One Tag to Bind Them All: Measuring Term Abstractness in Social Metadata

Dominik Benz*[1], Christian Körner*[2],
Andreas Hotho[3], Gerd Stumme[1], and Markus Strohmaier[2]

[1] Knowledge and Data Engineering Group (KDE), University of Kassel
{benz,stumme}@cs.uni-kassel.de
[2] Knowledge Management Institute, Graz University of Technology
{christian.koerner,markus.strohmaier}@tugraz.at
[3] Data Mining and Information Retrieval Group, University of Würzburg
hotho@informatik.uni-wuerzburg.de

Abstract. Recent research has demonstrated how the widespread adoption of collaborative tagging systems yields emergent semantics. In recent years, much has been learned about how to harvest the data produced by taggers for engineering light-weight ontologies. For example, existing measures of tag similarity and tag relatedness have proven crucial step stones for making latent semantic relations in tagging systems explicit. However, little progress has been made on other issues, such as understanding the different levels of tag generality (or tag abstractness), which is essential for, among others, identifying hierarchical relationships between concepts. In this paper we aim to address this gap. Starting from a review of linguistic definitions of word abstractness, we first use several large-scale ontologies and taxonomies as grounded measures of word generality, including Yago, Wordnet, DMOZ and WikiTaxonomy. Then, we introduce and apply several folksonomy-based methods to measure the level of generality of given tags. We evaluate these methods by comparing them with the grounded measures. Our results suggest that the generality of tags in social tagging systems can be approximated with simple measures. Our work has implications for a number of problems related to social tagging systems, including search, tag recommendation, and the acquisition of light-weight ontologies from tagging data.

Keywords: tagging, generality, measures, emergent semantics, folksonomies

1 Introduction

Since the advent of participatory web applications like Flickr[4], Youtube[5] or Delicious[6], social annotations (especially in the form of collaboratively created

* Both authors contributed equally to this work.
[4] http://www.flickr.com
[5] http://www.youtube.com
[6] http://www.delicious.com

keywords or *tags*) form an integral part of current approaches to collaborative knowledge management. Analyses of the structure of the resulting large-scale bodies of human-annotated resources have shown several interesting properties, especially regarding the presence of *emergent semantics*. Motivated by the vision of bridging the gap towards the Semantic Web, much has been learned in recent years about how to harvest the data produced by taggers for engineering light-weight ontologies. However, little progress has been made on other issues, such as understanding the different levels of tag generality (or tag abstractness), which is essential for e.g. identifying hierarchical relationships between concepts. While several methods of deriving taxonomies from tagging systems have been proposed, a systematic comparison of the underlying notion of abstractness is largely missing.

This paper aims to address this gap by presenting a systematic analysis of various folksonomy-derived notions of term abstractness. Starting from a review of linguistic definitions of word abstractness, we first use several large-scale ontologies and taxonomies as grounded measures of word generality, including Yago, Wordnet, DMOZ and WikiTaxonomy. Then, we introduce and apply several folksonomy-based methods to measure the level of generality of given tags. We evaluate these methods by comparing them with the grounded measures.

Our results show that the abstractness judgments by some of the measures under consideration come close to those of well-defined and manually built taxonomies. Furthermore, we provide empirical evidence that tag abstractness can be approximated by simple measures. The results of this research are relevant to all applications who benefit from a deeper understanding of tag semantics, e.g. ontology learning or clustering algorithms, tag recommendations systems or folksonomy navigation facilities. In addition, our results can help to alleviate the problem of varying "basic levels" in folksonomies [11] by matching more specific terms (used usually by domain experts) to more general ones.

This paper is structured as follows: At first we give an overview about related work, especially regarding term abstractness and emergent semantics (Section 2). This is followed by some basic notions in Section 3. In the subsequent section we give an overview of the introduced measures (section 4) and evaluate them in Section 5 with the help of established datasets as ground truth and a user study. Finally we conclude in Section 6 and point to future work.

2 Related Work

The first research direction relevant to this work has its roots in the analysis of the structure of collaborative tagging systems. Golder and Huberman [11] provided a first systematic analysis, mentioning among others the hypothesis of "varying basic levels" – according to which users use more specific tags in their domain of expertise. However, the authors only provided exemplary proofs for this hypothesis, lacking a well-grounded measure of tag generality. In the following, a considerable number of approaches proposed methods to make the implicit semantic structures within a folksonomy explicit [19, 13, 22, 2]. All of

the previous works comprise in a more or less explicit way methods to capture the "generality" of a tag (e.g. by investigating the centrality of tags in a similarity graph or by applying a statistical model of subsumption) – however, a comparison of the chosen methods has not been given. Henschel et al. [12] claim to generate more precise taxonomies by an entropy filter. In our own recent work [17] we showed that the quality of semantics within a social tagging system is also dependent on the tagging habits of individual users, Heymann [14] introduced another entropy-based tag generality measure in the context of tag recommendation.

From a completely different point of view, the question of which factors determine the generality or abstractness of natural language terms has been addressed by researchers coming from the areas of Linguistics and Psychology. The psychologist Paivio [20] published in 1968 a list of 925 nouns along with human concreteness rankings; an extended list was published by Clark [8]. Kammann [16] compared two definitions of word abstractness in a psychological study, namely imagery and the number of subordinate words, and concluded that both capture basically independent dimensions. Allen et al. [1] identify the generality of texts with the help of a set of "reference terms", whose generality level is known. They also showed up a correlation between a word's generality and its depth in the WordNet hierarchy. In their work they developed statistics from analysis of word frequency and the comparison to a set of reference terms. In [25], Zhang makes an attempt to distinguish the four linguistic concepts fuzziness, vagueness, generality and ambiguity.

