Evaluating Similarity Measures for Emergent Semantics of Social Tagging

Benjamin Markines[1,2]* Ciro Cattuto[2] Filippo Menczer[1,2]
Dominik Benz[3] Andreas Hotho[3] Gerd Stumme[3]

[1]School of Informatics, Indiana University, Bloomington, Indiana, USA
[2]Complex Networks Lagrange Laboratory, Institute for Scientific Interchange Foundation, Torino, Italy
[3]Knowledge and Data Engineering Group, University of Kassel, Germany

ABSTRACT

Social bookmarking systems and their emergent information structures, known as folksonomies, are increasingly important data sources for Semantic Web applications. A key question for harvesting semantics from these systems is how to extend and adapt traditional notions of similarity to folksonomies, and which measures are best suited for applications such as navigation support, semantic search, and ontology learning. Here we build an evaluation framework to compare various general folksonomy-based similarity measures derived from established information-theoretic, statistical, and practical measures. Our framework deals generally and symmetrically with users, tags, and resources. For evaluation purposes we focus on similarity among tags and resources, considering different ways to aggregate annotations across users. After comparing how tag similarity measures predict user-created tag relations, we provide an external grounding by user-validated semantic proxies based on WordNet and the Open Directory. We also investigate the issue of scalability. We find that mutual information with distributional micro-aggregation across users yields the highest accuracy, but is not scalable; per-user projection with collaborative aggregation provides the best scalable approach via incremental computations. The results are consistent across resource and tag similarity.

Categories and Subject Descriptors

H.1.1 [**Models and Principles**]: Systems and Information Theory; H.1.2 [**Models and Principles**]: User/Machine Systems; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.4 [**Information Storage and Retrieval**]: Systems and Software

General Terms

Algorithms, Design, Experimentation, Human Factors, Performance

Keywords

Social similarity, semantic grounding, ontology learning, Web 2.0

1. INTRODUCTION

We are transitioning from the "Web 1.0," where information consumers and providers are clearly distinct, to the so-called "Web 2.0" in which anyone can easily annotate objects (sites, pages, media, and so on) that someone else authored. These annotations take many forms such as classification, voting, editing, and rating. Social bookmarking systems [17] are increasingly popular among "Social Web" applications. Their emergent information organizations, based on free-form tag annotations, have become known as *folksonomies*.

From the perspective of critical Web applications such as search engines, we see this as the second major transition in the brief history of the Web. The first one occurred when researchers went beyond the textual analysis of content by taking into account the hyperlinks created by authors as implicit endorsements between pages, leading to effective ranking and clustering algorithms such as PageRank [6]. Now folksonomies grant us access to a more explicit and semantically richer source of social annotation. They allow us to extend the assessment of what a page is about from content analysis algorithms to the collective "wisdom of the crowd." If many people agree that a page is about programming then with high probability it is about programming even if its content does not include the word 'programming.' This bottom-up approach of collaborative content structuring is able to counteract some core deficiencies of knowledge management applications, such as the knowledge acquisition bottleneck. The usage of folksonomy induced information and the combination with Semantic Web technology is seen as the next transition towards the "Web 3.0."

The fact that collaborative tagging leverages large-scale human annotation of Web resources makes it a perfect candidate for bootstrapping Semantic Web applications. Hereby, the notion of *similarity* plays a crucial role. For example, keyword (tag) similarity supports navigation, keyword clustering, query expansion, tag recommendation and ontology learning; and resource (page/site) similarity supports result clustering, similarity search, ontology population and again page recommendation and navigation. Figure 1 illustrates three applications thereof.

Measures of semantic similarity between objects are naturally based upon a precise understanding of how the object space is structured. The inherent tripartite data structure of folksonomies (consisting of users, tags and resources) differs fundamentally from well-studied schemes like ontologies or the Web's link graph. Hence, a key question is how to extend and adapt traditional content and link analysis algorithms to folksonomies.

In this work, we focus on defining and analyzing semantic similarity relationships obtained from mining socially annotated data. As a large-scale evaluation of semantic relationships is a difficult task, we perform a two-step experimentation: First, we compare the ability of various tag similarity measures to predict user-created tag relations from the social bookmarking system `BibSonomy.org`, and second we provide an external grounding to reliable measures validated by user studies on large and open reference data sets. Our insights inform the choice of an appropriate measure, e.g. in a given application context.

*Corresponding author. Email: `bmarkine@cs.indiana.edu`

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

Contributions and Outline

The similarity notions that we wish to derive from folksonomies represent bottom-up, emergent, social semantic relationships obtained by aggregating the opinions of many users who are likely to have inconsistent knowledge and semantic representations. There are many ways to pursue such a goal and many open questions. For example, should a relationship be stronger if many people agree that two objects are related than if only few people do? Which weighting schemes regulate best the influence of an individual? How does the sparsity of annotations affect the accuracy of these measures? Are the same measures most effective for both resource and tag similarity? Which aggregation schemes retain the most reliable semantic information? Which lend themselves to incremental computation?

We address all of the above questions here by describing an evaluation framework to compare various general folksonomy-based similarity measures. The main contributions of this paper are:

- A general and extensive foundation for the formulation of similarity measures in folksonomies, spanning critical design dimensions such as the symmetry between users, resources, and tags; aggregation schemes; exploitation of collaborative filtering; and information-theoretic issues. Some of the measures considered have been introduced and investigated before (cf. [10]), but no systematic study including all dimensions of a folksonomy and all measures exists to date about their application to social similarity. (§ 3)
- An experimental assessment of the effectiveness of several similarity measures for both tags and resources. For the former, we establish a comparison with user-created tag relations to measure effectiveness. For both tags and resources, as a second step we gauge the similarity measures against reliable grounding measures validated by user studies on large and open reference data sets. This evaluation addresses several key limitations of traditional user based assessments. (§ 4)
- An analysis of the empirical evaluation results in the context of their scalability, in particular their viability for practical implementations in existing social bookmarking systems. A clear trade-off between effectiveness and efficiency is demonstrated and discussed. (§ 5)

