A

Evaluation of Folksonomy Induction Algorithms

MARKUS STROHMAIER, Graz University of Technology, Austria
DENIS HELIC, Graz University of Technology, Austria
DOMINIK BENZ, University of Kassel, Germany
CHRISTIAN KÖRNER, Graz University of Technology, Austria
ROMAN KERN, Know-Center Graz, Austria

Algorithms for constructing hierarchical structures from user-generated metadata have caught the interest of the academic community in recent years. In social tagging systems, the output of these algorithms is usually referred to as folksonomies (from folk-generated taxonomies). Evaluation of folksonomies and folksonomy induction algorithms is a challenging issue complicated by the lack of golden standards, lack of comprehensive methods and tools as well as a lack of research and empirical/simulation studies applying these methods. In this paper, we report results from a broad comparative study of state-of-the-art folksonomy induction algorithms that we have applied and evaluated in the context of five social tagging systems. In addition to adopting *semantic* evaluation techniques, we present and adopt a new technique that can be used to evaluate the usefulness of folksonomies for *navigation*. Our work sheds new light on the properties and characteristics of state-of-the-art folksonomy induction algorithms and introduces a new pragmatic approach to folksonomy evaluation, while at the same time identifying some important limitations and challenges of folksonomy evaluation. Our results show that folksonomy induction algorithms specifically developed to capture intuitions of social tagging systems outperform traditional hierarchical clustering techniques. To the best of our knowledge, this work represents the largest and most comprehensive evaluation study of state-of-the-art folksonomy induction algorithms to date.

Categories and Subject Descriptors: I2.6 [**Artificial Intelligence**]: Learning – Knowledge Acquisition; H3.7 [**Digital Libraries**]: System issues

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: folksonomies, taxonomies, evaluation, social tagging systems

ACM Reference Format:
Strohmaier, M., Helic, D., Benz, D., Körner C. and Kern, R., 2011. Evaluation of Folksonomy Induction Algorithms. ACM Trans. Intell. Syst. Technol. V, N, Article A (January YYYY), 22 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

In recent years, social tagging systems have emerged as an alternative to traditional forms of organizing information. Instead of enforcing rigid taxonomies with controlled vocabulary, social tagging systems allow users to freely choose so-called tags to annotate resources [Koerner et al. 2010; Strohmaier et al. 2010]. In related research, it has been suggested that social tagging systems can be used to acquire latent hierarchical structures that are rooted in the language and dynamics of the underlying user population [Benz et al. 2010; Heymann and Garcia-Molina 2006; Hotho et al. 2006; Cattuto et al. 2008]. The notion of

...

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

"folksonomies" - from folk-generated taxonomies - emerged to characterize this idea[1]. While a number of algorithms have been proposed to obtain folksonomies from social tagging data [Plangprasopchok et al. 2010a; Heymann and Garcia-Molina 2006; Benz et al. 2010], we know little about the nature of these algorithms, their properties and characteristics. Although measures for evaluating folksonomies exist (such as [Dellschaft and Staab 2006]), their scope is often narrow (i.e. focusing on certain properties only), and they have not been applied widely to state-of-the art folksonomy algorithms. This paper aims to address some of these shortcomings. In this work, we report results from (i) implementing 3 different classes of folksonomy induction algorithms (ii) applying them to 5 different tagging datasets and (iii) comparing them in a study by adopting an array of evaluation techniques. The main contribution of this paper is a broad evaluation of state-of-the-art folksonomy algorithms across different datasets using existing *semantic* evaluation techniques. An additional contribution of our work is the introduction and application of a new, pragmatic technique for folksonomy evaluation that allows to assess the usefulness of folksonomies for navigation. The results presented in this paper highlight some challenges of choosing among different folksonomy algorithms, but also lead to new insights about the properties and characteristics of existing folksonomy induction algorithms, and help to illuminate a path towards future, more effective, folksonomy induction algorithm designs and evaluations. To the best of our knowledge, this work represents the largest and most comprehensive comparative evaluation study of state-of-the-art folksonomy induction algorithms to date.

The paper is structured as follows: In section 2, we give a description of three classes of state-of-the art algorithms for folksonomy induction. Section 3 provides an introduction to semantic evaluation of folksonomies. We present a novel pragmatic (i.e. navigation-focused) approach to evaluating folksonomies in section 4. In section 5, we describe our experimental setup and in section 6 the results of conducting semantic and pragmatic evaluations are presented. Finally, we discuss implications and conclusions of our work.

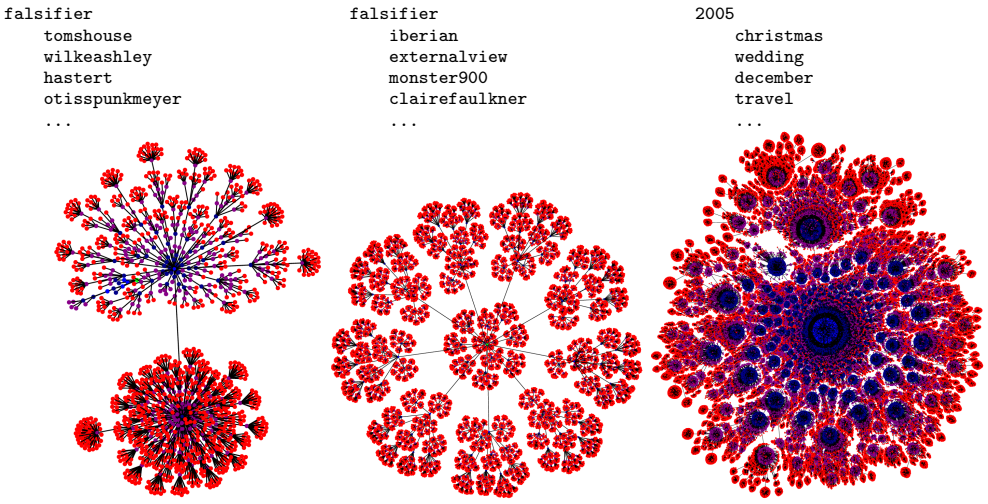
2. FOLKSONOMY INDUCTION ALGORITHMS

While different aspects of emergent semantics have been studied by the tagging research community (see, for example, [Angeletou 2010; Yeung et al. 2008; Au Yeung et al. 2009]), the common objective of folksonomy induction algorithms is to produce hierarchical structures ("folksonomies") from the flat-structured tagging data. Such algorithms analyze various evidence such as tag-to-resource networks [Mika 2007], tag-to-tag networks [Heymann and Garcia-Molina 2006], or tag co-occurrence [Schmitz et al. 2006] to learn hierarchical relations between tags. While further algorithms exist (such as [Li et al. 2007]), we have selected the following three classes of algorithms because (i) they were well documented and (ii) for their ease of implementation. Figures 1 and 2 illustrate exemplary folksonomies and folksonomy excerpts induced by these algorithms. In the following, we briefly describe each considered class of algorithms and how we have applied them in this paper.

2.1. Affinity Propagation

Frey and Dueck introduced Affinity Propagation (AP) as a new clustering method in [Frey and Dueck 2007]. A set of similarities between data samples provided in a matrix represents the input for this method. The diagonal entries (self-similarities) of the similarity matrix are called preferences and are set according to the suitability of the corresponding data sample to serve as a cluster center (called "exemplar" in [Frey and Dueck 2007]). Although it is not

[1]Different definitions for the term "folksonomy" exist in the literature (see for example [Yeung et al. 2008; Plangprasopchok et al. 2010b]). Without necessarily agreeing with either of these definitions, for practical matters, we adopt the view proposed by [Vander Wal 2007] and used by for example [Plangprasopchok et al. 2010b] from here on, where a folksonomy is understood as a "user-created bottom-up categorical structure [...]" [Vander Wal 2007].



(a) Affinity Propagation (b) Hierarchical K-Means (c) DegCen/Cooc

Fig. 1. Examples of folksonomies obtained from tagging data using (a) Affinity Propagation (b) Hierarchical K-Means and (c) Tag Similarity Graph (DegCen/Cooc) algorithms. The different algorithms produce significantly different folksonomies, their semantics or pragmatic usefulness for tasks such as navigation is generally unknown. The visualizations include the top five folksonomy levels of the Flickr dataset. DegCen/Cooc produces hierarchies where most general tags, e.g. 2005, christmas, december occupy the top hierarchy levels (the dataset includes tags from December 2005, see Section 5.1 for detailed dataset description). The color of the root node is green (the root node is visible in (a) and (b)), then the color gradient starts at blue for the top levels and proceeds to red for the lower levels. DegCen/Cooc produces hierarchies that are broader and include more nodes at top levels (more blue in (c)).

required to set a cluster number explicitly, the preference values correlate with the number of resulting clusters (lower preference values result in fewer clusters and vice versa).

In several iterations, AP exchanges messages between data samples to update their "responsibility" and "availability" values. Responsibility values reflect how well data samples serve as exemplars for other data, and the availability values show the suitability of other data samples to be the exemplars for specific data samples. Responsibility and availability are refined iteratively with a parameter λ as an update factor. A full description of AP is beyond the scope of this paper, we point the interested reader to [Frey and Dueck 2007] for further information.