3 Basic Notions

As stated above, the main intent of a term generality measure is to allow a differentiation of lexical entities $l_1, l_2, \ldots$ by their degree of abstractness (i.e. their ability to "bind" other tags). As a prerequisite for a formalization of this problem, we will first introduce a common terminology which allows us to refer to the usage of lexical entities in the context of taxonomies and collaborative tagging systems in a unified way.

Taxonomies, Core Ontologies and Lexicons First of all, according to [7] a *taxonomy* can also be regarded as a part of a *core ontology*, [3] $\mathbb{O} := (C, root, \geq_C, L^C, \mathcal{F})$, whereby C is a set of concept identifiers and *root* is a designated root concept for the partial order $\geq_C$ on C. $\geq_C$ is called concept hierarchy or *taxonomy*; if $c_1 \geq_C c_2 (c_1, c_2 \in C)$, then c_1 is a *superconcept* of c_2, and hence we assume c_1 to be more abstract or "general" than c_2. L^C is a set of lexical labels for concepts and a mapping relation $\mathcal{F}$ which associates concepts with their respective label. Please note that a concept c can be associated with one or more labels, i.e. $\forall l \in L^C : |\{l : (c, l) \in \mathcal{F}\}| \geq 1$. As an example, in scientific contexts the terms "article" and "paper" are often used synonymously, which would be reflected by $(c_1, paper) \in \mathcal{F}$ and $(c_1, article) \in \mathcal{F}$, given that c_1 is the concept

identifier of scientific articles. In the literature, one often defines a separate *lexicon* $\mathbb{L} = (L^C, \mathcal{F})$ and associates it with a core ontology [18]; but as it suffices for the context of this work, we assume the lexicon to be an integral part of the ontology itself for the sake of simplicity.

Folksonomies As an alternative approach to taxonomies, collaborative tagging systems have gained a considerable amount of attention. Their underlying data structure is called *folksonomy*; according to [15], 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. $Y \subseteq U \times T \times R$. An element $y \in Y$ is called a *tag assignment* or TAS. A *post* is a triple (u, T_{ur}, r) with $u \in U$, $r \in R$, and a non-empty set $T_{ur} := \{t \in T \mid (u, t, r) \in Y\}$. Intrinsically, concepts are not explicitly present within a folksonomy; however, the set of tags T contains lexical items similar to the vocabulary set L^C of a core ontology.

Term Graphs Both core ontologies and folksonomies introduce various kinds of relations among the lexical items contained in them. A typical example are *tag cooccurrence networks*, which constitute an aggregation of the folksonomy structure indicating which tags have occurred together. Generally spoken, these *term graphs* $\mathbb{G}$ can be formalized as weighted undirected graphs $\mathbb{G} = (L, E, w)$ whereby L is a set of vertices (corresponding to lexical items), $E \subseteq L \times L$ model the edges and $w\colon E \to \mathbb{R}$ is a function which assigns a weight to the edges. As an example, given a folksonomy (U, T, R, Y), one can define the post-based[7] *tag-tag cooccurrence graph* as $\mathbb{G}_{cooc} = (T, E, w)$ whose set of vertices corresponds to the set T of tags. Two tags t_1 and t_2 are connected by an edge, iff there is at least one post (u, T_{ur}, r) with $t_1, t_2 \in T_{ur}$. The *weight* of this edge is given by the number of posts that contain both t_1 and t_2, i.e. $w(t_1, t_2) := \text{card}\{(u, r) \in U \times R \mid t_1, t_2 \in T_{ur}\}$

As we will define term abstractness measures based on core ontologies, folksonomies and term graphs, we will commonly refer to them as *term structures* $\mathbb{S}$ in the remainder of this paper. $L(\mathbb{S})$ is a projection on the set of lexical items contained in $\mathbb{S}$. Based on the above terminology, we now formally define a term abstractness measure in the following way:

Definition 1. *A* term abstractness measure $\sqsupset^{\mathbb{S}}$ *based upon a term structure* $\mathbb{S}$ *is a partial order among the lexical items* L *present in* $\mathbb{S}$*, i.e.* $\sqsupset^{\mathbb{S}} \subseteq L(\mathbb{S}) \times L(\mathbb{S})$*. If* $(l_1, l_2) \in \sqsupset^{\mathbb{S}}$ *(or* $l_1 \sqsupset^{\mathbb{S}} l_2$*) we say that* l_1 *is more abstract than* l_2*.*

In the following, we will make frequent use of *ranking functions* $r\colon L(\mathbb{S}) \to \mathbb{R}$ for lexical items in order to define a tag abstractness measure; please note that a ranking function corresponds to a partial order according to $(l_1, l_2) \in \sqsupset^{\mathbb{S}} \Leftrightarrow r(l_1) > r(l_2)$. We will denote the resulting term abstractness measure as $\sqsupset_r^{\mathbb{S}}$.

[7] Other possibilities are resource-based and user-based cooccurrence; we use post-based cooccurrence in the scope of this work as it is efficiently computable and captures a sufficient amount of information.

