mining (cf. [47]). Therefore, they make our vision of utilizing mining to help to build the Semantic Web and to analyze it real. Hence, one long term goal is to use the weakly structured data of a folksonomy as data source for the Semantic Web. Further, convincing people to use a kind of “Semantic Bookmarking System” which is usable in the same easy way as the existing non-semantic versions is the vision and part of the future work. First steps in this direction with promising results were presented in this paper and we could show that it is possible to extract valuable information from folksonomies and to use data mining techniques to support user of social bookmarking systems.

A central phenomenon of the Web 2.0 is the contribution of many users distributed over the world but tied to a computer. The next step is to bring the web to mobile devices and to set up new services which do not only allow users to provide information but also to monitor their activities. This physical information will provide new kinds of data which allow for new services. A combination of the physical world with its small devices, sensors etc., the Web 2.0 look and feel, and the Semantic Web to connect everything will lead to the next generation of the Web.

Acknowledgements

This paper would not have been possible without the fruitful joint work with my colleagues from the KDE research group in Kassel, Germany, namely G. Stumme, C. Schmitz, R. Jäschke, D. Benz, and B. Krause. Thanks also to the collaborators A. Baldassarri, C. Cattuto, B. Ganter, V. Loreto, L. B. Marinho, F. Menczer, L. Schmidt-Thieme, and V. D. P. Servedio from outside the group. All papers summarized in this work are results of an inspiring collaboration with all of them.

Part of this research was funded by the European Union in the Nepomuk (FP6-027705) and Tagora (FET-IST-034721) projects.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: SIGMOD 1993: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207–216. ACM Press, New York (1993)
2. Benz, D., Hotho, A.: Position paper: Ontology learning from folksonomies. In: Hinneburg, A. (ed.) LWA 2007: Lernen - Wissen - Adaption, Halle, Workshop Proceedings (LWA), September 2007, pp. 109–112. Martin-Luther-University Halle-Wittenber (2007)
3. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems 30(1-7), 107–117 (1998)
4. Budanitsky, A., Hirst, G.: Evaluating WordNet-based measures of lexical semantic relatedness. Computational Linguistics 32(1), 13–47 (2006)
5. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic grounding of tag relatedness in social bookmarking systems. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 615–631. Springer, Heidelberg (2008)
6. Cattuto, C., Loreto, V., Pietronero, L.: Collaborative tagging and semiotic dynamics, arXiv:cs.CY/0605015 (May 2006)
7. Cattuto, C., Schmitz, C., Baldassarri, A., Servedio, V.D.P., Loreto, V., Hotho, A., Grahl, M., Stumme, G.: Network properties of folksonomies. AI Communications 20(4), 245–262 (2007)
8. Chandler, D.: Semiotics: The Basics, 2nd edn. Taylor & Francis, Abington (2007)
9. Cimiano, P., Hotho, A., Staab, S.: Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. Journal of Artificial Intelligence Research (JAIR) 24, 305–339 (2005)
10. de Saussure, F.: Course in General Linguistics. Duckworth, London [1916] (1983) (trans. Roy Harris)
11. Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P., Tomkins, A.: Visualizing tags over time. In: Proceedings of the 15th International WWW Conference (May 2006)
12. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery: An overview. In: Advances in Knowledge Discovery and Data Mining, pp. 1–34. MIT Press, Cambridge (1996)
13. Fellbaum, C. (ed.): WordNet: an electronic lexical database. MIT Press, Cambridge (1998)
14. Firth, J.R.: A synopsis of linguistic theory 1930-55. Studies in Linguistic Analysis (special volume of the Philological Society) 1952-59, 1–32 (1957)
15. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Heidelberg (1999)
16. Golder, S., Huberman, B.A.: The structure of collaborative tagging systems. Journal of Information Science 32(2), 198–208 (2006)
17. Gruber, T.R.: Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In: Guarino, N., Poli, R. (eds.) Formal Ontology in Conceptual Analysis and Knowledge Representation, Deventer, Netherlands. Kluwer, Dordrecht (1993)
18. Halpin, H., Robu, V., Shepard, H.: The dynamics and semantics of collaborative tagging. In: Proceedings of the 1st Semantic Authoring and Annotation Workshop (SAAW 2006), vol. 209. CEUR-WS (2006)

