Measuring Similarity in Large-scale Folksonomies

Giovanni Quattrone, Emilio Ferrara, Pasquale De Meo, Licia Capra

Abstract

Social (or folksonomic) tagging has become a very popular way to describe content within Web 2.0 websites. Unlike taxonomies, which overimpose a hierarchical categorisation of content, folksonomies enable end-users to freely create and choose the categories (in this case, tags) that best describe some content. However, as tags are informally defined, continually changing, and ungoverned, social tagging has often been criticised for lowering, rather than increasing, the efficiency of searching, due to the number of synonyms, homonyms, polysemy, as well as the heterogeneity of users and the noise they introduce. To address this issue, a variety of approaches have been proposed that recommend users what tags to use, both when labelling and when looking for resources. Whilst these techniques have proved to increase the accuracy of searches, we argue that they do so in unrealistic settings, where the datasets used during tests have been manipulated, typically computing a p-core that projects the folksonomy onto a much denser space. As we illustrate in this paper, real world folksonomies are characterized by power law distributions of tags instead, over which commonly used similarity metrics, including the Jaccard coefficient and the cosine similarity, fail to compute. We thus propose a novel metric, specifically developed to capture similarity in large-scale folksonomies, that is based on a mutual reinforcement principle: that is, two tags are deemed similar if they have been associated to similar resources, and vice-versa two resources are deemed similar if they have been labelled by similar tags. We offer an efficient realisation of this similarity metric, and assess its quality experimentally, by comparing it against cosine similarity, on three large-scale datasets, namely Bibsonomy, MovieLens and CiteULike.

1 Introduction

The rise of Web 2.0 has transformed users from passive consumers to active producers of content. This has exponentially increased the amount of information that is available to users, from videos on sites like YouTube and MySpace, to pictures on Flickr, music on Last.fm, blogs on Blogger, and so on. This content is no longer categorised according to pre-defined taxonomies (or ontologies). Rather, a new trend called *social* (or *folksonomic*) *tagging* has emerged, and quickly become the most popular way to describe content within Web 2.0 websites. Unlike taxonomies, which overimpose a hierarchical categorisation of content, folksonomies empower end users by enabling them to freely create and choose the tags that best describe a piece of information (a picture, a blog entry, a video clip, etc.). However, this freedom comes at a cost: since tags are informally defined, continually changing, and ungoverned, finding content of interest has become a main challenge, because of the number of synonyms, homonyms, polysemy, as well as the inevitable heterogeneity of users and the noise they introduce.

In order to assist users finding content of their own interest within this information abundance, new techniques, inspired by traditional recommender systems, have been developed: for example, whenever a user searches from some content using query tags $\{t_1, \ldots, t_m\}$, new tags $\{t_{m+1}, \ldots, t_{m+n}\}$ are being added to the query, based on their *similarity* to their original query tags. This is done to increase the chances of finding content of relevance in these extremely sparse settings. Various metrics have been used to compute the similarity among folksonomy entities, including, for instance, cosine similarity, Jaccard coefficient, and Pearson Correlation. Performance results demonstrate an increase in accuracy and coverage of searches when using these techniques; however, evaluation has been conducted on manipulated datasets, obtained by computing a p-core (see [7]) over the original folksonomy, so to obtain a much denser one ($p = 5$ in most cases). We argue that such manipulations alter the nature of real folksonomies, and indeed eliminate the problem, rather than solving it.

Unmodified real-world folksonomies are characterized by two key properties: the *power law* distribution of tags, and the *non-independence* of data. Empirical studies [4, 5] illustrate that tag usage in folksonomies follows a power law distribution; this means that, if we were select any two tags, the probability that the resources jointly labelled by them is non-zero is extremely low. As a result, computing tag similarity on un-modified folksonomies, using tra-

ditional metrics like cosine similarity, would almost always yield close-to-zero values, thus failing to support users in retrieving resources relevant to their queries. Furthermore, metrics like cosine assume that tags are semantically independent of each other; once again, this assumption does not hold in real folksonomies, where tags may be synonyms to each other.