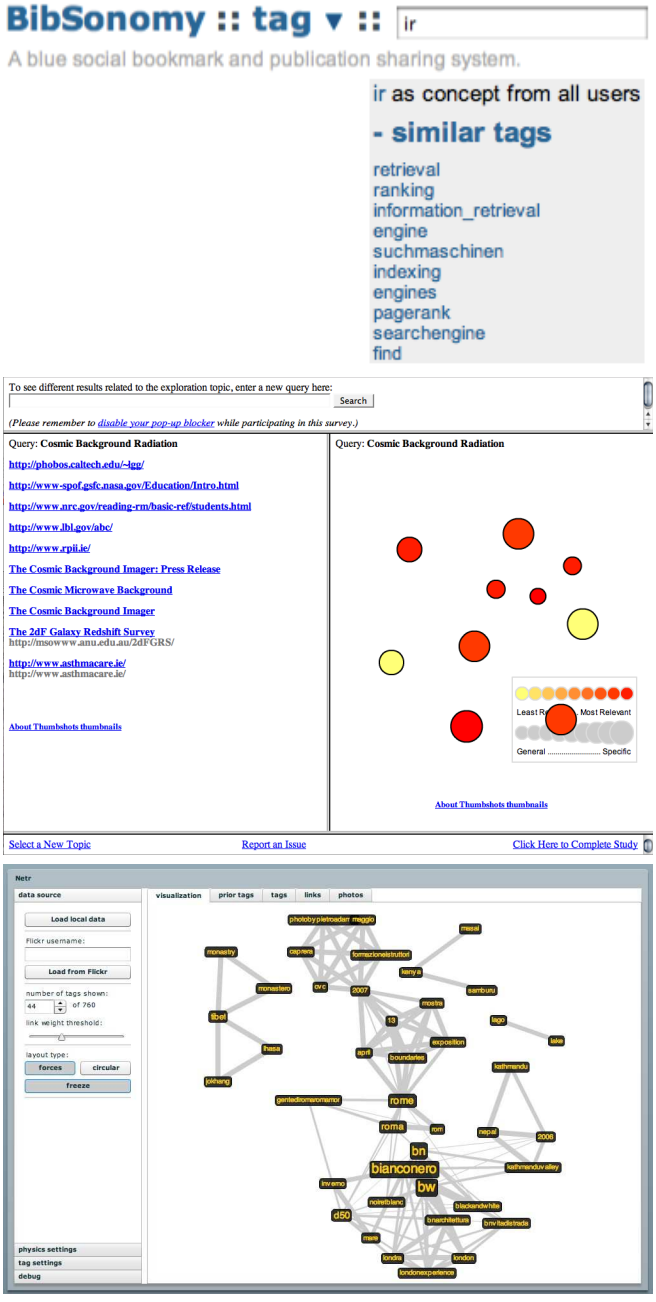

Figure 1: Three applications utilizing relationships between objects induced by social media. *Top:* Given a tag, the social bookmarking system `BibSonomy.org` suggests semantically similar tags. *Middle:* `GiveALink.org` leverages a similarity network of resources to visualize search results [13]. *Bottom:* The online tool at `netr.it` visualizes tag relationships generated from a user's `Flickr` profile.

2. BACKGROUND

Social bookmarking is a way to manage bookmarks online, for easy access from multiple locations, and also to share them with a community. There are many social bookmarking sites including popular ones such as `Del.icio.us`, `StumbleUpon.com`, `CiteULike.org`, `BibSonomy.org`, and too many others to list. A number of early social bookmarking tools and their functionalities were reviewed by Hammond *et al.* [17, 29].

While bookmarking and tagging may share several incentives [34], they are separate processes; many people bookmark without tagging [38]. The tagging approach has several limitations including lack of structure [3], lack of global coherence [38], polysemy, and word semantics [15]. Synonymy, the use of different languages, and spelling mistakes may force users to search through numerous tags. Navigation can be enhanced by suggesting tag relations grounded in content-based features [1].

Collaborative tagging is often contrasted with more traditional knowledge management approaches. Voss [46] provides evidence that the difference between controlled and free indexing blurs with sufficient feedback mechanisms. The weaknesses and strengths of these different metadata mechanisms are compared by Christiaens [11]. In prior work, we have contrasted peer-to-peer knowledge management with tagging approaches [41].

Measuring the relationships among tags or tagged resources is an active research area. Mika provides a model of semantic-social networks for extracting lightweight ontologies from `del.icio.us` [35]. Cattuto *et al.* use a variation of set overlap in the spirit of TF-IDF to build an adjacency matrix of Web resources [9]. Wu *et al.* [47] modify HITS to look for relationships between Web resources. Hotho *et al.* [23] convert a folksonomy into an undirected weighted network, used for computing a modified PageRank algorithm called FolkRank for ranking query results. Semantic networks

among tags have been built using co-occurrence [4, 44] and Jaccard's coefficient [18], also to reconstruct a concept hierarchy [19]. Relationships among users can also be extracted from tagging systems. Diederich and Iofciu use a cosine variant to compute similarities between users [12]. Others have proposed alternative approaches for extracting taxonomic relations or inferring global semantics from a folksonomy [43, 42, 16, 49]. In the context of social network analysis, Liben-Nowell and Kleinberg [27] explore several notions of node similarity for link prediction. Unlike all of the above literature, here we introduce a systematic analysis of a broad range of similarity measures that can be applied directly and symmetrically to build networks of users, tags, or resources.

Closely related to the task of measuring the relatedness of tags and resources is also the application domain of recommendations in folksonomies. The literature is still sparse. Existing work can be broadly divided into approaches that analyze the content of the tagged resources with information retrieval techniques (e.g. [36, 7]), approaches that use collaborative filtering methods based on the folksonomy structure (e.g. [48, 24]), and combinations of the two [21]. Sarwar *et al.* built networks using variations of cosine and correlation based similarity measures. Each type of network was exploited after assembly for investigating two collaborative filtering techniques [40]. Our approach differs in that we capture collaborative filtering in the similarity measure during network assembly. Several studies and algorithms consider social annotations as a means of improving Web search. Examples include studies to compare the content of social networks with search engines [37, 20] and enhancing search through the use of tags and rankings from social bookmarking systems [45, 33, 2]. In prior work, we addressed scalability with collaborative filtering when assembling a resource-by-resource similarity network in the social bookmarking system `GiveALink.org` [32]. This discussion extends our prior work.

The empirical evaluation in this paper leverages externally validated semantic similarity measures for Web resources and tags, each used as a grounding reference. In the case of resources, semantic similarity refers to the degree of relatedness between two Web sites or documents, as perceived by human subjects. Web directories such as the Open Directory Project (ODP, `dmoz.org`) provide user-compiled taxonomies of Web sites. Measures of semantic similarity based on taxonomies are well studied [14]. Maguitman *et al.* extended Lin's [28] information-theoretic measure to infer similarity from the structure of general ontologies, both hierarchical and non-hierarchical [31, 30]. Such a measure, validated by means of a user study, will serve as our grounding for resource similarity. Jiang and Conrath [25] developed a notion of distance in WordNet (`wordnet.princeton.edu`) that combines the taxonomic path length with an information-theoretic similarity measure by Resnik [39]. The Jiang-Conrath distance was validated experimentally by means of user studies as well as by its superior performance in the context of a spell-checking application [8]. Therefore it will serve as our grounding for tag similarity.

3. SIMILARITY FRAMEWORK

Before diving into the details of the similarity measures that we propose to explore, let us first review the representation that we assume for the annotations to be mined.

3.1 The Triple Annotation Representation

Our approach is based on the *triple* representation widely adopted in the Semantic Web community [23], which is closely related to the triadic context in formal concept analysis [26]. A folksonomy F is a set of triples. Each triple (u, r, t) represents user u annotating resource r with tag t. This is a highly flexible representation for which efficient data store libraries exist. Folksonomies are readily represented via triples; a post $(u, r, (t_1, \ldots, t_n))$ is transformed into a set of triples $\{(u, r, t_1), \ldots, (u, r, t_n)\}$. Note that hierarchical classifications can also be represented by triples by equating categories (or folders) with tags and applying inheritance relationships in a straightforward way; a classification (u, r, t) implies $\{(u, r, t), (u, r, t_1), \ldots, (u, r, t_n)\}$ for all ancestor classes t_i of t. Therefore the triple representation subsumes hierarchical taxonomies and folksonomies. As an example, Fig. 2 displays seven triples corresponding to a set of four posts by two users. In the following we use this running example to illustrate different definitions of similarity.

