Tag Relatedness in Image Folksonomies

Hatem Mousselly-Sergieh[*], Elöd Egyed-Zsigmond[†]
Gabriele Gianini[‡], Mario Döller[§], Jean-Marie Pinon[¶], Harald Kosch[‖]

Folksonomies – networks of users, resources, and tags – allow users to retrieve, organize and browse web contents. However, their advantages are limited mainly due to the noisiness of user provided tags. To overcome this issue, we propose an approach for characterizing related tags in folksonomies: we use tag co-occurrence statistics and Laplacian Score based feature selection in order to create empirical co-occurrence probability distribution for each tag; then we identify related tags on the basis of the dissimilarity between their distributions. To this purpose we introduce variant of the Jensen-Shannon Divergence, which is more robust to statistical noise. We experimentally evaluate our approach by using WordNet and compare it to a common tag-relatedness approach based on the cosine similarity. The results show the advantage of our approach over the competing method.

[*]Universität Passau, Innstr. 43, 94032 Passau, Germany
[†]Université Lyon, 20 Av. Albert Einstein, 69621 Villeurbanne, France
[‡]Università degli Studi di Milano, via Bramante 65, 26013 Crema (CR), Italy
[§]FH Kufstein, Andreas Hoferstr. 7, 6330 Kufstein, Austria
[¶]Université Lyon, 20 Av. Albert Einstein, 69621 Villeurbanne, France
[‖]Universität Passau, Innstr. 43, 94032 Passau, Germany

Contents

1 Introduction

Nowdays, collaborative tagging systems have become ubiquitous tools that allow users to add contents to the web, annotate them using tags, and share them. This creates complex networks of users, resources and tags which are commonly referred to as folksonomies. According to the degree of user collaboration, folksonomies are classified in two main categories: broad and narrow [1]. In broad folksonomies, e.g., del.icio.us[1], multiple users tag the same resources with a variety of terms; in narrow folksonomies, the tagging activity is mainly performed by the content creators. Image folksonomies like Flickr[2] belong to the latter category.

Tags simplify resource retrieval and browsing. Additionally, tagging allows users to annotate the same resources with several terms, which enables multifaceted organization. However, tagging suffers from several intrinsic issues: Mathes [2] points to two main issues of user-supplied tags: *ambiguity* and lack of synonym control, which is also known as *redundancy* [3]. Tag ambiguity arises when the same tag is used to indicate different meanings. Typical examples are word-sense ambiguity (e.g. the word "palm" in different context) and language ambiguity (e.g. "Gift" means *poison* in German and *present* in English) (for further details refer to [4]). On the other hand, tag redundancy emerges when different tags are used to describe the same thing. For instance using different syntactic forms to express the same thing (e.g. "New York" vs. "New-York") is very common among taggers.

To overcome these problems, researches worked on techniques for identifying related tags in folksonomies (e.g. [5, 6, 7]). The proposed solutions help to identify redundant tags and to resolve tag ambiguity by providing the needed context through groups of related tags.

Here a clarification is in order, about the use of the terms *similarity* and *relatedness*. Semantic *similarity* and semantic *relatedness* are two linked concepts but are not synonyms. The authors of [8] point out that semantic relatedness is a more general concept than semantic similarity: similar entities are semantically related via their similarity ("auto"-"car"), but non-similar entities may also be semantically related by meronymy ("hand"-"palm"), antinomy ("left"-"right"), rather than just frequent association. Applications typically require relatedness rather than similarity: for example, "leaf" and "hand" are cues which can be used to disambiguation of the term "palm".

Hereafter the term *dissimilarity* will be used as the opposite of *relatedness*.

Most research contributions adopt an existing tag-to-tag dissimilarity metrics, creates a tag dissimilarity matrix and then build over it a clustering algorithm: tags belonging to the same cluster will be assumed to correspond to the same meaning; distinct research contributions differ typically in the characteristics of the proposed clustering algorithm and in their performance measured for instance in terms of computational efficiency or in terms of the quality of the results.

So far, less research has focused on the dissimilarity measure used to create the tag dissimilarity matrix. Most approaches follow a simple procedure for creating the tag

[1] www.delicious.com (Accessed: 17/1/2014)
[2] www.flickr.com (Accessed: 17/1/2014)

dissimilarity matrix based on the cosine similarity of tag co-occurrence vectors. Despite the efficiency of the cosine method, we believe that more sophisticated dissimilarity metrics can significatively improve the tag clustering algorithms' quality of results.

The present paper investigates the effect of different (dis)similarity measures on identifying related tags in folksonomies. A key point of our method is a specific tag representation: we represent tags as empirical probability distribution. A tag empirical probability distribution is defined by the co-occurrence of the tag with other "special" tags, identified as features in the folksonomy.

In synthesis the method consists of the following steps: given a folksonomy in a typical representation of objects-tag associations

- First we determine the tags of the feature set, to this purpose we introduce a method based on the idea of Laplacian score for feature selection [9].
- Next, related tags are identified by calculating the distance between the corresponding probability distributions. For this purpose, we propose a new dissimilarity metrics based on the well-known Jensen-Shannon Divergence (JSD). The new metrics, called Adapted Jensen-Shannon Divergence (AJSD), takes into account the statistical fluctuations present in the empirical probability distributions and is more robust w.r.t. statistical noise than the bare JSD of the two empirical probability distributions.
- Finally we apply a standard clustering algorithm.

We experimentally evaluated the proposed approach and compared it to a common method for tag relatedness based on the cosine similarity. The results show the advantage of our approach.

The rest of the paper is organized as follows. In Section 2, the related work is reviewed. In Section 3 folksonomies are defined and the different options for building a tag's context are discussed. Our solution is presented in detail in Section 4 and the experimental evaluation is described in section 5. Section 6 concludes the paper and discusses the future work.

2 Related work

The definition of a tag relatedness metrics is an essential component for applications that depend on mining knowledge from collective user annotations. Conventionally, a tag relatedness metrics is used to create the tag dissimilarity matrix, which is used in a next step as input for a clustering algorithm to identify related tag groups.

