A Case Study on Emergent Semantics in Communities[⋆]

Elke Michlmayr

Women's Postgraduate College for Internet Technologies (WIT),
Institute for Software Technology and Interactive Systems
Vienna University of Technology, Austria
michlmayr@wit.tuwien.ac.at

Abstract. This paper delivers a case study on the properties of meta-data provided by a folksonomy. We provide the background about folksonomies and discuss to which extend the process of creating meta-data in a folksonomy is related to the idea of emergent semantics as defined by the IFIP 2.6 Working Group on Data Semantics. We conduct experiments to analyse the meta-data provided by the *del.icio.us* folksonomy and to develop a method for selecting subsets of meta-data that adhere to the principle of interest-based locality, which was originally observed in peer-to-peer environments. In addition, we compare data provided by *del.icio.us* to data provided by the DMOZ taxonomy.

Keywords: Emergent Semantics in Communities, Folksonomies, Online harvesting of Semantic Network Information
Tags: folksonomies p2p computer science

1 Introduction

Recently, lots of discussions were raised by the advent of a new user-centric approach to categorization called folksonomies. Most of the debate is focused on the relationship between folksonomies and other approaches to categorization, such as taxonomies or ontologies. Folksonomies are comprised of a large amount of publicly available meta-data about lots of items, e.g., bookmarks or images, which can be retrieved from a central server. These meta-data are created by the users of the system without any restrictions posed by the system. Hence, the meta-data are inconsistent by nature, but the system tolerates these inconsistencies and exploits them for computing similarities between the keywords used for annotation. Our interest in folksonomies arises from being in need of simulation data for peer-to-peer applications. The contribution of this work is twofold: First, we deliver an in-depth study of the properties of meta-data produced by

[⋆] This research has partly been funded by the Austrian Federal Ministry for Education, Science, and Culture, and the European Social Fund (ESF) under grant 31.963/46-VII/9/2002.

folksonomies. Second, we investigate how meta-data produced by folksonomies can serve as simulation data for peer-to-peer environments.

The paper is organized as follows. In Section 2 we provide the necessary background about folksonomies and discuss their strengths and weaknesses. In Section 3 we compare the behaviour of the participants in a folksonomy to that of peers in a peer-to-peer network, and we examine the relationship between emergent semantics and folksonomies. In Section 4 we describe the experiments conducted on the provided meta-data in order to analyse its properties, and we report on the results of these experiments. Finally, in Section 5, we sum up our findings.

2 Folksonomies and their characteristics

The term *folksonomies* refers to a class of multi-user applications that provide a simple categorization system. This system is used to organize items, e.g., bookmarks or images. Instead of managing them within the browser application or on the local hard disk, the items are sent to a central server and stored there, together with meta-data authored by the user. These meta-data are comprised of one or more keywords — so-called *tags* — which describe the item. The keywords can be chosen freely by the user. Unlike in other categorization systems, there is no controlled vocabulary that defines which terms can be used as keywords in the categorization process. Another difference to existing categorization systems is that all keywords lie within the same namespace. There is no intention and hence no possibility to build hierarchical relationships between different keywords. The system uses a very simple data model which is depicted in Figure 1.

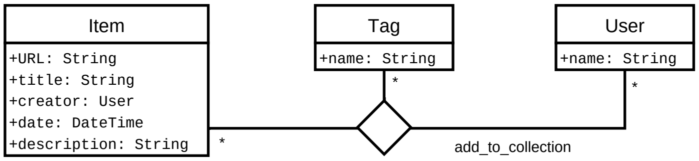

Fig. 1. Simple data model

The service provides each participant with his or her own Web page that shows the participant's item collection which contains all items together with the corresponding tags. The items are sorted in chronological order, showing the latest entry first. It is possible to filter the list of items by tag names to show only items that are annotated with a certain tag A. It is also possible to use another tag B to filter this list and retrieve all items that were annotated with both A and B.