In this paper, we propose a novel similarity metric that can be used to accurately quantify tag similarity in large-scale real-world folksonomies (Section 3). This similarity metric is computed following an iterative algorithm, grounded on a *mutual reinforcement* principle: that is, two tags are similar if they label similar resources, and vice-versa, two resources are similar if they have been labelled by similar tags. We describe an efficient realisation of this similarity metric (Section 4), and empirically quantify its quick convergence on three large-scale datasets, namely BibSonomy[1], MovieLens[2], and CiteULike[3]. We measure Precision and Recall of our metric, and compare it to cosine similarity, on different p-cores of the above datasets (Section 5). Our findings demonstrate that: on p-core of 5 (i.e., highly manipulated, dense datasets), the two metrics yields similarly-good results; however, when considering the original un-manipulated datasets, the performance of cosine degenerates, while our novel similarity metrics continues to provide high Precision and Recall. Section 6 covers related works on similarity measures, mainly applied to folksonomies. Finally, in Section 7 we draw our conclusions.

2   Background

In this section, we formally introduce some concepts that will be extensively used in the following, when presenting our approach. The first concept we consider is that of a folksonomy [13]:

Definition 2.1 *Let* $US = \{u_1, \ldots, u_{n_u}\}$ *be a set of users, let* $RS = \{r_1, \ldots, r_{n_r}\}$ *be a set of resource URIs and let* $TS = \{t_1, \ldots, t_{n_t}\}$ *be a set of tags. A folksonomy* F *is a tuple* $F = \langle US, RS, TS, AS \rangle$, *where* $AS \subseteq US \times RS \times TS$ *is a ternary relationship called* tag assignment set.

In this definition we do not make any assumption about the nature of resources; they could be URLs (like in Delicious), photos (as in Flickr), music files (as in Last.fm), documents (as in CiteULike), and so on.

According to Definition 2.1, a folksonomy F is a "three-dimensional" data structure whose "dimensions" are represented by users, tags and resources. In particular, an element $a \in AS$ is a triple $\langle u, r, t \rangle$, indicating that user u

---

[1] http://www.bibsonomy.org/
[2] http://www.movielens.org/
[3] http://www.citeulike.org/

labelled resource r with tag t. To simplify modeling and management of folksonomies, their inherent tripartite graph structure is often mapped into three *matrices*, whereby each matrix models one relationship at a time (i.e., between tags and resources, tags and users, and resources and users) [21]. In this paper, we adopt the same matrix-based representation. Specifically, being n_r, n_t and n_u the number of resources, tags and users respectively, we represent a folksonomy as the following three matrices:

- **TR** (Tag-Resource): a $n_t \times n_r$ matrix such that $\mathbf{TR}_{ij}$ is the number of times the tag i labelled resource j;

- **TU** (Tag-User): a $n_t \times n_u$ matrix such that $\mathbf{TU}_{ij}$ is the number of times the tag i has been used by user j;

- **RU** (Resource-User): a $n_r \times n_u$ matrix such that $\mathbf{RU}_{ij}$ is the number of times resource i has been labelled by the user j.

Tag similarity within a folksonomy can then be computed by looking at the resources these tags have been attached to. In particular, each tag t_i can be mapped onto a vector $\mathbf{t}_r(i)$ corresponding to the i-th row of **TR**. Given an arbitrary pair of tags t_i and t_j, their similarity $s(t_i, t_j)$ can be computed as the *cosine similarity* (CS) of the vectors $\mathbf{t}_r(i)$ and $\mathbf{t}_r(j)$:

$$s(t_i, t_j) = \frac{\langle \mathbf{t}_r(i), \mathbf{t}_r(j) \rangle}{\sqrt{\langle \mathbf{t}_r(i), \mathbf{t}_r(i) \rangle} \sqrt{\langle \mathbf{t}_r(j), \mathbf{t}_r(j) \rangle}} \quad (1)$$

being $\langle \cdot, \cdot \rangle$ the usual *inner product* in $\mathbb{R}^{n_r}$.