Based on [Frey and Dueck 2007], the authors of [Plangprasopchok et al. 2010a] have introduced an adaption of affinity propagation to infer folksonomies from social tagging data. The authors incorporated structural constraints directly into the global objective function of affinity propagation, so that a tree evolves naturally from execution. In this paper, we follow a simpler approach by applying the original AP recursively in a bottom-up manner. In a first step, the top 10 Cosine similarities (pruned for memory reasons) between the tags in a given dataset serve as the input matrix, and the minimum of those serves as preference for all data samples. Then, AP produces clusters by selecting examples with associated data samples. If the ratio between the number of clusters and the data samples is between 3 and 15 (which we use as an adjustable parameter), then the result will be retained, otherwise another run with lower (too many clusters have been selected) or higher preference values (too few clusters have been selected) will be executed. Finally, the centroids of the clusters are calculated by using the sum of the connected data samples normalized to unit length. Now the Cosine similarities between the centroids serve as the

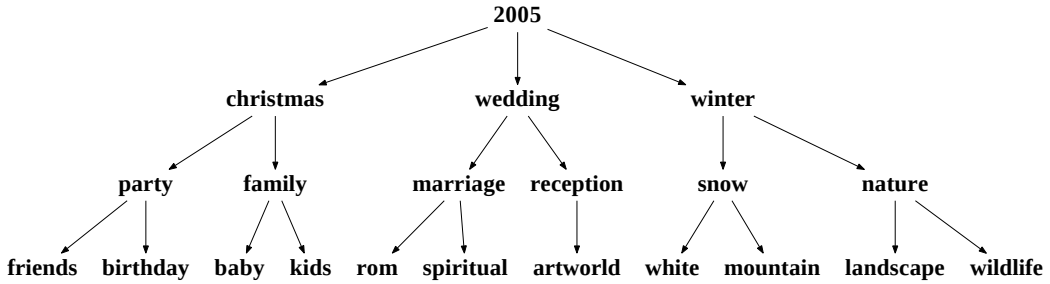


Fig. 2. A small excerpt from the DegCen/Cooc Flickr folksonomy containing the root node and selected other top level nodes.

input matrix for the next run of affinity propagation. This approach is executed until the top-level is reached.

Since our objective is to construct a tag hierarchy where each node represents a unique tag, a tag in each cluster is used as a label. The label is selected by choosing the nearest tag to the centroid. Furthermore, this tag is removed from the actual tags contained in the leaf cluster and is not used as a representative in lower hierarchy levels. We set the AP parameter λ_0 to 0.6 with increasing values depending on the iteration count (i) ($\lambda_i = \lambda_{i-1} + (1.0 - \lambda_0) * i/i_{max}$). AP will terminate after either a maximum of 5000 iterations (i_{max}) or if the exemplars of clusters are stable for at least 10 iterations.

2.2. Hierarchical K-Means

In [Dhillon et al. 2001], the authors introduce an adaption to the k-means algorithm for textual data by optimizing the Cosine similarity instead of Euclidean distance [Dhillon et al. 2001], while [Zhong 2005] introduced an efficient version of an online spherical k-means. Without going into detail, these adaptations allow an online version to be at least as fast as a batch spherical k-means with better results. We utilize k-means iteratively in a top-down manner to build a tag hierarchy. Basically, in the first step, the whole input data set is used for clustering the data into 10 clusters. Our decision to use k=10 was motivated by a desire to capture cognitive limitations of users who are interacting with folksonomies (e.g. to capture the limited ability of users to navigate hierarchies with 100s of children nodes). Clusters containing more than 10 connected samples are further partitioned while ones with less than 10 samples are considered as leaf clusters. However, since a cluster set of 11 samples would also be partitioned into 10 clusters we introduced a special case to give some freedom to the clustering process for these border cases by setting the cluster number to the maximum of 10 or number of data samples divided by 3 what would result in 3 clusters in case of 11 samples. The tag representing a node is selected by taking the nearest tag to the centroid. Furthermore, this tag is removed from the actual tags contained in a cluster and which are further clustered in the next step, if there are more than 10 samples left.

2.3. Generality in Tag Similarity Networks

In [Heymann and Garcia-Molina 2006], the authors introduce an algorithm as an alternative to producing hierarchical structures from tagging data by means of hierarchical clustering approaches. The input for the algorithm is a so-called *tag similarity network* – an unweighted network where each tag is a node in the network, and two nodes are linked to each other if their similarity is above a predefined similarity threshold. In the simplest case, the threshold is defined through tag overlap – if the tags do not overlap in at least one resource then they

Table I. Statistical Properties of the Induced Folksonomies

	BibSonomy		*CiteULike*		*Delicious*		*Flickr*		*LastFM*	
	br	*dia*	*br*	*dia*	*br*	*dia*	*br*	*dia*	*br*	*dia*
Affprop	3.36	6	3.42	4	3.61	5	3.23	4	3.99	4
Clo/Cos	2.21	13	2.1	16	2.46	12	2.25	13	2.05	10
Deg/Cooc	8.04	8	6.82	9	8.14	9	9.17	7	6.25	7
KMeans	3.78	6	3.81	10	3.71	36	3.81	5	3.72	17
Random	10	2	10	3	10	2	10	4	10	3

Source: Statistical properties of the induced folksonomies by all proposed methods. *br* depicts the average branching factor, computed over all non-leaf nodes (For comparison, the branching factors of the reference taxonomies are: WordNet 4.86, Yago 14.32, Wikitaxonomy 48.82). *dia* depicts the full network diameter, based on 500 randomly selected nodes (For comparison, the diameters of the reference taxonomies are: WordNet 7, Yago 7, Wikitaxonomy 2).

are not linked to each other in the tag similarity network. The second prerequisite for the algorithm is the ranking of nodes in a descending order according to how central the tags are in the tag similarity network. In particular, this ranking produces a generality order where the most general tags from a dataset are in the top positions. The algorithm starts by a single node tree with the most general tag as the root node. The algorithm then proceeds by iterating through the generality list and adding each tag to the tree – the algorithm calculates the similarities between the current tag and each tag currently present in the tree and adds the current tag as a child to its most similar tag. The authors describe their algorithm as extensible as they leave the possibility to apply different similarity, as well as different centrality measures. The presented algorithm adopts cosine similarity and closeness centrality, and we denote this algorithm henceforth CloCen/Cos.

In [Benz et al. 2010], the authors describe an extension of the algorithm presented in [Heymann and Garcia-Molina 2006]. Generally, this new algorithm is based on principles similar to Heymann's algorithm – but the new algorithm applies tag co-occurrence as the similarity measure and the degree centrality as the generality measure (DegCen/Cooc). In particular, the algorithm executes an extensive preprocessing of the dataset e.g. to remove synonym tags or to resolve ambiguous tags. For this paper, we study both published variations of these algorithms: CloCen/Cos and DegCen/Cooc. For reasons of simplicity, we skipped preprocessing of the dataset and only applied the alternative similarity and centrality measures. Table I summarizes some statistical properties of all resulting folksonomies.

In the following, we briefly discuss and present state-of-the art evaluation techniques for folksonomy algorithms. Specifically, we review current *semantic* evaluation techniques in section 3 and present a new *pragmatic* approach to folksonomy evaluation in section 4.

3. SEMANTIC EVALUATION

A universal measure for evaluating the overall semantic quality of a taxonomy is difficult to envision or design for several reasons. First, taxonomies are usually constructed not only for the mere purpose of representing knowledge, but they are often targeted towards a particular application. Depending on the application type and the anticipated user population, different aspects of the taxonomy will be more or less important, which should be reflected in any evaluation approach. If, for example, there exists a direct interface where humans work with the taxonomy, the quality of the lexical labels to describe concepts will be more important than in a case where the taxonomically captured knowledge serves only as an input for an automatic process. Hence there exist different evaluation paradigms focusing on assessing different aspects of a learned taxonomy. In general, we can broadly distinguish between three evaluation paradigms (also c.f. [Dellschaft and Staab 2006; Brank et al. 2006]):

Table II. Statistical Properties of the Reference Datasets

	#concepts	#relations	#labels
Wordnet	79,690	81,866	141,391
Yago	244,553	249,465	206,418
Wikitaxonomy	2,445,974	4,447,010	2,445,974

Source: Statistical properties of the reference datasets used in the semantic evaluation.

— *Application-centered*: When taxonomies are engineered towards a certain application, a natural measure of taxonomy quality would be the performance improvement achieved by using different taxonomies as input. A requirement hereby is the existence of measures to compare the achieved results. Though this paradigm reflects clearly the actual "utility" of a taxonomy, a problematic issue is how to disentangle the influence of the input taxonomy from other application parameters.
— *Human Assessment*: This paradigm relies on the judgement of human experts how well an taxonomy meets a set of predefined criteria. Hereby it is obviously an important question on which criteria to agree. This paradigm can be expected to provide valuable assessments of taxonomy quality at a high cost due to the heavy involvement of human interaction.
— *Reference-based*: The prerequisite of this methodology is the existence of a "gold-standard", to which the learned taxonomy can be compared. The gold standard can be an taxonomy itself, but also e.g. a set of documents covering the domain in question. The key issues hereby are how to assess the quality of the gold-standard itself, and the establishment of valid comparison measures.

Comparing the paradigms, [Dellschaft 2005] concludes that only reference-based methods are practically feasible for large-scale and frequent evaluation. We will adopt this approach as an initial semantic evaluation methodology in the scope of our work. As an additional check for the validity of this methodology in the context of our work, we also performed a human subject experiment where we asked human subjects to judge the quality of a subset of learned hierarchical relationships. We will now first provide details on the reference-based evaluation, and then explain the setup of our human subject experiment.