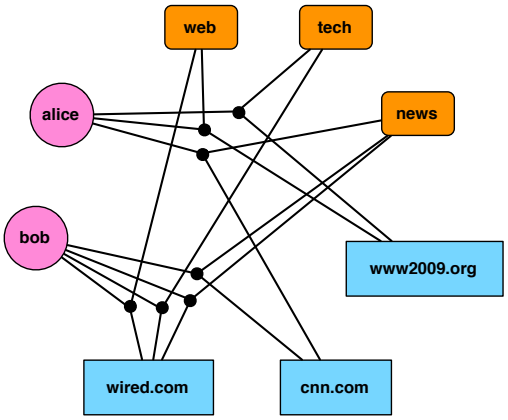

Figure 2: Example folksonomy. Two users (`alice` and `bob`) annotate three resources (`cnn.com`, `www2009.org`, `wired.com`) using three tags (`news`, `web`, `tech`). The triples (u, r, t) are represented as hyper-edges connecting a user, a resource and a tag. The 7 triples correspond to the following 4 posts: (`alice`, `cnn.com`, {`news`}), (`alice`, `www2009.org`, {`web`, `tech`}), (`bob`, `cnn.com`, {`news`}), (`bob`, `wired.com`, {`news`, `web`, `tech`}).

We will define similarity measures $\sigma(x, y)$ where x and y can be two resources (pages, media, etc.) or tags (keywords, phrases, categories, etc.). Since measures for similarity and relatedness are not well developed for three-mode data such as folksonomies, we consider various ways to obtain two-mode views of the data. In particular, we consider two-mode views in which the two dimensions considered are dual — for example, resources and tags can be dual views if resources are represented as sets of tags and vice-versa, or if tags are represented as vectors of resources and vice-versa. We focus on the development of information-theoretic similarity measures, which take into account the information/entropy associated with each item.

3.2 Aggregation Methods

In reducing the dimensionality of the triple space, we necessarily lose correlation information. Therefore the aggregation method is critical for the design of effective similarity measures; poor aggregation choices may negatively affect the quality of the similarity by discarding informative correlations.

As mentioned above, we can define similarity measures for each of the three dimensions (users, resources, tags) by first aggregating across one of the other dimensions to obtain a two-mode view of the annotation information. For evaluation purposes, we focus here on resource-resource and tag-tag similarity, for which we have reference data as empirical grounding. Therefore we aggregate across users, and obtain dual views of resources and tags, yielding dual definitions for resource and tag similarity. To keep the notation a bit simpler, let us make explicit the dimension of users along which we aggregate, even though the discussion can be extended in a straight-

forward way to aggregate across tags or resources. Below we consider four approaches to aggregate user information.

3.2.1 Projection

The simplest aggregation approach is to project across users, obtaining a unique set of (r, t) pairs. If the triples are stored in a database relation F, this corresponds to the projection operator in relational algebra: $\pi_{r,t}(F)$. Another way to represent the result of aggregation by simple projection is a matrix with binary elements $w_{rt} \in \{0, 1\}$ where rows correspond to resources (as binary vectors, or sets of tags) and columns corresponds to tags (as binary vectors, or sets of resources). All similarity measures are then derived directly from this set information. As an example, the projected binary matrix for the folksonomy of Fig. 2 is reported below. Given a resource and a tag, a 0 in the corresponding matrix element means that no user associated that resource with that tag, whereas a 1 means that at least one user has performed the indicated association.

	news	web	tech
cnn.com	1	0	0
www2009.org	0	1	1
wired.com	1	1	1

3.2.2 Distributional

A more sophisticated form of aggregation stems from considering distributional information associated with the set membership relationships between resources and tags. One way to achieve distributional aggregation is to make set membership fuzzy, i. e., weighted by the Shannon information (log-odds) extracted from the annotations. Intuitively, a shared tag may signal a weak association if it is very common. Thus we will use the information of a tag (resp. resource) x defined as $-\log p(x)$ where $p(x)$ is the fraction of resources (resp. tags) annotated with x.

Another approach is to count the users who agree on a certain resource-tag annotation while projecting across users. This yields a set of *frequency-weighted* pairs (r, t, w_{rt}) where the weight w_{rt} is the number of users tagging r with t. Such a representation corresponds to a matrix with integer elements w_{rt}, where rows are resources vectors and columns are tag vectors. For the folksonomy of Fig. 2, such a matrix is reported below. Similarity measures are then derived directly from the weighted representation.

	news	web	tech
cnn.com	2	0	0
www2009.org	0	1	1
wired.com	1	1	1

We will use both of the above distributional aggregation schemes, as appropriate for different similarity measures. The fuzzy set approach is appropriate when we want to perform row/column normalization of tag/resource probabilities to prevent very popular items from dominating the similarity. Other measures such as the dot product depend naturally on weighted vector representations.

3.2.3 Macro-Aggregation

By analogy to micro-averaging in text mining, distributional aggregation can be viewed as "micro-aggregation" if we think of users as classes. Each annotation is given the same weight, so that a more active user would have a larger impact on the weights and consequently on any derived similarity measure. In contrast, macro-aggregation treats each user's annotation set independently first, and then aggregates across users. In relational terms, we can select the triples involving each user u and then project, yielding a set of pairs for u: $\{(r, t)_u\} = \pi_{r,t}(\sigma_u(F))$. This results in per-user binary matrices of the form $w_{u,rt} \in \{0, 1\}$. For the example folksonomy of Fig. 2, we report below the matrices for the user `alice` (top) and `bob` (bottom).

	news	web	tech
cnn.com	1	0	0
www2009.org	0	1	1
wired.com	0	0	0

	news	web	tech
cnn.com	1	0	0
www2009.org	0	0	0
wired.com	1	1	1

The per-user binary matrix representations $w_{u,rt} \in \{0, 1\}$ are used to compute a "local" similarity $\sigma_u(x, y)$ for each pair of objects (resources or tags) x and y. Finally, we macro-aggregate by voting, i. e., by summing across users to obtain the "global" similarity

$$\sigma(x, y) = \sum_u \sigma_u(x, y). \quad (1)$$

Macro-aggregation does not have a bias toward users with many annotations. However, in giving the same importance to each user, the derived similarity measures amplify the relative impact of annotations by less active users. It is an empirical question which of these biases is more effective.

3.2.4 Collaborative

Macro-aggregation lends itself to explore the issue of collaborative filtering in folksonomies. Thus far, we have only considered feature-based representations. That is, a resource is described in terms of its tag features and vice-versa. If two objects share no feature, all of the measures defined on the basis of the above aggregation schemes will yield a zero similarity. In collaborative filtering, on the other hand, the fact that one or more users vote for (or in our case annotate) two objects is seen as implicit evidence of an association between the two objects. The more users share a pair of items, the stronger the association. We want to consider the same idea in the context of folksonomies. If many users annotate the same pair of resources, even with different tags, the two resources might be related. Likewise, if many users employ the same pair of tags, the two tags might be related even if they share no resources.

Macro-aggregation incorporates the same idea by virtue of summing user votes, if we assign a non-zero local similarity $\sigma_u(x, y) > 0$ to every pair of objects (x, y) present in u's annotations, irrespective of shared features. This is accomplished by adding a feature-independent local similarity to every pair (x, y) of resources or tags. In practice we can achieve this by adding a special "user tag" (resp. "user resource") to all resources (resp. tags) of u. This way all of u's items have at least one annotation in common.