Cosine similarity has been successfully applied in the context of Information Retrieval [18]. Within a folksonomy, Equation 1 states that the similarity score of a pair of tags is high if they *jointly co-occur* in labelling the same subset of resources. However, two key properties of folksonomies, that is, *(i)* the power law distribution of tags and *(ii)* their non-independence, causes Equation 1 to yield very poor results in this domain, as we shall discuss next.

Power Law in Tag Distribution. Let us consider a real-world folksonomy like BibSonomy. BibSonomy [12, 14] is a social bookmarking service in which users are allowed to tag both URLs and scientific papers. A power law distribution of tags on scientific references. In particular, resources were described by no more than 5 different tags (roughly 81%), and usually less than 3 (roughly 58%). A small portion of frequently adopted tags used to bookmark scientific references, and a long tail of tags (roughly 81%) being used less than 5 times.

Following the above observations, matrix **TR** is rather sparse; thus, if we were to select any pair of tags t_i and t_j, most of the components of the corresponding vectors $\mathbf{t}_r(i)$ and $\mathbf{t}_r(j)$ would be 0 and, therefore their inner product would be close to 0. The cosine similarity between *any* t_i

and t_j would therefore be almost 0, regardless of the initial choice of t_i and t_j. Such counter-intuitive result is a result of the inadequacy of cosine similarity to capture properties of tags in large-scale real folksonomies.

Non-Independence of Tags. Cosine similarity implicitly assumes that the components of the vectors appearing in Equation 1 are *independent* of each other. Such an assumption does not often hold true. For instance, consider a folksonomy consisting of two resources r_1 and r_2, representing two different scientific papers, both discussing about folksonomies. Suppose that the paper associated with r_2 is an extension of the paper associated with r_1. Finally, assume to bookmark the resource r_1 with the tag $t_1 =$ "folksonomy" and to bookmark the resource r_2 with the tag $t_2 =$ "social tagging". In this case, the similarity between t_1 and t_2 computed according to Equation 1 would be 0, even if t_1 and t_2 should result similar each other. The mutual similarity between t_1 and t_2 can be assessed only if we consider the non-independence of the resources they label.

3 Approach Description

In this section, we present a new definition of tag (and resource) similarity, that is particularly suited to quantify similarity of elements (be them tags of resources) in datasets characterized by power law distribution and non-independence of data. Our definition of similarity relies on the *mutual reinforcement principle*:

> *Two tags are similar if they label similar resources, and conversely, two resources are similar if they are labelled by similar tags.*

In the following, we shall derive a mathematical formula to compute tag and resource similarity on the basis of the principle stated above. After this, we shall illustrate why our formula is able to effectively address the power law and non-independence challenges.

We designed an *iterative algorithm* to compute the similarity score. In the base case, given a pair of tags $\langle t_a, t_b \rangle$ and a pair of resources $\langle r_a, r_b \rangle$, we define the *tag similarity* $st^0(t_a, t_b)$ and the *resources similarity* $sr^0(r_a, r_b)$ as follows:

$$st^0(t_a, t_b) = \delta_{ab} \qquad sr^0(r_a, r_b) = \delta_{ab} \qquad (2)$$

being δ_{ab} the *Kronecker symbol*[4]. Equation 2 reflects the fact that, in the initial step, each tag (resp., resource) is similar only to itself and it is dissimilar to all other tags (resp., resources).