4 Measures of Tag Generality

Based on the notions defined above, we will now introduce a set of ranking functions r which are supposed to order lexical items within a folksonomy $\mathbb{F}$ by their degree of abstractness, inducing a partial order $\sqsupseteq_r^{\mathbb{F}}$ among the set of tags.[8] The measures are partially based on prior work in related areas, and build on different intuitions. One commonality they all share is that none of them considers the textual content of a tag itself (e.g. with linguistic methods). All measures operate solely on the folksonomy structure itself or on a derived term network, making them language-independent.

Frequency-based measures A first natural intuition is that more abstract tags are simply used more often, because there exist more resources which they describe – as an example, the number of "computer"s in the world is much larger than the number of "notebook"s; hence one might assume that within a folksonomy, the tag "computer" is used more often than the tag "notebook". We capture this intuition in the abstractness measure $\sqsupseteq_{freq(t)}^{\mathbb{F}}$ induced by the ranking function *freq* which counts the number of tag assignments according to $freq(t) = \text{card}\{(u, t', r) \in Y : t = t'\}$

Entropy-based measures Another intuition stems from information theory: Entropy measures the degree of uncertainty associated with a random variable. Considering the application of tags as a random process, one can expect that more general tags show a more even distribution, because they are probably used at a relatively constant level to annotate a broad spectrum of resources. Hence, more abstract terms will have a higher entropy. This approach was also used by Heymann [14] to capture the "generality" of tags in the context of tag recommendation. We adapt the notion from there and define

$$entr(t) = -\sum_{t' \in cooc(t)} p(t'|t) \log p(t'|t) \tag{1}$$

whereby $cooc(t)$ is the set of tags which cooccur with t, and $p(t'|t) = \frac{w(t',t)}{\sum_{t'' \in cooc(t)} w(t'',t)}$ (with $w(t', t)$ being the cooccurrence weight defined in Section 3). $entr(x)$ induces the term abstractness measure $\sqsupseteq_{entr}^{\mathbb{F}}$.

Centrality Measures In network theory the centrality of a node $v \in V$ in a network G is usually an indication of how important the vertex is [24]. Applied to our problem at hand, centrality can also be contemplated as a measure of abstractness or generality, following the intuition that more abstract terms are also more "important". We adopted three standard centralities (degree, closeness, betweenness). All of them can be applied to a term graph $\mathbb{G}$, leaving us

[8] Note that all term abstractness measures based on real-value ranking functions are by construction total orders, but this is not mandatory.

with three measures $\sqsupset^{\mathbb{G}}_{dc}$, $\sqsupset^{\mathbb{G}}_{bc}$ and $\sqsupset^{\mathbb{G}}_{cc}$ as follows: *Degree centrality* simply counts the number of direct neighbors $d(v)$ of a vertex v in a graph $G = (V, E)$:

$$dc(v) = \frac{d(v)}{|V| - 1} \tag{2}$$

According to *betweenness centrality* a vertex has a high centrality if it can be found on many shortest paths between other vertex pairs:

$$bc(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3}$$

Hereby σ_{st} denotes the number of shortest paths between s and t and $\sigma_{st}(v)$ is the number of shortest paths between s and t passing through v. As its computation is obviously very expensive, it is often approximated [4] by calculating the shortest paths only between a fraction of points. Finally, a vertex ranks higher according to *closeness centrality* the shorter its shortest path length to all other reachable nodes is:

$$cc(v) = \frac{1}{\sum_{t \in V \setminus v} d_G(v, t)} \tag{4}$$

$d_G(v, t)$ denotes hereby the geodesic distance (shortest path) between the vertices v and t.

Statistical Subsumption Schmitz et.al. [22] applied a statistical *model of subsumption* between tags when trying to infer hierarchical relationships. It is based on the assumption that a tag t subsumes another tag t' if $p(t|t') > \xi$ and $p(t'|t) < \xi$ for a suitable threshold ξ. For measuring generality, the number of subsumed tags can be seen as an indicator of abstractness – the more tags a tag subsumes the more general it is:

$$subs(t) = \text{card}\{t' \in T : p(t|t') > \xi) \wedge p(t'|t) < \xi\} \tag{5}$$

5 Evaluation

In order to assess the quality of the tag abstractness measures $\sqsupset^{\mathbb{F}}_{freq}$, $\sqsupset^{\mathbb{F}}_{entr}$, $\sqsupset^{\mathbb{G}}_{dc}$, $\sqsupset^{\mathbb{G}}_{bc}$, $\sqsupset^{\mathbb{G}}_{cc}$ and $\sqsupset^{\mathbb{F}}_{subs}$ introduced above, a natural approach is to compare them against a ground truth. A suitable grounding should yield reliable judgments about the "true" abstractness of a given lexical item. Of special interest are hereby taxonomies and concept hierarchies, whose hierarchical structure typically contains more abstract terms like "entity" or "thing" close to the taxonomy root, whereby more concrete terms are found deeper in the hierarchy. Hence, we have chosen a set of established core ontologies and taxonomies, which cover each a rather broad spectrum of topics. They vary in their degree of controlledness – WordNet (see below) on the one hand being manually crafted by language experts, while the Wikipedia category hierarchy and DMOZ on the other hand are built in a much less controlled manner by a large number of motivated web users. In the following, we first briefly introduce each dataset; an overview about their statistical properties can be found in Table 1.

Table 1: Statistical properties of the datasets used in the evaluation.