19. Hammond, T., Hannay, T., Lund, B., Scott, J.: Social Bookmarking Tools (I): A General Review. D-Lib Magazine 11(4) (April 2005)
20. Harris, Z.S.: Mathematical Structures of Language. Wiley, New York (1968)
21. Heymann, P., Koutrika, G., Garcia-Molina, H.: Fighting spam on social web sites: A survey of approaches and future challenges. IEEE Internet Computing 11(6), 36–45 (2007)
22. Hotho, A.: Social bookmarking. In: Back, A., Gronau, N., Tochtermann, K. (eds.) Web 2.0 in der Unternehmenspraxis: Grundlagen, Fallstudien und Trends zum Einsatz von Social Software, pp. 26–38. Oldenbourg Verlag, München (2008)
23. Hotho, A., Benz, D., Jäschke, R., Krause, B., (eds.): ECML PKDD Discovery Challenge 2008 (RSDC 2008). Workshop at 18th Europ. Conf. on Machine Learning (ECML 2008) / 11th Europ. Conf. on Principles and Practice of Knowledge Discovery in Databases, PKDD 2008 (2008)
24. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: BibSonomy: A social bookmark and publication sharing system. In: Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures, pp. 87–102. Aalborg University Press, Aalborg (2006)
25. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 411–426. Springer, Heidelberg (2006)
26. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Trend detection in folksonomies. In: Avrithis, Y., Kompatsiaris, Y., Staab, S., O'Connor, N.E. (eds.) SAMT 2006. LNCS, vol. 4306, pp. 56–70. Springer, Heidelberg (2006)
27. Illig, J.: Machine learnability analysis of textclassifications in a social bookmarking folksonomy. Bachelor thesis, University of Kassel, Supervisor: Andreas Hotho, Kassel (2008)
28. Illig, J., Hotho, A., Jäschke, R., Stumme, G.: A comparison of content-based tag recommendations in folksonomy systems. In: Postproceedings of the International Conference on Knowledge Processing in Practice (KPP 2007). Springer, Heidelberg (2009) (to appear)
29. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. CoRR, cmp-lg/9709008 (1997)
30. Jäschke, R., Hotho, A., Schmitz, C., Ganter, B., Stumme, G.: Discovering shared conceptualizations in folksonomies. Web Semantics: Science, Services and Agents on the World Wide Web 6(1), 38–53 (2008)
31. Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., Stumme, G.: Tag recommendations in social bookmarking systems. AI Communications 21(4), 231–247 (2008)
32. Jäschke, R., Marinho, L.B., Hotho, A., Schmidt-Thieme, L., Stumme, G.: Tag recommendations in folksonomies. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 506–514. Springer, Heidelberg (2007)
33. Kosala, R., Blockeel, H.: Web mining research: A survey. SIGKDD Explorations 2(1), 1–15 (2000)
34. Krause, B., Schmitz, C., Hotho, A., Stumme, G.: The anti-social tagger - detecting spam in social bookmarking systems. In: Proc. of the Fourth International Workshop on Adversarial Information Retrieval on the Web, pp. 61–68. ACM, New York (2008)

35. Lehmann, F., Wille, R.: A triadic approach to formal concept analysis. In: Ellis, G., Rich, W., Levinson, R., Sowa, J.F. (eds.) ICCS 1995. LNCS, vol. 954, pp. 32–43. Springer, Heidelberg (1995)
36. Lund, B., Hammond, T., Flack, M., Hannay, T.: Social Bookmarking Tools (II): A Case Study - Connotea. D-Lib Magazine 11(4) (April 2005)
37. Mathes, A.: Folksonomies – Cooperative Classification and Communication Through Shared Metadata (December 2004), `http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html`
38. Mika, P.: Ontologies Are Us: A Unified Model of Social Networks and Semantics. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 522–536. Springer, Heidelberg (2005)
39. Patashnik, O.: BibTeXing (Included in the BIBTEX distribution) (1988)
40. Salton, G.: Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley Longman Publishing Co. Inc., Boston (1989)
41. Schmitz, C., Hotho, A., Jäschke, R., Stumme, G.: Mining association rules in folksonomies. In: Batagelj, V., Bock, H.-H., Ferligoj, A., Ziberna, A. (eds.) Data Science and Classification (Proc. IFCS 2006 Conference) Studies in Classification, Data Analysis, and Knowledge Organization, pp. 261–270. Springer, Heidelberg (2006)
42. Staab, S., Santini, S., Nack, F., Steels, L., Maedche, A.: Emergent semantics. Intelligent Systems, IEEE [see also IEEE Expert] 17(1), 78–86 (2002)
43. Staab, S., Studer, R. (eds.): Handbook on Ontologies. International Handbooks on Information Systems. Springer, Heidelberg (2004)
44. Steels, L.: The origins of ontologies and communication conventions in multiagent systems. Autonomous Agents and Multi-Agent Systems 1(2), 169–194 (1998)
45. Stützer, S.: Lernen von Ontologien aus kollaborativen Tagging-Systemen. Master thesis, University of Kassel, Supervisor: Andreas Hotho, Kassel (2009)
46. Stumme, G.: A finite state model for on-line analytical processing in triadic contexts. In: Ganter, B., Godin, R. (eds.) ICFCA 2005. LNCS (LNAI), vol. 3403, pp. 315–328. Springer, Heidelberg (2005)
47. Stumme, G., Hotho, A., Berendt, B.: Semantic web mining - state of the art and future directions. Journal of Web Semantics 4(2), 124–143 (2006)
48. Tonkin, E., Guy, M.: Folksonomies: Tidying up tags? D-Lib 12(1) (2006)
49. Wetzker, R., Umbrath, W., Said, A.: A hybrid approach to item recommendation in folksonomies. In: ESAIR 2009: Proceedings of the WSDM 2009 Workshop on Exploiting Semantic Annotations in Information Retrieval, pp. 25–29. ACM, New York (2009)
50. Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In: Rival, I. (ed.) Ordered sets, pp. 445–470, Reidel (1982)