At the k-th step, let $st^{k-1}(t_a, t_b)$ (resp., $sr^{k-1}(r_a, r_b)$) be the tag (resp., resource) similarity between the tags t_a

---

[4]We recall that the Kronecker symbol δ_{ab} is equal to 1 if a and b coincide and 0 otherwise.

and t_b (resp., resources r_a and r_b). We apply the following rules to update $st^{k-1}(t_a, t_b)$ (resp., $sr^{k-1}(r_a, r_b)$):

$$st^k(t_a, t_b) = \frac{ST^k(t_a, t_b)}{\sqrt{ST^k(t_a, t_a)} * \sqrt{ST^k(t_b, t_b)}} \qquad (3)$$

$$sr^k(r_a, r_b) = \frac{SR^k(r_a, r_b)}{\sqrt{SR^k(r_a, r_a)} * \sqrt{SR^k(r_b, r_b)}} \qquad (4)$$

where:

$$ST^k(t_a, t_b) = \sum_{i,j=1}^{n_r} \mathbf{TR}_{ai} * \Psi_{ij} * sr^{k-1}(r_i, r_j) * \mathbf{TR}_{bj} \qquad (5)$$

$$SR^k(r_a, r_b) = \sum_{i,j=1}^{n_t} \mathbf{TR}_{ia} * \Psi_{ij} * st^{k-1}(t_i, t_j) * \mathbf{TR}_{jb} \qquad (6)$$

Here Ψ_{ij} is equal to 1 if $i = j$ and it is equal to ψ if $i \neq j$, where ψ (called *propagation factor*) is a value belonging to the interval $[0, 1] \in \mathbb{R}$.

Equations 3–4 rely on the following intuitions. Given a pair of tags $\langle t_a, t_b \rangle$, at the k iteration, we consider *all pair of resources* $\langle r_i, r_j \rangle$ in our folksonomy and we take their similarity $sr^{k-1}(r_i, r_j)$ into account to compute $st^k(t_a, t_b)$. In particular, we compute a *weighted sum* of all the similarity values $sr^{k-1}(r_i, r_j)$, where the weights reflect the *strength of the association* between the tag t_a and the resource r_i, and the tag t_b and the resource r_j. As a consequence, the higher the similarity between r_i and r_j, the higher the contribution of the association between the tag t_a and the resource r_i, and the tag t_b and the resource r_j. Finally, the term Ψ_{ij} is instrumental to give higher relevance to tags that labelled the very *same* resources, w.r.t. the fact that they labelled two *similar* (but different) resources.

Note that, in the special case in which $\psi = 0$, our method does not depend on k and Equations 3–4 reduce to the cosine similarity formulation. In fact, in this particular case, all the contributions $sr^{k-1}(r_i, r_j)$ and $st^{k-1}(r_i, r_j)$ are disregarded when $i \neq j$, and are taken into consideration only when $i = j$. Since all contributions $sr^{k-1}(r_i, r_i)$ and $st^{k-1}(r_i, r_i)$ are equal to 1 by definition, it follows that Equations 3–4 reduce to the cosine similarity formulation.

Equations 3–4 are able to effectively address the power law and non-independence of data challenges we outlined above. In fact:

- In the computation of tag (resp., resource) similarity, we leverage on the similarity of all pairs of resource (resp., tag) similarities. As a consequence, unlike cosine similarity, we do not restrict ourselves to consider only the resources jointly labelled by two tags (resp., the tags jointly labelling two resources), which can be few, but we *iteratively propagate* similarity scores by considering all the pairs of similar resources jointly labelled by the two tags (resp., all the pairs of similar

tags jointly labelling two resources). In this way we are able to face the power law occurring in tag usage.

- In our definition of similarity, if two tags label similar, *even if not coincident*, resources their similarity score will be greater than 0, whereas the cosine similarity would return 0. As a consequence, our similarity method takes into account forms of correlation among pairs of resources and/or tags rather than assuming their independence.

4 Realization

From a computational standpoint, Equations 3–4 could entail a large overhead for two reasons:

- From a theoretical standpoint, our approach may need an infinite number of iterations. As a consequence, we need a stopping criterium allowing us to safely terminate the execution of Equations 3–4 after a finite (and low) number of iterations.

- Equation 3 (resp., Equation 4) requires the computation of n_r^2 resource-resource (resp,. n_t^2 tag-tag) similarities, at each k-th step. This could make our similarity measure inapplicable in practical cases, because each iteration requires exactly $n_r^2 \times n_t^2$ computations.