3.1. Reference-based Evaluation

When adopting a reference-based evaluation paradigm, it is a non-trivial task to judge the similarity between a learned concept hierarchy and a reference hierarchy, especially regarding the absence of well-established and universally accepted evaluation measures. This typically requires researchers to find answers to at least two crucial questions: (i) Which reference (gold-standard) ontology to choose, and (ii) which measure to use to compute the similarity between the learned and the gold-standard ontology. In order to support a comparative evaluation of all the folksonomy induction algorithms presented earlier, we have chosen a set of rather general reference datasets – i.e. taxonomies derived from WordNet, Yago and Wikipedia (see below). The reason for that lies in the significant vocabulary overlap that we found between the folksonomies and these reference datasets. Other reference datasets, such as MusicMoz[2] or the ACM Classification Schema[3], did not produce sufficient vocabulary overlap for comprehensive evaluation. Particular vocabulary matching scores are presented in Table IV. In the following, we briefly describe each of the gold standard taxonomies used in our work, and then proceed with the presentation of the evaluation measures adopted by this paper.

[2] `http://musicmoz.org/`
[3] `http://www.acm.org/about/class/`

— *WordNet* [Miller 1995] is a structured lexical database of the English language. It contains roughly 203.000 terms grouped into 115.400 synsets. Among the synsets, several relations are defined; one of the most important ones is the taxonomic relation. As a first gold-standard, we extracted the taxonomic hierarchy among synsets in WordNet.
— *Yago* [Suchanek et al. 2007] 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 II), 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[4] among concepts.
— The *"Wikitaxonomy"* [Ponzetto and Strube 2007] is the third dataset used for evaluation. This large scale domain independent taxonomy[5] 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 II), but this comes at the cost of a lower level of manual control.

Starting from several gold-standard taxonomies, the next task is to judge the similarity between a learned taxonomy $\mathcal{F}$ and a reference taxonomy $\mathcal{T}$. Finding a universally applicable, valid similarity score for two (possibly very large) hierarchical structures is non-trivial. Yet, a number of useful measures have been proposed by past research. Dellschaft et al. [Dellschaft and Staab 2006] for example propose two measures, i.e. *taxonomic precision* and *taxonomic recall* for this purpose. The basic idea is hereby to find a concept c present in both taxonomies, and then to extract a characteristic excerpt (consisting e.g. from the sub- and super-concepts) from both taxonomies, i.e. $ce(c, \mathcal{F})$ and $ce(c, \mathcal{T})$. If both excerpts are very similar, then the location of the concept c in both taxonomies is similar. Hence, taxonomic precision and recall have a local part tp and tr, respectively, according to:

$$tp(c, \mathcal{F}, \mathcal{T}) = \frac{|ce(c, \mathcal{F}) \cap ce(c, \mathcal{T})|}{|ce(c, \mathcal{F})|} \qquad tr(c, \mathcal{F}, \mathcal{T}) = \frac{|ce(c, \mathcal{F}) \cap ce(c, \mathcal{T})|}{|ce(c, \mathcal{T})|}$$

Then, all local values are summed up over the concept overlap between both structures according to:

$$TP(\mathcal{F}, \mathcal{T}) = \frac{1}{|C_{\mathcal{F}} \cap C_{\mathcal{T}}|} \sum_{c \in C_{\mathcal{F}} \cap C_{\mathcal{T}}} tp(c, \mathcal{F}, \mathcal{T})$$

Whereby $C_{\mathcal{F}}$ denotes the set of concepts in the learned folksonomy and $C_{\mathcal{T}}$ the set of concepts of the reference taxonomy. TR is computed analogously. Finally the taxonomic F-measure is computed as the harmonic mean of taxonomic precision and recall according to $TF(\mathcal{T}, \mathcal{F}) = \frac{2 \cdot TP(\mathcal{T}, \mathcal{F}) \cdot TR(\mathcal{T}, \mathcal{F})}{TP(\mathcal{T}, \mathcal{F}) + TR(\mathcal{T}, \mathcal{F})}$.

The same idea underlies the measure of *taxonomic overlap* proposed by Maedche [Maedche 2002]; its local and global part are computed according to:

$$to(c, \mathcal{F}, \mathcal{T}) = \frac{ce(c, \mathcal{F}) \cap ce(c, \mathcal{T})}{ce(c, \mathcal{F}) \cup ce(c, \mathcal{T})}$$

[4] `http://www.mpi-inf.mpg.de/yago-naga/yago/subclassof.zip` (v2008-w40-2)
[5] `http://www.h-its.org/english/research/nlp/download/wikitaxonomy.php`

$$TO(\mathcal{F},\mathcal{T}) = \frac{1}{|c \in C_{\mathcal{F}} \cap C_{\mathcal{T}}|} \sum_{c \in C_{\mathcal{F}} \cap C_{\mathcal{T}}} to(c,\mathcal{F},\mathcal{T})$$

In all cases, an important aspect is the composition of the characteristic excerpt *ce*. A common approach is to choose the *semantic cotopy* [Maedche 2002], which consists of all sub- and superconcepts of a given concept c and *the concept itself*. Because all local measures *tp*, *tr* and *to* are based on the intersection of excerpts, adding the concept c to each cotopy leads to a "trivial hit" - i.e. $ce(c,\mathcal{F}) \cap ce(c,\mathcal{T}) >= 1$ in all cases. This has an especially strong effect when the average size of the excerpts is small, which happens e.g. in rather shallow hierarchies. We first used the semantic cotopy as characteristic excerpt, but with limited success - because especially the randomly generated folksonomies were strongly favored by this method due to their inherent shallow structure. For this reason, we used another excerpt, i.e. the *common semantic cotopy* (as defined in [Dellschaft and Staab 2006]). It basically contains the sub- and superconcepts of c which are present in both taxonomies, but *excluding the concept c itself*. This choice eliminates the problematic "trivial hit", leading to much more useful results.

While these measures have not been applied widely, they are theoretically sound and interesting. This makes them promising candidates for the folksonomy evaluation study at hand. We will adopt all measures for our evaluation, i.e. taxonomic precision, recall, F1-measure and overlap. As an additional check for the validity of these measures, we performed a small human subject experiment, which will be introduced next.

3.2. Evaluation by human assessment

Although the human ability to interpret and integrate information in a meaningful way can surely be seen as superior to current automatic approaches, the task of evaluating the "quality" of a learned hierarchical structure remains challenging even for skilled subjects. Especially the manual comparison of two (potentially very large and complex) taxonomies will probably not lead to consistent and reproducable evaluation results. For this reason, we have chosen a simpler approach targeted towards the assessment of the consistency of each learned taxonomy. Our basic idea hereby was to sample a subset of all direct taxonomic subsumption pairs from a learned hierarchy, and then to let humans judge if (and if yes - how) the two contained terms are related. We used a web interface to present each human subject one term pair (A, B) at a time, asking *"What's the relation between the two terms A and B?"*. As an answer, the subject could choose between selecting one of the following options:

(1) *A is the same as B.*
(2) *A is a kind of B.*
(3) *A is a part of B.*
(4) *A is somehow related to B.*
(5) *A is not related to B.*
(6) *I don't know the meaning of A or B.*

In order to allow as many meaningful answers as possible from a broad audience, we performed an a-priori filtering of the term pairs by a list of "common" words, namely the 5.000 nouns which were used most often in the Brown corpus[6]. We only kept those pairs (A, B) as candidates for the study where both terms A and B were present in this list of popular nouns.

[6]This corpus was compiled in 1960 and contains roughly 2 million words from a general set of English texts (see `http://khnt.aksis.uib.no/icame/manuals/brown/`)

The intuition behind this approach is that a "better" taxonomy will yield a lower percentage of pairs being judged as unrelated. The reason why we allowed for a further distinction of relations (i.e. "same as", "kind of", "part of" and "somehow related") is that we do not expect our analyzed algorithms to produce exclusively semantically sharp taxonomic (i.e. "kind of") relations. Our semantic evaluation methodology will be complemented by pragmatic evaluation measures, which are introduced next.

4. PRAGMATIC EVALUATION

While semantic evaluation of hierarchical structures in social tagging systems has received some attention in the literature, pragmatic (i.e. task-oriented) evaluation represents a rather new aspect [Helic and Strohmaier 2011; Helic et al. 2011]. In the following, we introduce a novel way to evaluate the usefulness of folksonomies for user tasks in social tagging systems.

One way of assessing the suitability of folksonomies in supporting user tasks is to assess their usefulness for searching or navigating social tagging systems. Following this line of thought, we can measure the extent to which a folksonomy aids a user in navigating the system. This is the approach employed in this paper. Instead of observing real user behavior, our method of choice is simulation, mainly because current tagging systems do not adopt folksonomy-based navigational support yet and simulation provides us with better experimental control and thus makes it possible to evaluate different folksonomy constructing algorithms across multiple datasets. In the following, we shortly introduce our simulation model and its theoretical background.

