Stop Thinking, Start Tagging: Tag Semantics Emerge from Collaborative Verbosity

Christian Körner[*]
Knowledge Management Institute
Graz University of Technology
Graz, Austria
christian.koerner@tugraz.at

Dominik Benz[*]
Knowledge & Data Engineering Group
University of Kassel
Kassel, Germany
benz@cs.uni-kassel.de

Andreas Hotho
Data Mining and Information Retrieval Group
University of Würzburg
Würzburg, Germany
hotho@informatik.uni-wuerzburg.de

Markus Strohmaier
Knowledge Management Institute and Know-Center
Graz University of Technology
Graz, Austria
markus.strohmaier@tugraz.at

Gerd Stumme
Knowledge & Data Engineering Group
University of Kassel
Kassel, Germany
stumme@cs.uni-kassel.de

ABSTRACT

Recent research provides evidence for the presence of emergent semantics in collaborative tagging systems. While several methods have been proposed, little is known about the factors that influence the evolution of semantic structures in these systems. A natural hypothesis is that the quality of the emergent semantics depends on the pragmatics of tagging: Users with certain usage patterns might contribute more to the resulting semantics than others. In this work, we propose several measures which enable a *pragmatic* differentiation of taggers by their degree of contribution to emerging semantic structures. We distinguish between *categorizers*, who typically use a small set of tags as a replacement for hierarchical classification schemes, and *describers*, who are annotating resources with a wealth of freely associated, descriptive keywords. To study our hypothesis, we apply semantic similarity measures to 64 different partitions of a real-world and large-scale folksonomy containing different ratios of categorizers and describers. Our results not only show that 'verbose' taggers are most useful for the emergence of tag semantics, but also that a subset containing only 40 % of the most 'verbose' taggers can produce results that match and even outperform the semantic precision obtained from the whole dataset. Moreover, the results suggest that there exists a causal link between the pragmatics of tagging and resulting emergent semantics. This work is relevant for designers and analysts of tagging systems interested (i) in fostering the semantic development of their platforms, (ii) in identifying users introducing "semantic noise", and (iii) in learning ontologies.

Categories and Subject Descriptors: H.1.2 [Information Systems]: Models and Principles [Human information processing] H.1.m [Information Systems]: Models and Principles H.3.5 [Information Storage and Retrieval]: Online Information Services[Web-based services] H.5.3 [Information Interfaces and Presentation]:Group and Organization Interfaces[Collaborative computing, Web-based interaction]

General Terms: Algorithms, Human Factors

Keywords: folksonomies, tagging, user characteristics, semantics, pragmatics

1. INTRODUCTION

Folksonomies are the core data structure of collaborative tagging systems. They are large-scale bodies of lightweight annotations provided by their user communities. Clearly, every user is following his own terminology and is only willing to a very small extent (if at all) to follow any naming conventions. Nevertheless, there is evidence for the presence of emergent semantics in such collaborative tagging systems, mainly based on tags and the folksonomical relationships between them [8, 39]. While several methods have achieved promising results for capturing emergent semantics in folksonomies (e. g., [7, 26, 36, 33, 19]), little is known about the factors that influence the evolution of semantics in these systems.

A natural hypothesis is that emergent *semantics* in folksonomies are influenced by the *pragmatics* of tagging, i. e., the tagging practices of individuals: users with certain usage patterns (cf. [14]) might contribute more to the resulting semantics than others. For example: one may assume that users who follow an 'ontology-engineering style' of tagging — i. e., users who try to maintain a "clean vocabulary" with no redundancy – contribute more to the structure of a folksonomy, which is blurred by other users who are not following this approach. However, we will show in this paper that this is *not* the case.

To this end, we will distinguish between two types of users in a folksonomy, called *categorizers* and *describers*, following the approach in [34]. Categorizers typically use a well-defined set of tags as a replacement for hierarchical classification schemes, while describers are annotating resources with a wealth of freely associated, descriptive keywords. We use a number of measures focused on capturing tagging pragmatics and approximating the membership of a user to either of the two types. These *pragmatic* measures will be used to partition a tagging dataset into subsets on which we

[*]Both authors contributed equally to this work.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

apply *semantic* measures [7] in order to study potential effects of tagging pragmatics on tag semantics.

Our results not only show that particular users contribute more to emerging semantics than others, but also that the "collaborative verbosity" of a fraction of *describers* can achieve and even outperform semantic precision levels obtained from the entire dataset. In summary, our results suggest that a key factor for users to be effective contributors to aggregated semantic structures is their tagging verbosity. In addition, our work provides first empirical evidence that the emergent semantics of tags in folksonomies are influenced by the pragmatics of tagging, i. e., the tagging practices of individual users.

The results of this work are relevant for researchers who want to analyze folksonomies for ontology learning purposes. For example, users who introduce "semantic noise" and hinder the semantic evolution can be identified and excluded from the data based on pragmatic measures that capture individual tagging styles of users. The proposed methods can also be used to improve and inform the design of ontology learning algorithms.

The paper is organized as follows: In section 2 we provide an overview about folksonomies and their emergent tag semantics. Section 3 deals with measures aimed at capturing different aspects of tagging pragmatics. This is followed by section 4 covering the semantic implications of tagging pragmatics in which we describe the conducted experiments and present a discussion of our results. Subsequently we give an overview of the related work (section 5). We discuss our results in the context of ontology learning and related tasks in section 6, where we also point to future work.

2. EMERGENT TAG SEMANTICS

Since the advent of folksonomies as a part of the "Web 2.0" paradigm, large corpora of human-annotated content have attracted the interest of researchers from different disciplines. In particular, there has been the early idea to study the semantics of folksonomies, e. g., work by Mika [30] or Golder and Huberman [14]. Later, more and more approaches arose to "harvest" the semantics of a folksonomy (see the section on related work for details). In many of these approaches, distributional measures were used to infer semantic relations among tags. However, in most cases the choice of these measures was done on a rather ad-hoc basis without a deeper knowledge of the semantic characteristics of each measure. A first systematic analysis which *kind* of semantic relations are returned by different measures was done by us in [7, 26]. The semantic grounding procedure presented there confirms the assumption that distributional tag relatedness measures are an appropriate means to capture the emerging semantic structures between tags in folksonomies. As our presented analysis makes strongly use of this work, we recall it here in greater detail.

2.1 Folksonomy model

In the following we will use the definition of folksonomy provided in [21]:

Definition A *folksonomy* is a tuple $\mathbb{F} := (U, T, R, Y)$ where U, T, and R are finite sets, whose elements are called *users*, *tags* and *resources*, respectively. Y is a ternary relation between them, i. e., $Y \subseteq U \times T \times R$. The elements $y \in Y$ are called *tag assignments* (TAS). A *post* is a triple (u, T_{ur}, r) with $u \in U$, $r \in R$, and a non-empty set $T_{ur} := \{t \in T \mid (u, t, r) \in Y\}$.

Furthermore, we denote the (tag) *vocabulary* of a user as $T_u := \{t \in T \mid \exists r : (u, t, r) \in Y\}$. This represents the set of distinct tags a user has used at least once. Analogously we define $R_u := \{r \in R \mid \exists t : (u, t, r) \in Y\}$ as the set of resources a given user has tagged.