Core ontology	$\|C\|$	$\mid \geq_C \mid$	$\|L^C\|$	$\|\mathcal{F}\|$	$\mid \sqsupset^{\mathbb{0}} \mid$
WORDNET	79,690	81,866	141,391	141,692	2,028,925
YAGO	244,553	249,465	206,418	244,553	2,078,788
WIKI	2,445,974	4,447,010	2,445,974	2,445,974	13,171,439
DMOZ	767,019	767,019	241,910	767,019	5,210,226

Folksonomy	$\|U\|$	$\|T\|$	$\|R\|$	$\|Y\|$
DEL (Delicious)	667,128	2,454,546	18,782,132	140,333,714

Term Graphs	$\|T\|$	$\|E\|$
COOC	892,749	38,210,913
SIM	10,000	405,706

5.1 Grounding Datasets

WordNet [9] is a semantic lexicon of the English language. In WordNet, words are grouped into *synsets*, sets of synonyms that represent one concept. Among other relations, the *is-a* relation connects a *hyponym* (more specific synset) to a *hypernym* (more general synset). A synset can have multiple hypernyms, so that the graph is not a tree, but a directed acyclic graph. In order to allow for comparison with the other grounding datasets, we focussed on the noun subsumption network[9]. As it consists of several disconnected hierarchies, it is useful to add a fake top-level node subsuming all the roots of those hierarchies, making the graph fully connected and allowing a relative abstractness judgment between all contained pairs of nouns.

Yago [23] is a large ontology which was derived automatically from Wikipedia and WordNet. Manual evaluation studies have shown that its precision (i.e. the percentage of "correct" facts) lies around 95%. It has a much higher coverage than WordNet (see Table 1), because it also contains named entities like people, books or products. The complete ontology contains 1.7 million entities and 15 million relations; as our main interest lies in the taxonomy hierarchy, we restricted ourselves to the contained *is-a* relation[10] among concepts.

WikiTaxonomy [21] is the third dataset used for evaluation. This large scale domain independent taxonomy[11] was derived by evaluating the semantic network between Wikipedia concepts and labeling the relations as *isa* and *notisa*, using methods based on the connectivity of the network and on lexico-syntactic patterns. It contains by far the largest number of lexical items (see Table 1), but this comes at the cost of a much lower level of manual controlledness.

DMOZ[12] (also known as the open directory project or ODP) is an open content directory for links of the World Wide Web. Although it is hierarchically structured, it differs from the above-mentioned datasets insofar as its internal link structure does not always reflect a sub-concept/super-concept relationship. Despite this fact, we we included the DMOZ category hierarchy as a grounding

[9] `http://wordnet.princeton.edu/wordnet/download/` (v2.1)
[10] `http://www.mpi-inf.mpg.de/yago-naga/yago/subclassof.zip` (v2008-w40-2)
[11] `http://www.h-its.org/english/research/nlp/download/wikitaxonomy.php`
[12] `http://www.dmoz.org/`

dataset because it was built for a similar purpose like many collaborative bookmarking services (namely organizing WWW references). In addition, some of its top level categories (like "arts" or "business") are described by rather abstract terms.

5.2 Tagging Dataset

In order to test the performance of our proposed term abstractness measures, we used a dataset crawled from the social bookmarking system Delicious in November 2006.[13] From the raw data, we first derived the *tag-tag cooccurrence graph* $COOC = (T', E_{cooc}, w_{cooc})$. Two tags t_1 and t_2 are connected by an edge, iff there is at least one post (u, T_{ur}, r) with $t_1, t_2 \in T_{ur}$. The edge weight is given by $w_{cooc}(t_1, t_2) := \text{card}\{(u, r) \in U \times R \mid t_1, t_2 \in T_{ur}\}$. In order to exclude cooccurrences introduced by chance and to enable an efficient computation of the centrality measures, we removed all tags from the resulting graph with a degree of less than 2.

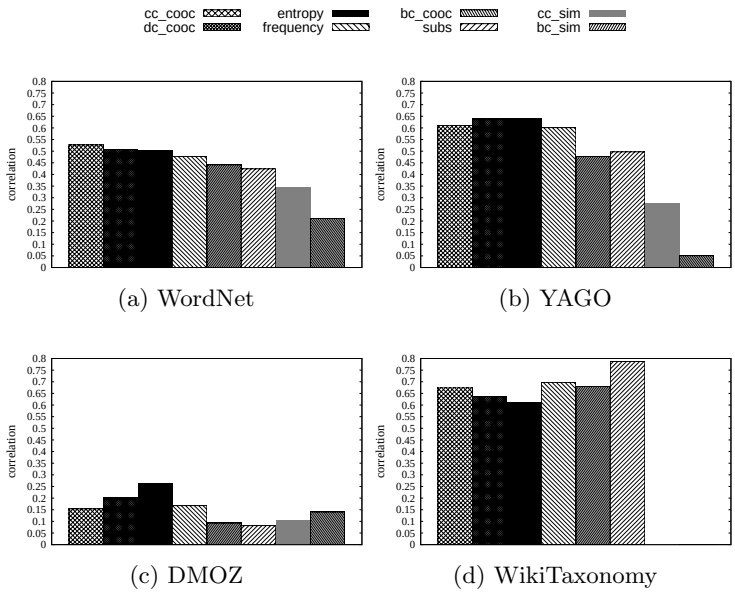
In a similar way to [13], we also derived a *tag-tag similarity graph* $SIM = (T'', E_{sim}, w_{sim})$ by computing the Resource-Context-Similarity described in [5]. The latter is based on a frequency-based representation of tags in the vector space of all resources, in which similarity is computed by the cosine similarity. Because rarely used tags have very sparse vector representations, we restricted ourselves to the 10,000 most frequently used tags. Based on the resulting pairwise similarity values, we added an edge (t_1, t_2) to the edge list E_{sim} when the similarity was above a given threshold $min\_sim = 0.04$. This threshold was determined by inspecting the distribution of all similarity values. Table 1 summarizes the statistics of all tagging datasets.