4.1. Greedy Search and Network Navigability

One of the research questions attracting a lot of interest in the field of networks is the relation between network structure and function, such as the relation between the structure and routing function of a network. Ever since the "small world" experiment [Milgram 1967] conducted by Stanley Milgram, researchers have been intrigued by the *routing efficiency* or *navigability* question in social networks – how people are able to find unknown people who are, potentially, geographically and socially distant to themselves. The key aspect of this question is the *absence of the global knowledge of the network* – people know only their friends and therefore *posses only the local knowledge of the network* but are still able to find unknown people. Similar navigability has been observed in other real networks such as metabolic or neural networks, or has been an important design goal for engineers of communicational networks such as the Internet or different peer-to-peer networks (see e.g. [Adamic et al. 2001]). Researchers identified the concept of *similarity between nodes* [Watts et al. 2002; Menczer 2002; Leicht et al. 2006] or more generally the concept of *distance between nodes* [Kleinberg 2000a; 2000b; 2001; Watts et al. 2002; Adamic and Adar 2005] as an important aspect of establishing networking navigability. Combining the notion of distance between nodes with the algorithmic term of *greedy routing* [Kleinberg 2000b], Kleinberg theoretically explained network navigability [Kleinberg 2000a; 2001] in the following way: nodes use distance to select the next node in a routing session and the greedy algorithm selects the adjacent node closest (with the smallest distance) to the current destination node. The algorithm and its applications have been studied in the recent literature, see e.g. [Kleinberg 2006].

In [Serrano et al. 2008] the authors abstract the notion of distance as introduced by Kleinberg to a *hidden distance* between nodes. Hidden distances define a *hidden metric spaces* which governs not only routing in the network but also the network formation and emergence of network structural properties such as power-law degree distributions and high node clustering. The authors connect observable emergent structural properties of a network with its navigability by defining a region of navigable networks in two dimensional space with clustering-coefficient [Watts and Strogatz 1998] and power-law exponent as dimensions. On the other hand, a hidden metric space is also a geometric entity in which nodes are

identified by their co-ordinates in it – distance between nodes is their geometric distance in that particular metric space. An interesting research question is the structure of such hidden metric spaces that underlie observable networks. In [Boguñá et al. 2009], the authors introduce a model with the circle as a hidden metric space and show its effects on routing in the global airport network. In [Krioukov et al. 2010] the authors discuss hyperbolic geometry as a hidden metric space whereas in [Boguñá et al. 2010] the authors apply hyperbolic geometry as a model of the hidden metric space of the Internet and design a novel greedy Internet routing algorithm.

The relation between Kleinberg's node distance and the recent work on hidden metric spaces can easily be established. In Kleinberg's model, the nodes are organized into a hierarchy according to their similarity – the distance between two nodes corresponds then to the height of their least common ancestor in that hierarchy [Kleinberg 2001] (Adamic [Adamic and Adar 2005] and Watts [Watts et al. 2002] have similar distance definitions that are also based on the node distance in one or more hierarchies). Hyperbolic geometry, as well as a hierarchy, distribute distances exponentially – it is, therefore, possible to approximate a hyperbolic metric space by a tree [Krioukov et al. 2010].

4.2. Pragmatic Evaluation Method

In the first step of our folksonomy evaluation, we generate tag-to-tag networks from different tagging datasets. We adopt a model of a tagging dataset as a tripartite hypernetwork with $V = R \cup U \cup T$, where R is the resource set, U is the user set, and T is the tag set [Cattuto et al. 2007; Schmitz et al. 2006; Ramezani et al. 2009]. An annotation of a particular resource with a particular tag produced by a particular user is a hyperedge (r, t, u), connecting three nodes from these three disjoint sets. Such a tripartite hypernetwork can be mapped onto three different bipartite networks connecting users and resources, users and tags, and tags and resources, or onto e.g. tag-to-tag networks. For different purposes it is often more practical to analyze one or more of these networks. For example, in the context of ontology learning, the bipartite networks of users and tags has been shown to be an effective projection [Mika 2007]. In this paper, we focus on navigating the tag-to-tag network (based on a tag-to-resource network), to mimic a tag-based user navigation task.

In the second step, we construct different folksonomies from a number of tagging datasets, where we apply the algorithms that we have introduced in Section 2.

In the final step, we adopt a folksonomy as a particular incarnation of a hidden metric space. We simulate greedy routing through the observable tag-to-tag network querying the folksonomy for node distances – the idea is that greedy routing will be more successful if the co-ordinates imposed by a folksonomy are closer to the real hidden metric space of the network in question. We quantify the quality of a folksonomy by measuring the success rate of the greedy algorithm (the number of successfully reached destination nodes divided by the total number of routing sessions), and by the stretch, which is the ratio of the average greedy hops to average shortest paths (this measure tells us how longer are greedy paths as compared to global shortest paths). The measures are similar to those introduced in [Boguñá et al. 2010]. In addition to the global values calculated in [Boguñá et al. 2010], we calculate the measures for each observable shortest path in the networks. The folksonomies that perform better, i.e. folksonomies where the success rate is higher better reflect the underlying hidden metric space and therefore are more suitable for instructing greedy routing. Stretch value is a control value – achieving values close to 1 means that folksonomies are good at finding shortest paths quickly, i.e. in an almost optimal way. On the other side, high strecth values, e.g. 2 or more would mean that greedy search takes often sub-optimal paths and that the folksonomy in question does not represent the hidden metric space optimally. Without making assumptions about actual user behavior, we can conclude theoretically that better performing folksonomies would provide a better navigation support

Table III. Datasets

	BibSonomy	*CiteULike*	*Delicious*	*Flickr*	*LastFM*
Tags	56,424	347,835	380,979	395,329	281,818
Links	2,003,986	27,536,381	39,808,439	17,524,927	84,787,780

Source: Statistical properties of the tag-tag-networks derived from five social tagging systems.

for users. We leave the task of testing whether this conclusion also holds in practice, e.g. with actual user behavior, to future work.

5. EXPERIMENTAL SETUP

In our experiments, we apply 4 folksonomy induction algorithms (from 3 distinct classes) to five different social tagging systems yielding 20 different folksonomies. We evaluate these 20 folksonomies on a semantic level against 3 reference datasets, and on a pragmatic level against the task of navigation on the underlying tag network structure. The detailed experimental setup is presented next.

5.1. Datasets

Data from the following social tagging systems was used as an empirical basis (see Table III for an overview):

Dataset BibSonomy. This dataset[7] contains nearly all 916,495 annotations and 235,340 unique resources (scientific articles) from a dump of BibSonomy [Hotho et al. 2006] until 2009-01-01. The tag-tag network comprises 56,424 nodes and 2,003,986 links.
Dataset CiteULike. This dataset contains 6,328,021 annotations and 1,697,365 unique resources (scientific articles) and is available online[8]. The tag-tag network consists of 347,835 tags and 27,536,381 links.
Dataset Delicious. This dataset is an excerpt from the PINTS experimental dataset[9] containing a systematic crawl of Delicious and Flickr in 2006 and 2007. We extracted all data from November 2006. The resources in this dataset are Web addresses. The tag-tag network consists of 380,979 tags and 39,808,439 links.
Dataset Flickr. This dataset is also an excerpt from the PINTS Flickr crawls. It contains the data from December 2005. The resources in Flickr are user-generated photos. The tag-tag network consists of 395,329 tags and 17,524,927 links.
Dataset LastFm. This dataset is from [Schifanella et al. 2010]. It contains annotations that were crawled from the last.fm website in the first half of 2009. The resources in this dataset are songs, artists and albums. The tag-tag network consists of 281,818 tags and 84,787,780 links.

5.2. Semantic Evaluation

While our reference-based semantic evaluation adopts the measures presented in Section 3.1, for the human subject experiment we first extracted all subsumption pairs containing "common" terms (as described also in Section 3) present in each folksonomy induced from the Flickr dataset. We focussed on this dataset because its scores in the reference-based evaluation were comparatively high, and data from this system was used in related work on folksonomy induction before [Plangprasopchok et al. 2010a]. From the resulting sets of candidate pairs, we randomly selected 25 pairs for each folksonomy induction algorithm under consideration, leading to 125 term pairs. As a control condition, we also added 25

[7]http://www.kde.cs.uni-kassel.de/ws/dc09/
[8]http://www.citeulike.org/faq/data.adp
[9]https://www.uni-koblenz.de/FB4/Institutes/IFI/AGStaab/Research/DataSets/PINTSExperimentsDataSets/

term pairs randomly sampled from one of our reference hierarchies (namely the WordNet noun taxonomy), leading to a total number of 150 term pairs to be judged for each of our subjects. We then sent a link[10] pointing to the online study to students and staff from our two IT departments. In summary, 27 persons took part in the evaluation. Because some of them did not completely finish the rating of all pairs, we received 3,381 votes, including 249 "don't know" choices – leading to a total of 3,132 useful answers for our study. In order to consider only pairs for which we have a sufficient amount of votes, we only included those tag pairs for which at least 18 subjects had provided useful answers. This left us with a final set of 128 term pairs. For each term pair, we computed the fraction of each possible judgement, and averaged these values subsequently over each folksonomy induction algorithm. Figure 4 shows the results. Apart from this, pragmatic evaluation was adopted in the following way:

5.3. Pragmatic Evaluation Using Greedy Search

With greedy search we model and then simulate navigation in tagging systems. We select 100,000 resource nodes uniformly at random from the bipartite tag-to-resource tagging network. Each of these nodes represents a *starting node* for decentralized search, modeling an arbitrary user entry page into the system (e.g. a landing page from a search engine, the latest resource from a news feed, homepage, or similar). We assume that users who come to the tagging system would *explore* the system to find one or more related topics or resources of current interest. To model this, we select another resource node from the tagging network uniformly at random. Tags associated with the second resource are both related to each other (they overlap at least at the second resource) and represent a collection of related resources that a user might be interested in. We define the set of resources connected by those tags as *target nodes* for the greedy search. The goal of the agent is to find a short path from the starting node to one of the target nodes in the search pair.