2.2 Semantic grounding of tag relatedness measures

As stated above, the notion of tag relatedness is a crucial aspect of emerging semantics in folksonomies. One way of defining it is to map the tags to a thesaurus or lexicon like Roget's thesaurus[1] or WordNet [12],[2] and to measure relatedness by means of existing semantic measures. Another option is to define measures of relatedness directly on the network structure of the folksonomy. A reason why distributional measures in folksonomies are used in addition to mapping tags to a thesaurus is the observation that the vocabulary of folksonomies often includes community-specific terms that are not included in lexical resources.

In our previous work [7] we identified several possibilities to measure tag relatedness directly in a folksonomy. Most of them use statistical information about different types of *co-occurrence* between tags, resources and users. Other approaches adopt the *distributional hypothesis* [13, 17], which states that words found in similar contexts tend to be semantically similar.

More specifically we have analyzed five measures for the relatedness of tags: the *co-occurrence count*, three context measures which capture distributional information by computing the *cosine similarity* [32] in the vector spaces spanned by users, tags, and resources, and *FolkRank* [21], a graph-based measure that is an adaptation of PageRank to folksonomies.

We observed in our experiments in [7] that the tag and resource context measures performed best, by comparing them to thesaurus-based measures based on WordNet. This indicates that the distributional hypothesis [13, 17] does not only influence the human judgment of semantic similarity [29], but also folksonomy-based distributional measures. To provide a semantic grounding of our folksonomy-based measures, we mapped the tags of a large-scale del.icio.us dataset to synsets of WordNet and used the semantic relations of WordNet to infer corresponding semantic relations in the folksonomy. In WordNet, we measured the similarity by using a similarity measure (JCN from here on) by Jiang and Conrath [23] that has been validated in previous user studies and applications [5].

We discovered that the context measure based on cosine similarity in a vector space that is spanned by the tags yielded an almost optimal performance at an acceptable level of computational complexity. This distributional measure is defined as follows.

The Tag Context Similarity (*TagCont*) is computed in the vector space $\mathbb{R}^T$, where, for tag t, the entries of the vector $\vec{v}_t \in \mathbb{R}^T$ are defined by $v_{tt'} := w(t, t')$ for $t \neq t' \in T$, where the weight w is the co-occurrence count , and $v_{tt} = 0$. The reason for giving weight zero between a node and itself is that we want two tags to be considered related when they occur in a similar context, and not when they occur together. TagCont is determined by using the cosine measure, a measure customary in Information Retrieval [32]: If two tags t_1 and t_2 are represented by $\vec{v}_1, \vec{v}_2 \in \mathbb{R}^T$, their cosine similarity is defined as: $\text{cossim}(t_1, t_2) := \cos \measuredangle(\vec{v}_1, \vec{v}_2) = \frac{\vec{v}_1 \cdot \vec{v}_2}{||\vec{v}_1||_2 \cdot ||\vec{v}_2||_2}$. The cosine similarity is thus independent of the length of the vectors. As in our case the vectors contain only positive entries, its value ranges from 0 (for totally orthogonal vectors) to 1 (for vectors pointing into the same direction).

By studying the taxonomic path lengths in WordNet and the number of up and down edges on the paths, we further observed that pairs of tags which had been determined as closest pairs ac-

[1] http://www.gutenberg.org/etext/22

[2] http://wordnet.princeton.edu

cording to the cosine measure and which had a path distance of 2 in WordNet were significantly more frequently siblings[3] in WordNet than pairs determined with other measures. This implied that even if the cosine measure was not able to provide an immediate synonym, it still often provided a similar tag which was on an equal level of abstraction.

In [26] we have studied further measures of tag relatedness. We discovered there that mutual information gain is yielding even more precise results. However, the quadratic complexitymakes a frequent application to numerous large-scale folksonomy subsets (as needed in our case) infeasible. Given that TagCont has been proven to make meaningful judgements of semantic tag relatedness (as shown in [7]), we use it in the remainder of this paper as a measure for emergent tag semantics.

To complement the presented semantic measures, the next section will introduce and discuss measures aimed at capturing pragmatic aspects of tagging.

3. PRAGMATICS OF TAGGING

In addition to research on emergent semantics in folksonomies, the research community has developed an interest in usage patterns of tagging, such as why and how users tag. Early work by for example Golder and Huberman [14], and later Marlow et al [27], has identified different usage patterns among users. Further work provides evidence that different tagging systems afford different tag usage and motivations [18, 16]. More recent work shows that even within the same tagging system, motivation for tagging between individual users varies greatly [34]. These observations have led to the formulation of the hypothesis that the *emergent properties of tags in tagging systems — and their usefulness for different tasks — are influenced by pragmatic aspects of tagging* [18]. If this was the case, different tagging practices and motivations would effect the processes that yield emergent semantics. This would mean that in order to assess the usefulness of methods for harvesting semantics from folksonomies, we would need to know whether these methods produce similar results across different user populations characterized by different tagging practices and driven by different motivations for tagging. Given these implications, it is interesting to explore *whether and how emergent semantics of tags are influenced by the pragmatics of tagging*.

3.1 Tagging motivation

Previous work such as [27, 16] and [18] suggests that a distinction between at least two types of user motivations for tagging is interesting: On one hand, users can be motivated by categorization (in the following called *categorizers*). These users view tagging as a means to categorize resources according to some (shared or personal) high-level conceptualizations. They typically use a rather elaborated tag set to construct and maintain a navigational aid to the resources for later browsing. On the other hand, users who are motivated by description (so called *describers*) view tagging as a means to accurately and precisely describe resources. These users tag because they want to produce annotations that are useful for later searching and retrieval. Developing a personal, consistent ontology to navigate to their resources is not their goal. Table 1 gives an overview of characteristics of the two different types of users, based on [34]. While these two types make an ideal distinction, tagging in the real world is likely to be motivated by a combination of both. A user might maintain a few categories while pursuing a description approach for the majority of resources and vice versa, or additional categories might be introduced over time. Second, the distinction between categorizers and describers is a distinction based on the pragmatics of tagging, and not related to tag semantics. One implication of that is that it would be perfectly plausible for the same tag (for example "`java`") to be used by both describers and categorizers, and serve both functions at the same time — for different users. In other words, the same tag might be used as a category or a descriptive label. Thereby tagging pragmatics represent an additional perspective on folksonomical data, and yet it can be expected to have effects on the emergent semantics of tags. For example, it is reasonable to assume that the tags produced by describers are more descriptive than tags produced by categorizers. If this was the case, algorithms focused on utilizing tags for ontology learning would benefit from knowledge about the users' motivation for tagging.

Table 1: Two Types of Taggers

	Categorizer	*Describer*
Goal of Tagging	later browsing	later retrieval
Change of Tag Vocabulary	costly	cheap
Size of Tag Vocabulary	limited	open
Tags	subjective	objective

3.2 Measures of tagging pragmatics

Because the motivation behind tagging is difficult to measure without direct interaction with users, we use this distinction as an inspiration for the definition of the following surrogate measures for pragmatic aspects of tagging only.

3.2.1 Vocabulary size

$$vocab(u) = |T_u| \tag{1}$$

The *vocabulary size* (as proposed by for example [14] or [27]) reflects the number of tags found in a user's tag vocabulary T_u. Describers would likely produce an open set of tags with a unlimited and dynamic tag vocabulary while categorizers would try to keep their vocabulary limited and would need far fewer tags. A deficit of this measure is that it does not reflect on the total number of annotated resources, which are considered in the next measure.