Subsequently, we computed all term abstractness measures introduced in the previous chapter based on *DEL*, *COOC* and *SIM*, i.e. $\sqsupset^{DEL}_{freq}$, $\sqsupset^{DEL}_{entr}$, $\sqsupset^{COOC}_{dc}$, $\sqsupset^{COOC}_{bc}$, $\sqsupset^{COOC}_{cc}$, $\sqsupset^{SIM}_{bc}$, $\sqsupset^{SIM}_{cc}$ and $\sqsupset^{\mathbb{F}}_{subs}$.

5.3 Direct Evaluation Metric

As stated above, our grounding datasets contain information about concept subsumptions. If a concept c_1 subsumes concept c_2 (i.e. $(c_1, c_2) \in \geq_C$), we assume c_1 to be more abstract than c_2; as the taxonomic relation is transitive, we can infer $(c_1, c_2), (c_2, c_3) \in \geq_C \Rightarrow (c_1, c_3) \in \geq_C$ and hence that c_1 is also more abstract than c_3. In other words, thinking of the taxonomic relation as a directed graph, a given concept c is more abstract than all other concepts contained in the subgraph rooted at c. As we are interested in abstractness judgments about lexical items, we can consequently infer that concept labels for more abstract concepts are more abstract themselves. However, hereby we are facing the problem of polysemy: A given lexical item l can be used as a label for several concepts

[13] The data set is publicly available at `http://www.uni-koblenz-landau.de/koblenz/fb4/AGStaab/Research/DataSets/PINTSExperimentsDataSets/index_html`



(a) WordNet (b) YAGO

(c) DMOZ (d) WikiTaxonomy

Fig. 1: Grounding of each introduced term abstractness measure $\sqsupset^{\mathbb{S}}$ against four ground-truth taxonomies. Each bar corresponds to a term abstractness measure; the y-axis depicts the gamma correlation as defined in Equation 7. (Values for `cc_sim` and `bc_sim` in (d) are -0.05 and -0.005, resp.)

of different abstractness levels. Consequently, l has "several" abstractness levels, depending in which context it is used. As a most simple approach, which removes possible effects of word sense disambiguation techniques, we "resolve" ambiguity in the following way: The abstractness measure $\sqsupset^{\mathbb{O}} \subseteq L^C \times L^C$ on the vocabulary of a core ontology $\mathbb{O}$ is constructed according to

$$(l_1, l_2) \in \sqsupset^{\mathbb{O}} \Leftrightarrow (c_1, l_1) \in \mathcal{F} \wedge (c_2, l_2) \in \mathcal{F} \wedge (c_1, c_2) \in \geq_C \tag{6}$$

whereby $\mathcal{F}$ is the label assigment relation defined in Section 3 . Due to the polysemy effect described above, $\sqsupset^{\mathbb{O}}$ is not necessarily a partial order, as it may contain cycles. But despite this fact, $\sqsupset^{\mathbb{O}}$ contains the complete information which terms $l_i \in L^C$ are more abstract than other terms $l_j \in L^C$ according to the taxonomy of $\mathbb{O}$. Hence we can use it as a "ground truth" to judge the quality of a given term abstractness measure $\sqsupset^{\mathbb{S}}$.

We are interested how well $\sqsupset^{\mathbb{O}}$ correlates to $\sqsupset^{\mathbb{S}}$; picking up the idea of the *gamma rank correlation* [6], we define *concordant* and *discordant* pairs between $\sqsupset^{\mathbb{S}}$ and $\sqsupset^{\mathbb{S}}$ as follows: a pair of terms l and k is called concordant w.r.t. two partial orderings $\sqsupset, \sqsupset_*$, if they agree on it, i.e. $(l \sqsupset k \wedge l \sqsupset_* k) \vee (k \sqsupset l \wedge k \sqsupset_* l)$. It is called discordant if they disagree, i.e. $(l \sqsupset k \wedge k \sqsupset_* l) \vee (k \sqsupset l \wedge l \sqsupset_* k)$. Note that there may exist pairs which are neither concordant nor discordant.

Based on these notions, the gamma rank correlation is defined as

$$CR(\sqsupset, \sqsupset_*) = \frac{|C| - |D|}{|C| + |D|} \tag{7}$$

whereby C and D denote the set of concordant and discordant pairs, respectively.