We use *length of the shortest path* as the reference point in the evaluation. This reflects a typical scenario of navigation in tagging systems – the user will explore the tagging system by navigating to find relevant topics and resources *as quickly as possible*, i.e., with as few clicks as possible. We calculate the global shortest path between nodes from each search pair using breadth first search. If there is no global path between nodes from a pair (i.e. when one of the target nodes does not belong to the giant component) then this node is removed from future calculations.

The folksonomy is applied as a hidden metric space to provide the distance between nodes. Although search starts at a resource node, as soon as the first tag is selected, the search becomes a search in the tag-to-tag network. Search is considered successful if the algorithm finds at least one of the target tags. To model users behavior during navigation we apply the following strategy: if the agent arrives at a certain node for the second time, the search stops and is counted as a failure (no backtracking) – this mimics the situation where a user arrives at a tag that he has already visited, and then decides to, e.g., switch to the search field or to leave the system. The success rate s of the greedy search thereby provides an answer to the question of the pragmatic suitability of a folksonomy to support navigation. In addition to the success rate we calculate so-called stretch τ [Krioukov et al. 2010] with $\overline{h}$ (average greedy hop length) and $\overline{l}$ (average shortest path length) as:

$$\tau = \frac{\overline{h}}{\overline{l}} \tag{1}$$

To obtain a baseline (a lower bound) for the performance of a particular folksonomy, we also apply a random folksonomy as a hidden metric space. The results of applying semantic and pragmatic evaluation are introduced next.

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

Table IV. Lexical Overlap Among Concepts

	BibSonomy	*CiteULike*	*Delicious*	*Flickr*	*LastFM*
WordNet	8,680	22,380	21,830	23,480	10,790
Yago	5,970	14,180	13,620	13,770	6,450
Wikitaxonomy	11,280	33,430	40,270	37,950	14,900
ACM	170	400	n/a	n/a	n/a
MusicMoz	n/a	n/a	n/a	n/a	330

Source: Lexical overlap among concepts present in the folksonomies and taxonomies. The values are approximated, as some folksonomy induction algorithms led to slight variations of the overlap, but to a negligible amount (+/- 100 concepts). WordNet, Yago and Wikitaxonomy exhibit significant overlap with nodes from the learned folksonomies, while ACM Classification System and MusicMoz exhibit little overlap.

6. RESULTS

6.1. Results of Semantic Evaluation

As a first result, we present the vocabulary overlap between concepts present in the folksonomies and those in selected reference datasets (see Table IV). While the overlap is significant for WordNet, Yago and Wikitaxonomy, it is extremely small for ACM and MusicMoz. Due to the small overlap, we discarded ACM and MusicMoz from all subsequent investigations, and focused our evaluations on WordNet, Yago and Wikitaxonomy.

Figure 3 displays the results of the reference-based semantic evaluation. On the y-axis of each figure, the similarity between each folksonomy and a reference gold-standard taxonomy is depicted. We measure similarity using different measures, including taxonomic precision (TP), taxonomic recall (TR), taxonomic F1-measure (TF) and taxonomic overlap (TO). As explained in Section 3.1, all these measures are based on the comparison of "characteristic excerpts" from both hierarchical structures. The local values are then summed up and averaged into a global value.

At a first glance, the results from our experiments convey a consistent picture: Taking the taxonomic F1-measure (black bars) as an example, one can observe that across almost all experimental conditions the folksonomies induced by generality-based methods (Clo/Cos and Deg/Cooc in the figures) outperform the clustering-based ones (Affprop and Kmeans). A similar distribution is found for the other measures (TP, TR and TO). In all cases, the folksonomy induced by the random algorithm performs worst and yields a similarity score of close to zero.

A slight exception to these first observations are the folksonomies induced from the LastFM dataset (lowermost row), for which e.g. affinity propagation slightly outperforms the generality-based Clo/Cos algorithm. However, the general level of similarity is much lower for all folksonomies based on this dataset. We attribute this to the fact that the LastFM dataset has a relatively strong topical focus, i.e. the tagging of music-related items like songs, artists or albums. Our choice of gold-standard taxonomies, however, was targeted towards topically more general hierarchies in order to enable a comparison across diffferent datasets. Our results suggest that this choice makes sense for thematically "general-purpose" tagging systems like BibSonomy, CiteULike, Delicious or Flickr, but is less well-suited for more specific ones like LastFM. We also experimented with domain-specific taxonomies like the ACM Computing classification system[11] which might be better suitable for BibSonomy and CiteULike, as well as with a music genre taxonomy derived from MusicMoz[12] fitting obviously to LastFM – but due to the relatively small lexical overlap, we also had limited success to this end. Hence we will focus in the remaining discussion of the results on our more general datasets (topmost four rows).

[11] `http://http://www.acm.org/about/class/`
[12] `http://www.musicmoz.org`

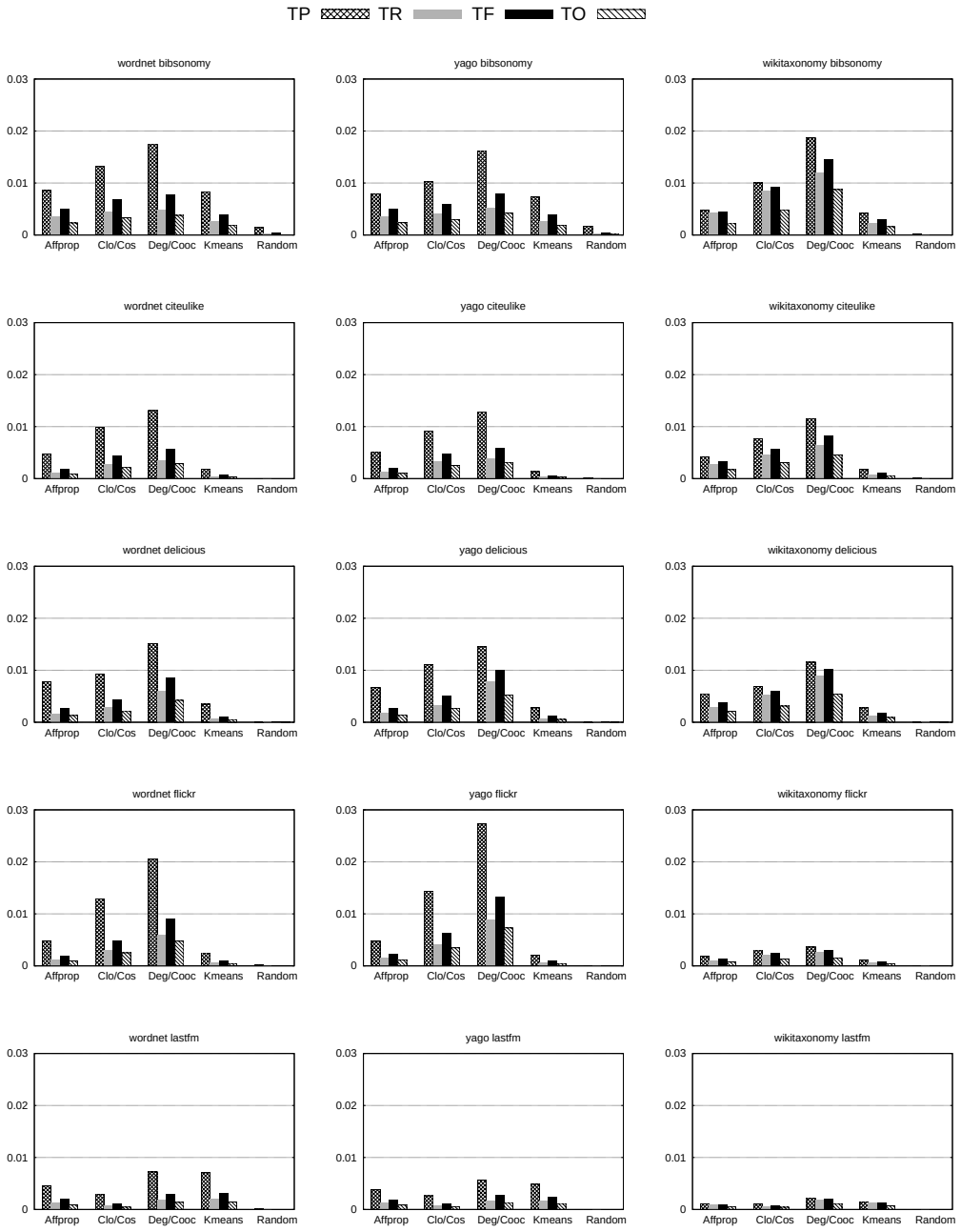


Fig. 3. Results of the reference-based semantic evaluation. The figures depict the similarity of each learned folksonomy based on five datasets (rows: BibSonomy, CiteULike, Delicious, Flickr, LastFm) to three general-purpose gold-standard taxonomies (columns: WordNet, Yago, Wikitaxonomy). by each algorithm under consideration. Similarity is depicted on the y-axis and is measured by the taxonomic precision (TP), taxonomic recall (TP), taxonomic F1-measure (TF) and the taxonomic overlap (TO); see Section 3 for an explanation of the measures. In all cases, higher values indicate stronger similarity to the gold-standard and hence a better performance of the respective algorithm.