Fortunately, there are two important results making our similarity measure applicable and entailing the same complexity level as cosine similarity. The first result can be stated by the theorem showed and proved in the appendix Appendix[5] which affirms that the sequences $st^k(t_a, t_b)$ and $sr^k(r_a, r_b)$ defined as in Equations 3–4 converge.

This theorem ensures that, after a certain number of iterations, Equations 3–4 converge to stable values. During experimentation conducted on three real folksonomies (see Section 5.1), we empirically found that convergence was achieved after as little as five iterations, thus suggesting that our similarity measure is applicable in practical cases.

Furthermore, Equations 3–4 can be defined, without any loss of generality, as a simple matrix product (such as in cosine similarity). Specifically, let st^k and sr^k be the tag-tag and resource-resource similarity matrices respectively, with $st^0 = I_t$ and $sr^0 = I_r$; here $st^0 = I_t$ (resp., $sr^0 = I_r$) is the $n_t \times n_t$ (resp., $n_r \times n_r$) *identity matrix*. If we indicate with the symbol "∘" the *Hadamard matrix product*[6] [8], at the k-th step, the st^k and sr^k matrices can be computed as:

$$st^k = ST^k \circ DT^k \qquad (7)$$

$$sr^k = SR^k \circ DR^k \qquad (8)$$

---

[5]See http://tinyurl.com/proof-seke2011.

[6]Given two matrices $\mathbf{A}$ and $\mathbf{B}$ of the same dimensions, the *Hadamard product* $\mathbf{A} \circ \mathbf{B}$ is a matrix of the same dimensions of $\mathbf{A}$ and $\mathbf{B}$ and it is defined as follows: $(\mathbf{A} \circ \mathbf{B})_{ij} = \mathbf{A}_{ij} \cdot \mathbf{B}_{ij}$

where:

$$\mathbf{ST}^k = \mathbf{TR} \times \left(\mathbf{\Psi_r} \circ sr^{k-1} \right) \times \mathbf{TR}^t \qquad (9)$$

$$\mathbf{SR}^k = \mathbf{TR}^t \times \left(\mathbf{\Psi_t} \circ st^{k-1} \right) \times \mathbf{TR} \qquad (10)$$

$$\mathbf{DT}_{ab}^k = \frac{1}{\sqrt{\mathbf{ST}_{aa}^k}\sqrt{\mathbf{ST}_{bb}^k}} \qquad (11)$$

$$\mathbf{DR}_{ab}^k = \frac{1}{\sqrt{\mathbf{SR}_{aa}^k}\sqrt{\mathbf{SR}_{bb}^k}} \qquad (12)$$

In the above equations, we have indicated with $\mathbf{\Psi}_r$ (resp., $\mathbf{\Psi}_t$) a square matrix $n_r \times n_r$ (resp., $n_t \times n_t$) where all the elements are set to ψ, with the exception of the diagonal where the elements are set to 1; the symbol $\mathbf{TR}^t$ represents the transpose of matrix $\mathbf{TR}$. We have thus reduced the computational complexity of each iterative step from $n_r^2 \times n_t^2$ to a simple matrix product; this reduction, coupled with the empirical observation that 5 iterative steps are sufficient to find convergence, makes our similarity metrics suitable in practical contexts. The last question that needs answering is how effective (in terms of Precision and Recall) our similarity metric is w.r.t. traditional ones like cosine. We answer this question next.

5 Experiments

In order to evaluate the performance of our similarity measure, we built a prototype in Java and MySQL and we conducted experiments using three well known social tagging websites: Bibsonomy, CiteULike, and MovieLens. The experiments we carried out aimed to answer the following question:

> *If we consider any two tags t_i and t_j belonging to a folksonomy, is our similarity measure capable of accurately assessing the extent to which they are related (similar) each other? And can it do so even when such tags have been drawn from the long tail of low popularity tags?*

5.1 The Dataset

To answer the above question, we conducted experiments on the following three datasets.