3.2.2 Tag/resource ratio (trr)

$$trr(u) = \frac{|T_u|}{|R_u|} \tag{2}$$

This measure relates the vocabulary size with the total number of annotated resources. Taggers who use lots of different tags for their resources would score higher values for this measure than users that use fewer tags. Due to the limited vocabulary, a categorizer would likely achieve a lower score on this measure than a describer who employs a theoretically unlimited vocabulary. The equation above shows the formula used for this calculation where R_u represents the resources which were annotated by a user u. What this measure does not reflect on is the average number of assigned tags per post. This is considered next.

3.2.3 Average tags per post (tpp)

$$tpp(u) = \frac{\sum^{r} |T_{ur}|}{|R_u|} \tag{3}$$

This measure quantifies how many tags a user applies to a resource on average. Taggers who usually apply lots of tags to their re-

[3] An example for this are the tags '`java`' and '`python`'.

sources get higher scores by this measure than users who use few tags during the annotation process. Describers would score high values for this measure because of their need for detailed and verbose tagging. In contrast categorizers would score lower values because they try to annotate their resources in an efficient way.

3.2.4 Orphan ratio

$$orphan(u) = \frac{|T_u^o|}{|T_u|}, T_u^o = \{t||R(t)| \leq n\}, n = \left\lceil \frac{|R(t_{max})|}{100} \right\rceil \quad (4)$$

As a final measure, we introduce the *orphan ratio* of users to capture the degree to which users produce *orphaned tags*. Orphaned tags are tags that users assign to just a few resources. The *orphan ratio* thus captures the percentage of items in a user's vocabulary that represent such orphaned tags. T_u^o denotes the set of orphaned tags in a user's tag vocabulary T_u (based on a threshold n). The threshold n is derived from each user's individual tagging style in which t_{max} denotes the tag that was used the most. $|R(t)|$ denotes the number of resources which are tagged with tag t by user u. The measure ranges from 0 to 1 where a value of 1 identifies users with lots of orphaned tags and 0 identifies users who maintain a more consistent vocabulary. Considering the categorizer - describer paradigm this would mean that categorizers tend more towards values of 0 because orphaned tags would introduce noise to their personal taxonomy. For a describer's tag vocabulary, this measure would produce values closer to 1 due to the fact that describers tag resources in a verbose and descriptive way, and do not mind the introduction of orphaned tags to their vocabulary.

3.3 Properties of measures

While these measures of tagging pragmatics were inspired by the dichotomy between categorizers and describers, we do not require them to accurately capture this distinction. Another aspect is that these measures might not only capture intrinsic user characteristics, but can also be influenced by e.g. elements of user interfaces (such as recommenders). What is important in the light of our hypothesis is that all of *the above measures are independent of semantics* — they capture *usage patterns* of tagging (the pragmatics of tagging) only. This allows us to explore a potential link between tagging pragmatics and the emergent semantics of tags.

4. SEMANTIC IMPLICATIONS OF TAGGING PRAGMATICS

As detailed in Sec. 2.2, the distributional hypothesis states that words used in similar contexts tend to have similar meanings. As tags in a folksonomy can be regarded as natural language entities, a crucial question is how to identify an adequate context for capturing their semantics. However, given the massive amounts of data available in social tagging systems, the question is not only to identify a *valid* context, but also to identify the *minimal* context which retains the relevant structures while allowing for efficient computation. As human annotators are the creators of implicit semantic structures, an important aspect hereby is which users should be included in an optimal context composition. Following our discussion in the prior section, our hypothesis is that individual tagging pragmatics can play an important role for selecting "productive" users. The question is whether the categorizers — who follow the ontology engineering principle of a clean vocabulary — or the describers — who provide more descriptions to their resources — are the more "productive" ones.

Table 2: del.icio.us dataset statistics.

dataset	$\lvert T\rvert$	$\lvert U\rvert$	$\lvert R\rvert$	$\lvert Y\rvert$
full	10,000	511,348	14,567,465	117,319,016
min100res	9,944	100,363	12,125,476	96,298,409

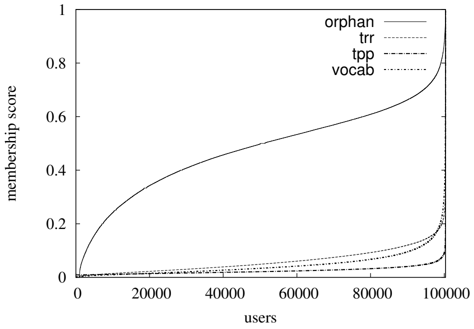


Figure 1: Distribution of the membership scores for each introduced measure of tagging motivation (orphan ratio, tag/resource ratio, tags per post and vocabulary size), computed for the 100,393 users present in our del.icio.us dataset (x-axis). Values close to 0 on the y-axis indicate strong categorizers, while values close to one 1 point to describer users. All measures were normalized to the interval $[0, 1]$.

In order to answer this question, our strategy is to analyze the suitability of each of our previously introduced pragmatic measures to assemble a (preferentially small) subset of users which provides a sufficient context to harvest emergent tag semantics. The general idea hereby is to start at both ends of the scale with the "extreme" categorizers and describers, and then to subsequently add more users (in the order given by the respective measure). In each step, we check how well the folksonomy partition defined by the current user subset serves as a basis to compute semantically related tags. For the latter, we revert to the tag context relatedness measure that has shown to produce valid results (cf. Sec. 2). The assumption hereby is that this TagCont measure will yield more closely related tags when better implicit semantic structures are present. Hence, this whole procedure allows us to assess the quality of the emergent semantics and finally the degree to which tagging pragmatics have influenced its evolution.

4.1 Experiments

The goal of our experiments is to quantify the influence of individual tagging practices on emergent tag semantics in a folksonomy. We will first provide details on our dataset and then explain each experimentation step before discussing the results.

4.1.1 Description of the dataset

In order to validate our hypothesis on real-world data, we used a dataset crawled from the social bookmarking system del.icio.us in November 2006.[4] In total, data from 667,128 users of the del.icio.us community were collected, comprising 2,454,546 tags, 18,782,132 resources, and 140,333,714 tag assignments. As our experimental methodology involves the comparison with semantically related tags obtained from the full dataset, we need to ensure that the quality of those is high. Because the applied tag relatedness measure is based on the co-occurrence of tags with other tags, the inherent sparseness of infrequent tags makes them less useful for our purpose. Hence, we stick to our dataset containing the 10,000 most

[4] All data sets used in this study are publicly available at http://www.kde.cs.uni-kassel.de/benz/papers/2010/www.html

frequent tags of del.icio.us, and to the resources/users that have been associated with at least one of those tags. We will refer to the resulting folksonomy as the *full* dataset (see Table 2).

In order to eliminate noise introduced by our measures misjudging new users, we furthermore removed all users having less than 100 resources in their collection. The reason behind this is that e. g., the tag/resource ratio is not very informative in the case of a new user with very few resources. Interestingly, our result shows that removing this "long tail" of new (or inactive) users already increases the quality of the learned semantic relations. Details of this observation will be discussed in Section 4.3. We will denote the resulting dataset as *min100res* (see Table 2).