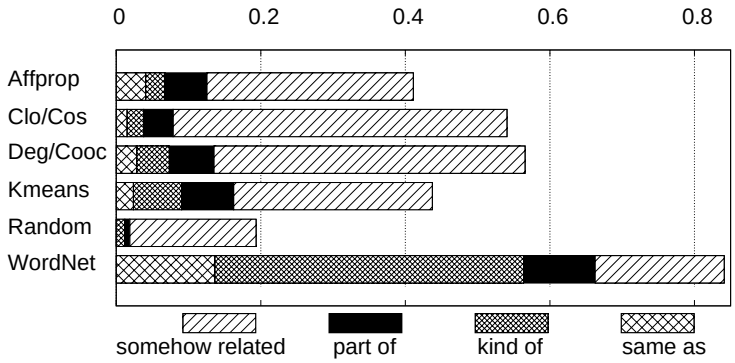


Fig. 4. Results of the semantic evaluation performed by a user study. The upper five horizontal bars correspond each to a folksonomy induced on the Flickr dataset by each algorithm under consideration; the lowest bar depicts a control condition based upon the WordNet noun taxonomy. The different patterns correspond to the average fraction of choices the human subjects have made when presented with a sample of subsumption pairs from each hierarchy (see Section 3.2).

A conclusion that can be drawn from these empirical results is that the clustering techniques we investigated seem to produce folksonomies which exhibit a smaller degree of similarity to gold-standard taxonomies than techniques based on term generality. Especially the folksonomies produced by taking degree centrality as generality measure and co-occurrence as similarity measue seem to resemble most closely to the reference taxonomies. This is an interesting observation, especially regarding that these measures are computationally much more lightweight compared to e.g. closeness centrality, cosine similarity or elaborate clustering mechanisms. We have also tried other parameter settings for K-means (different k's) and did not observe a substantial difference.

When comparing the clustering techniques, it seems that affinity propagation has a slight advantage over kmeans, however to a much lesser extent than the difference to the generality-based methods. An open question which remains is how to interpret the absolute similarity values, or in other words: Is e.g. a score of 0.02 captured by the taxonomic F1-measure an indication of a "strong" similarity between the learned and the reference taxonomy? Due to the complexity and the size of the involved structures, it is difficult to make a clear decision to this end. Because the values are averaged over the complete concept overlap, it is possible that some branches are very similar, while others are not. In order to facilitate a better understanding of the "true" quality of the learned folksonomies, we also performed a small-scale human subject experiment, whose results will be discussed next.

Figure 4 summarizes the results of this experiment involving the human assessment of folksonomies induced on the Flickr dataset. The topmost five rows correspond to the algorithms used, while the lowermost row is a control condition based on the WordNet noun taxonomy. The values on the y-axis depict the average fraction of choices for each possible answer - as an example, among all judgements on subsumption pairs produced by affinity propagation, the average fraction of "part of" answers was roughly 5,8% (0.058, black part of the uppermost bar). Please note that only "positive" answers are included in this plot (i.e. answers stating that there is a meaningful relation among two terms). However, the percentage of "negative" answers (i.e. explicit statements by the users that two terms are not related) can be deduced from the figure by subtracting the sum of positive votes from 1. As an example, for affinity propagation we received a fraction of roughly 59% (0.59, topmost row, average) of "not related" answers for each pair. So as a short statement, one can say that the "longer" the bars are, the higher is the quality of the corresponding folksonomy.

To start with the lower and upper bounds, the folksonomy produced by the random algorithm performs worst - all "positive" relation judgements are adding up to roughly

0.2. On the contrary, the control judgements on the WordNet noun taxonomy sum up to $\tilde{0}.82$, including a large portion ($\tilde{0}.42$) of "kind of" answers. So as a first observation, we can say that the random folksonomy was judged to be the worst and the WordNet noun taxonomy was judged to be the best hierarchy – which confirms our intuition and validates our experimental methodology. In between these bounds, the sum of positive votes seems to confirm the impression from the reference-based evaluation: Again, the two generality-based methods yield a higher percentage of positive votes compared to the two clustering approaches. Despite this fact, taking a more detailed look one can also see that the percentage of "kind of" and "part of" votes (which are semantically more precise compared to "somehow related") is highest for the KMeans clustering algorithm. This could of course be an artifact of sampling, but could also point towards a greater semantic precision of the folksonomy induced by KMeans clustering. However, taking a closer look at the "somehow related" pairs, it turns out that despite their lesser degree of semantic preciseness, the obtained relations can still be useful especially for organizational purposes of a category hierarchy (e.g. "pot / stove"). In light of these observations, the results of the human subject experiment can be seen as a confirmation of the validity of the measures we used in our reference-based evaluation setting.

So in summary, the results of our semantic evaluation suggest that the generality-based algorithms we analysed lead to folksonomies which capture a higher amount of meaningful semantics compared to the ones obtained by clustering algorithms. This insight will now be complemented by the results of the pragmatic evaluation.

6.2. Results of Pragmatic Evaluation

The results of pragmatic evaluation are depicted in Figure 5. As a baseline, we perform exploratory navigation with a randomly generated folksonomy to obtain a lower bound. We can assert that the cause why an agent using a random folksonomy as hidden metric space finds considerable short paths is because tagging networks are highly connected and have a low effective diameter (< 3.5) [Helic et al. 2010]. Due to high link density, the majority of tags are connected by *multiple* short paths. That means that even if the agent takes a single non-optimal or wrong link towards the destination tag, with high probability there exists an alternative link which also leads to the destination tag. In particular for the (global) shortest path of 2, an agent using a random folksonomy is considerably successful in finding short path – regardless of the first tag selected, that tag is in the majority of cases linked to the destination tag. However, as the path towards the destination becomes longer (≥ 3) the ability of an agent using a random folksonomy as hidden metric space deteriorates. The LastFM dataset exhibits even more extreme behavior in this respect – since tags in this dataset are music genres, the overlap in certain resources seems to be extremely high. However, for agents it is possible to find the shortest paths or alternative short paths with the given folksonomies. Across all datasets, we see that agents using folksonomies produced by the introduced algorithms find significantly shorter paths than when using a random folksonomy.

Structurally, the hierarchies generated with K-Means are typically unbalanced. We performed additional experiments to introduce a balancing factor to resolve these structural issues to obtain more balanced clusters. However, preliminary results show that this approach improves the success rate of greedy search only marginally (the success rate could be improved by 1% for the BibSonomy dataset), and thereby does not seem to have a significant impact on the validity of our results.

A problem with Aff. Prop. seems to be the choice of the cluster representative. In the current implementation, the cluster representative is chosen by taking the nearest sample to the centroid. As the similarities in tagging datasets are often small and sparse, the similarities between cluster members are equal, and thus the selection of the cluster representative is completely arbitrary. The same issues seem to influence the construction of the hierarchy