In our case, $\sqsupset_*$ is not a partial ordering, but only a relation – which means that in the worst case, a pair l, k can be concordant and discordant at the same time. As is obvious from the definition of the gamma correlation (see Eq. 7), such inconsistencies lead to a lower correlation. Hence, our proposed method of "resolving" term ambiguity by constructing $\sqsupset^{\mathbb{O}}$ according to Eq. 6 leads to a lower bound of correlation. Figure 1 summarizes the correlation of each of our analyzed measures, grounded against each of our ground truth taxonomies. First of all, one can observe that the correlation values between the different grounding datasets differ significantly. This is most obvious for the DMOZ hierarchy, where almost all measures perform only slightly better than random guessing. A slight exception is the entropy-based abstractness measure $\sqsupset^{\mathbb{F}}_{entropy}$, which in general gives greater than 0.25 across all datasets. Another relatively constant impression is that the centrality measures based on the tag similarity graph (*cc_sim* and *bc_sim*) show a smaller correlation than the other measures. The globally best correlations are found for the WikiTaxonomy dataset, namely by the subsumption-model-based measure *subs*. Apart from that, the centrality measures based on the tag cooccurrence graph and the frequency-based measure show a similar behavior.

5.4 Derived measures

The grounding approach of the previous section gave a first impression of the ability of each measure to predict term abstractness judgments explicitly present in a given taxonomy. This methodology allowed only for an evaluation based on term pairs between which a connection exists in $\sqsupset^{\mathbb{O}}$, i.e. pairs where l_1 is either a predecessor or a successor of l_2 in the term subsumption hierarchy. However, our proposed measures make further distinctions among terms between which no connection exists within a taxonomy (e.g. the *freq* states that the most frequent term t is more abstract than *all* other terms). This phenomenon can probably also be found when asking humans – e.g. if one would ask which of the terms "art" or "theoretical computer science" is more abstract, most people will probably choose "art", even though both words are not connected by the is-a relation in (at least most) general-purpose taxonomies.

In order to extend our evaluation to these cases, we derived two straightforward measures from a taxonomy which allow for a comparison of the abstractness level between terms occurring in disconnected parts of the taxonomy graph. Because this approach goes beyond the explicitly encoded abstractness information, the question is justified to which extent it makes sense to compare the generality of completely unrelated terms, e.g. between "waterfall" and "chair". Besides our own intuition, we are not aware of any reliable method to determine

Table 2: Results from the user study.

Category	Number of classifications
One tag more general	41
Same level	11
Not comparable	154
Do not know one or two tags	3

when humans perceive the abstractness of two terms as *comparable* or not. For this reason, we validated the derived measures – namely (i) the shortest path to the taxonomy root and (ii)the number of subordinate terms – by an experiment with human subjects.

Shortest path to taxonomy root As stated above, most taxonomies are built in a top-down fashion, whereby more abstract terms are more likely to occur closer to the taxonomy root. Hence, a natural candidate for judging the abstractness of a term is to measure its distance to the root node. This corresponds to a ranking function $sp\_root(l)$, which ranks the terms l contained in a taxonomy in ascending order by the length of the shortest path between *root* and l.

Number of subordinate terms Another measure is inspired by Kammann et al. [16], who stated that "*the abstractness of a word or a concept is determined by the number of subordinate words it embraces[...]*". Given a taxonomy $\mathbb{O}$ and its comprised term subsumption relation $\sqsupset^{\mathbb{O}}$, we can easily determine the number of "sub-terms" by $subgraph\_size(l) = \text{card}\{(l, l') \in \sqsupset^{\mathbb{O}}\}$. We are aware that this measure is strongly influenced e.g. by fast-evolving domains like e.g. "mobile computing", whose rapid growth along with a strong expansion of the included vocabulary might lead to an overestimation of its abstractness level. This is another motivating reason for the user study presented in the next paragraph.

Validation by user study In order to check whether $sp\_root(l)$ and $subgraph\_size(l)$ correspond to human judgments of term abstractness, we performed an exemplary user study with 12 participants[14]. As a test set, we drew a random sample of 100 popular terms occurring in each of our datasets; for each term, we selected 3 candidate terms, taking into account cooccurrence information from the folksonomy *DEL*. The resulting 300 term pairs were shown to the each subject via a web interface[15], asking them to label the pair by one of 5 options (see Table 2)

We calculated Fleiss' κ [10] to get a closer look at the agreement of the study participants. In our experiment, $\kappa = 0.2836$ is indicating fair agreement. Table 2 shows the results of the number of classifications given that an agreement of 6 or more participants signalizes significant agreement. The relatively high number

[14] students and staff from two IT departments

[15] `http://www.kde.cs.uni-kassel.de/benz/generality_game.html`

Table 3: Accuracy of the taxonomy-derived abstractness measures.

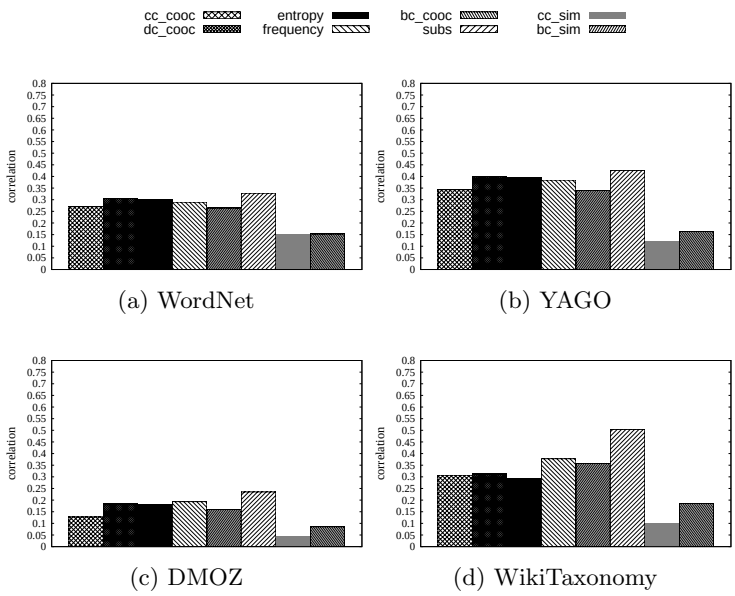
	Wordnet	Yago	DMOZ	WikiTaxonomy
sp_root	0.94	0.42	0.88	0.45
subgraph_size	0.94	0.96	0.8	0.87