These services are of value to the individual for managing items, but more important is that all participants of a folksonomy cooperate by allowing public access to their item collection and the associated meta-data. Public availability has two advantages. First, all Web users have access to the annotated item collections. Second, since all meta-data are stored at one single server, it is possible to analyse and aggregate it without having additional communication costs in terms of bandwidth. Aggregations are performed for both items and tags. Information about items can be aggregated since the participants individually and independently create meta-data about the same items. For each item, there exists a Web page where all tags that an item was annotated with are shown, together with the total number of participants that used that tag to annotate the item (see Figure 2). In addition, the service lists all participants that included the item in their collection and shows the total number of these participants. The total number shows the popularity of a certain item. Aggregation for tags is possible because the participants use the same tags to annotate different items. For each tag, there exists a Web page with a list of all items at least once annotated with this tag. The result of these aggregations is a network of related concepts. At the data level, folksonomies are undirected weighted graphs that can be seen from different perspectives. When viewing items as the nodes of the network, two nodes are connected if they are annotated with the same tags. These kinds of connections are weighted. The more often the same tags were used, the higher the weight of the edge. When viewing the participants as the nodes of the network, the edges are built by those items that are shared between the participants. These graphs are exploited as input for an algorithm that computes the relatedness of tags. Each Web page containing aggregated information about a certain tag also shows the tag names of related tags as computed by the algorithm. For example, the algorithm used by the service *Flickr* [14] computes *tent*, *fire*, *hiking* as related tags for tag *camping*.

43	physics	15	mathematics	7	library	5	article	3	ai
41	science	10	articles	6	preprint	4	computer	3	study
27	research	10	journal	6	books	4	arxiv		
23	math	9	archive	6	programming	4	literature		
19	papers	8	biology	5	cs	4	toread		
18	reference	7	eprint	5	academic	4	computerscience		

Fig. 2. Tag distribution used to annotate a sample del.icio.us item

This approach to categorization is different to the top-down approach that is employed in traditional categorization systems, e.g. taxonomies or ontologies. Ontologies provide domain-specific vocabularies that describe the conceptual elements of a domain and the relationship between these elements, such as *is-a*-relationships or *part-of*-relationships. Creating an ontology requires careful analysis of what kind of objects and relations can exist in that domain [5]. This analysis is done by domain experts together with information architects who need to reach consensus about the exact meaning of objects and relations. After the ontology has been developed and put in place, the data items are categorized according to the chosen categorization scheme. The same top-down approach is used in the database world. Before the actual data comes in, a schema is built that for that data. The schema defines which entities exists and how they are related. A third example is object-oriented modelling, where an instance can not exist without being a member of a class. The model defines the hierarchy of objects and their relationships. In a folksonomy, each participant uses a certain tag with his or her personal meaning in mind. There is no controlled vocabulary or ontology that defines the meaning of tags. Everybody has the possibility to express his or her opinions about the categorization of a certain object. This freedom of choice induces all problems controlled vocabularies try to avoid. The participants can use either the singular or the plural form of a term. Hence, they create two different tags with exactly the same meaning. There is no synonym control; hence different terms that refer to the same concept are in use. A special problem are keywords that consist of two terms: some participants create one tag by combining the terms with underscores or hyphens, or by creating a compound word with no separating character in between, while others create two tags, one for each term. The bottom-up approach to categorization avoids the necessity to reach consensus about the most appropriate categorization of a certain object. Semantic reconciliation is performed by the magnitude of participants that added meta-data to the folksonomy. The tags that are used most often to annotate a certain item express the opinion of the majority. Thus, the utility of a folksonomy is directly proportional to its number of participants and the amount of meta-data produced by them.

In addition to the already mentioned shortcomings of folksonomies that are caused by not using controlled vocabularies, a major weakness lies in the user interface. While browsing and filtering items by tag names is supported very well, searching for a certain item by name is impossible. Another shortcoming of folksonomies is that they can easily be spammed. Malicious participants can abuse the system by adding items of their interests and by assigning lots of tags for these items. There are a number of existing services like *del.icio.us* [8] for bookmarks, *Flickr* [14] for images, *Connotea* [9] for references to scientific literature, and others. A review of the existing services is provided in [7]. Services for bookmarks are also called *social bookmarking services*. Those services provide convenient interfaces to their participants, e.g. the possibility to add an item using a special link containing JavaScript code — a so-called bookmarklet — that transfers the URL to be added to the server. While typing in keywords, the participant is shown a list of the keywords he or she already used before. If the bookmark is already stored in the system, a list of popular keywords for that bookmark is shown as well. Those two lists are facilities that assist the participants in choosing appropriate keywords. After storing the bookmark, he or she is automatically redirected to the newly added Web page. More information about the general ideas behind folksonomies can be found in [3] and [7]. In [6], the major differences between folksonomies and taxonomies are discussed and some statistical information about the tag distribution of *del.icio.us* [8] is presented.