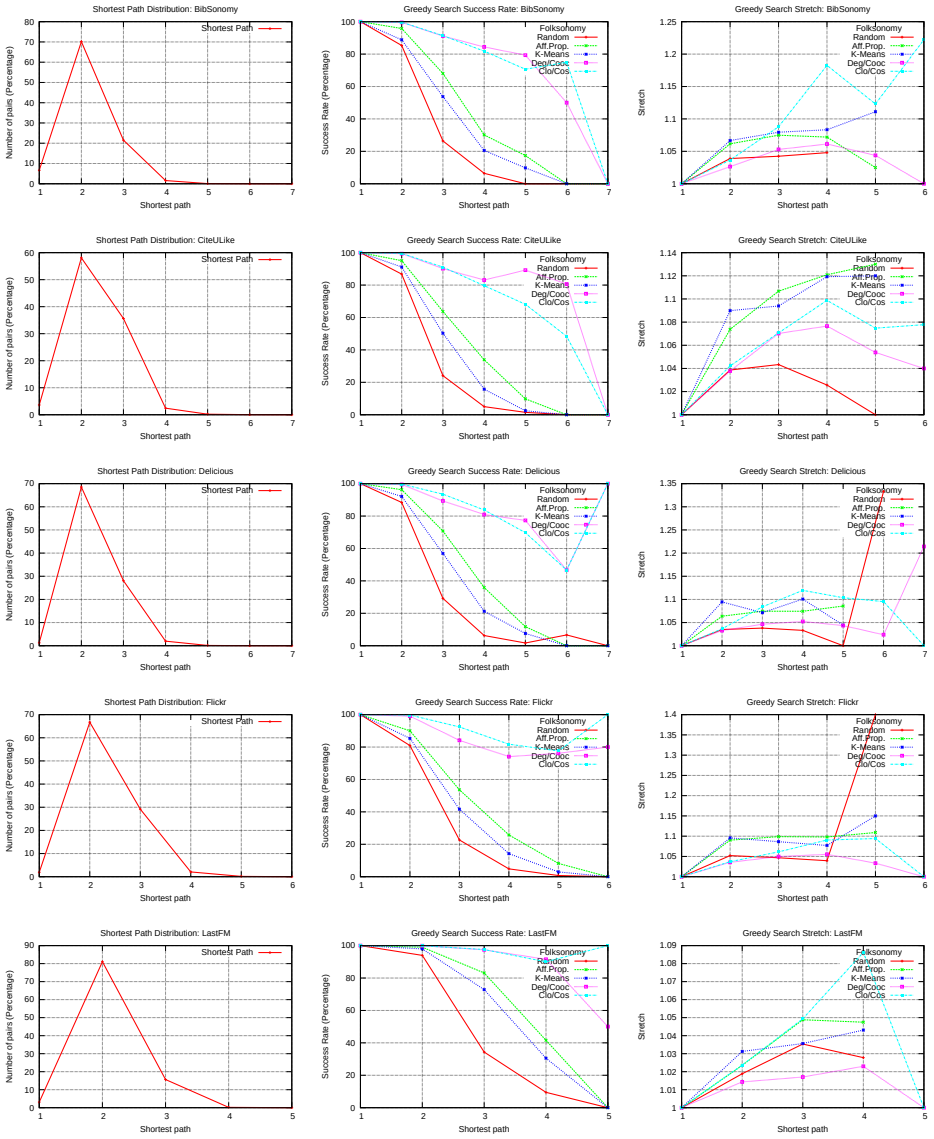


Fig. 5. Shortest path distribution, success rate and stretch of greedy search per observable shortest path in the analyzed datasets. The left column depicts the distance between two tags with global knowledge (shortest path), the middle column shows the number of times an agent finds a short path with local knowledge only (success rate) and the right column plots the penalty incurred by using local knowledge as opposed to global knowledge (stretch). Consistently, over all datasets the tag similarity network folksonomies (Deg/Cooc and Clo/Cos) outperform other folksonomies, in particular for longer shortest paths. Affinity propagation and K-Means perform better than the random folksonomy. The stretch values are consistently close to 1 – if successful, greedy search finds the destination node, on average, in almost optimal number of steps. Slightly higher stretch values for shortest paths longer than 4 come from a higher dispersion about the mean at those shortest path lengths (as the shortest path distributions are strongly skewed with peaks at 2 there are only few shortest paths longer than e.g. 4 and the dispersion becomes higher).

Table V. Overall success rate s and stretch τ for analyzed datasets

	Random		Aff.Prop.		K-Means		Deg/Cooc		Clo/Cos	
	s	τ	s	τ	s	τ	s	τ	s	τ
BibS.	0.723	1.038	0.890	1.063	0.809	1.067	0.975	1.034	0.976	1.052
CiteU.	0.627	1.038	0.824	1.085	0.748	1.090	0.957	1.052	0.960	1.055
Delicious	0.702	1.035	0.878	1.067	0.808	1.088	0.962	1.037	0.976	1.055
Flickr	0.626	1.051	0.781	1.092	0.713	1.092	0.942	1.040	0.972	1.047
LastFM	0.847	1.020	0.965	1.028	0.940	1.032	0.995	1.015	0.995	1.029

Source: Overall success rate s and stretch τ for analyzed datasets. Existing algorithms produce folksonomies that are more useful for navigation than a random baseline folksonomy. Folksonomies obtained by tag similarity network methods (Deg/Cooc and Clo/Cos) perform better in supporting navigation than folksonomies obtained by hierarchical clustering methods (Aff.Prop. and K-Means). Stretch is in all cases close to 1 meaning that all of the observer folksonomies are applicable as a hidden metric space.

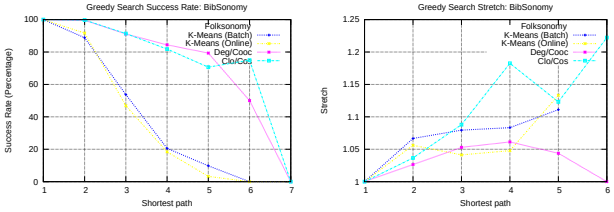


Fig. 6. Success rate and stretch of greedy search per observable shortest path in the BibSonomy dataset with batch and online K-Means folksonomies. Two variants of K-Means perform similarly and are clearly out-performed by the generality based folksonomy algorithms.

that is based on the similarity between the centroids of the previous execution steps. One possible remedy for this could be to use an average similarity of connected data samples. An advantage of Aff. Prop. is that on the upper hierarchical levels, the algorithm produces broader structures than, for example, K-Means, which seems to make them more suitable for navigation.

Summarizing, hierarchical clustering methods seem to lack additional information about the dataset as given by the tag similarity network and centrality ranking. Note that while [Heymann and Garcia-Molina 2006] came to a similar conclusion based on intuition, our paper provides an empirical justification for this.

There are no significant differences in performance of DegCen/Cooc and CloCen/Cos combinations of the centrality and similarity measures. We have performed additional experiments and produced folksonomies by combining betweenness centrality and co-occurrence as well as closeness centrality and co-occurrence. The choice of centrality or similarity measure does not significantly influence performance. Any combination of these two measures perform similar.

6.2.1. K-Means Variations. To further investigate the reasons for a bad performance of hierarchical clustering algorithms we produced folksonomies with another variation of the K-Means algorithm. The results presented so far have been produced by the K-Means algorithm operating in batch mode. We also produced K-Means folksonomies operating in online mode (i.e. incremental additions). Figure 6 shows the simulation results with the BibSonomy dataset. There are no significant differences in the performance of those two algorithm variations. The simulations with other datasets produce comparable results – across all datasets there are no significant differences in performance between the batch and online K-Means algorithm variations.

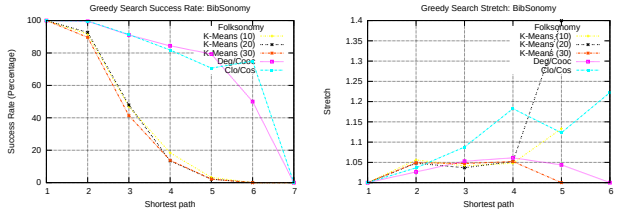


Fig. 7. Success rate and stretch of greedy search per observable shortest path in the BibSonomy dataset. K-Means folksonomies have been produced with different cluster sizes. Again, all K-Means folksonomies perform similarly and are left behind by generality based folksonomy algorithms.

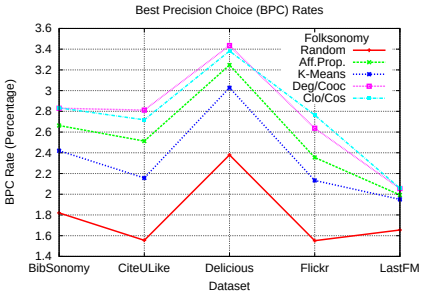


Fig. 8. Best Precision Choice (BPR) Rates for different datasets and algorithms. The figure presents the precision of tags presented to a greedy agent at step n-1 (the last step before the agent reaches his target). Consistent with other pragmatic evaluation metrics, the folksonomies obtained from tag similarity networks achieve higher precision values than other approaches (k-means and Affinity Propagation).

6.2.2. Cluster Size. We also investigated the effects of cluster size (the choice of k) on the performance of K-Means folksonomies. To that end, we produced further folksonomies using online K-Means with $k = 20$ and $k = 30$ as the cluster size. Figure 7 shows the results of pragmatic evaluation for our BibSonomy dataset with varying k. The plot shows that there is little or no influence on the performance of folksonomies generated with different k's. Comparable results have been obtained on our other datasets, which suggests that the choice of k - within the limits of our experiments - is not a confounding factor for evaluation.

6.2.3. Alternative Pragmatic Metrics. In addition to success rate and stretch, the usefulness of folksonomies can be evaluated with other pragmatic evaluation metrics as well. For example: When a user navigates to a target tag, we can calculate how hard it is for the user to identify the target resources among the tags presented to her in the last step of navigation. The last step of navigation represents a situation where the user is just one-click away from her set of target resources. A useful metric to capture this problem is precision, which we define as the number of relevant resources divided by the number of total resources presented to the user for a given tag. We define relevant resources as all resources tagged by all of the target tags. We can simulate this by letting a greedy agent navigate the tagging system, and then calculate the fraction of greedy paths which go trough the best precision tag for all algorithms and datasets at step n-1 (the last step before the agent reaches his target). The outcome of this experiment is presented in Figure 8. Our results are consistent with the results obtained from applying our other pragmatic evaluation metrics (stretch, success rate): The folksonomies obtained from tag similarity networks outperform the other approaches.

7. CONCLUSIONS

To the best of our knowledge, this paper represents the most comprehensive attempt at evaluating state-of-the-art folksonomy induction algorithms both empirically and via simulation to date. Based on a review of existing measures for semantic folksonomy evaluation, we have selected a subset and applied it to 20 folksonomies created from 5 social tagging system datasets. The results of our semantic evaluation show that folksonomy induction algorithms specifically developed for social tagging systems outperform algorithms based on traditional hierarchical clustering mechanisms consistently across most datasets. However, the results of the reference-based evaluation have shown to be somewhat sensitive towards the composition of the characteristic excerpts used by existing taxonomy similarity measures. Our particular composition of excerpts however painted a clearer picture of the usefulness of different folksonomy induction algorithms. An assessment of the induced folksonomies by human subject confirmed the validity of our reference-based evaluation. In addition, we have presented a new pragmatic evaluation method that compared the 20 folksonomies from a navigation-oriented perspective. The results obtained from pragmatic evaluation are consistent with the semantic evaluation: Again, generality-based approaches tailored towards the characteristics of social tagging systems show a superior performance compared to clustering algorithms. In summary, our work sheds new light on the properties and characteristics of state-of-the-art folksonomy induction algorithms and introduced a new pragmatic approach to folksonomy evaluation, while at the same time identifying some important limitations and challenges of evaluating hierarchical structures in information systems in general.