of "not comparable" judgments show that even with our elaborate filtering, the task of differentiating abstractness levels is quite difficult. Despite this fact, our user study provided us with a well-agreed set of 41 term pairs, for which we got reliable abstractness judgments. Denoting these pairs as $\sqsupset_{manual}$, we can now check the accuracy of the term abstractness measures introduced by *sp_root* and *subgraph_size*, i.e. the percentage of correctly predicted pairs. Table 3 contains the resulting accuracy values. From our sample data, it seems that the subgraph size (i.e. the number of subordinate terms) is a more reliable predictor of human abstractness judgments. Hence, we will use it for a more detailed grounding of our folksonomy-based abstractness measures.

The ranking function *subgraph_size* naturally induces a partial order $\sqsupset^{\mathbb{O}}_{subgraph\_size}$ among the set of lexical items present in a core ontology $\mathbb{O}$. In order to check how close each of our introduced term abstractness measures correlate, we computed the *gamma correlation coefficient* [6] between the two partial orders (see Eq. 7). Figure 2 shows the resulting correlations. Again, the correlation level between the datasets differs, with DMOZ having the lowest values. This is consistent with the first evaluation based solely on the taxonomic relations (see Figure 1). Another consistent observation is that the measure based on the tag similarity network (*bc_sim* and *cc_sim*) show the weakest performance. The globally best value is found for the subsumption model, compared to the WikiTaxonomy (0.5); for the remaining conditions, almost all correlation values lie in the range between 0.25 and 0.4, and correlate hence weakly.

5.5 Discussion

Our primary goal during the evaluation was to check if folksonomy-based term abstractness measures are able to make reliable judgments about the relative abstractness level of terms. A first consistent observation is that measures based on frequency, entropy or centrality in the tag cooccurrence graph do exhibit a correlation to the abstractness information encoded in gold-standard-taxonomies. One exception is DMOZ, for which almost all measures exhibit only very weak correlation values. We attribute this to the fact that the semantics of the DMOZ topic hierarchy is much less precise compared to the other grounding datasets; as an example, the category `Top/Computers/Multimedia/Music_and_Audio/Software/Java` does hardly imply that `Software` "is a kind of" `Music_and_Audio`. WordNet on the contrary subsumes the term *Java* (among others) under taxonomically much more precise parents: `[...] > communication > language > artifical language > programming language > java` The same holds for Yago, and the WikiTaxonomy was also built with a strong focus on *is-a* relations [21]. This is actually an interesting observation: Despite the fact that both DMOZ and Delicious were built for similar



(a) WordNet (b) YAGO

(c) DMOZ (d) WikiTaxonomy

Fig. 2: Grounding of the term abstractness measure $\sqsupset^{\mathbb{S}}$ against $\sqsupset^{\mathbb{O}}_{subgraph\_size}$ derived from four ground-truth taxonomies. Each bar corresponds to a term abstractness measure; the y-axis depicts the gamma correlation as defined in Equation 7.

purposes (namely organizing WWW references), the implicit semantics within Delicious resembles more closely to well-established semantic repositories than to the bookmark-folder-inspired hierarchical organization scheme of DMOZ.

Another consistent observation is that abstractness measures based on tag similarity graphs (as used e.g. by [13]) perform worst through all experimental conditions. This is consistent with observations in our own prior work [5], where we showed that distributional similarity measures (like the one used in this paper or by [13]) induce connections preferably among tags having the same generality level. On the contrary, applying e.g. centrality measures to the "plain" tag cooccurrence graph yield better results. Hence, a justifiable conclusion is that tag-tag cooccurrence encodes a considerable amount of "taxonomic" information.

But this information is not solely present in the cooccurrence graph – also a probabilistic model of subsumption [22] yields good results in some conditions, especially when grounding against the taxonomy-derived *subgraph_size* ranking. We attribute this to the fact that both measures (the subsumption model and the subgraph size) are based on the same principle, namely that a term is more general the more other terms it subsumes.

Apart from that, even the simplest approach of measuring term abstractness by the mere frequency (i.e. the number of times a tag has been used) already exhibits a considerable correlation to our gold-standard taxonomies. This has an

interesting application to the *popularity/generality problem*: Our results point in the direction that popular tags are on average more abstract (or more general) than less frequently used ones. In summary, the interpretation of our results can be condensed in two statements: First, folksonomy-based measures of term abstractness do exhibit a partially strong correlation to well-defined semantic repositories; and second, the abstractness level of a given tag can be approximated well by simple measures.