The work [5] proposes a tag relatedness measure which is based on tag co-occurrence counts. In that approach, the co-occurrence of each tag pair is computed and a cut-off threshold is used to decide whether two tags are related. The cut-off threshold is determined using the first and the second derivatives of the tag co-occurrence curve. Finally, tag clusters are built by providing the computed tag similarity matrix as input to a spectral bisection clustering algorithm.

Gemmell and coauthors [6, 10] propose an agglomerative approach for tag clustering. For that purpose, they presente a tag relatedness measure based on the idea of *term frequency-inverse document frequency* (TF-IDF): in their approach the resources take the role of documents while the tags take the role of terms: each tag is represented as a vector of tag-frequency-inverse resource frequency and the similarity between two tags is defined by the cosine similarity between the tag vectors.

For their tag clustering approach, the authors of [11] propose a tag relatedness measure based on tag co-occurrence counts. First, the tags are organized in a co-occurrence matrix with the columns and the rows corresponding to the tags. The entries of the matrix represent the number of times two tags were used together to annotate the same resource. Next, each tag is represented by a co-occurrence vector and the similarity between two tags is calculated by applying the cosine measure on the corresponding vectors.

Simpson and coauthors [12] propose a tag relatedness approach which uses the Jaccard measure to normalize tag co-occurrences. The tags are then organized in a co-occurrence graph, which is then fed to an iterative divisive clustering algorithm to identify clusters of related tags.

The tag relatedness measure presented in [7] is based on a graph-theoretical metrics. Tags are organized in a graph with the edges weighted according to the structural similarity between the nodes: tags that have a large number of common neighbors are considered related.

Weinberger and coauthors [4] propose a statistical approach for identifying ambiguous tags based on the Kullback-Leibler (KL) divergence. For this purpose, a representation for each tag is created based on the co-occurrence with top frequent tags in the folksonomy.

All the above works start by exploiting tag co-occurrence counts to define their tag relatedness metrics. Subsequently, either a simple threshold for tag co-occurrences [5, 12] or the cosine measure are used to identify similar tags [6, 10, 11]. The present work with respect to the literature brings original contributions in mainly two respects:

- although we use the same representation for tags as probability distributions as done in [4], our method deals also with statistical fluctuations in the created probability distributions and propose extension for the well-known Jensen-Shannon Divergence;
- to best of our knowledge, this work is the first to deal with the problem of feature selection for building tag co-occurrence vectors: we propose a solution based on the method of Laplacian score for feature selection and demonstrate its advantage for tag relatedness measures.

3 Folksonomies and Tag Relatedness

A folksonomy F can be defined as a tuple $F = \{T, U, R, A\}$ [13] where T is the set of tags contributed by a set of users U to annotate a set of resources R, while A is a ternary assignment relation, i.e. $A \subseteq U \times T \times R$: a triple $(t, u, r) \in A$ captures the fact that a tag t has been used by user u to tag the resource r. We say that two tags $t_1, t_2 \in T$ co-occur if they are used by one or more users to describe the same resource $r \in R$.

3.1 Vector representations for a tag

By counting co-occurrences with the other tags we can define for each tag an histogram of empirical frequencies, and use the corresponding vector $v(t)$ as a representation of the tag itself: in this way the tag becomes a vector of the real space $\mathbb{R}^{|T|}$, indicated in short by $\mathbb{R}^T$. This representation of a tag is called *tag-context*.

More formally, in the $\mathbb{R}^T$ representation each tag $t \in T$ is defined as a vector $v(t) \in \mathbb{R}^T$, so that the entries t of the vector $v(t)$ correspond to the set of unique tags in the folksonomy and the value of each entry correspond to the measure of co-occurrence of t with the tag associated with that entry: this measure can be given in the form of a count or of a frequency (i.e. count of co-occurrences over total count of occurrences). Notice that in the latter case the vector $v(t)$ is an *empirical probability distribution*; this fact will be used later, during the definition of the tag dissimilarity.

Indeed, this idea can be generalized also to the other dimensions of the folksonomy: a tag can be represented as a vector in one of three possible real vector spaces: $\mathbb{R}^T$, $\mathbb{R}^U$ and $\mathbb{R}^R$, or in a combination of them [14].

The second kind of tag representation, $\mathbb{R}^U$, is called *user-context*. The entries of the tag vectors $v(t) \in \mathbb{R}^U$ correspond to the unique users in the folksonomy. The value of an entry related to a user $u \in U$ indicates how often u has used t in his annotation activity.

The third kind of tag representation, $\mathbb{R}^R$, is the *resource-context*. The entries of the tag vector $v(t) \in \mathbb{R}^R$ correspond to the unique resources in the folksonomy. The value of an entry related to a resource $r \in R$ corresponds to the number of times in which t was used to annotate r.

In the present paper we will use only the tag-context, for reasons which will be clarified hereafter.

3.2 Standard definitions of tag relatedness

Approaches for tag relatedness use one (or more) of the above mentioned vector space representations to characterize the related tags. This is done by generating the chosen vector representation of the two tags in all the possible, or relevant, tag pairs and then calculating the *cosine similarity* of each pair.

Hence for two tags $t_1, t_2 \in T$, which are represented by the vectors $v(t_1)$ and $v(t_2)$,

in a vector representation, the relatedness, $sim(t_1, t_2)$, can be defined by:

$$sim(t_1, t_2) = cos(v(t_1), v(t_2)) = \frac{v(t_1) \cdot v(t_2)}{||v(t_1)|| \cdot ||v(t_2)||} \quad (1)$$

The importance of each of the mentioned vector space representations for identifying related tags differs according to the type of the folksonomy, i.e., narrow or broad. In this paper, we focus on image folksonomies which are usually narrow folksonomies. Hence, *user-context* have a limited value in identifying related tags due to the low user interaction. As for the *resource-context*, there are two reasons which make it unsuitable for identifying related tags in image folksonomies. First, in image folksonomies it is unlikely that the same tag will be applied multiple times to describe the same photo. Second, whereas with textual resources further occurrences of the tags can be acquired by analyzing the associated textual context for images there is not in general an associated textual context. For those reasons, in this work we restrict to *tag-context*. Tag-context provides however rather rich information about the pattern of tag usage in the folksonomies.