4.1.2 Experimental setup

In order to assess the capability of each of our measures to predict "productive" users, we followed an incremental approach: For each of our measures $m \in \{\mathit{orphan,vocab,trr,tpp}\}$, we first created a list L_m of all users $u \in U$ sorted in ascending order according to $m(u)$. All our measures yield low values for categorizers, while giving high scores to describers. This means that e.g. the first user in the orphan ratio list (denoted as $L_{\mathit{orphan}}[1]$) is assumed to be the most extreme categorizer, while the last one ($L_{\mathit{orphan}}[k], k = |U|$) is assumed to be the most extreme describer. Figure 1 depicts the obtained distribution of membership scores for each ordered list L_{tpp}, L_{trr}, L_{orphan} and L_{vocab}. An observation which can be made in this figure is that the distribution of the orphan measure differs clearly from the other three measures. This implies that the orphan ratio seems to be able to make more fine-grained distinction between users. However, our results did not exhibit a positive impact on the resulting semantics; rather contrary, the orphan ratio performs often worse than the other measures (see section 4.2 for details).

Because we are interested in the minimum amount of users needed to provide a valid context, we start at both ends of L and extract two folksonomy partitions CF_1^m and DF_1^m based on 1% of the "strongest" categorizers ($\mathit{Cat}_1^m = \{L_m[i] \mid i \leq 0.01 \cdot |U|\}$) and describers ($\mathit{Desc}_1^m = \{L_m[i] \mid i \geq 0.99 \cdot |U|\}$). $CF_1^m = (CU_1^m, CT_1^m, CR_1^m, CY_1^m)$ is then the sub-folksonomy of F induced by Cat_1^m, i. e., it is obtained by $CU_1^m := \mathit{Cat}_1^m$, $CY_1^m := \{(u,t,r) \in Y | u \in \mathit{Cat}_1^m\}$, $CT_1^m := \pi_2(CY_1^m)$, and $CR_1^m := \pi_3(CY_1^m)$. The sub-folksonomy DF_1^m is determined analogously.

As a next step, we took the first extracted partition CF_1^m as input to extract semantic tag relations, in the way described in Section 2.2. We check whether the data produced by a very small subset of "extreme" categorizers already suffices to compute meaningful semantic relations. More specifically, for each tag $t \in CT_1^m$, we computed its most similar tag t_{sim} according to the tag context relatedness defined in [7]. We then looked up each resulting pair (t, t_{sim}) in WordNet and measured – whenever both t and t_{sim} were present — the Jiang-Conrath distance $\mathit{JCN}(t, t_{\mathit{sim}})$ between both words (see Sec. 2.2). After that we took the average JCN distance of all mapped tag pairs as an indicator of the quality of emergent semantic structures contained in CF_1^m:

$$\mathit{JCN}_{\mathit{avg}}(CF_1^m) = \frac{\sum_{t \in CT_1^m} \mathit{JCN}(t, t_{\mathit{sim}})}{\mathit{wn\_pairs}(CT_1^m)}$$

Here, $\mathit{wn\_pairs}(DT_1^m)$ denotes the number of tag pairs (t, t_{sim}) (i. e., a tag and its most similar tag) for which both t and t_{sim} are present in WordNet. The corresponding describer partition DF_1^m was processed in the same manner.

As discussed in Sec. 2.2, we use the Jiang-Conrath distance as an indicator of the "true" semantic relatedness between tags. However, in order to avoid the dependency of our results on a single measure of semantic similarity, we also measured the *taxonomic path length* for each mapped tag pair (t, t_{sim}) between the two synsets s_1 and s_2 containing t and t_{sim}, respectively.[5] This measure counts the number of nodes in the WordNet subsumption hierarchy along the shortest path between s_1 and s_2. We noticed that the judgements of both measures (JCN and taxonomic path length) were almost perfectly correlated throughout our experimentation; for this reason, we will stick to the JCN distance in the remainder of this paper, because it has been shown to be a better surrogate for the human perception.

We repeated this overall procedure for each of our measures $m \in \{\mathit{orphan,vocab,trr,tpp}\}$ and for the following user fractions i:

$$i \in \{1, 2, 3, \ldots, 24, 25, 30, 40, 50, 60, 70, 80, 90\}$$

As we keep adding users while incrementing i, it is important to notice that the size of the resulting "sub-folksonomy" is growing towards the size of the full dataset i. e., $DF_{100}^m = CF_{100}^m = F$. Another important aspect is the fact that users are added in descending order of their membership degree in the respective user class: This means that CF_1^m contains users u who score high on measure m, while e. g., CF_{50}^m contains a more mixed population. "Mixed" in this context means that there exist users in CF_{50}^m which are to a certain degree assumed to exhibit describer characteristics as measured by m. This implies that the distinction between both user groups is blurred while incrementing i. In other words, one can also read these partitions from the other side, namely that CF_{90}^m contains all users *except* 10% of the most extreme describers.

So in summary, we created 64 partitions for each of our 4 measures (32 categorizer + 32 describer), summing up to a total of 256 sub-folksonomies, each being extracted by a different composition of users according to their tagging characteristics. Before presenting our results on the most suitable partitions for extracting semantic tag relations, we discuss upper and lower bounds. As we measured the quality of an extracted relation between two tags t and t_{sim} by its Jiang-Conrath distance within WordNet, a lower bound can be identified by computing the pairwise JCN distance between all tags $t \in T$ and averaging over the minimum distance found for each tag:

$$\mathit{JCN}_{\mathit{lower}}(F) = \frac{\sum_{t \in T} \min_{t_{\mathit{sim}} \in T} \mathit{JCN}(t, t_{\mathit{sim}})}{\mathit{wn\_pairs}(T)}$$

As an upper bound we assume that the respective folksonomy subset does not contain any inherent semantics and hence only randomly related tags are returned by our measure. We simulate this by defining a random relatedness function $\mathit{rand}(t)$, which returns a randomly selected tag $t_{\mathit{sim}} \in T, t_{\mathit{sim}} \neq t$. The upper bound is then:

$$\mathit{JCN}_{\mathit{upper}}(F) = \frac{\sum_{t \in T} \mathit{JCN}(t, \mathit{rand}(t))}{\mathit{wn\_pairs}(T)}$$

For the del.icio.us dataset it turned out that $\mathit{JCN}_{\mathit{upper}} \approx 15.834$ and $\mathit{JCN}_{\mathit{lower}} \approx 0.758$. Please recall that JCN is a semantic *distance* measure — which means a low JCN distance corresponds to a high degree of semantic relatedness.

As seen later (cf. Figure 2), none of our experimental conditions (including the full dataset) came close to the lower bound. There are (at least) two explanations for this. Firstly, the lower bound was determined independently of a sub-folksonomy of the full dataset. It would be interesting to determine that sub-folksonomy that provides the optimal average Jiang-Conrath distance. Then one could check how far it is away from this optimum, and one could try to

[5] If t and t_{sim} were present in more than one synset, we took the shortest possible path.

learn a classifier for this target dataset. Unfortunately, the computation of this sub-folksonomy requires the consideration of all subsets of the user set U and is thus computationally unfeasible.