6 Conclusions

In this paper, we performed a systematic analysis of folksonomy-based term abstractness measures. To this end, we first provided a common terminology to subsume the notion of term abstractness in folksonomies and core ontologies. We then contributed a methodology to compare the abstractness information contained in each of our analyzed measures to established taxonomies, namely WordNet, Yago, DMOZ and the WikiTaxonomy. Our results suggest that centrality and entropy measures can differentiate well between abstract and concrete terms. In addition, we have provided evidence that the tag cooccurence graph is a more valuable input to centrality measures compared to tag similarity graphs in order to measure abstractness. Apart from that, we also shed light on the *tag generality vs. popularity* problem by showing that in fact, popularity seems to be a fairly good indicator of the "true" generality of a given tag. These insights are useful for all kinds of applications who benefit from a deeper understanding of tag semantics. As an example, tag recommendation engines could take generality information into account in order to improve their predictions, or folksonomy navigation facilities could offer a new direction of browsing towards more general or more specific directions. Finally, our results inform the design of algorithms geared towards making the implicit semantics in folksonomies explicit.

As next steps, we plan to apply our measures to identify generalists and specialists in social tagging systems. A possible hypothesis hereby is that specialists use a more specific vocabulary whereas generalists rely mainly on abstract tags.

Acknowledgments. We would like to thank Dr. Denis Helic and Beate Krause for fruitful discussions during the creation of this work. The research presented in this work is in part funded by the Know-Center, the FWF Austrian Science Fund Grant P20269, the European Commission as part of the FP7 Marie Curie IAPP project TEAM (grant no. 251514), the WebZubi project funded by BMBF and the VENUS project funded by Land Hessen.

References

1. Allen, R., Wu, Y.: Generality of texts. In: Digital Libraries: People, Knowledge, and Technology. Lecture Notes in Computer Science, Springer, Heidelberg (2010)
2. Benz, D., Hotho, A., Stumme, G.: Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge. In: Proc. of WebSci2010. Raleigh, NC, USA (2010)

3. Boszak, E. et al: KAON – Towards a Large Scale Semantic Web. In: Proc. EC-Web 2002, LNCS, vol. 2445, pp. 304–313. Springer (2002)
4. Brandes, U., Pich, C.: Centrality estimation in large networks. I. J. Bifurcation and Chaos 17(7), 2303–2318 (2007)
5. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic analysis of tag similarity measures in collaborative tagging systems. In: Proc. of the 3rd Workshop on Ontology Learning and Population (OLP3). pp. 39–43. Patras, Greece (July 2008)
6. Cheng, W., Rademaker, M., Baets, B.D., Hüllermeier, E.: Predicting partial orders: Ranking with abstention. In: ECML/PKDD. LNCS, vol. 6321. Springer (2010)
7. Cimiano, P., Hotho, A., Staab, S.: Learning concept hierarchies from text corpora using formal concept analysis. Journal of Artificial Intelligence Research (JAIR) 24, 305–339 (2005)
8. Clark, J., Paivio, A.: Extensions of the Paivio, Yuille, and Madigan (1968) norms. Behavior Research Methods, Instruments, & Computers 36(3), 371 (2004)
9. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press (1998)
10. Fleiss, J., et al.: Measuring nominal scale agreement among many raters. Psychological Bulletin 76(5), 378–382 (1971)
11. Golder, S., Huberman, B.A.: Usage patterns of collaborative tagging systems. Journal of Information Science 32(2), 198–208 (April 2006)
12. Henschel, A., Woon, W.L., Wächter, T., Madnick, S.: Comparison of generality based algorithm variants for automatic taxonomy generation. In: Proc. of IIT09. pp. 206–210. IEEE Press, Piscataway, NJ, USA (2009)
13. Heymann, P., Garcia-Molina, H.: Collaborative creation of communal hierarchical taxonomies in social tagging systems. Tech. Rep. 2006-10, CS dep. (April 2006)
14. Heymann, P., Ramage, D., Garcia-Molina, H.: Social tag prediction. In: SIGIR '08: Proc. of the 31st annual Int'l ACM SIGIR conference on Research and development in information retrieval. pp. 531–538. ACM (2008)
15. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. In: The Semantic Web: Research and Applications. LNAI, vol. 4011, pp. 411–426. Springer (2006)
16. Kammann, R., Streeter, L.: Two meanings of word abstractness. Journal of Verbal Learning and Verbal Behavior 10(3), 303 – 306 (1971)
17. Körner, C., Benz, D., Hotho, A., Strohmaier, M., Stumme, G.: Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In: Proc. of WWW2010. pp. 521–530. ACM (2010)
18. Maedche, A.: Ontology Learning for the Semantic Web. Kluwer Academic Publishing, Boston (2002)
19. Mika, P.: Ontologies are us: A unified model of social networks and semantics. In: International Semantic Web Conference. pp. 522–536. LNCS, Springer (2005)
20. Paivio, A., Yuille, J.C., Madigan, S.A.: Concreteness, imagery, and meaningfulness values for 925 nouns. Journal of Experimental Psychology 76 (1968)
21. Ponzetto, S.P., Strube, M.: Deriving a large-scale taxonomy from wikipedia. In: AAAI. pp. 1440–1445. AAAI Press (2007)
22. Schmitz, P.: Inducing ontology from flickr tags (2006)
23. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A Core of Semantic Knowledge. In: 16th international World Wide Web conference (WWW 2007) (2007)
24. Wasserman, S., Faust, K.: Social network analysis: Methods and applications. Cambridge Univ Pr (1994)
25. Zhang, Q.: Fuzziness - vagueness - generality - ambiguity. Journal of Pragmatics 29(1), 13 – 31 (1998)