A case study on the service *Connotea* that provides a folksonomy for sharing meta-data about scientific literature can be found in [9].

3 Folksonomies and peer-to-peer environments

In Section 3.1, we explain why folksonomies can be used for retrieving simulation data for peer-to-peer networks. In Section 3.2, we briefly introduce the idea of emergent semantics and its relationship to folksonomies.

3.1 User behavior

Folksonomies are centralized services that heavily rely on the aggregation of meta-data. These aggregations are possible because the meta-data reside on a single server. For example, it is easy to determine all participants who share a certain information item. In a peer-to-peer environment, there is no central server and all peers store their information items at the local hard disk. Aggregating data consumes network resources. Knowing which peers share a certain information item is a non-trivial task for which it is necessary to track how the items are replicated within the network [4].

Although the architectures of folksonomies (centralized) and peer-to-peer networks (distributed) are completely different, the important point is that the behaviour of participants in a folksonomy is comparable to the behaviour of peers in an unstructured peer-to-peer network. All participants act autonomously and there is no central authority coordinating them. All participants provide information items to others that can be browsed and retrieved. Since the meta-data produced by folksonomies are publicly available and can be easily retrieved from one central server, folksonomies are suitable for retrieving test data for peer-to-peer applications. The available data can be used for modelling peers and their content distribution. Folksonomies do not provide any data about queries and query distribution. As a by-product of this paper, the data gathered during the experiments described in Section 4 will be used as a test suite for an algorithm for query routing in peer-to-peer networks described in [10].

3.2 Do folksonomies provide emergent semantics?

The term *emergent semantics* was defined by Aberer et al. in [2]. In this work, the authors discuss semantic interoperability for loosely coupled information sources and observe that a-priori agreements on concepts, e.g., the use of ontologies, are not appropriate in ad-hoc situations, because there is no possibility for the communicating peers to anticipate all interpretations. Instead, a semantic handshake protocol is suggested that allows negotiations between pairs of peers to reach an agreement over the meaning of models, e.g., by local schema mapping [1]. In order to save network resources, these negotiations are local interactions whenever possible. Global agreements are obtained by aggregating local agreements.

Semantic interoperability is constructed incrementally by lots of negotiations which are influenced by the context of existing global agreements.

Letting aside the major differences that stem from the fact that folksonomies operate in a centralized environment, there are some ideas from emergent semantics that can be found in folksonomies. As suggested in [2], folksonomies are self-organized systems. Both approaches (1) do not force their users to commit themselves to an existing ontology, (2) rely on lots of small interactions as well as on (3) aggregation of the results of these interactions, and both (4) construct their global properties incrementally. Another main distinction is that while the interactions in emergent semantics are initiated in order to reach consensus, in a folksonomy there are only information-exchanging acts that to not lead to a definite agreement. In summary, folksonomies employ an approach that is similar to emergent semantics, but address a simplified problem because a centralized architecture exists and because all participants rely on the same schema.

4 Experiments and Results

In this section we report on the experiments conducted on data retrieved from a social bookmarking service. In Section 4.1, the experimental setup is described. In Section 4.2, we show a method for data selection and present statistics about the retrieved test data sets. In Section 4.3 we evaluate if it is possible to join test data sets. In Section 4.4 we compare two different sets of categorization data for the same items. Finally, in Section 4.5 we analyse the impact of a bookmark's popularity.

4.1 Experimental setup and test data

The test data used in the following experiments was gathered by downloading[1] selected bookmarks from *del.icio.us* [8], which is one of the most successful social bookmarking services having more than 55.000 users. For the implementation of the downloading routines, Perl scripts were used. We kept a list of all already retrieved URLs to prevent multiple downloading of the same information. In addition, error handling facilities were necessary since internal server errors of the service occurred frequently. In order not to take up too many resources from the service, we used a delay of five seconds between each subsequent request. All downloaded data was saved to text files with very simple formats. If the routine encountered a bookmark that was included in the bookmark collection of more than hundred participants, all tags and their distribution for the bookmark were additionally retrieved. Different test data was selected for each experiment. The test data suites are described in the following sections and available upon request.

4.2 Data selection

In the first experiment, we want to find a feasible method for selecting a subset of the provided data in which the principle of *interest-based locality* [12] can be

[1] The data was downloaded between June 21 and June 30, 2005.