In the next sections, we present our approach for identifying related tags by analyzing their co-occurrence patterns in the corresponding folksonomies. We also provide experimental evaluation using as a reference a widely used approach based on the cosine similarity of tag co-occurrence vectors.

4 Tag Relatedness Approach

We propose a tag relatedness approach using the *tag-context* representation.

Fig. 4 provides a generic description of the procedure we follow to identify related tags. We start from a folksonomy, represented by the included tags and the associated resources. Next, feature selection is applied to extract a set of important tags, called *feature set*. In order to isolate the feature set, we propose a feature selection approach based on the Laplacian score (LS) method [9]. After that, for each unique tag in the folksonomy a probability distribution is created based on the co-occurrence of that tag with the elements of the feature set. Finally, the relatedness between two tags is determined based on the dissimilarity between their probability distributions. To calculate this dissimilarity, we propose a new metric, called Adapted Jensen-Shannon Divergence (AJSD), based on the well-known Jensen-Shannon Divergence (JSD) [15], but adapted so as to make it robust w.r.t. statistical fluctuations present in the empirical probability distributions.

We expose the details of each phase in the upcoming subsections.

4.1 Feature Selection for Tag Relatedness

Identifying related tags in a folksonomy is an all-pairs-similarity-search problem (APSS) [16] since each tag has to be compared to all other tags in the folksonomy. Given the set of $|T|$ tags and considering that each tag is represented by a d dimensional vector, the naive approach would compute the similarity between all tag pairs in $O(|T|^2 \cdot |d|)$

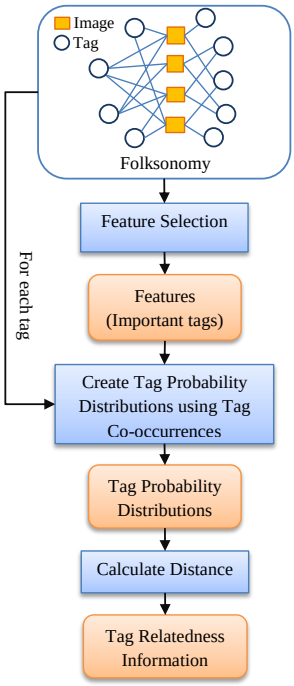

Figure 1: The work-flow of the proposed tag relatedness approach

time. In the case of tag-context approach where $d = |T|$ the algorithm will have a complexity $O(|T|^3)$. For large folksonomies, performing such computations is impractical. However, the computational cost can be reduced if the tags are represented in reduced vector space, i.e, $\mathbb{R}^{\mathcal{F}}$ where $\mathcal{F} \subset T$ and $|\mathcal{F}| \ll |T|$. Of course, in this case, the challenge is to provide a feature selection approach which can maintain, if not improve, the quality of the tag relatedness measure.

A simple approach to build the feature set $\mathcal{F}$, is to select a subset of the most frequent tags in the folksonomy (e.g. [14, 4]). This technique has some effectiveness, but a main issue: the most frequent tags may have almost uniform co-occurrence patterns with most other tags in the folksonomy; in this case, all tags would be considered related to each other. Hence, a more sophisticated approach for identifying $\mathcal{F}$ is required.

A possible solution for this challenge is provided by the Laplacian Score feature selection method [9]. LS is a technique based on a graph-theoretical metrics, for identifying good features for clustering problems: this makes it suitable also for tag relatedness approaches, which aim eventually at finding clusters, i.e., groups of related tags.

4.1.1 The Laplacian Score technique

The Laplacian Score (LS) technique for feature selection is based on Laplacian Eigenmaps [17] and Locality Preserving Projections [9] techniques. Those techniques allow to represent a dataset, whose points are characterized by a high dimensionality, by means of a lower dimensional representation, implicitly based on a low dimensional sub manifold of the whole space: those techniques postulate that such a manifold exist and that it can be represented efficiently in terms of a small subset of the data-points (those will be the selected features).

This schema fits the problem at hand: the keywords are the points of our dataset; they are represented initially by high-dimensional vectors (the co-occurrence frequency histograms with all the other keywords). The results of the application of the method described hereafter confirm ex-post the soundness of the assumptions.

To compute the LS of a dataset, the data-points are first organized in a weighted undirected graph, in which nodes correspond to data points and an edge is drawn between two nodes if they are close to one another according to some predefined similarity measure (such as the cosine measure); edges are weighted proportionally to the similarity between the connected data points. The Laplacian (matrix) L of such a graph is a square matrix defined by the difference of the degree matrix and the adjacency matrix (see below) of the graph: intuitively, the Laplacian matrix is a discrete analog of the Laplacian operator in multi-variable calculus and serves a similar purpose by measuring to what extent a graph differs at one vertex from its values at nearby vertices. Thanks to such a measure, one can define the Laplacian score for each individual vertex (the less it differs from the neighbors the higher its score) and consequently choose those points who turn out to have the highest scores, as representative features.

The feature selection algorithm and estimation for the solution of the objective function are summarized in the following steps (more details can be found in [9]):

1. For the set of n data points a k-nearest-neighbor graph is generated. In that graph, an edge between two data points x_i and x_j is drawn if the points are close to each other, i.e., if x_i belongs to the set of k nearest neighbors of x_j and vice versa.

2. The edges between close nodes are then weighted according to a similarity function. To calculate the similarity, there are several options: common measures are the cosine similarity (Equation 1) and the Gaussian similarity which is defined as:

$$S_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2u}} \tag{2}$$

where x_i, x_j are two data points and u is a free parameter that can be determined experimentally. Pairwise similarities of the data points are then combined into a similarity matrix S.

3. For a feature f, defined as a vector over the data points, let:

$$\tilde{f} = f - \frac{f^T D\mathbb{1}}{\mathbb{1}^T D\mathbb{1}}\mathbb{1} \tag{3}$$

where $\mathbb{1} = [1...1]^T$ and $D = diag(S\mathbb{1})$, i.e. D is a diagonal matrix in which each diagonal entry d_{ii} corresponds to the sum of the entries of the column i in the similarity matrix S.