Secondly, WordNet is built by language experts with the goal to capture *all* existing senses of a given word. Given two tags t_1 and t_2, our JCN implementation searched for the smallest possible distance between *any* two senses of each tag. By doing so for all possible pairs of tags $t \in T$, the probability is quite high to find two quite closely related (or even equal) senses. Contrary to that, the technophile bias of the user population of del.icio.us leads to some usage-induced relations which are not reflected well within WordNet; as an example, the most related tag to `doom` in a folksonomy subset was `quake`, leading to a large JCN distance of ≈ 18.08, while the optimal distance was found between `doom` and `will` with ≈ 1.88. This observation does not invalidate the procedure of semantic grounding as a whole, because we *do* find matching semantics in both systems. The same approach has also been taken in previous publications focused on measures for semantic relatedness [7].

4.2 Results

In Figures 2(a) and 2(b) we present the results of our analysis of the different sub-folksonomies which were created in each of our 256 experimental conditions.

The horizontal axis displays the percentage of included users; the vertical axis displays the average JCN distance obtained from computing semantically related tags based on the respective partition. The dashed line at the bottom of each figure represents the level of semantic precision obtained from the full dataset.

A first impression is — in all diagrams, independently of the selection strategy — that mass matters: the average JCN distance decreases and hence the results get better while more users are included. This equally holds for the random selection strategy (solid line, $+$). In other words, the more people contribute to a collaborative tagging system, the higher is the quality of the semantic tag relations which can be obtained from the folksonomy structure they produce. This matches the intuition that a sufficient "crowd" is necessary to facilitate the emergence of the "wisdom of the crowds".

However, the obvious differences between the two Figures 2(a) and 2(b) suggests that the composition of the crowd also seems to make a difference: When incrementally adding users ordered from categorizers to describers (starting from the left of Figure 2(a)), all resulting folksonomy partitions yield systematically weaker semantic precisions compared to adding users in random order (solid line, $+$). This effect can be observed most clearly for the vocabulary size measure *vocab* (dotted line, ▲), which judges users as categorizers when the size of their tag vocabulary is small (see Sec. 3.2.1). Only after the addition of 90 % of all users in this order, the quality of the inherent semantics are on the same level of randomly selected 90 %. The other measures — with an exception of the tags per post ratio (dotted line, •) which will be discussed later — show a very similar behavior, namely the tag/resource ratio (dotted line, ■) and the orphan ratio (dotted line, $*$).

When incrementally building sub-folksonomies starting from describer users (Figure 2(b)), we see a completely different picture: most measures start on the same or even on a slightly higher level of contained semantics compared to adding users in a random order. Beginning from roughly 10 % included users, all sub-folksonomies yield better results than the random case. In addition, after having added 40 % of the users in the order of the tag/resource ratio (dotted line, □), we can even observe a first improvement of the results compared with the full dataset. This implies that a bit less than the "better half" of the complete folksonomy population produces equally precise semantic structures compared to the whole unfiltered "crowd". This improvement increases and reaches its maximum after adding 70 % of all users, before it decreases again to the global level.

Table 3: Statistical properties of selected folksonomy partitions. %t denotes the fraction of the tags from the complete dataset included in the respective partition; %w denotes the number of similar tag pairs (t, t_{sim}) found in WordNet for the respective partition divided by the number of mapped pairs from the whole dataset. For the entire dataset, $|T| = 9944$ and *wn_pairs*$(T) = 4335$.

	DF_i^{trr}		DF_i^{tpp}		DF_i^{orphan}		DF_i^{vocab}	
i	%t	%w	%t	%w	%t	%w	%t	%w
1	0.93	1.03	0.96	1.01	0.97	1.02	0.98	1.04
3	0.96	1.02	0.98	1.02	0.99	1.01	0.99	1.03
5	0.97	1.02	0.99	1.02	0.99	1.02	0.99	1.03
10	0.97	1.03	0.99	1.02	1.00	1.02	0.99	1.01
20	0.98	1.02	0.99	1.00	1.00	1.03	0.99	1.01
50	0.98	1.02	1.00	1.00	1.00	1.00	1.00	1.01
70	0.99	1.01	1.00	1.00	1.00	1.00	1.00	1.00

	CF_i^{trr}		CF_i^{tpp}		CF_i^{orphan}		CF_i^{vocab}	
i	%t	%w	%t	%w	%t	%w	%t	%w
1	0.56	0.48	0.44	0.00	0.48	0.59	0.27	0.18
3	0.86	0.77	0.74	0.23	0.78	0.77	0.59	0.44
5	0.94	0.83	0.87	0.49	0.89	0.88	0.76	0.59
10	0.97	0.90	0.95	0.80	0.95	0.95	0.91	0.78
20	0.99	0.95	0.97	0.88	0.97	0.98	0.97	0.88
50	1.00	1.00	0.98	0.96	0.98	1.01	0.98	0.95
70	1.00	1.00	0.98	0.98	0.99	0.99	0.98	0.98

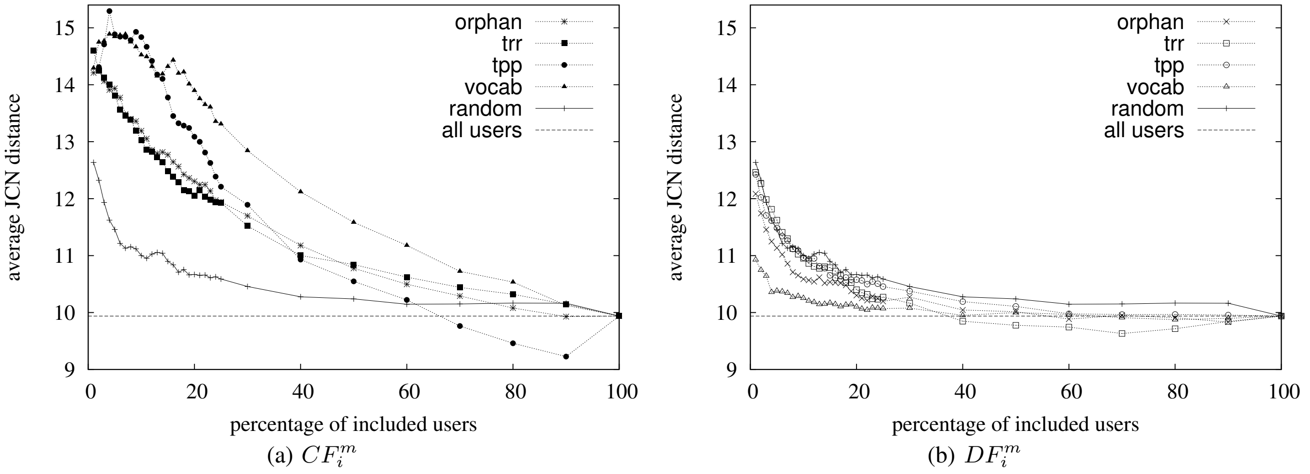
Especially for very small partitions (roughly ≤ 20 %), users selected in descending order by their vocabulary size yield the best results (dotted line, △). Interestingly, this effect is inverse when adding users the other way round (dotted line, ▲, in Fig. 2(a)): Even quite a large number of users with small vocabularies perform considerably worse than most other folksonomy partitions. This means that scale still matters, as the quality almost constantly increases while adding users; but the "collaborative verbosity" of a small subset of users with large vocabularies seems to lead to much richer inherent semantics than the contributions of a larger set of more "tight-lipped" users.