Test set Nr.	1	2	3	4
Number of participants	551	155	248	280
Number of items	17575	5709	8861	10237
Number of unique items in % of all items	12855 73,14 %	4311 75,51 %	6045 68,22 %	6643 64,89 %
Number of popular items	5691	2393	3691	4207
Number of unique popular items in % of all unique items	2301 17,90 %	1217 28,23 %	1479 24,47 %	1483 22,23 %
Average number of items per user (Max: 50)	31,9	36,83	35,73	36,56
Average number of popular items per user	10,32	14,79	13,61	13,50

Table 1. Properties of the test sets

observed. This principle was originally observed in peer-to-peer environments. It means that if a participant A has a particular piece of content participant B is interested in, it is likely the case that the other information items stored by participant A are also of interest to participant B. As already discussed in Section 3, the user behaviour in a folksonomy is comparable to the user behaviour in peer-to-peer networks. Thus, we assume that interest-based locality can also be observed in folksonomies. Given the interfaces for data retrieval provided by *del.icio.us*, two different methods for data selection are possible:

- Select a certain tag and retrieve all bookmarks for this tag
- Select a certain bookmark and retrieve all participants that store this bookmark

Since we want to retrieve data from participants that form a community by sharing a common interest, and sharing the same bookmark is a stronger connection than sharing the same tag, we decided to choose the second option. First, a random bookmark b was chosen as a starting point. In the second step, the participant names of all participants that store b in their bookmark collection were retrieved. In the third step, the bookmark collections of all of these participants were downloaded. The fifty entries of each participant's bookmark collection which were added latest were included. For participants storing less than fifty entries, all existing entries were considered. Using the procedure described above, four test sets of different size were collected. The four random bookmarks[2] were chosen to be bookmarks that refer to Web sites containing information about diverse topics.

The properties of the test sets are described in Table 1. First of all, we can see that the total number of bookmarks of a test set is proportional to the number of participants that store bookmark b. Since the percentage of unique

[2] test set 1: b is `http://del.icio.us/url/06df5507a27ab5aa297fbb7748374df6`,
test set 2: b is `http://del.icio.us/url/463f3f6f9ce9471fef7f9edb881ad2d7`,
test set 3: b is `http://del.icio.us/url/245d7b2a49a80771da9a4d3a02d539c3`,
test set 4: b is `http://del.icio.us/url/c745432a483a84037c90e08d79f7c306`

All items shared ...	by > 10	by 5 - 10	by 4	by 3	by 2	not shared
Test set 1	0,41 %	1,84 %	1,34 %	2,76 %	9,09 %	84,56 %
Test set 2	0,16 %	1,83 %	1,37 %	2,85 %	9,21 %	84,57 %
Test set 3	0,58 %	2,45 %	1,70 %	2,96 %	8,68 %	83,62 %
Test set 4	0,90 %	2,60 %	1,55 %	3,10 %	8,63 %	83,22 %
Average	0,51 %	2,18 %	1,49 %	2,92 %	8,90 %	84,00 %

Popular items shared ...	by > 10	by 5 - 10	by 4	by 3	by 2	not shared
Test set 1	1,91 %	8,43 %	6,04 %	9,87 %	23,55 %	50,20 %
Test set 2	0,49 %	6,08 %	3,94 %	9,37 %	20,46 %	59,65 %
Test set 3	2,50 %	8,32 %	3,92 %	8,92 %	17,58 %	58,76 %
Test set 4	3,84 %	8,36 %	4,18 %	8,90 %	18,07 %	56,64 %
Average	2,19 %	7,80 %	4,52 %	9,27 %	19,92 %	56,30 %

Table 2. Distribution of bookmarks if (a) considering all bookmarks (top), or (b) considering popular bookmarks only (bottom)

bookmarks in each test set ranges from 64,89 % to 75,51 %, the number of unique bookmarks in a test set is not proportional to the total number of bookmarks. There are only small differences in the average number of bookmarks included in a participant's collection (Min: 31,9, Max: 36,83). The number of participants in a test set has a small impact on this number: In test set 1, which is by far the biggest of all sets, the average number of bookmarks per participant is higher than in the other test sets. Our decision to consider only the first fifty entries of each bookmark collection was based on the assumption that a high percentage of all *del.icio.us* participants stores more than fifty entries. We can observe from the retrieved data that this is not the case. On average, only 1,15 percent of participants in each set own a collection which is comprised of fifty bookmarks (and probably more that we did not retrieve). A bookmark is defined to be a popular bookmark if it is included in the bookmark collection of more than hundred *del.icio.us* participants. On average, a third of all bookmarks fall in the category of popular bookmarks. As can be seen in test set 1 and test set 2, the smaller the test set in terms of number of participants, the higher the amount of popular bookmarks per user.