4. Let $L = D - S$ be the Laplacian matrix of the similarity graph [18]. The Laplacian Score of the feature f is then computed as:

$$\mathcal{L}(f) = \frac{\tilde{f}^T L \tilde{f}}{\tilde{f}^T D \tilde{f}} \tag{4}$$

5. The final *feature set* $\mathcal{F}$ contains those features with a Laplacian Score greater than a predefined threshold θ:

$$\mathcal{F} = \{\ f \mid \mathcal{L}(f) > \theta\ \} \tag{5}$$

In our case, the data points as well as the features correspond to the tags of the folksonomy. That is, each tag in the *tag-context* representation, i.e, $v(t) \in \mathbb{R}^T$ defines a data point as well as a feature vector at the same time.

4.1.2 An illustrative example

To clarify how the Laplacian score algorithm can be applied to select important features in a folksonomy, consider the tag co-occurrence matrix shown in Figure 2. The column and the rows of the matrix corresponds to the tags while the entries correspond to the co-occurrence counts of the tag pairs, as observed in the folksonomy. The co-occurrence of a tag with itself is set to zero. In this example the tags "France" and "Paris" occur most. Furthermore, both tags show uniform occurrence patterns with the other tags.

	France	*Paris*	*Tower*	*Eiffel*	*Sky*	*City*
France	0	30	30	30	30	30
Paris	30	0	20	20	20	20
Tower	30	20	0	20	5	5
Eiffel	30	20	20	0	10	10
Sky	30	20	5	10	0	5
City	30	20	5	10	5	0

Figure 2: A sample tag co-occurrence matrix

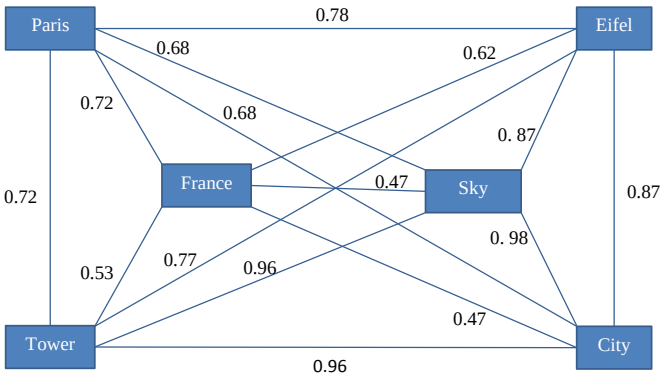

Figure 3: Similarity graph for the data points corresponding to the rows of the matrix shown in Figure 2. The nodes corresponds to the tags with edges weighted using the cosine similarity

Although the data point set and the feature set contain identical elements, for sake of clarity, here we make a distinction between them by denoting the data points by x_i and the features by f_i.

Data points, as well as features, can be derived directly from the rows and columns of the co-occurrence matrix, respectively. For example the data-point vector corresponding to the tag "France" is given by $x_{\text{France}} = (0, 30, 30, 30, 30, 30)$, while the feature vector corresponding to the tag Tower is given by $f_{\text{Tower}} = (30, 20, 0, 20, 5, 5)^T$.

In the next step, we create a weighted nearest-neighbor graph for data-points (step 1 and 2 of the algorithm). Due the small number of data points, we use a complete graph (instead of a nearest-neighbor graph) and chose the cosine similarity to weight the edges (Fig. 3). Next, the graph is mapped into a similarity matrix S. Additionally, the diagonal matrix D as well as the Laplacian of the graph L are calculated (Fig. 4).

$$S = \begin{pmatrix} 0 & 0.72 & 0.53 & 0.62 & 0.47 & 0.47 \\ 0.72 & 0 & 0.72 & 0.78 & 0.68 & 0.68 \\ 0.53 & 0.72 & 0 & 0.77 & 0.96 & 0.96 \\ 0.62 & 0.78 & 0.77 & 0 & 0.87 & 0.87 \\ 0.47 & 0.68 & 0.96 & 0.87 & 0 & 0.98 \\ 0.47 & 0.68 & 0.96 & 0.87 & 0.98 & 0 \end{pmatrix}$$

$$D = \begin{pmatrix} 2.81 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3.58 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3.93 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3.91 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3.97 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3.97 \end{pmatrix}$$

$$L = D - S = \begin{pmatrix} 2.81 & -0.72 & -0.53 & -0.62 & -0.47 & -0.47 \\ -0.72 & 3.58 & -0.72 & -0.78 & -0.68 & -0.68 \\ -0.53 & -0.72 & 3.93 & -0.77 & -0.96 & -0.96 \\ -0.62 & -0.78 & -0.77 & 3.91 & -0.87 & -0.87 \\ -0.47 & -0.68 & -0.96 & -0.87 & 3.97 & -0.98 \\ -0.47 & -0.68 & -0.96 & -0.87 & -0.98 & 3.97 \end{pmatrix}$$

Figure 4: The similarity matrix S, the diagonal matrix D and the Laplacian matrix L as generated from the nearest neighbor graph of Figure 3

Feature	Laplacian Score
f_{Sky}	-0.07
f_{City}	-0.07
f_{Tower}	-0.09
f_{France}	-0.14
f_{Eiffel}	-0.16
f_{Paris}	-0.23

Table 1: The feature vectors ordered according to their importance (Laplacian score) from most to least important

Now, we have all the needed information which enables us to calculate the Laplacian score for the features (tags) of our example according to equation (4). Table 1 shows the features and the corresponding LS scores in increasing order of importance. As we can see, the features "City" and "Sky" are considered more important by the LS algorithm than "France" and "Paris". This is because, the tags "Paris" and "France" have uniform co-occurrence patterns with all other tags, consequently, their influence on identifying groups of related data points is negligible or even biasing.

It is important to mention that the presented example is not representative enough, however, it gives an idea about the way in which the Laplacian score algorithm can be applied so as to discover important tags in folksonomies. Furthermore, it shows a main characteristic of the LS algorithm, namely its ability to determine the importance of the tags independently of their frequency of occurrence as well as to discover features of uniform co-occurrence patterns and reducing their importance.