Prior to macro-averaging, u's local similarity σ_u for each pair must be computed in such a way that the special annotations yield a small but non-zero contribution. This requires a revision of the information-theoretic similarity measures. For illustration, consider adding the special tag t^*_u to all resources annotated by u. The probability of observing tag t^*_u associated with any of u's resources is one, therefore the fact that two resources share t^*_u carries no information value (Shannon's information is $-\log p(t^*_u|u) = -\log 1 = 0$). Let us redefine user u's odds of tag (resp. resource) x as

$$p(x|u) = N(u, x)/(N(u) + 1) \quad (2)$$

where $N(u, x)$ is the number of resources (resp. tags) annotated by u with x, while $N(u)$ is the total number of resources (resp. tags) annotated by u. This way, $-\log p(t^*_u|u) = -\log[N(u)/(N(u) + 1)] > 0$. Below we imply this construction in the definitions of the similarity measures with collaborative aggregation.

Table 1: Summary of similarity measures by aggregation methods. Entries refer to equation numbers.

Measure	Project.	Distrib.	Macro	Collaborative
Matching	3	4	1, 5	1, 5, 2
Overlap	6	7	1, 8	1, 8, 2
Jaccard	9	10	1, 11	1, 11, 2
Dice	12	13	1, 14	1, 14, 2
Cosine	15	16	1, 17	1, 17, 2
M.I.	18, 19	18, 20	1, 21, 19	1, 21, 19, 2

3.3 Similarity Measures

We wish to evaluate several information-theoretic, statistical, and practical similarity measures. Table 1 summarizes the measures defined below. Each of the aggregation methods requires revisions and extensions of the definitions for application to the folksonomy context, i. e., for computing resource and tag similarity from triple data.

Recalling that all measures are symmetric with respect to resources and tags, we simplify the notation as follows: x represents a tag or a resource and X is its vector representation. For example, if x is a resource, X is a vector with tag elements w_{xy}. If x is a tag, X is a vector with resource elements w_{xy} (note we do not switch the subscript order for generality). For projection aggregation, the binary vector X can be interpreted as a set and we write $y \in X$ to mean $w_{xy} = 1$ and $|X| = \sum_y w_{xy}$. Analogously for a single user u, $y \in X^u$ means $w_{u,xy} = 1$ and $|X^u| = \sum_y w_{u,xy}$. We will use σ to refer to all similarity measures, and σ_u to refer to all those similarity measures that are based on a single user u and are to be macro-aggregated (Equation 1).

3.3.1 Matching

The *matching similarity* measure is defined, for the projection case, as

$$\sigma(x_1, x_2) = \sum_y w_{x_1 y} w_{x_2 y} = |X_1 \cap X_2|. \quad (3)$$

As an example, below we report the resulting similarity matrices for the resources and the tags of Fig. 2:

	cnn.com	www2009.org	wired.com
cnn.com	-	0	1
www2009.org	0	-	2
wired.com	1	2	-

	news	web	tech
news	-	1	1
web	1	-	2
tech	1	2	-

The distributional version of the matching similarity is

$$\sigma(x_1, x_2) = -\sum_{y \in X_1 \cap X_2} \log p(y). \quad (4)$$

This and the other measures use the p definition of § 3.2.2. For the example case of Fig. 2, the resources and the tags have the following probabilities: $p(\texttt{cnn.com}) = 1/3$ (out of 3 tags, cnn.com is associated with 1 tag only, news), $p(\texttt{www2009.org}) = 2/3$, $p(\texttt{wired.com}) = 1$ (i.e., no information content about tags), $p(\texttt{news}) = 2/3$ (out of 3 resources, news is associated with 2 of them, cnn.com and wired.com), $p(\texttt{web}) = 2/3$, $p(\texttt{tech}) = 2/3$. This yields the following similarity matrices for resources and tags (numeric values were truncated at the second decimal place):

	cnn.com	www2009.org	wired.com
cnn.com	-	0	0.41
www2009.org	0	-	0.81
wired.com	0.41	0.81	-

	news	web	tech
news	-	0	0
web	0	-	0.41
tech	0	0.41	-

Notice how the similarity of news with both web and tech is zero in the distributional case, whereas it is non-zero in the projection case above. This is due to the fact that the tag news shares only one resource, wired.com, with both web and tech. wired.com has zero information content for tags, as it is associated with all of them. Thus it gives no contribution to tag similarities.

In the case of macro and collaborative aggregation, an analogous definition applies to local (per-user) matching similarity:

$$\sigma_u(x_1, x_2) = -\sum_{y \in X_1^u \cap X_2^u} \log p(y|u). \quad (5)$$

For the case of Fig. 2, we need to compute the conditional probabilities for the two users. For alice, we have: $p(\texttt{cnn.com}|\texttt{alice}) = 1/3$, $p(\texttt{www2009.org}|\texttt{alice}) = 2/3$, $p(\texttt{wired.com}|\texttt{alice}) = 0$, $p(\texttt{news}|\texttt{alice}) = 1/2$ (news is associated with one of the two resources alice has annotated), $p(\texttt{web}|\texttt{alice}) = 1/2$, $p(\texttt{tech}|\texttt{alice}) = 1/2$. For bob: $p(\texttt{cnn.com}|\texttt{bob}) = 1/3$, $p(\texttt{www2009.org}|\texttt{bob}) = 0$, $p(\texttt{wired.com}|\texttt{bob}) = 1$, $p(\texttt{news}|\texttt{bob}) = 1$, $p(\texttt{web}|\texttt{bob}) = 1/2$, $p(\texttt{tech}|\texttt{bob}) = 1/2$. We compute the similarity matrices as we did above, separately for users alice and bob, and then we sum them to obtain the aggregated similarity matrices below:

	cnn.com	www2009.org	wired.com
cnn.com	-	0	1.10
www2009.org	0	-	0
wired.com	1.10	0	-

	news	web	tech
news	-	0	0
web	0	-	0
tech	0	0	-

Notice how computing per-user similarities and then aggregating over users produces more sparse similarity matrices than aggregating over users first. In the example of Fig. 2, due to the tiny size of the folksonomy, the consequences are extreme: The contribution of user alice to both matrices is zero, and for tag similarities this is also true for user bob, so that all entries of the aggregated tag similarity matrix are zero.

Collaborative filtering is able to extract more signal when aggregating similarities over users, as it exposes the similarity that is implicit in the user context. In our example, when we modify the probabilities of tags and resources as described in § 3.2.4, we find for alice: $p(\texttt{cnn.com}|\texttt{alice}) = 1/4$, $p(\texttt{www2009.org}|\texttt{alice}) = 1/2$, $p(\texttt{wired.com}|\texttt{alice}) = 0$, $p(\texttt{news}|\texttt{alice}) = 1/3$, $p(\texttt{web}|\texttt{alice}) = 1/3$, $p(\texttt{tech}|\texttt{alice}) = 1/3$. For bob: $p(\texttt{cnn.com}|\texttt{bob}) = 1/4$, $p(\texttt{www2009.org}|\texttt{bob}) = 0$, $p(\texttt{wired.com}|\texttt{bob}) = 3/4$, $p(\texttt{news}|\texttt{bob}) = 2/3$, $p(\texttt{web}|\texttt{bob}) = 1/3$, $p(\texttt{tech}|\texttt{bob}) = 1/3$. The probabilities of the (per-user) dummy tag t^* and dummy resource r^* used in the construction of § 3.2.4 are: $p(t^*_{\text{alice}}|\texttt{alice}) = 2/3$, $p(r^*_{\text{alice}}|\texttt{alice}) = 3/4$, $p(t^*_{\text{bob}}|\texttt{bob}) = 2/3$, $p(r^*_{\text{bob}}|\texttt{bob}) = 3/4$. The resulting similarity matrices for collaborative aggregation are:

	cnn.com	www2009.org	wired.com
cnn.com	-	0.41	0.81
www2009.org	0.41	-	0
wired.com	0.81	0	-

	news	web	tech
news	-	0.86	0.86
web	0.86	-	1.56
tech	0.86	1.56	-

Notice how collaborative filtering recovers non-zero values for the tag similarities.