Bibsonomy. Bibsonomy is a social bookmarking website promoting the sharing of both scientific reference and general URL. We downloaded a snapshot of the website in June 2009, containing bookmarks made between January 1999 and June 2009.

CiteULike. CiteULike is a social bookmarking website that aims to promote and develop the sharing of scientific references amongst researchers. CiteULike enables scientists to organize their libraries with freely chosen tags which produce a folksonomy of academic interests. CiteULike

Dataset	Users	Resources	Tags	Bookmarks
Bibsonomy	4,696	578,587	147,076	648,924
CiteULike	57,053	1,928,302	401,620	2,281,609
MovieLens	4,009	7,601	15,240	55,484

Table 1: Datasets' features

Propagation	Bibsonomy	CiteULike	MovieLens
$\psi = 0$	0.100638896	0.057922233	0.075126961
$\psi = 0.15$	0.128318833	0.063290603	0.112358995
$\psi = 0.3$	0.139761842	0.070652236	0.115026291
$\psi = 0.6$	0.140748308	0.079320913	0.115534133

Table 2: Precision values in our datasets

Propagation	Bibsonomy	CiteULike	MovieLens
$\psi = 0$	0.100625714	0.057864697	0.075143054
$\psi = 0.15$	0.128373044	0.063273342	0.110927464
$\psi = 0.3$	0.139939546	0.070634975	0.119373901
$\psi = 0.6$	0.140429576	0.079303652	0.119949092

Table 3: Recall values in our datasets

runs a daily process which produces a snapshot summary of what articles have been posted by whom and with what tags up to that day. We downloaded one such archive in November 2009, containing bookmarks made between November 2004 to November 2009.

MovieLens. MovieLens is a rate-based recommendation website that suggests to users movies they might like. We downloaded such dataset in January 2009, containing bookmarks made from December 2005 to January 2009.

Table 1 summarizes the features of the involved datasets.

5.2 Simulation Setup

Our experimental investigation aimed to quantify, in each of the above datasets, the extent to which our similarity measure was capable of identifying related tags, especially when tags were drawn from the long tail. To investigate this, for each dataset of Table 1 has been used as follows. We split it into two different sets, called *test set* and *train set*. Each train set was composed of 90% random bookmarks taken from the involved dataset; we used these bookmarks for training purposes. Test sets contained the remaining 10% of bookmarks which were used for testing. Each bookmark in a test set has then been used as a query; specifically, if the number of tags in such bookmark was large enough, then these were split into two different sets – if possible of the same size – called $tSet_Q$ (*query tag set*) and $tSet_E$ (*expected tag set*). In our experiments, a bookmark was considered large enough if it had at least 3 tags associated. Tags composing $tSet_Q$ were used to query the train set; in particular, we selected from the train set the k tags most similar to tags belonging to $tSet_Q$, according to two metrics: the one we proposed in Section 3, and cosine similarity, which we used as benchmark. We denote this set as $tSet_R$ (*result tag set*). The value of k was chosen equal to the size of the expected set in such a way that $tSet_R$ and $tSet_E$ had the same size. Finally, we compared $tSet_R$ with $tSet_E$: the higher the overlap between $tSet_R$ and $tSet_E$, the more effective the similarity measure in identifying related tags. This follows the intuition that, if a user associated a set of tags to a certain resource, such tags are related to each other (that is, $tSet_E$ contains tags related to those contained in $tSet_Q$).

To quantitatively evaluate our similarity measure, we computed two metrics commonly used in Information Retrieval, namely Precision and Recall [1]:

$$Precision = \frac{|tSet_R \cap tSet_E|}{|tSet_R|} \qquad (13)$$

$$Recall = \frac{|tSet_R \cap tSet_E|}{|tSet_E|} \qquad (14)$$

We computed Precision and Recall values for each test bookmark; we repeated this process 10 times over different train and test splits of the datasets. The results we present next are averages of such runs.

5.3 Results

Tables 2 and 3 shows values of Precision and Recall we obtained by applying our similarity measure on the datasets of Table 1, for different values of ψ (see Equations 3–4). The benchmark is our similarity measure with $\psi = 0$, that is, the case in which our similarity measure degenerates into cosine similarity.