4.2 Tag Probability Distribution

In this processing phase, each tag in the folksonomy is given a representation in terms of an empirical probability distribution. For this purpose, we quantify the co-occurrences of a given tag with each of the elements of the feature set. Recall the notation of the folksonomy $F = \{T, U, R, A\}$ and let $\Re : T \rightarrow \wp(R)$ be a function from the set of tags to the power set of the resource set, that maps a given *tag* to the *set* of resources which are annotated with it. That means, for a tag $t \in T$ we have:

$$\Re(t) = \{\ r \mid r \in R\ \wedge\ \exists u \in U\ \wedge\ \exists (u, t, r) \in A\ \} \tag{6}$$

The measure of co-occurrence of two tags can defined by the function $C : T \times T \rightarrow \mathbb{N}$, given by:

$$C(t_i, t_j) = |\Re(t_i) \cap \Re(t_j)| \tag{7}$$

Equation (7) means that the measure $C(t_i, t_j)$ of co-occurrence of two tags corresponds to the number of resources which are annotated by *both* of them.

To create an empirical probability distribution for a tag t, the co-occurrences of t with each feature $f \in \mathcal{F}$ are counted so as to obtain a histogram in the variable f.

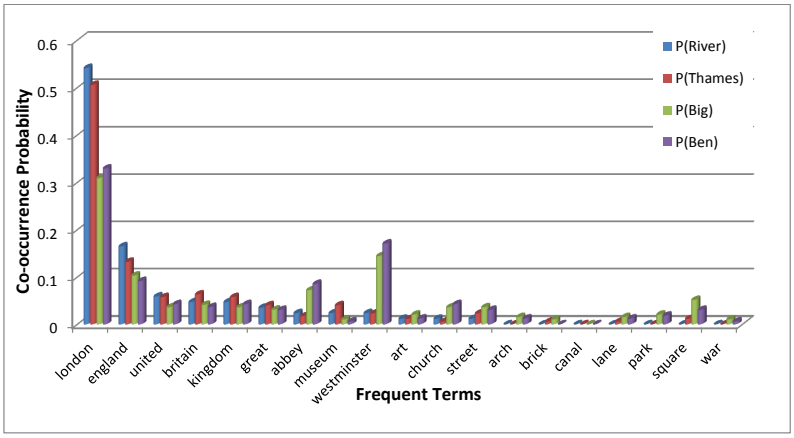

Figure 5: Empirical probability distributions of four tags (River, Thames, Big and Ben) which were used to annotate images taken in London. Each distribution consists of several histogram channels corresponding to the elements of a feature set (x-axis). The value of a histogram channel is given by the normalized co-occurrence of each of the four tags with the corresponding element from the feature set

Then, by normalizing this histogram, with the total number of co-occurrences of t with the elements of the set $\mathcal{F}$, a vector representing the empirical co-occurrence probability distribution $P(f|t)$ for the tag t with the elements $f \in \mathcal{F}$ is obtained:

$$P(f|t) = \frac{C(t,f)}{\sum_{f \in \mathcal{F}} C(t,f)} \tag{8}$$

where C is the tag-to-tag co-occurrence function given in equation (7). Each entry f of the vector $P(\,f\,|\,t\,)$ corresponds to the set of unique tags in the folksonomy which have been designated as features in the previous phase – the *feature tags* – while the value $P(\,f\,|\,t\,)$ of each entry corresponds to the measure of co-occurrence of t with the *feature tag* associated with that entry; The empirical probability distribution of the tag t over the complete set of features $\mathcal{F}$ can be denoted in short by $P(\mathcal{F}|t)$. Figure 5 shows sample segments of the empirical probability distributions corresponding to the tags "River", "Thames", "Big" and "Ben", which have been used to annotate photos taken in the city of London. The x-axis corresponds to the elements of the feature set, which, in this example, consists of a subset of the most frequent tags in the associated folksonomy. Each feature is represented by a histogram channel while the value of the channel (y-axis) corresponds to the normalized co-occurrence counts – equation (8) – with each of the four tags. Note that, the tags "Big" and "Ben" show a similar co-occurrence behavior. The same holds also for the tags "River" and "Thames".

4.3 Dissimilarity Metrics

At this point of the procedure, in order to determine if two tags are related, the dissimilarity between their corresponding empirical co-occurrence probability distributions must be computed. In the literature, the Jensen-Shannon Divergence (JSD) [15] is a widely used metrics which has shown to outperform other measures [19]. It is based on Kullback-Leibler Divergence (KLD) [20], however, it is symmetric and has always a finite value.

Since the presented tag probability distributions are created from samples (ideally drawn from the true distribution), and are necessarily affected by statistical fluctuations, we propose an extension of the standard JSD measure, called Adapted Jensen-Shannon Divergence (AJSD), based on a Maximum Likelihood estimate of the JSD which both takes into account fluctuations and provides a measure of the statistical error of the results.

Before introducing the new metric, we review the KLD and JSD approaches to calculate the distance between probability distribution. Let us consider two tags $t_1, t_2 \in T$ and the corresponding empirical co-occurrence probability distributions $P(\mathcal{F} \mid t_1)$ and $P(\mathcal{F} \mid t_2)$ over the feature set $\mathcal{F} = \{f_1, ..., f_m\}$. We can simplify the notation – by omitting at the same time the feature sets from the arguments – as follows: $P(\mathcal{F}) \equiv P(\mathcal{F} \mid t_1)$ and $Q(\mathcal{F}) \equiv P(\mathcal{F} \mid t_2)$; the values of P and Q at a specific feature $f \in \mathcal{F}$, will hereafter be represented simply by $P(f)$ and $Q(f)$, respectively.

The most typical metrics for dissimilarity between two probability distributions is the Kullback-Leiber divergence D_{KL}, defined as follows:

$$D_{KL}(P||Q) = \sum_{f \in \mathcal{F}} P(f) \log \frac{P(f)}{Q(f)} \tag{9}$$

Notice that the expression $D_{KL}(P||Q)$ is asymmetric in its arguments, i.e in general $D_{KL}(P||Q) \neq D_{KL}(Q||P)$. This problem can be solved by adopting, as a definition of divergence, a symmetrized version of the previous expression:

$$D_{SKL}(P||Q) = \frac{1}{2} \left\{ \sum_{f \in \mathcal{F}} P(f) \log \frac{P(f)}{Q(f)} + \sum_{f \in \mathcal{F}} Q(f) \log \frac{Q(f)}{P(f)} \right\} \tag{10}$$

However the KL divergences become infinite as soon as either P or Q vanish in one point of the support set, due to the denominators in the logarithm arguments of the two terms. This problem can be fixed by using the Jensen-Shannon (JS) Divergence, which is given by the following equation

$$D_{JS}(P||Q) = \frac{1}{2} \sum_{f \in \mathcal{F}} \left(P(f) \log \frac{2P(f)}{P(f) + Q(f)} + Q(f) \log \frac{2Q(f)}{P(f) + Q(f)} \right) \tag{11}$$

which differs from the SKL divergence of equation (10) in that the denominator of the logarithm's argument consists now in the arithmetic average $\frac{P(f)+Q(f)}{2}$ of the two functions.