One could suspect now that this comparison is not completely fair: Especially when selecting users with small vocabularies, the question is to which extent semantic relations *can* be present at all in the data. In other words: If the aggregated small vocabularies of a subset of categorizers result in a considerably smaller global vocabulary compared to aggregating more verbose users, then the probability to find semantically close tags would consequently be much lower. In the worst case, the vocabulary would be so small that the "right partner" for a given tag *does not exist.*

In order to eliminate this concern, we counted the size of the collective tag vocabulary for each sub-folksonomy. In addition, we measured how many tag pairs (t, t_{sim}) could be mapped to WordNet during the computation of the JCN distance. By doing this we want to make sure that the average semantic distance is computed roughly over the same number of tag pairs. Table 3 summarizes some selected statistics relative to the complete dataset.[6]

The first observation is that in all partitions based on describers (upper half of the table) the global vocabulary is almost completely contained ($\geq$ 93 %). For partitions larger than 20 %, this value raises to 98 %. The same holds for the fraction of tag pairs mapped

[6] We did not include the statistics for every partition for space reasons; missing values can be interpolated from the given examples.



(a) CF_i^m (b) DF_i^m

Figure 2: Average Jiang-Conrath distance between pairs of semantically related tags computed from different folksonomy partitions. The partitions were created based on user subsets as determined by different pragmatic measures (orphan ratio, tag/resource ratio, tags per post, vocabulary size). Each datapoint corresponds to a "sub-folksonomy" CF_i^m (a) / DF_i^m (b) with $i = 1, 2, \ldots, 25, 30, 40, 90$ (from left to right in both cases). The x-axis denotes the percentage of all folksonomy users included in the subset, and the y-axis depicts the quality of the semantic tag relations obtained from the respective partition by means of the JCN distance. In Figure 2(a), users were added ordered from categorizers to describers, and in Figure 2(b) ordered in the reverse direction. (Note: Empirical lower-/upper bounds are $\approx 0.758/15.83$, respectively; cf. Sec. 4.1.2.)

to WordNet. On the first sight, values > 1 might appear counter-intuitive here. The explanation is the following: It can happen that for a given tag t that its most similar tag t_{sim} based on the complete dataset it not present in WordNet, but its most similar tag t'_{sim} based on a particular partition is contained. A high percentage of mapped tags does not imply better semantics per se (as the two mapped tags can still be semantically distant); but the comparison of different sub-folksonomies is more meaningful when they both allow for a roughly equal number of mapped pairs. As expected, the coverage observed for the describer-based case is not as complete for the categorizer-based excerpt: For very small samples, the collective tag pool is in fact small. However, this effect is mitigated already for samples of 3 %; and starting from roughly 10-20 % sample size, a sufficient global vocabulary exists (≈ 97 %). This means that the comparison in general is performed on a fair basis, because the underlying vocabulary sizes are comparable.

Our results suggest that sub-folksonomies based on describers contain more precise inherent semantic structures than partitions based on categorizers. However, there seems to be a limitation with this observation: Inspecting the curve for the *tpp* measure on the right side of Figure 2(a), one can observe that the most precise semantic relations among all experimental conditions are found after the addition of 90 % of the categorizers according to this measure. As stated above, this partition can also be read from the other side and corresponds to a removal of 10 % of the most extreme describers. As the *tpp* measure captures the average numbers of tags per post, there seems to be a number of "ultra-taggers" who use a large number of tags per post (many spammers, typically more than 9 tags per post in our case) have detrimental effects on the global tag semantics. In other words, removing these users seems to eliminate "semantic noise", leading to more precise tag semantics.

4.3 Discussion and implications

Recent research demonstrated that the collective output of tagging systems can be used for harvesting emergent semantic structures from the web [35, 33, 7]. Our results show that the effectiveness of current semantic measures for tag relatedness are influenced by factors originating outside of the semantic realm. On small data samples (up to 40 % of users in our dataset), we have singled out a group of users (categorizers) that has particularly detrimental effects on the performance of current semantic measures compared to random sampling. At the same time, describers (based on the tags-per-resource measure) consistently outperform random sampling, and can level and even outperform the results achieved on the entire dataset with as little as 40 % of users. This suggests that methods for harvesting *semantics* from samples of tagging systems can be made more effective when utilizing knowledge about the *pragmatics* of tagging, considering individual user behavior. For analysts of small data samples who wish to improve semantic relatedness measures, this would mean focusing on those users that use tagging systems in a verbose 'Stop Thinking, Start Tagging' fashion. With increasing sample sizes (>50 % of users), we can observe that adding more categorizers does not produce significantly better results. However, when adding more describers, we see significant improvements in performance until we hit an accuracy limit at approximately 90 % of users. This suggests that rewarding verbose taggers comes with limitations itself: The most verbose taggers (in our case: mostly spammers) negatively influence the results as well.

The practical implications of our results concern mainly two questions: (i) What is the minimum amount of users needed to produce meaningful tag semantics in collaborative tagging systems and how can these users be selected? (ii) Does the quality of emerging tag semantics increase with the available amount of data, or can it be improved by eliminating "semantic noise"?

A main contribution of our analysis lies in the observation that tagging pragmatics, i. e., individual tagging characteristics, play an important role in both cases. The experiments described above reveal that not all users contribute equally to emerging semantics; we could show that a relatively small subset of describers yields significantly better results than a group of categorizers. Figure 3

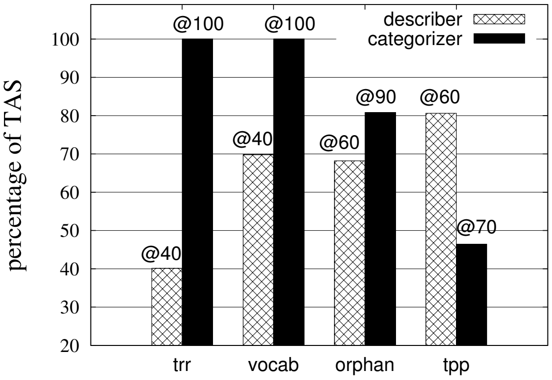


Figure 3: Minimum size of the folksonomy partitions created by each measure sufficient to reach the semantic precision of the complete dataset. The y-axis denotes the percentage of tag assignments contained in the smallest folksonomy partition which reached the global semantic precision; the labels above the bars depict the percentage of users the respective subfolksonomies are based on.

summarizes the minimum sizes of the folksonomy partitions identified by each of our introduced measures necessary to reach the level of semantic precision for the entire dataset. The white bars correspond to sampling users ordered from describers to categorizers (Fig. 2(b)) while the black bars correspond to sampling users ordered in the opposite direction (Fig. 2(a)). The number on top of each bar displays the user fraction needed to reach the global semantic precision; the y-axis depicts the size of the respective subfolksonomy relative to the complete one.

In general, most describer-based selection strategies create smaller folksonomies which produce meaningful semantics. The "smallest" one consists of 40 % describers according to the trr measure, responsible for roughly 40 % of all tag assignments. However, the observation that uncontrolled verbosity is not a good thing is confirmed by the fact that removing 30 % of the most extreme describers according to the tags-per-post measure (rightmost black bar) also creates a comparatively small and semantically precise partition. According to Figure 3, two adequate strategies for creating the smallest possible scaffolding for global tag semantics can be identified: (1) include roughly half of the users with a high tag/resource ratio, and (2) remove roughly one third of "ultra-taggers" identified by a large average number of tags per post.