From the analysis of Tables 2 and 3 we can draw the following main observation: in large scale folksonomies, classical approaches – such as cosine similarity ($\psi = 0$) – have difficulties finding similarity relationships among the tags belonging to the long tail, as their Precision and Recall is lower than those achieved with our iterative approach for any value of ψ. In fact, the considered datasets are characterized by a very long and prominent tail of low popularity tags; in these real cases, out iterative measure of similarity produces Precision/Recall that is approximately 40% better than cosine similarity for BibSonomy and CiteULike, and approximately 50% better for MovieLens.

6 Related Work

In the last few years, folksonomies have been the subject of extensive research. An interesting survey on the characteristics of folksonomies can be found in [4]. One of the first investigations into the characteristics of folksonomies has been presented by Mathes [20]: in that work, the author discusses advantages (e.g., simplicity of use) and disadvantages (e.g., ambiguity, synonyms) of folksonomies, and

investigates the community aspects behind folksonomies (e.g., underlying motivations explaining its success, how cooperation works, how sharing thoughts reflects on users), on two scenarios, Flickr[7] and Delicious[8].

Despite their easy-of-use, the lack of structure that characterises folksonomies makes it difficult to browse and find relevant content. To tackle this issue, the research community has been actively researching techniques to support information retrieval. Approaches in this area have followed one of two streams: they have either tried to empirically derive an ontology from the underlying folksonomy, or they have tried to apply graph-exploration techniques on the folksonomy itself.

Lambiotte [16] and Mika [21], for example, were the first to extend the classic bipartite model of tag-resource towards a tripartite model, which takes into account both users (as actors), tags (as concepts) and resources (as instances); they showed that, by applying this model to Delicious, a lightweight ontology could be extracted from the underlying folksonomy. Similarly, [10] used similarity metrics to reconstruct a concept hierarchy.

Hotho et al. [13, 22] followed a different approach instead: they presented a formal model, which converts a folksonomy into an undirected weighted graph, and coupled it with a new search algorithm, namely "FolkRank", based on the well-known seminal "PageRank" [2]. They applied this algorithm to Delicious, and showed how it can be used as a tag recommender system. Other extensions of recommender systems to folksonomy structures have been explored [23, 15, 11]; some of these have been assessed against one of the datasets we adopted in this study, namely BibSonomy [12, 15, 14].

All the above approaches rely on a similarity measure to quantify tag relatedness. Measures which have been often used in the literature include the Jaccard coefficient [9], the cosine similarity [6], and a number of improvements over it [17, 24]. Liu et al. [17] dwelt further into the problem of computing similarities in folksonomies; in particular, they questioned the common assumption that text categorization can be mapped onto orthogonal spaces, due to problems of synonyms and ambiguities (as already figured out by [20]). They then devised an improved similarity metric ("SNOS", Similarity equations in the Non-Orthogonal Space) which is optimized for comparing objects mapped onto non-orthogonal spaces, considering a principle of "mutual reinforcement" from which we drew inspiration in this work. They proved the convergence of this technique and experimentally investigated the performance of SNOS on synthetic datasets, such as the formerly called MSN search engine (now, Bing[9]). Their novel metric was shown to out-

perform the classic cosine similarity, if applied to the context of finding similar queries. Some of their findings are here extended to the domain of folksonomies.

Similarity measures have often been evaluated on different datasets, making it difficult to assess their relative advantages and disadvantage in different domains. Furthermore, they have often been applied to manipulated datasets, making the comparison even more difficult. Indeed, in order to critically compare them, an evaluation framework has recently been proposed [19], with the aim of providing support to systematically compare several tag similarity measures, using data from Delicious [3]. This work contributes to the assessment of the suitability of similarity measures to scenarios characterized by power-law distribution of tags and non-independence of data, showing how traditional measures like cosine do not work, and proposing an alternative, iterative measure that provides good accuracy instead.