3.3.2 Overlap

Projection-aggregated *overlap similarity* is defined as

$$\sigma(x_1, x_2) = \frac{|X_1 \cap X_2|}{\min(|X_1|, |X_2|)} \tag{6}$$

while distributional overlap is given by

$$\sigma(x_1, x_2) = \frac{\sum_{y \in X_1 \cap X_2} \log p(y)}{\max(\sum_{y \in X_1} \log p(y), \sum_{y \in X_2} \log p(y))}. \tag{7}$$

Local overlap for macro and collaborative aggregation is

$$\sigma_u(x_1, x_2) = \frac{\sum_{y \in X_1^u \cap X_2^u} \log p(y|u)}{\max(\sum_{y \in X_1^u} \log p(y|u), \sum_{y \in X_2^u} \log p(y|u))}. \tag{8}$$

3.3.3 Jaccard

Jaccard similarity aggregated by projection is

$$\sigma(x_1, x_2) = \frac{|X_1 \cap X_2|}{|X_1 \cup X_2|}. \tag{9}$$

Distributional Jaccard similarity is defined by

$$\sigma(x_1, x_2) = \frac{\sum_{y \in X_1 \cap X_2} \log p(y)}{\sum_{y \in X_1 \cup X_2} \log p(y)} \tag{10}$$

while the macro and collaborative versions are based on

$$\sigma_u(x_1, x_2) = \frac{\sum_{y \in X_1^u \cap X_2^u} \log p(y|u)}{\sum_{y \in X_1^u \cup X_2^u} \log p(y|u)}. \tag{11}$$

3.3.4 Dice

The projected *Dice coefficient* is

$$\sigma(x_1, x_2) = \frac{2|X_1 \cap X_2|}{|X_1| + |X_2|} \tag{12}$$

with its distributional version defined as

$$\sigma(x_1, x_2) = \frac{2\sum_{y \in X_1 \cap X_2} \log p(y)}{\sum_{y \in X_1} \log p(y) + \sum_{y \in X_2} \log p(y)} \tag{13}$$

and the macro and collaborative Dice built upon

$$\sigma_u(x_1, x_2) = \frac{2\sum_{y \in X_1^u \cap X_2^u} \log p(y|u)}{\sum_{y \in X_1^u} \log p(y|u) + \sum_{y \in X_2^u} \log p(y|u)}. \tag{14}$$

3.3.5 Cosine

Cosine similarity with binary projection is given by

$$\sigma(x_1, x_2) = \frac{X_1}{\sqrt{|X_1|}} \cdot \frac{X_2}{\sqrt{|X_2|}} = \frac{|X_1 \cap X_2|}{\sqrt{|X_1| \cdot |X_2|}}. \tag{15}$$

For the distributional version of the cosine, it is natural to use the frequency-weighted representation:

$$\sigma(x_1, x_2) = \frac{X_1}{||X_1||} \cdot \frac{X_2}{||X_2||} = \frac{\sum_y w_{x_1 y} w_{x_2 y}}{\sqrt{\sum_y w_{x_1 y}^2}\sqrt{\sum_y w_{x_2 y}^2}}. \tag{16}$$

The macro and collaborative aggregation versions are based on local cosine

$$\sigma_u(x_1, x_2) = \frac{X_1^u}{\sqrt{|X_1^u|}} \cdot \frac{X_2^u}{\sqrt{|X_2^u|}} = \frac{|X_1^u \cap X_2^u|}{\sqrt{|X_1^u| \cdot |X_2^u|}} \tag{17}$$

where in the collaborative case the construction of § 3.2.4 is applied without need of log-odds computations.

3.3.6 Mutual Information

The last measure we consider is *mutual information*. With projection and distributional aggregation we define

$$\sigma(x_1, x_2) = \sum_{y_1 \in X_1} \sum_{y_2 \in X_2} p(y_1, y_2) \log \frac{p(y_1, y_2)}{p(y_1)p(y_2)} \tag{18}$$

where for the projection case the probabilities $p(y)$ are defined in the usual manner (§ 3.2.2), and the joint probabilities $p(y_1, y_2)$ are also based on resource/tag (row/column) normalization:

$$p(y_1, y_2) = \frac{\sum_x w_{xy_1} w_{xy_2}}{\sum_x 1}. \tag{19}$$

With distributional aggregation, the joint probabilities must be matrix rather than row/column normalized; we compute fuzzy joint probabilities from the weighted representation:

$$p(y) = \frac{\sum_x w_{xy}}{\sum_{r,t} w_{rt}}, \quad p(y_1, y_2) = \frac{\sum_x \min(w_{xy_1}, w_{xy_2})}{\sum_{r,t} w_{rt}} \tag{20}$$

where min is a fuzzy equivalent of the intersection operator. Finally, macro and collaborative aggregation of local mutual information use

$$\sigma_u(x_1, x_2) = \sum_{y_1 \in X_1^u} \sum_{y_2 \in X_2^u} p(y_1, y_2|u) \log \frac{p(y_1, y_2|u)}{p(y_1|u)p(y_2|u)} \tag{21}$$

where simple and joint probabilities are resource/tag (row/column) normalized for each user's binary representation, and collaborative mutual information uses the construction and probability definition of § 3.2.4.

3.3.7 Other Measures

For space reasons we omit discussion of other measures we experimented with. These include distributional versions of matching, overlap, Dice, and Jaccard similarity with matrix-normalized probabilities based on the weighted representation. They did not perform as well as the measures defined above.

4. EVALUATION

`BibSonomy.org` is a social bookmark and publication management system. For our analysis, we used a benchmark dataset from December 2007, which is available on the BibSonomy site.[1] We focused on the bookmark part of the system. The BibSonomy snapshot that we used contains $128,500$ bookmarks annotated by $1,921$ users with $58,753$ distinct tags. We focus on resource similarity and tag similarity, aggregating across users as the third dimension of our annotation data.

4.1 Predicting Tag Relations

`BibSonomy.org` allows users to input directed relations such as `tagging` → `web2.0` between pairs of tags. These relationships are suitable for, but not limited to, defining *is-a* relationships. The semantics can thus be read as "`tagging` is a `web2.0`" or "`tagging` is a subtag of `web2.0`" [22]. The most straightforward evaluation of our similarity measures therefore consists in using them to predict user-defined relations between tags. Such a prediction task requires that we set some threshold on the similarity

[1] `http://www.bibsonomy.org/faq#faq-dataset-1`

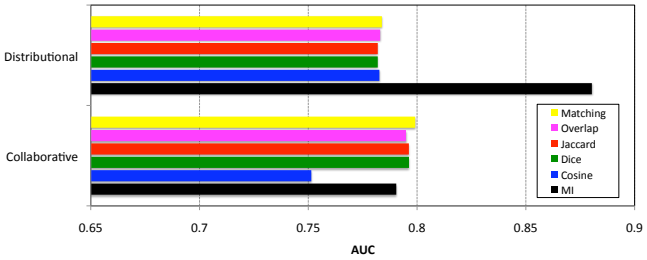

Figure 3: Areas under ROC curves (AUC) for tag relation predictions based on similarity measures with distributional and collaborative aggregation. In an ROC curve, the true positive rate is plotted against the false positive rate as a function of similarity thresholds. A good similarity measure can select many true positives with few false positives, yielding a higher AUC.

values, such that a similarity above threshold implies a prediction that two tags are related and vice-versa. To determine which similarity measures are more effective predictors, we plot in Fig. 3 the areas under ROC curves for a couple of aggregation methods. These results suggest that mutual information outperforms the other measures with distributional aggregation. For collaborative aggregation it is difficult to establish a clear ranking between the measures.