The next interesting question to ask is whether, and to which extent we can even infer *more precise* semantics when removing users. Figure 4 displays the obtained semantic precision (y-axis) plotted against the amount of tag assignments removed when removing users according to different selection strategies. The first and most simple strategy is to remove the "long tail" of users with less than 100 resources in their collection. This already eliminates roughly 18 % of the data, while interestingly slightly improving the semantic precision. One cannot conclude from that that the long tail of users does not contain valuable information at all. But with regard to *popular* tags (recall that we restricted our dataset to the top 10.000 tags), a valid first insight is that the long tail of inactive users can be discarded during the computation of semantic tag relations.

As discussed before, our results indicate that categorizers also have a detrimental effect on the quality of the emerging structures. Removing 30 % of them as determined by the tag/resource ratio leads to a further improvement in semantic precision. The best result in all of our experimental conditions however was reached by eliminating 10 % of the extreme describers according to the tags-per-post measure. Those "hyper-active" users (in our case mostly spammers as confirmed by manual inspection) generate roughly 40 % of the global amount of tag assignments. Spammers typically use a large number of semantically disjoint tags to attract other users and to bias search engines towards their posted URLs. Unsurprisingly, they are not very helpful for creating meaningful tag relations. Rather the contrary is the case: we can see in our results that spammers introduce significant semantic noise — a removal of them leads to an overall improvement in accuracy of the resulting semantic structures. Turning the tables around, this insight can of course also be useful for spammer detection itself — but because our dataset does not contain explicit spammer labels for each user, determining the exact ratio of spammers detected by each of our pragmatic measures is subject to future work.

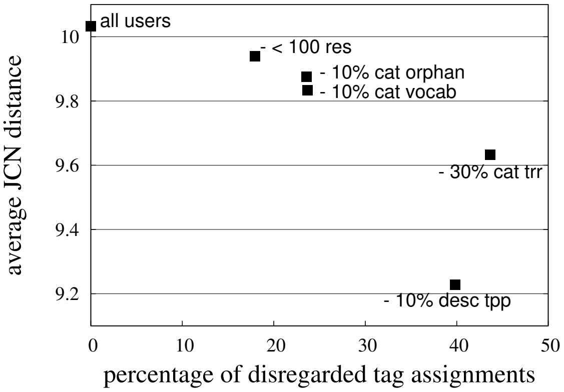


Figure 4: Improvement of semantic precision by removing users from the complete dataset. The y-axis depicts the semantic precision of the (sub-)folksonomies, while the x-axis denotes the percentage of tag assignments which were disregarded by removing certain users. The label at each data point describes which users were removed.

4.4 Generalization on other datasets

In order to exclude the possibility that the implications mentioned above are influenced by characteristics from the del.icio.us dataset, we repeated the experimental procedure described in section 4.1.2 on a dataset from January 2010 of our own social bookmarking system BibSonomy[7]. It contained 17,777 users, 10,000 tags and 4,520,212 resources connected by 34,505,061 TAS. Space does not permit a detailed presentation of the results; but in general, all measures exhibited a very similar behavior as observed for the del.icio.us dataset in Figures 2(a) and 2(b). Especially the practical implications discussed in Section 4.3 were valid in a nearly identical way for the BibSonomy data: 30% of describers according to the trr measure were sufficient to reach the semantic precision of the whole dataset, and removing 20% of describers according to the tpp measure led to the best overall semantics.

5. RELATED WORK

There is series of research discussing folksonomies from a formal [30] and informal [28] perspective. First quantitative analysis of folksonomies are provided in [14] and the underlying structure is

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

analyzed in [8]. Tag-based metrics for resource distance have been introduced in [6]. [1] gives evidence that social annotations are a potential source for generating semantic metadata.

Many publications on folksonomies introduce measures for tag relatedness, e. g., [19, 33]. However, the choice of a specific measure of relatedness is often made without justification and often it appears to be rather ad hoc. Which context information captures the meaning of tags best has been addressed by [38]. Questions that have not been addressed previously include which users contribute to what extent to emergent semantics in folksonomies, and to what extent are tag semantics influenced by tagging pragmatics. In [7] we performed first analysis on different kinds of relatedness measures and different types of semantic relationships. In the paper at hand, we investigate different measures to characterize users and their level of contribution to the semantics of a folksonomy. To the best of our knowledge, no other analysis in the literature addresses the interrelation between pragmatic aspects of tagging (namely user characteristics) and their semantic implications for tag relatedness.

[25] generalizes standard tree-based measures of semantic similarity to the case where documents are classified in the nodes of an ontology with non-hierarchical components. The measures introduced there were validated by means of a user study. [31] analyses distributional measures of word relatedness and compares them with measures of semantic relatedness in thesauri like WordNet. In [26] we provide a systematic analysis of a broad range of similarity measures that can be applied directly and symmetrically to build networks of users, tags, or resources and to compute similarities between these entities.

A task which depends heavily on quantifying tag relatedness is that of tag recommendation in folksonomies. In the last years, a lot of research activities can be observed as two ECML PKDD discovery challenges [20, 11] were based on this topic. Existing work in general can be broadly divided in approaches that analyze the content of the tagged resources with information retrieval techniques [4] and approaches that use collaborative filtering methods based on the folksonomy structure [37]. An example of the latter class of approaches is [22]. Relatedness measures also play a role in assisting users who browse the contents of a folksonomy. [3] shows that navigation in a folksonomy can be enhanced by suggesting tag relations grounded in content-based features.

A considerable number of investigations are motivated by the vision of "bridging the gap" between the Semantic Web and Web 2.0 by means of ontology-learning procedures based on folksonomy annotations. [30] provides a model of semantic-social networks for extracting lightweight ontologies from del.icio.us. Other approaches for learning taxonomic relations from tags are provided by [19, 33]. Another branch of research is concerned with the enrichment of folksonomies by including data from existing semantic repositories and ontologies [2]. [24] proposes an RDFS model to formalize the meaning of tags relative to other tags. [15] presents a generative model for folksonomies and also addresses the learning of taxonomic relations. [39] applies statistical methods to infer global semantics from a folksonomy. The results of our paper are especially relevant to inform the design of such learning methods.

6. CONCLUSION AND FUTURE WORK

In this paper, we analyzed the influence of individual tagging practices in collaborative tagging systems on the emergence of global tag semantics. After proposing a number of statistical measures to assign users to two broad classes of categorizers and describers, we systematically built folksonomy partitions by incrementally adding users from each class. We then judged the quality of the emergent semantics contained in each of these "sub-folksonomies" by means of semantically grounded tag relatedness measures. Apart from the observation that adding more users is beneficial in many — but not all — cases, our results reveal a dependence of the obtained semantic structures on the different partitions. In general, the collaborative verbosity of describers provides a better basis for harvesting meaningful tag semantics. However, this observation comes with a limitation: The most verbose taggers (in our case mostly spammers) negatively influenced semantic accuracy. From a practical perspective, the pragmatic measures can be used to select a comparatively small subset of users which produce tag relations of equal or better quality than the entire set of users. In addition, the measures can facilitate improvement of the global semantic precision by eliminating users that introduce "semantic noise". Experiments with an additional dataset corroborate the assumption that our findings can be generalized to other collaborative tagging systems.