4.3.1 Adapted Jensen-Shannon Divergence (AJSD)

If, as in our case, the probabilities P and Q are not available, we have an estimate of them through a finite sample represented in the form of a histogram for P and a histogram for Q. In this case the divergence computed on the histograms is a random variable. This variable, under appropriate assumptions, can be used to compute an estimate of the divergence between P and Q using error propagation under a Maximum Likelihood (ML) approach, as illustrated hereafter.

For P and Q consider that the channels at a point (feature) f of the corresponding histograms are characterized by the number of co-occurrences with f, denoted as k_f and h_f respectively. We define the following measured frequencies where:

$$x_f \equiv k_f/n \quad y_f \equiv h_f/m \tag{12}$$

Here, $n = \sum_f k_f$ and $m = \sum_f h_f$ are the sum of counts for the first and second histogram, respectively. When the number of co-occurrences is high enough (large n and m), the quantities x_f and y_f can be considered to have normal distributions around the true probabilities $P(f)$ and $Q(f)$ respectively. Consequently, the *measured* JSD, denoted as d, can be considered as a stochastic variable defined as a function of the two normal variables x_f and y_f. By substituting x_f and y_f in Equation 11 we get:

$$d = \frac{1}{2}\sum_{f\in\mathcal{F}}\left(x_f \log\frac{2x_f}{x_f+y_f} + y_f \log\frac{2y_f}{x_f+y_f}\right) \tag{13}$$

The value of this expression does not correspond, in general, to the maximum likelihood (ML) estimate of JSD since the variances of the terms in the sum are unequal. In order to find the maximum likelihood estimate $\hat{d}$ of the divergence, we need to proceed through error propagation as in the following steps:

1. Thanks to the normality condition stated above, the ML estimate of $P(f)$ corresponds to $x_f = k_f/n$ with the variance given in a first approximation by $\sigma^2_{P(f)} = k_f/n^2$. Similarly, the ML estimate of $Q(f)$ is $y_f = h_f/m$ with the variance given by $\sigma^2_{Q(f)} = h_f/m^2$.

2. We represent the individual addendum term in the sum expression of equation (13) as a random variable z_f:

$$z_f \equiv x_f \log\frac{2x_f}{x_f+y_f} + y_f \log\frac{2y_f}{x_f+y_f} \tag{14}$$

   If the two variables x_f and y_f are independent, the variance propagation at the first order is given by:

$$\sigma^2(z_f) \simeq \left(\frac{\partial z_f}{\partial x_f}\right)^2\sigma^2(x_f) + \left(\frac{\partial z_f}{\partial y_f}\right)^2\sigma^2(y_f) \tag{15}$$

$$\simeq \log^2\frac{2x_f}{x_f+y_f}\sigma^2(x_f) + \log^2\frac{2y_f}{x_f+y_f}\sigma^2(y_f) \tag{16}$$

The variance $\sigma^2(z_f)$ can be easily calculated by substituting the quantities of step 1 in the equation (16).

3. Define the (statistical) precision w_f (to be used later as a weight) as: $w_f \sim \frac{1}{\sigma^2(z_f)}$. Then, the maximum likelihood estimate of the quantity d of equation (13) is given by the following weighted sum:

$$\hat{d} = \frac{\sum_f w_f z_f}{\sum_f w_f} \equiv D_{AJSD}(P||Q) \tag{17}$$

With the variance given by:

$$\sigma^2(\hat{d}) = \frac{1}{\sum_f w_f} \tag{18}$$

We use $\hat{d}$ as Adapted Jensen-Shannon Divergence (AJSD). Note that, due to the statistical fluctuations in the samples, AJSD gives, in general, values greater than zero even when two samples are taken from the same distribution, i.e. even when the true divergence is zero. However, by weighting the terms according to their (statistical) precision, the scores produced by AJSD are expected to provide better estimate of the divergence than JSD does (see next section).

5 Evaluation

5.1 Dataset

To evaluate the performance of the proposed tag relatedness approach we performed several experiments on a folksonomy extracted from Flickr. The folksonomy corresponds to images taken in the area of London[3]. To avoid bulk tagging we restricted the dataset to one image per user. The final dataset contains around 54,000 images with 4,776 unique tags occurring more than 10 times and a total of 544,000 tag assignments.

5.2 Qualitative Insight

For each of the 4,776 unique tags in the dataset, we identified its most related tags. Table 2 shows sample tags (first column) with the corresponding related tags ordered according to their degree of relatedness from left to right. The related tags are obtained by the cosine (COS), JSD and AJSD measures, respectively, and by using the top 2000 Laplacian features. First, one can notice the overlap among the groups of related tags corresponding to the same initial tag. That is, because the tag relatedness measures use the same context, namely the tag-context. Second, we have recognized that, in general, the groups of related tags which are identified by AJSD have a higher cardinality than

[3]Dataset and code: https://sites.google.com/site/hmsinfo2013/home/software

Initial Tag	Method	Related Tags
Airport	COS	Heathrow, KLM, duty, check, airports, runway
	JSD	Heathrow, runway, African, international, ramp
	AJSD	Heathrow, ramp, departures, president, restaurants
Car	COS	automobile, Citroen, driving, rolls, pit, wreck
	JSD	cars, classic, motor, Sunday, Ford, Mini, BMW, driving
	AJSD	cars, classic, Sunday, Ford, Mini, BMW, driving, Caterham, pit
Garden	COS	Covent, jardin, ING
	JSD	flower, gardens, rose, Covent, jardin
	AJSD	flower, gardens, Covent, jardin, pots, Nicholson, rocks
Thames	COS	path, Kingston, river, mud, embankment, Sunbury, shore
	JSD	river, path, Kingston, riverside, Greenwich, ship, embankment
	AJSD	river, water, riverside, path, Kingston, Greenwich, embankment
Music	COS	musician, bands, records, fighting, acoustic
	JSD	concert, rock, stage, festival, pop, jazz, song, records
	AJSD	concert, rock, festival, stage, pop, jazz, Simon, song
Olympics	COS	triathlon, men's
	JSD	Olympic, men's, arena, venue, women's, athlete
	AJSD	Olympic, men's, center, athlete, women's, venue, game, triathlon