Next, we analyse the distribution of bookmarks in the test sets. All unique bookmarks of a test set were considered. The results of this analysis are shown in Table 2. For each bookmark, we determined the total number of participants that store it in their collection. It turns out that the distributions of unique bookmarks share equal properties in each test set. On average, only 0,51 % of the bookmarks are stored by more than ten participants. 2,18 % are stored by a group of five to ten participants. 1,49 % are stored by four participants. 2,92 % are stored by three participants. 8,90 % are stored by 2 participants. The percentage of bookmarks that are not shared, but stored in only one participant's collection, is nearly equal in each test and on average 84 %. This is a very high value that shows that our data retrieval method as described above is not sufficient for selecting subsets of the *del.icio.us* data which conform to the principle of interest-based locality. Hence, we want to know if interest-based locality can be observed when considering only the popular bookmarks. As can be seen in the bottom of Table 2, in this case the percentage of bookmarks that are present in only one collection lowers to 56,3 %. On average, 19,92 % percent of all popular bookmarks are shared by two participants, 9,27 % are shared by three participants, 4,52 % by four participants, 7,8 % are shared by between five and ten participants, and 2,19 % by more than ten participants.

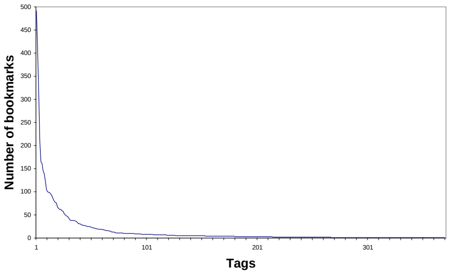

Fig. 3. Top tag distribution (ranked plot, linear scale) in set 1

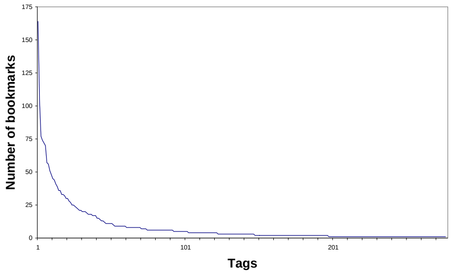

Fig. 4. Top tag distribution (ranked plot, linear scale) in set 2

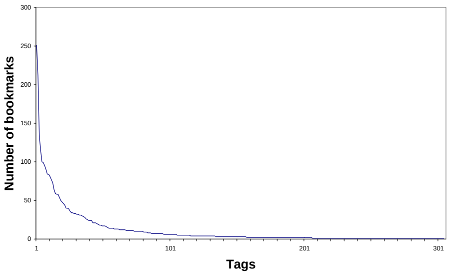

Fig. 5. Top tag distribution (ranked plot, linear scale)in set 3

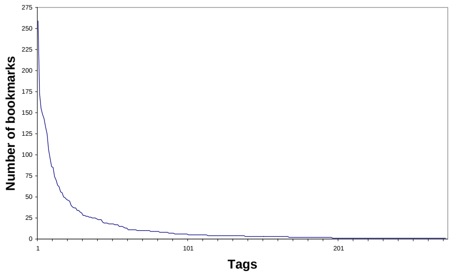

Fig. 6. Top tag distribution (ranked plot, linear scale) in set 4

Assuming that bookmarks that refer to Web sites with similar topics are annotated with the same tags, we now consider the tags associated with the popular bookmarks. For each bookmark, the top tag that was used by most participants was considered. The distributions of the top tags for all popular bookmarks are shown in Figure 3 to Figure 6. As can be seen from these figures, the distribution curves for all four test sets show equal properties. There is a long tail in each curve that reveals that there are many top tags that are included only once. The reason for that lies in the diversity of bookmark collections. Too many