This evaluation approach has some important limitations:

- While folksonomies contain many tags, available user data about tag relations is very sparse. For example we considered 2000 tags (4×10^6 candidate relations) and found among these only 142 tag relations provided by BibSonomy users. With such little labeled data, assessments are bound to be noisy.
- Similarity values are broadly distributed, spanning several orders of magnitude. The tag relation prediction task forces us to turn this high resolution data into binary assessments, potentially losing precious information. The results are very sensitive to small changes in the similarity threshold; for example increasing the threshold from 0 to 10^{-7} decreases the false positive rate from 1 to less then 0.1. Such sensitivity suggests that fine-grained information is critical, and negatively affects the reliability of the evaluation results.
- Although there is no such requirement, users' tag relations usually focus on hierarchical relationships and thus may miss many potentially strong non-hierarchical relations. For example, we may have `python` → `programming` and `perl` → `programming` but no relation between `python` and `perl`. This may unfairly penalize our measures.
- Finally, user data is only available for tag relations while we would like to also evaluate the resource similarity measures.

To address these limitations, we look to an alternative evaluation approach that, while still based (indirectly) on user data, allows us to access a much larger pool of high resolution similarity assessments for both tags and resources. For each, we need a reliable external source of similarity data as a grounding reference to evaluate the effectiveness of the various proposed similarity measures.

4.2 Evaluation via External Grounding

Given a similarity measure to be evaluated, we want to assess how well it approximates the reference similarity measures. Since different similarity measures have different distributional properties, we turn to a non-parametric analysis that only looks at the *ranking* of the pairs by similarity rather than the actual similarity values. This reflects the intuition that while it is problematic for someone to quantify the similarity between two objects, it is natural to rank pairs on the basis of their similarities, as in "a chair is more similar to a table than to a computer." Indeed in most applications we envision (e.g. search, recommendation) ranking is key: Given a particular tag or resource we want to show the user a set of most similar objects. We thus turn to *Kendall's* τ correlation between the similarity vectors whose elements are pairs of objects. We compute τ efficiently with Knight's $O(N \log N)$ algorithm as implemented by Boldi *et al.* [5]. A higher correlation is to be interpreted as a better agreement with the grounding and thus as evidence of a better similarity measure. Of course only a subset of the tags or resources in any social bookmarking system are found in a reference similarity dataset, so we cannot use the grounding to measure similarity in general. We can however use the reference similarities to evaluate our proposed measures, which can in turn be applied to any pair of objects.

4.3 Tag Similarity

4.3.1 *WordNet Grounding*

We use the WordNet term collection for the semantic grounding of the tag similarity measures. In particular we rank tag pairs by their Jiang-Conrath distance [25], which combines taxonomic and information-theoretic knowledge. This WordNet distance measure is an appropriate reference as it was validated experimentally [8, 10].

For our evaluation of tag similarity, we focus on the subset of the BibSonomy annotations whose tags overlap with the WordNet dataset. This subset comprises $17,041$ tags, or about 29% of the total number of tags in the BibSonomy dataset. Similarities are computed between all pairs of tags in this set, however it was not possible to use the full annotation data from the folksonomy due to the time complexity of the similarity computations in conjunction with the dimensionality of the underlying vector space. The issues of time complexity and scalable computation of similarity are discussed below in § 5. Let us first evaluate the effectiveness of the measures by limiting the analysis to the most popular resources. More specifically we select, among the tags in the overlap subset, those associated with the 2000 most frequent resources, i.e., those resources that appear in the largest number of triples across the entire folksonomy. We then compute the similarities using all the folksonomy annotations relative to these top tags, disregarding less used, noisier tags. As an illustration, let us consider the tags of Fig. 2 and the similarities among them as extracted from the WordNet grounding as well as two of our measures. Below we show the ranked similarities and the resulting τ values.

Rank	WordNet	Distrib. Jaccard	Distrib. MI
1	`tech-web`	`news-tech`	`news-web`
2	`news-web`	`tech-web`	`tech-web`
3	`news-tech`	`news-web`	`news-tech`
τ	1	1/3	2/3

4.3.2 *Results*

Figure 4 plots Kendall's τ correlation between each measure introduced in § 3 and the WordNet reference. As a baseline we computed τ with a randomly generated ranking of the tag similarities. Among micro-aggregated measures, distributional information does not have any positive impact on accuracy compared to the simpler binary representation stemming from projection. Consistently with Fig. 3, mutual information is by far the most accurate measure of tag similarity. Matching, overlap, Dice and Jaccard do not differ significantly from each other.

Macro-aggregation is the worst-performing aggregation method, with the exception of matching (which almost equals the micro-averaged version) and mutual information (which outperforms all of the micro-averaged measures except mutual information itself).

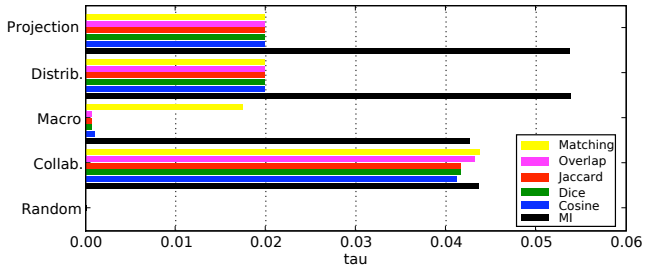

Figure 4: Tag-tag similarity accuracy, according to Kendall's τ correlations between the similarity vectors generated by the various measures and the reference similarity vector provided by the WordNet grounding measure. All similarity measures perform significantly better than the randomly generated set of similarities ($\tau = 10^{-4}$).

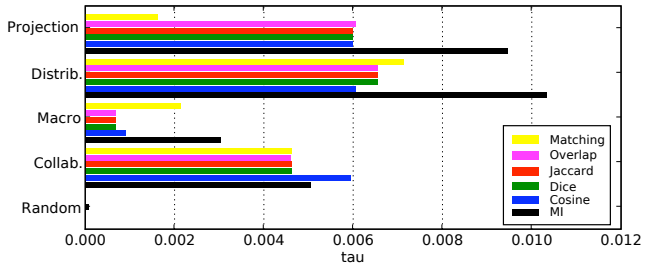

Figure 5: Resource-resource similarity accuracy, according to Kendall's τ correlations between the similarity vectors generated by the various measures and the reference similarity vector provided by the ODP grounding measure. All similarity measures perform significantly better than the randomly generated set of similarities ($\tau = 8 \times 10^{-5}$).

Collaborative aggregation provides a large boost to accuracy for tags. Each of the collaborative measures outperforms all of the others, except mutual information. These results underscore the critical semantic information contained in single-user annotations. Combining these individually induced tag relations by collaborative aggregation yields globally meaningful semantic tag relations.

4.4 Resource Similarity

4.4.1 ODP Grounding

We use the URL collection of the Open Directory Project for the semantic grounding of the resource similarity measures. In particular we rely on Maguitman *et al.*'s graph-based similarity measure [31, 30], which extends Lin's hierarchical similarity [28] by taking non-hierarchical structure into account. The ODP graph similarity is an appropriate reference because it was shown to be very accurate through a user study [30].