Table 2: Sample tags with the corresponding most related tags

their counterparts which are identified using JSD and the cosine approaches (e.g. Car, Garden in Table 2). The reason for this is that AJSD generates non-zero similarity even if the two tags have different sample distributions (section 4.3.1).

To investigate the effect of feature selection, we applied the Laplacian score method on the dataset to identify the most important tags. To generate the tag graph we set the number of nearest neighbors to 10 and used the Gaussian similarity function with t = 1. Fig. 6 shows a plot of the top tags according to LS against the number of occurrences of the tag (frequency). Additionally, the plot illustrates the most frequent tags in the folksonomy (italic). According to LS, the importance of a tag is determined according to its graph-preserving power and not according to its frequency. For example a tag like *potter* which is much less frequent than the tag *england* has a higher Laplacian score, thus, considered as more important. This can be explained since the folksonomy contains images taken in London, thus, it is very likely that most images will be tagged with the word *england* disregarding their contents. Correspondingly, *england* should have a kind of uniform co-occurrence with all other tags in the folksonomy. Therefore, it is less discriminative (has a low LS) than a more specific tag

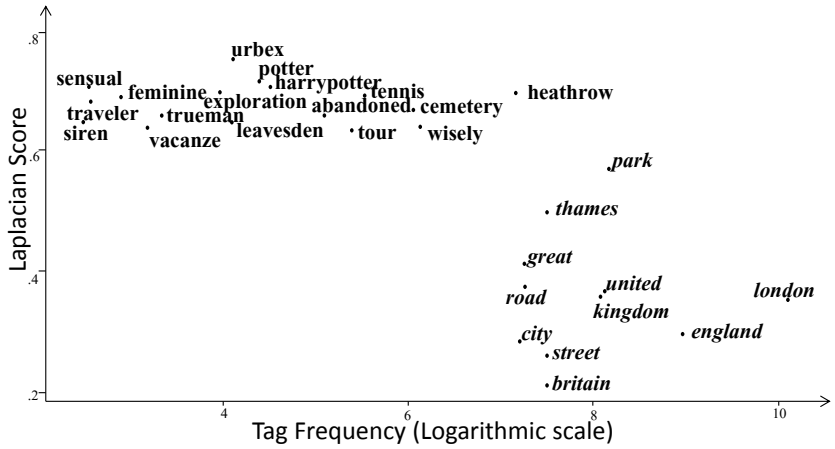

Figure 6: Tags importance (Laplacian Score) vs. tag frequency

like *potter* which expected to have non-uniform tag co-occurrence distribution.

5.3 Semantic Grounding using WordNet

To provide a quantitative evaluation, we performed additional experiments using WordNet[4]. WordNet has been used by several works as a tool for semantically grounding tag relatedness measures [14, 21, 22]. The goal is to assess how a given tag relatedness measure approximates a reference measure. For our study, we used the Jinag & Conrath (JCN) distance measure as a reference since it showed a high correlation with human judgment [23]. Initially, a gold standard dataset was created by extracting most similar tag pairs from our dataset according to WordNet and by applying JCN measure. After that, the relatedness between the tag pairs of the gold standard is calculated according to our tag relatedness approach as well as the cosine method. To evaluate the effectiveness of LS feature selection, we performed several experiments using different thresholds on the number of top LS features. Furthermore, we compared the performance of LS to frequency based features selection (FRQ).

The performance of the tag relatedness measures is determined according to the average JCN distance over the collection of most related tag pairs as identified by each of the investigated methods. Figure 7 shows the average JCN distance for the most similar tag pairs (y-axis). The x-axis corresponds to the number of the features. The compared methods include the three measures JSD, AJSD and Cosine (COS) combined with the two features selections approaches, namely the Laplacian score (LS) and the

[4]http://wordnet.princeton.edu/ (Accessed: 17/1/2014)

frequency based approach (FRQ) which identifies the features by simply selecting the top N most frequent tags. The number of tag pairs which have correspondences in WordNet varies according to the applied similarity method. The average number of recognized WordNet pairs is 975 per method with a standard deviation of 81,6. The standard error in estimating the average JCN distance depends also on the similarity method. However, we observed close values in the range [0.15,0.19].

LS leads to lower average JCN distance than FRQ for all similarity measures and disregarding the number of features (Figure 7). Moreover, LS enables reducing the dimension of co-occurrence vector/probability distribution while preserving the quality of the identified similar tag pairs. For instance, a minimum JCN distance can be achieved when the top 1,500 Laplacian features (around 31% of total unique tags) are used to perform the calculation. Finally, regarding the distance measures, AJSD produces shorter JCN distances than JSD which in turn performs better than the cosine measure (Figure 7).

Since the distributional properties of the investigated measures can be different, we followed the evaluation method described in [22]. Thereby, the performance of two tag relatedness measure can be compared according to how they rank the most similar tag pairs generated by each of them. To do this, the correlation between the rankings of each tag relatedness approach and the corresponding rankings of the reference measure (here JCN) is calculated. A suitable measure is provided by *Kendall* τ rank correlation coefficient.

Figure 8 shows that the performance of the tag relatedness measures based on Kendall correlation is in correspondence with our observations when JCN is used for the evaluation. AJSD combined with LS provides a higher correlation with WordNet than JSD and COS . By Using AJSD, we can even reduce the dimension of the probability distribution to 80% (the top 1,000 LS tags) while getting the best correlation with WordNet. Moreover, the frequency features selection have a much negative impact on the cosine approach. COS-FRQ is negatively correlated with WordNet as long as the number of features is below 3,000. In contrast, LS leads to a positive correlation factor in all cases.