7 Conclusions

In this paper, we have shown that real world folksonomies are characterized by power law distributions of tags and non-independence of data. Under these conditions, traditional similarity measures like cosine similarity fail to capture tags relatedness. To remedy this, we have proposed a novel metric, specifically developed to capture similarity in large-scale folksonomies, that is based on the mutual reinforcement principle: that is, two tags are deemed similar if they have been associated to similar resources, and vice-versa two resources are deemed similar if they have been labelled by similar tags. We have described an efficient realisation of this similarity metric, and assessed its quality experimentally, by comparing it against cosine similarity, on three large-scale datasets, namely Bibsonomy, MovieLens and CiteULike.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley Longman, 1999.

---

[7] http://www.flickr.com/
[8] http://www.delicious.com/
[9] http://www.bing.com/

[2] 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, 1998.

[3] C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In *Proceedings of the 7th International Conference on The Semantic Web*, pages 615–631. Springer-Verlag, 2008.

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

[5] P. De Meo, G. Quattrone, and D. Ursino. Exploitation of semantic relationships and hierarchical data structures to support a user in his annotation and browsing activities in folksonomies. *Information Systems*, 34(6):511–535, 2009.

[6] J. Diederich and T. Iofciu. Finding communities of practice from user profiles based on folksonomies. In *Proceedings of the 1st International Workshop on Building Technology Enhanced Learning solutions for Communities of Practice*, pages 288–297, 2006.

[7] J. Gemmell, T. Schimoler, M. Ramezani, and B. Mobasher. Adapting K-Nearest Neighbor for Tag Recommendation in Folksonomies. *CEUR Workshop Proceedings*, Vol. 528, 2009.

[8] G.H. Golub and C.F. Van Loan. *Matrix Computations, third edition.* Johns Hopkins University Press, 1996.

[9] Y. Hassan-Montero and V. Herrero-Solana. Improving tag-clouds as visual information retrieval interfaces. In *International Conference on Multidisciplinary Information Sciences and Technologies*, pages 25–28, 2006.

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

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

[12] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. BibSonomy: A social bookmark and publication sharing system. In *Proceedings of the First Conceptual Structures Tool Interoperability Workshop*, pages 87–102, 2006.

[13] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. *The Semantic Web: Research and Applications*, pages 411–426, 2006.

[14] R. Jäschke, A. Hotho, C. Schmitz, and G. Stumme. Analysis of the publication sharing behaviour in BibSonomy. *Conceptual Structures: Knowledge Architectures for Smart Applications*, pages 283–295, 2007.

[15] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag Recommendations in Folksonomies. In *Proceedings of the 11th European conference on Principles and Practice of Knowledge Discovery in Databases*, pages 506–514. Springer-Verlag, 2007.

[16] R. Lambiotte and M. Ausloos. Collaborative Tagging as a Tripartite Network. *Lecture notes in computer science*, pages 1114–1117, 2006.

[17] N. Liu, B. Zhang, J. Yan, Q. Yang, S. Yan, Z. Chen, F. Bai, and W.Y. Ma. Learning similarity measures in non-orthogonal space. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 334–341. ACM, 2004.

[18] C.D. Manning, P. Raghavan, and H. Schutze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[19] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th international conference on World Wide Web*, pages 641–650. ACM, 2009.

[20] A. Mathes. Folksonomies-cooperative classification and communication through shared metadata. *Computer Mediated Communication*, 2004.

[21] P. Mika. 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, 2007.

[22] C. Schmitz, A. Hotho, R. Jäschke, and G. Stumme. Mining association rules in folksonomies. *Data Science and Classification*, pages 261–270, 2006.

[23] Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: Collaborative tag suggestions. In *Proceedings of Collaborative Web Tagging Workshop at 15th International World Wide Web Conference*, 2006.

[24] B. Zhang, H.J. Zeng, W.Y. Ma, Z. Chen, N. Liu, and J. Yan. Method and system for determining similarity of items based on similarity objects and their features, May 12 2009. US Patent 7,533,094.