Item present in ...	one set	two sets	Item present in ...	one set	two sets	three sets
1 and 2	92,28 %	7,72 %	1,2, and 3	86,27 %	10,12 %	3,61 %
1 and 3	89,71 %	10,29 %	1,2, and 4	86,96 %	9,53 %	3,51 %
1 and 4	90,55 %	9,45 %	1,3, and 3	84,32 %	11,58 %	4,10 %
2 and 3	87,47 %	12,53 %	2,3, and 4	82,03 %	12,50 %	5,48 %
2 and 4	87,88 %	12,12 %				
3 and 4	85,09 %	14,91 %				

Table 3. Bookmarks present in more than one test set when comparing (a) two test sets (left), or (b) three test sets (right)

of them contain items about topics that are not related to the topics of the other items of the collection. Hence, even when considering popular bookmarks only, it is not possible to retrieve test sets that conform to the principle of interest-based locality without any further preparation of the data. The necessary preparations consist of removing all items that cause the tail of the top tag distribution.

4.3 Joining data

In this analysis, we want to find out if it is possible to join test sets in order to create one bigger test set out of them. The question is if there are any connections between the data and to which extend the test sets are overlapping. Two kinds of overlaps are possible. The first is the overlap of participants, where one participant is present in more than one of the data subsets. Analysing the distribution of participants, the four test sets are nearly disjoint. The majority of participants (96.04 %) is included in only one test set. 3.96 % of participants were included in two sets. No participant was included in three or all four test sets. The second possibility for overlaps is that bookmarks can be present in more than one set. Table 3 shows the results of comparing the test sets to each other. When comparing two test sets, on average 11,17 % are included in both sets. When comparing three test sets, on average 4,18 % are in all three sets and 10,93 % in two sets. When joining all four sets, 82,33 % of bookmarks are present in one set, 11,72 % in two sets, 3,63 % in three, and 2,32 % are present in all four sets. Hence, it is not possible to use the method described in Section 4.2 to retrieve data and to join these sets to create bigger sets. The overlap between the sets is too small.

4.4 Data semantics

In this experiment, we compare the meta-data provided by the *del.icio.us* folksonomy to an already existing annotated bookmark collection built by the *DMOZ Project* [11]. This project is an effort of a community of volunteers to build a taxonomy for Web pages and to categorize Web pages according to this taxonomy. Since the DMOZ project that has already been used for simulating user

```
URL http://arxiv.org/
DMOZ Top/Science/Physics/Publications
DMOZ Top/Science/Math/Publications
DMOZ Top/Science/Math/Publications/Online_Texts/Collections
DMOZ Top/Science/Publications/Archives/Free_Access_Online_Archives
ID 19aa8ff1e9e2a06677ab34f3f2a5b0c8
TITLE arXiv.org e-Print archive
TAGS physics:43;science:41;research:27;math:23;papers:19;reference:18;ma
thematics:15;journal:10;articles:10;archive:9;biology:8;eprint:7;library
:7;preprint:6;books:6;programming:6;cs:5;article:5;academic:5;computer:4
;arxiv:4;literature:4;toread:4;computerscience:4;ai:3;study:3;
```

Fig. 7. A sample entry containing meta-data from both sources

behaviour in a peer-to-peer network [13], it is interesting for us to know to which extend the meta-data provided by the DMOZ project is similar to the meta-data provided by *del.icio.us*. For conducting this experiment, we downloaded the RDF dump[3] of the structure and of the contents of the DMOZ directory and stored it in a relational database for performance reasons. After that, a database lookup for each popular bookmark included in the test sets described in Section 4.2 was performed to check if the bookmark is included in the DMOZ contents as well. Each time the lookup routine encountered a hit, the bookmark and its meta-data from both sources were appended to a text file with a very simple format (see Figure 7 for an example). If a bookmark was assigned more than one DMOZ topic, we considered all of them. Two observations we made while performing this task are worth mentioning:

- The intersection of *del.icio.us* and the DMOZ directory is rather small. Although only popular bookmarks were used, only 25 % of the bookmarks were also included in the contents of the DMOZ directory.
- Nearly 50 % of those bookmarks that are present both in both sources are instances of subtopics of the DMOZ topic `Top/Computers`.