ACKNOWLEDGMENTS

This work is in part funded by an FWF Austrian Science Agency grant (P20269) and the PUMA project funded by the German Research Foundation (DFG).

REFERENCES

Adamic, L. and Adar, E. 2005. How to search a social network. *Social Networks 27,* 3, 187 – 203.

Adamic, L. A., Lukose, R. M., Puniyani, A. R., and Huberman, B. A. 2001. Search in power-law networks. *Physical Review E 64,* 4, 046135 1–8.

Angeletou, S. 2010. Semantic enrichment of folksonomy tagspaces. In *International Semantic Web Conference (ISWC'08)*. Springer, 889–894.

Au Yeung, C., Gibbins, N., and Shadbolt, N. 2009. Contextualising tags in collaborative tagging systems. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*. ACM, 251–260.

Benz, D., Hotho, A., and Stumme, G. 2010. Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge. In *Proc. of the 2nd Web Science Conference (WebSci10)*. Web Science Trust, Raleigh, NC, USA.

Boguñá, M., Krioukov, D., and Claffy, K. C. 2009. Navigability of complex networks. *Nature Physics 5*, 74–80.

Boguñá, M., Papadopoulos, F., and Krioukov, D. 2010. Sustaining the Internet with hyperbolic mapping. *Nature Communications 1*, 62.

Brank, J., Madenic, D., and Groblenik, M. 2006. Gold standard based ontology evaluation using instance assignment. In *Proceedings of the 4th Workshop on Evaluating Ontologies for the Web (EON2006)*. CEUR-WS, Edinburgh, Scotland.

Cattuto, C., Benz, D., Hotho, A., and Stumme, G. 2008. Semantic grounding of tag relatedness in social bookmarking systems. In *The Semantic Web – ISWC 2008, Proc. of International Semantic Web Conference 2008*, A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, and K. Thirunarayan, Eds. LNAI Series, vol. 5318. Springer, Heidelberg, 615–631.

Cattuto, C., Schmitz, C., Baldassarri, A., Servedio, V. D. P., Loreto, V., Hotho, A., Grahl, M., and Stumme, G. 2007. Network properties of folksonomies. *AI Commun. 20,* 4, 245–262.

Dellschaft, K. 2005. Measuring the similarity of concept hierarchies and its influence on the evaluation of learning procedures. M.S. thesis, Institute for Computer Science, University of Koblenz-Landau, Germany.

Dellschaft, K. and Staab, S. 2006. On how to perform a gold standard based evaluation of ontology learning. In *Proceedings of ISWC-2006 International Semantic Web Conference*. Springer, LNCS, Athens, GA, USA.

Dhillon, I., Fan, J., and Guan, Y. 2001. Efficient clustering of very large document collections. In *Data Mining for Scientific and Engineering Applications*, R. Grossman, C. Kamath, and R. Naburu, Eds. Kluwer Academic Publishers, Heidelberg.

Frey, B. J. J. and Dueck, D. 2007. Clustering by passing messages between data points. *Science 315,* 5814, 972–976.

Helic, D. and Strohmaier, M. 2011. Building directories for social tagging systems. In *20th ACM Conference on Information and Knowledge Management (CIKM 2011), Glasgow, UK*.

Helic, D., Strohmaier, M., Trattner, C., Muhr, M., and Lerman, K. 2011. Pragmatic evaluation of folksonomies. In *20th International World Wide Web Conference (WWW2011), Hyderabad, India, March 28 - April 1, ACM*.

Helic, D., Trattner, C., Strohmaier, M., and Andrews, K. 2010. On the navigability of social tagging systems. In *Proc. of 2010 IEEE International Conference on Social Computing*. IEEE Computer Society, Los Alamitos, CA, USA, 161–168.

Heymann, P. and Garcia-Molina, H. 2006. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Stanford InfoLab. April.

Hotho, A., Jaeschke, R., Schmitz, C., and Stumme, G. 2006. Folkrank: A ranking algorithm for folksonomies. In *Proc. FGIR 2006*. Gesellschaft Für Informatik, Bonn, Germany, 111–114.

Hotho, A., Jäschke, R., Schmitz, C., and Stumme, G. 2006. Bibsonomy: A social bookmark and publication sharing system. In *Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures*, A. de Moor, S. Polovina, and H. Delugach, Eds. Aalborg University Press, Aalborg, Denmark, 87–102.

Kleinberg, J. 2000a. The small-world phenomenon: an algorithm perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*. STOC '00. ACM, New York, NY, USA, 163–170.

Kleinberg, J. 2006. Complex networks and decentralized search algorithms. In *International Congress of Mathematicians (ICM)*. European Mathematical Society Publishing House, Zürich, Switzerland, 1019–1044.

Kleinberg, J. M. 2000b. Navigation in a small world. *Nature 406,* 6798, 845.

Kleinberg, J. M. 2001. Small-world phenomena and the dynamics of information. In *Advances in Neural Information Processing Systems (NIPS) 14*. MIT Press, Cambridge, MA, USA, 2001.

Koerner, C., Benz, D., Strohmaier, M., Hotho, A., and Stumme, G. 2010. Stop thinking, start tagging - tag semantics emerge from collaborative verbosity. In *Proc. of the 19th International World Wide Web Conference (WWW 2010)*. ACM, Raleigh, NC, USA.

Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., and Boguñá, M. 2010. Hyperbolic geometry of complex networks. *Phys. Rev. E 82,* 3, 036106.

Leicht, E. A., Holme, P., and Newman, M. E. J. 2006. Vertex similarity in networks. *Phys. Rev. E 73,* 2, 026120.

Li, R., Bao, S., Yu, Y., Fei, B., and Su, Z. 2007. Towards effective browsing of large scale social annotations. In *Proc. of the 16th international conference on World Wide Web, WWW '07*. ACM, New York, NY, USA, 952.

Maedche, A. 2002. *Ontology Learning for the Semantic Web*. Kluwer Academic Publishing, Boston.

Menczer, F. 2002. Growing and navigating the small world web by local content. *Proc. Natl. Acad. Sci. USA 99,* 22, 14014–14019.

Mika, P. 2007. Ontologies are us: A unified model of social networks and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web 5,* 1, 5–15.

Milgram, S. 1967. The small world problem. *Psychology Today 1*, 60–67.

Miller, G. A. 1995. Wordnet: A lexical database for english. *Communications of the ACM 38,* 1, 39–41.

Plangprasopchok, A., Lerman, K., and Getoor, L. 2010a. From saplings to a tree: Integrating structured metadata via relational affinity propagation. In *Proceedings of the AAAI workshop on Statistical Relational AI*. AAAI, Menlo Park, CA, USA.

Plangprasopchok, A., Lerman, K., and Getoor, L. 2010b. Growing a tree in the forest: Constructing folksonomies by integrating structured metadata. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 949–958.

Ponzetto, S. P. and Strube, M. 2007. Deriving a large-scale taxonomy from wikipedia. In *AAAI* (2007-09-05). AAAI Press, Menlo Park, CA, USA, 1440–1445.

Ramezani, M., Sandvig, J., Schimoler, T., Gemmell, J., Mobasher, B., and Burke, R. 2009. Evaluating the impact of attacks in collaborative tagging environments. In *Computational Science and Engineering, 2009. CSE '09. International Conference on*. Vol. 4. IEEE Computer Society, Los Alamitos, CA, USA, 136 –143.

Schifanella, R., Barrat, A., Cattuto, C., Markines, B., and Menczer, F. 2010. Folks in folksonomies: social link prediction from shared metadata. In *Proc. of the third ACM international conference on Web search and data mining, WSDM '10*. ACM, New York, NY, USA, 271–280.

Schmitz, C., Hotho, A., Jöschke, R., and Stumme, G. 2006. Mining association rules in folksonomies. In *Data Science and Classification: Proc. of the 10th IFCS Conf., Studies in Classification, Data Analysis, and Knowledge Organization*. Springer, Berlin, Heidelberg, 261–270.

Serrano, M. A., Krioukov, D., and Boguñá, M. 2008. Self-similarity of complex networks and hidden metric spaces. *Phys. Rev. Lett. 100,* 7, 078701.

Strohmaier, M., Koerner, C., and Kern, R. 2010. Why do users tag? Detecting users' motivation for tagging in social tagging systems. In *International AAAI Conference on Weblogs and Social Media (ICWSM2010), Washington, DC, USA, May 23-26*. AAAI, Menlo Park, CA, USA.

Suchanek, F. M., Kasneci, G., and Weikum, G. 2007. Yago: A Core of Semantic Knowledge. In *16th international World Wide Web conference (WWW 2007)*. ACM Press, New York, NY, USA.

Vander Wal, T. 2007. Folksonomy coinage and definition, 2007.

Watts, D. J., Dodds, P. S., and Newman, M. E. J. 2002. Identity and search in social networks. *Science 296*, 1302–1305.

Watts, D. J. and Strogatz, S. H. 1998. Collective dynamics of small-world networks. *Nature 393,* 6684, 440–442.

Yeung, C., Gibbins, N., and Shadbolt, N. 2008. A k-nearest-neighbour method for classifying web search results with data in folksonomies. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*. Vol. 1. IEEE, 70–76.

Zhong, S. 2005. Efficient online spherical k-means clustering. *IJCNN 5*, 3180–3185 vol. 5.