For our evaluation of resource similarity, we focus on the subset of the BibSonomy annotations whose resources overlap with the ODP. This subset comprises $3,323$ resources, or about 2.6% of the total unique URLs in the BibSonomy dataset. Similarities are computed between all pairs of resources in this set, using the full annotation data from the folksonomy.

4.4.2 Results

Figure 5 plots the Kendall's τ correlation between each measure introduced in § 3 and the ODP reference. The baseline was computed using a random ranking of resources similarities, as we did for tags. Distributional aggregation yields the best performance. However, with the exception of the matching similarity measure, the distributional information does not seem to have a very large impact. Mutual information is again by far the most accurate measure. Overlap, Dice and Jaccard do not differ significantly from each other.

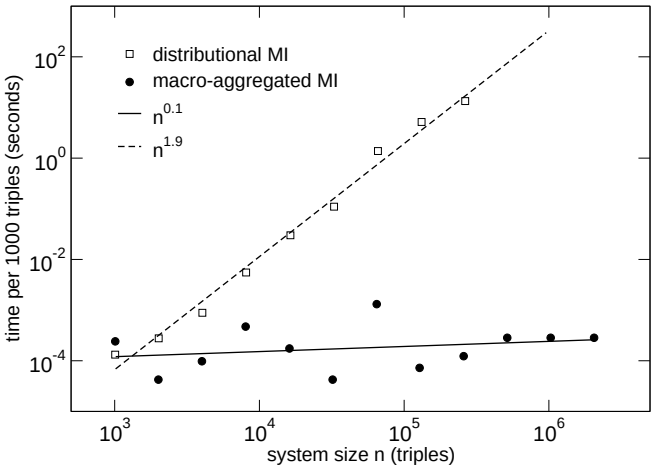

Figure 6: Scalability of the mutual information computation of resource similarity for different aggregation methods. We measured the CPU time necessary to update the similarities after a constant number of new annotations are received, as a function of system size n. Best polynomial fits $time \sim n^{\alpha}$ are also shown.

While macro-aggregation is the worst-performing aggregation method, collaborative aggregation greatly improves accuracy. In particular, cosine performs best in the collaborative setting. These results again suggest that collaborative filtering captures important semantic information in the folksonomy; the fact that two resources are annotated by the same users is telling about the relationship between these resources, beyond any tags they share. These results are consistent with experiments performed on another data set for a small subset of the measures explained here [32].

5. DISCUSSION AND SCALABILITY

The results outlined above for resource and tag similarity allow us to draw a few consistent observations: First, mutual information is the measure that best extracts semantic similarity information from a folksonomy. Mutual information considers conditional probabilities between two objects extracting the most data among the evaluated measures from an information theory point of view. We interpret this as the most fine-grained approach because we are not projecting out any information on the graph. Second, macro-aggregation is less effective than micro-aggregation. One interpretation is that since user data is necessarily more sparse, macro-aggregation adds noise by giving equal importance to each user. In other words, the user does not seem to be as good a "unit" of knowledge aggregation in a folksonomy as finer-grained individual annotation.

In spite of macro-aggregation's shortcomings, collaborative filtering extracts so much useful information about folksonomy relationships that it cannot be ignored. Especially for tag similarity, collaborative aggregation compensates for almost all the loss due to the noise of macro-aggregation. It seems therefore important for folksonomy-derived similarity measures to capture this form of social information, which differs from the more obvious notions of similarity based on shared features. Indeed we show there is useful information in annotation data even if we do away with tags when computing resource similarity and vice-versa (i.e., removing resources when computing tag similarity).

Another reason to consider macro-aggregation in general, and collaborative aggregation in particular, is related to the issue of scalability. As mentioned above, the computations of micro-aggregated

similarities have taxing computational complexity. Mutual information, the most effective measure, is also the most expensive. Having a look at its definition (Eq. 18), it is obvious that in its computation all possible combinations of attribute pairs for two given objects are involved. This implies a quadratic complexity — while e.g. the cosine similarity (Eq. 16) only runs in a linear fashion through the attribute overlap of two objects.

From a practical perspective, we consider as *scalable* those measures that can be updated to reflect new annotations at a pace that can keep up with the stream of incoming annotations. Suppose that some constant number c of new annotations arrive per unit time. The similarity of all pairs of tags/resources affected by each new annotation must therefore be updated in constant time $1/c$.

The problem with distributional aggregation is that similarities must be recomputed from scratch as frequency weights are updated. This is not scalable because its complexity clearly grows with the size of the system (e. g., number of triples). On the other hand, macro and collaborative aggregation allow for incremental computation because each user's representation is maintained separately. When a new annotation arrives from user u, only u's contribution σ_u need be updated. Such updates may be scalable.

An average-case complexity analysis is problematic due to the long-tailed distributions typical of folksonomies; quantities like average densities and average overlap are not necessarily characteristic of the system, given the huge fluctuations associated with broad distributions. Therefore we turn to an empirical analysis to examine how update complexity scales with system size. Figure 6 compares the computation of mutual information between resources using two aggregation methods: distributional (micro) and macro (or collaborative) aggregation. This is representative because mutual information is consistently the best measure with both micro and macro aggregation, and the second best with collaborative aggregation. As the plot shows, micro-aggregation scales almost quadratically with system size, while macro-aggregated similarity can be updated in almost constant time. Therefore macro and collaborative aggregation measures compensate a loss in accuracy with a huge scalability gain.

6. CONCLUSIONS

In summary, we have discussed a general and extensive foundation for the formulation of similarity measures in folksonomies, spanning critical design dimensions such as the symmetry between users, resources, and tags; aggregation schemes; exploitation of collaborative filtering; and information-theoretic issues. Experiments with resource and tag similarity alike have pointed to folksonomy-based mutual information measures as the best at extracting semantic associations from social annotation data.

The question of scalability has highlighted a critical trade-off between accuracy and complexity. Although some social aggregation methods achieve good accuracy in a non-scalable way, measures based on collaborative aggregation of annotations achieve competitive quality while minimizing computation thanks to incremental updates. This leads to the best performance/cost trade-off; we underscore the key role of scalability for the practical viability of similarity computations in existing social bookmarking systems.

Other similarity measures that we have not yet explored include matrix-normalized mutual information with binary projection aggregation and the integration of collaborative filtering with distributional aggregation.

The similarity measures analyzed in this paper can readily be employed to support many Social and Semantic Web applications, such as tag clustering for ontology construction and learning, query expansion, and recommendation. Our group has begun the use of these similarity measures for visualizing relationships among resources in search query results [13]. Another straightforward application of the socially induced similarity is to enrich Web navigation for knowledge exploration. These techniques can lead to a possible synergy between traditional and socio-semantic Web technologies.

Acknowledgments

We are grateful to Heather Roinestad for the ODP similarity network, Mark Meiss for his implementation of Kendall's τ, and Luis Rocha for helpful comments on fuzzy notions of joint probability that contributed to our extension of mutual information. This work was performed during FM's sabbatical at the ISI Foundation in Torino, Italy. We acknowledge support from the ISI and CRT Foundations. This work has been partly supported by the TAGora project (FP6-IST5-34721) funded by the European Commission's Future and Emerging Technologies program, and by NSF Grant No. IIS-0811994.

7. REFERENCES

[1] M. Aurnhammer, P. Hanappe, and S. L. Integrating collaborative tagging and emergent semantics for image retrieval. In *Proc. WWW Collaborative Web Tagging Workshop*, 2006.