In total, the test data for this experiment consists of 788 bookmarks together with all corresponding DMOZ topics and all tags and their numbers from *del.icio.us*. All DMOZ topics were considered except of the subtopics of topic `Top/World`, which is the branch in the DMOZ hierarchy that builds the top concept for multi-lingual categories not defined in the English language. All DMOZ topic names were converted to lower-case characters. Underscores and hyphens were removed from both topic and tag names. To overcome the problem that singular and plural versions of tags are in use, in case the last character of a tag or topic name was the letter `s`, it was removed (e.g., `computers` was changed to `computer`). Some DMOZ topics have 26 subtopics for each letter from A to Z in order to categorize items by the first letter of their name. Such topics consisting of only one character were removed from the topic path. The topic `Top` was removed from each topic path. Since the leaf entry from the DMOZ topic path is the one that most exactly categorizes a bookmark, we sorted the topic paths to

[3] available at `http://rdf.dmoz.org`

	1st	2nd	3rd	4th	5th	6th	7th to 11th
Top tag	9,44 %	15,94 %	12,67 %	4,72 %	3,28 %	1,72 %	0,81 %
Top 3 tags	20,37 %	27,55 %	21,58 %	14,29 %	12,23 %	6,21 %	2,30 %
Top 5 tags	28,32 %	34,81 %	27,72 %	19,75 %	16,42 %	11,03 %	3,69 %
Top 10 tags	37,38 %	44,53 %	35,94 %	27,08 %	25,91 %	18,28 %	6,25 %
Top 15 tags	44,30 %	52,45 %	43,17 %	34,16 %	32,12 %	26,55 %	8,93 %
All tags	52,99 %	62,55 %	52,48 %	46,34 %	44,34 %	40,34 %	14,73 %

Table 4. Comparison of categorization data from both sources, considering 1, 3, 5, 10, 15, or all tags for a given bookmark.

their reverse order. For example, the topic path shown in Figure 7 is converted to `publications physics science`.

On average, the topic path length of all DMOZ topics prepared as described above is 4,67. The average number of tags per bookmark is 24,59. The following method is employed for comparing topics to tags. Topics are used as a reference and tags are compared to them. If more than one topic is assigned to a bookmark, a separate comparison for each topic is performed. One comparison consists of several lookups, one for each entry of a topic path, e.g., for `publications physics science` three lookups are performed. The result of a lookup is either true in case of a match, or false. The results of this comparison are shown in Table 4. It turns out that the leaf entries of the topic path match more often than the top entries. This is not surprising, since the top entries are very general, e.g., `Computers`, `Arts`, or `Science` and only a few *del.icio.us* participants will use general terms to describe their items. When taking into account only the top tag, which is the one that most participants used for annotating, the highest percentage of matches is 15,94 % for the second entry of each topic path. It can be seen that the values rise linearly. The more tags are taken into account, the better the results. The fairest comparison is that of the top five tags, since this is the average number of the topic path length. In this case, the highest percentage of matches is 34,81 % for the second entry of each topic path.

In summary, it can be seen that the terms used for categorization are very different in both sources. Even when comparing all tags to each topic path entry and hence conducting on average 24 comparisons of tags to one single topic, there is no match in 37,45 % to 85,27 % of the cases.

4.5 Popularity of items

In the last experiment, our assumption is that users are more interested in a certain bookmark if it is already included in many other bookmark collections than if it is included in only a few. Hence, bookmarks that are already popular will become even more popular. In particular, we want to know if the list of popular bookmarks[4] which shows those bookmarks that most users added to

[4] available at `http://del.icio.us/popular/`

their collections recently has an impact on the popularity of a bookmark. We observed this list several times for the time span of a day and collected a snapshot every 10 minutes. These snapshots are comprised all listed bookmark's URLs and the total number of persons that included it in their bookmark collection at the given point in time. For clarity, Figure 8 and Figure 9 show only those bookmarks that were present in the list of popular bookmarks for the complete time of observation. Analysing these data, one can see in Figure 8 that the assumption is true to some extend, since those curves that are higher increase a little faster than those that are low, but the differences are not significant as we expected.