A main implication of our work is the presentation of first empirical evidence for a causal link between the pragmatics of tagging (individual tagging practices) and the emergent semantics of tags. This link is *not* dependent on our choice for a particular semantic relatedness measure, because 1) the chosen Jiang-Conrath distance has been shown to best reflect human judgements of semantic relatedness in previous validation studies [5] and 2) our experiments with alternative measures for semantic relatedness have produced similar results (cf. section 4.1).

This finding has a number of interesting implications for related areas of research: 1) While our results focus on semantic relatedness, it appears plausible that other semantic tasks, such as hypo/hypernym detection, exhibit similar effects. We argue that a general link between tagging pragmatics and tag semantics could yield new ways of thinking and new algorithm designs for learning ontologies from folksonomies. 2) Current tag recommender algorithms tap into semantic relations between tags in order to recommend tags to users. Our results suggest that knowledge about why and how users tag could help to further improve the performance of tag recommender systems. 3) Utilizing tag recommenders to influence tagging behavior and to direct the evolution of folksonomies towards more precise emergent semantics seems to represent an exciting and promising area for future work.

7. ACKNOWLEDGMENTS

The research presented in this work is in part funded by the Know-Center, the FWF Austrian Science Fund Grant P20269, the WebZubi project funded by BMBF and the VENUS project funded by Land Hessen.

8. REFERENCES

[1] H. S. Al-Khalifa and H. C. Davis. Exploring the value of folksonomies for creating semantic metadata. 2007.

[2] S. Angeletou. Semantic enrichment of folksonomy tagspaces. *The Semantic Web - ISWC 2008*, pages 889–894, 2009.

[3] M. Aurnhammer, P. Hanappe, and L. Steels. Integrating collaborative tagging and emergent semantics for image retrieval. In *Proc. WWW2006, Collaborative Web Tagging Workshop*, May 2006.

[4] C. H. Brooks and N. Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *WWW06: Proc. of the 15th Int'l Conference on World Wide Web*, pages 625–632, New York, NY, USA, 2006.

[5] A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

[6] C. Cattuto, A. Baldassarri, V. D. P. Servedio, and V. Loreto. Emergent community structure in social tagging systems. *Advances in Complex Physics*, 2007. Proc. of the European Confeence on Complex Systems ECCS2007.

[7] C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic grounding of tag relatedness in social bookmarking systems. In *The Semantic Web – ISWC 2008, Proc.Intl. Semantic Web Conference 2008*, volume 5318 of *LNAI*, pages 615–631, Heidelberg, 2008. Springer.

[8] C. Cattuto, C. Schmitz, A. Baldassarri, V. D. P. Servedio, V. Loreto, A. Hotho, M. Grahl, and G. Stumme. Network properties of folksonomies. *AI Communications Journal, Special Issue on Network Analysis in Natural Sciences and Engineering*, 20(4):245–262, 2007.

[9] D. Chandler. *Semiotics: The Basics*. Taylor & Francis, second edition, 2007.

[10] F. de Saussure. *Course in General Linguistics*. Duckworth, London, [1916] 1983. (trans. Roy Harris).

[11] F. Eisterlehner, A. Hotho, and R. Jäschke, editors. *ECML PKDD Discovery Challenge 2009 (DC09)*, volume 497 of *CEUR-WS.org*, Sept. 2009.

[12] C. Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998.

[13] J. R. Firth. A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1952-59:1–32, 1957.

[14] S. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, April 2006.

[15] H. Halpin, V. Robu, and H. Shepard. The dynamics and semantics of collaborative tagging. In *Proc. of the 1st Semantic Authoring & Annotation Workshop (SAAW)*, 2006.

[16] T. Hammond, T. Hannay, B. Lund, and J. Scott. Social bookmarking tools (i): A general review. *D-Lib Magazine*, 11(4), April 2005.

[17] Z. S. Harris. *Mathematical Structures of Language*. Wiley, New York, 1968.

[18] M. Heckner, M. Heilemann, and C. Wolff. Personal information management vs. resource sharing: Towards a model of information behaviour in social tagging systems. In *Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, San Jose, CA, USA, May 2009.

[19] P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, CS dep., April 2006.

[20] A. Hotho, D. Benz, R. Jäschke, and B. Krause, editors. *ECML PKDD Discovery Challenge 2008 (RSDC'08)*. Workshop at 18th Europ. Conf. on Machine Learning (ECML'08) / 11th Europ. Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'08), 2008.

[21] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In *The Semantic Web: Research and Applications*, volume 4011 of *LNAI*, pages 411–426, Heidelberg, 2006. Springer.

[22] R. Jäschke, L. B. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in folksonomies. In *Proc. PKDD 2007*, volume 4702 of *Lecture Notes in Computer Science*, pages 506–514, Berlin, Heidelberg, 2007.

[23] J. J. Jiang and D. W. Conrath. Semantic Similarity based on Corpus Statistics and Lexical Taxonomy. In *Proc. of the International Conference on Research in Computational Linguistics (ROCLING)*. Taiwan, 1997.

[24] F. Limpens, F. Gandon, and M. Buffa. Collaborative semantic structuring of folksonomies. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, 1:132–135, 2009.

[25] A. G. Maguitman, F. Menczer, F. Erdinc, H. Roinestad, and A. Vespignani. Algorithmic computation and approximation of semantic similarity. *World Wide Web*, 9(4):431–456, 2006.

[26] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *18th International World Wide Web Conference*, pages 641–641, April 2009.

[27] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HT'06: Proc. of the 17th conference on Hypertext and Hypermedia*, pages 31–40, New York, NY, USA, 2006.

[28] A. Mathes. Folksonomies – Cooperative Classification and Communication Through Shared Metadata, December 2004.

[29] S. Mcdonald and M. Ramscar. Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proc. of the 23rd Annual Conference of the Cognitive Science Society*, pages 611–6, 2001.

[30] P. Mika. Ontologies are us: A unified model of social networks and semantics. In *International Semantic Web Conference*, LNCS, pages 522–536. Springer, 2005.

[31] S. Mohammad and G. Hirst. Distributional measures as proxies for semantic relatedness. Submitted for publication, http://ftp.cs.toronto.edu/pub/gh/Mohammad+Hirst-2005.pdf.

[32] G. Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[33] P. Schmitz. Inducing ontology from Flickr tags. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, May 2006.

[34] M. Strohmaier, C. Körner, and R. Kern. Why do users tag? detecting users' motivation for tagging in social tagging systems. Technical report, Knowledge Management Institute - Graz University of Technology, 2009.

[35] H. Wu, M. Zubair, and K. Maly. Harvesting social knowledge from folksonomies. In *HYPERTEXT '06: Proc. of the seventeenth conference on Hypertext and hypermedia*, pages 111–114, New York, NY, USA, 2006. ACM Press.

[36] X. Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. In *WWW '06: Proc. of the 15th international conference on World Wide Web*, pages 417–426, New York, NY, USA, 2006. ACM.

[37] Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: Collaborative tag suggestions. In *Proc. of the Collaborative Web Tagging Workshop at the WWW 2006*, Edinburgh, Scotland, May 2006.

[38] C. A. Yeung, N. Gibbins, and N. Shadbolt. Contextualising tags in collaborative tagging systems. In *HT '09: Proc. of the 20th ACM conference on Hypertext and hypermedia*, pages 251–260, New York, NY, USA, 2009. ACM.

[39] L. Zhang, X. Wu, and Y. Yu. Emergent semantics from folksonomies: A quantitative study. *Journal on Data Semantics VI*, 2006.