[2] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *Proc. 16th Intl. Conf. on World Wide Web*, pages 501–510, 2007.

[3] J. Bar-Ilan, S. Shoham, A. Idan, Y. Miller, and A. Shachak. Structured vs. unstructured tagging — a case study. In *Proc. WWW Collaborative Web Tagging Workshop*, 2006.

[4] G. Begelman, P. Keller, and F. Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *Proc. WWW Collaborative Web Tagging Workshop*, 2006.

[5] P. Boldi, M. Santini, and S. Vigna. Do your worst to make the best: Paradoxical effects in pagerank incremental computations. *Internet Mathematics*, 2(3):387–404, 2005.

[6] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, April 1998.

[7] C. Brooks and N. Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *Proc. 15th Intl. Conf. on World Wide Web*, pages 625–632, 2006.

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

[9] C. Cattuto, A. Baldassarri, V. D. P. Servedio, and V. Loreto. Emergent community structure in social tagging systems. *Advances in Complex Systems*, 11:597–608, 2008.

[10] C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic grounding of tag relatedness in social bookmarking systems. In *Proc. ISWC 2008*, volume 5318 of *LNCS*, pages 615–631, Karlsruhe, Germany, 2008.

[11] S. Christiaens. Metadata mechanisms: From ontology to folksonomy ... and back. In *Proc. On the Move to Meaningful Internet Systems Workshop*, LNCS. Springer, 2006.

[12] J. Diederich and T. Iofciu. Finding communities of practice from user profiles based on folksonomies. *Proc. 1st Intl. Workshop on Building Technology Enhanced Learning Solutions for Communities of Practice*, 2006.

[13] J. Donaldson, M. Conover, B. Markines, H. Roinestad, and F. Menczer. Visualizing social links in exploratory search. In *Proc. 19th Conf. on Hypertext and Hypermedia (HT)*, pages 213–218, 2008.

[14] P. Ganesan, H. Garcia-Molina, and J. Widom. Exploiting Hierarchical Domain Structure to Compute Similarity. *ACM Trans. Inf. Syst.*, 21(1):64–93, 2003.

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

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

[17] T. Hammond, T. Hannay, B. Lund, and J. Scott. Social Bookmarking Tools (I): A General Review. *D-Lib Magazine*, 11(4), April 2005.

[18] Y. Hassan-Montero and V. Herrero-Solana. Improving tag-clouds as visual information retrieval interfaces. In *Proc. Intl. Conf. on Multidisciplinary Information Sciences and Technologies*, 2006.

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

[20] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *Proc. Intl. Conf. on Web Search and Data Mining (WSDM)*, pages 195–206, New York, NY, USA, 2008. ACM.

[21] P. Heymann, D. Ramage, and H. Garcia-Molina. Social tag prediction. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 531–538, New York, NY, USA, 2008. ACM.

[22] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. BibSonomy: A social bookmark and publication sharing system. In *Proc. Conceptual Structures Tool Interoperability Workshop at the 14th Intl. Conf. on Conceptual Structures*, pages 87–102, 2006.

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

[24] R. Jäschke, L. B. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in folksonomies. In *Proc. 11th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, volume 4702 of *LNCS*, pages 506–514, Berlin, Heidelberg, 2007. Springer.

[25] J. J. Jiang and D. W. Conrath. Semantic Similarity based on Corpus Statistics and Lexical Taxonomy. In *Proc. Intl. Conf. on Research in Computational Linguistics (ROCLING)*, 1997.

[26] F. Lehmann and R. Wille. A triadic approach to formal concept analysis. In G. Ellis, R. Levinson, W. Rich, and J. F. Sowa, editors, *Conceptual Structures: Applications, Implementation and Theory*, volume 954 of *LNCS*. Springer, 1995.

[27] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proc. 12th Intl. Conf. on Information and Knowledge Management (CIKM)*, pages 556–559, New York, NY, USA, 2003. ACM.

[28] D. Lin. An information-theoretic definition of similarity. In J. W. Shavlik, editor, *Proc. 15th Int.l Conf. on Machine Learning (ICML)*, pages 296–304, Madison, Wisconson, USA, 1998. Morgan Kaufmann.

[29] B. Lund, T. Hammond, M. Flack, and T. Hannay. Social Bookmarking Tools (II): A Case Study - Connotea. *D-Lib Magazine*, 11(4), April 2005.

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

[31] A. G. Maguitman, F. Menczer, H. Roinestad, and A. Vespignani. Algorithmic detection of semantic similarity. In *Proc. 14th Intl. World Wide Web Conference*, pages 107–116, 2005.

[32] B. Markines, H. Roinestad, and F. Menczer. Efficient assembly of social semantic networks. In *Proc. 19th Conf. on Hypertext and Hypermedia (HT)*, pages 149–156, 2008.

[33] B. Markines, L. Stoilova, and F. Menczer. Social bookmarks for collaborative search and recommendation. In *Proc. AAAI*, 2006.

[34] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proc. 17th Conf. on Hypertext and Hypermedia (HT)*, pages 31–40, New York, NY, USA, 2006. ACM Press.

[35] P. Mika. Ontologies are us: A unified model of social networks and semantics. In Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, editors, *Proc. Intl. Semantic Web Conf.*, volume 3729 of *LNCS*, pages 522–536. Springer, 2005.

[36] G. Mishne. Autotag: a collaborative approach to automated tag assignment for weblog posts. In *Prox. 15th Intl. Conf. on World Wide Web*, pages 953–954. ACM Press, 2006.

[37] K. P. G. A. Mislove and P. Druschel. Exploiting social networks for internet search. In *Proc. 5th Workshop on Hot Topics in Networks*, Irvine, CA, 2006.

[38] J. C. Paolillo and S. Penumarthy. The social structure of tagging internet video on del.icio.us. In *Proc. 4oth Hawaii Intl. Conf. on System Sciences (HICSS)*, Losa Alamitos, CA, 2007. IEEE Press.

[39] P. Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proc. IJCAI XI*, pages 448–453, 1995.

[40] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In *Proc. Intl. Conf. on World Wide Web*, pages 285–295, 2001.

[41] C. Schmitz, A. Hotho, R. Jäschke, and G. Stumme. Kollaboratives wissensmanagement. In T. Pellegrini and A. Blumauer, editors, *Semantic Web - Wege zur vernetzten Wissensgesellschaft*, pages 273–290. Springer, 2006.

[42] C. Schmitz, A. Hotho, R. Jäschke, and G. Stumme. Mining association rules in folksonomies. In *Data Science and Classification: Proc. of the 10th IFCS Conf.*, pages 261–270, Berlin, Heidelberg, 2006. Springer.

[43] P. Schmitz. Inducing ontology from Flickr tags. In *WWW Collaborative Web Tagging Workshop*, May 2006.

[44] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proc. 17th Intl. Conf. on World Wide Web*, pages 327–336, 2008.

[45] L. Stoilova, T. Holloway, B. Markines, A. Maguitman, and F. Menczer. GiveALink: Mining a Semantic Network of Bookmarks for Web Search and Recommendation. In *Proc. KDD Workshop on Link Discovery: Issues, Approaches and Applications*, 2005.

[46] J. Voss. Tagging, folksonomy & co - renaissance of manual indexing? Technical report, arXiv:cs/0701072, 2007.

[47] H. Wu, M. Zubair, and K. Maly. Harvesting social knowledge from folksonomies. In *Proc. 17th Conf. on Hypertext and Hypermedia (HT)*, pages 111–114, 2006.

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

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