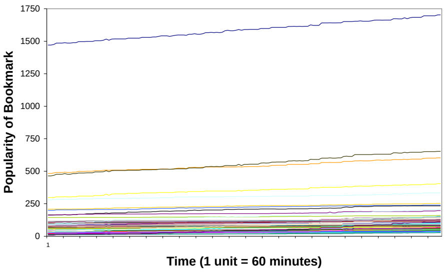

Fig. 8. Popular bookmarks on Tuesday, 28th of June

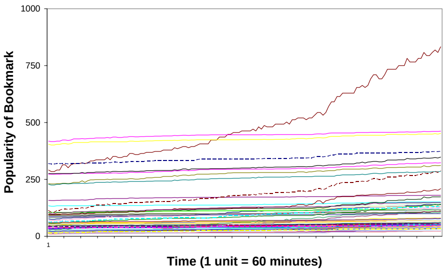

Fig. 9. Popular bookmarks on Thursday, 23th of June

Figure 9 shows that there are other factors that have more impact on the popularity of bookmarks over time than being included in the list of popular bookmarks. There is one particular bookmark with its popularity rising very quickly while all other bookmarks show the same performance as in Figure 8. Hence, the reason for this significant increase is not caused by being included the list of popular bookmarks.

5 Conclusion

In this paper, a method for selecting subsets of the meta-data provided by a folksonomy that adhere to the principle of interest-based locality was developed. The resulting data can be applied for simulating peers and their contents in a peer-to-peer environment. The properties of the test sets that were retrieved by using this method were analysed and discussed in order to prove that the proposed method selects subsets that have similar properties. Comparing the meta-data produced by the folksonomy to meta-data created by the DMOZ open directory project at the data level revealed that there are major differences between them. Finally, we showed that centrally provided lists of popular items have only small influences on the properties of these items.

Acknowledgements

We thank Marion Murzek and Gerhard Kramler for helpful comments on earlier versions of this paper.

References

1. K. Aberer, P. Cudre-Mauroux, and M. Hauswirth. The Chatty Web: Emergent Semantics Through Gossiping. In *Procceedings of the 12th International World Wide Web Conference (WWW 2003)*, May 2003.
2. K. Aberer, P. Cudre-Mauroux, A. M. Ouksel, T. Catarci, M.-S. Hacid, A. Illarramendi, V. Kashyap, M. Mecella, E. Mena, E. J. Neuhold, O. D. Troyer, T. Risse, M. Scannapieco, F. Saltor, L. de Santis, S. Spaccapietra, S. Staab, and R. Studer. Emergent Semantics Principles and Issues. In *Procceedings of the 9th International Conference on Database Systems for Advanced Applications (DASFAA 2004)*, March 2004.
3. Adam Mathes. Folksonomies – Cooperative Classification and Communication Through Shared Metadata. `http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html`, December 2004.
4. M. Bernauer, G. Kappel, and E. Michlmayr. Traceable Document Flows. In *Proceedings of the 2nd International Workshop on Web Semantics (WebS), DEXA2004*, September 2004.
5. B. Chandrasekaran, J. R. Josephson, and R. Benjamins. What are Ontologies, and Why Do We Need Them? *IEEE Intelligent Systems and Their Applications*, 14(1):20–26, 1999.
6. Clay Shirky. Ontology is Overrated: Categories, Links, and Tags. `http://shirky.com/writings/ontology_overrated.html`, 2005.
7. T. Hammond, T. Hannay, B. Lund, and J. Scott. Social Bookmarking Tools (I): A General Review. *D-Lib Magazine*, 11(5), April 2005.
8. Joshua Schachter. Del.icio.us Social Bookmarking Service. `http://del.icio.us/`, 2005.
9. B. Lund, T. Hammond, M. Flack, and T. Hannay. Social Bookmarking Tools (II): A Case Study – Connotea. *D-Lib Magazine*, 11(4), April 2005.
10. E. Michlmayr, S. Graf, W. Siberski, and W. Nejdl. Query Routing with Ants. In *Proceedings of the 1st Workshop on Ontologies in P2P Communities, ESWC2005*, May 2005.
11. Netscape Communications Corporation. DMOZ Open Directory Project. `http://www.dmoz.org`, 2005.
12. K. Sripanidkulchai, B. Maggs, and H. Zhang. Efficient Content Location Using Interest-Based Locality in Peer-to-Peer Systems. In *Proceedings of IEEE INFOCOM 2003*, April 2003.
13. C. Tempich, S. Staab, and A. Wranik. REMINDIN': Semantic Query Routing in Peer-to-Peer Networks based on Social Metaphors. In *Proceedings of the 13nd International World Wide Web Conference (WWW2004)*, May 2004.
14. Yahoo! Company. Flickr Online Photo Management Service. `http://flickr.com/`, 2005.