6 Conclusion

In this paper we introduced a tag relatedness approach based on the representation of the tag data in terms of co-occurrence vectors, which differs form the current approaches in terms of two elements: 1) we used the Laplacian Score feature selection in order to reduce the dimension of the representation and had each tag correspond to a histogram with a limited number of channels 2) we compared the different tags/histograms by a metrics derived as a maximum likelihood estimate of the Jensen-Shannon Divergence.

As a reference for validation we used the WordNet dataset and the Jinag and Conrath distance (JCN). Our adapted JSD metrics (AJSD) displays a better performance of the original JSD metrics: it discovers tag pairs of smaller WordNet (JCN) distances and of higher correlation with WordNet. Furthermore, both AJSD and JSD perform better than the cosine measure.

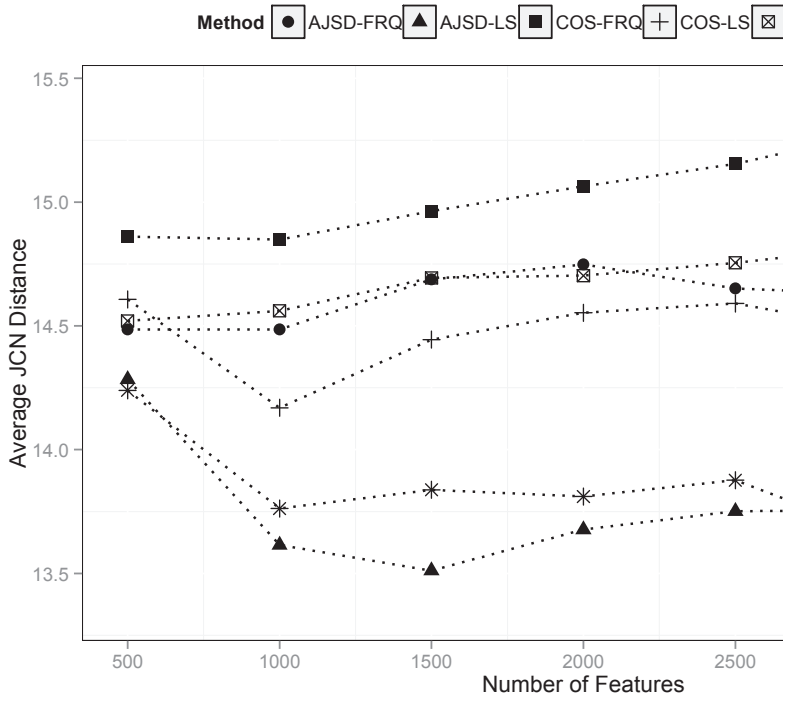

Figure 7: Average JCN achieved AJSD, JSD and COS. These measure are investigated using LS feature selection and the FRQ method and using increasing number of features

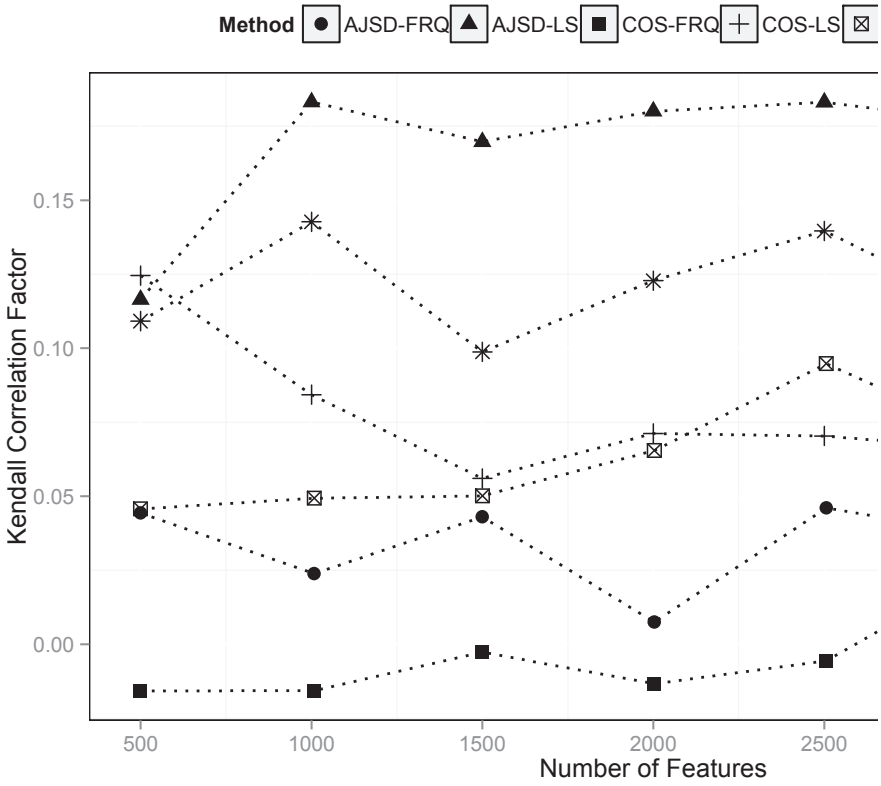

Figure 8: Kendall τ correlation achieved AJSD, JSD and COS. These measure are investigated using LS feature selection and the FRQ method and using increasing number of features

In future work, we will work on improving the performance of our approach by determining the best parameter values for the Laplacian Score. Also, we aim at evaluating the performance of our approach by integrating it into a tag recommendation system.

Acknowledgements. This work was partly supported by the UFI (Université Franco-Italienne) program Vinci2011, project C4-9 and by the UFA/DHA (Université Franco-Allemande / Deutch-Französische Hochschule) through the project PICS/MDPS. One of the authors (GG) was supported by the CNRS c.n. 30022/2011.

References

[1] T. Vanderwal, "Off the top: Folksonomy entries." http://vanderwal.net/random/category.php?cat=153, 2010. Accessed: 17/1/2014.

[2] A. Mathes, "Folksonomies-cooperative classification and communication through shared metadata," *Computer Mediated Communication*, vol. 47, no. 10, 2004.

[3] J. Gemmell, M. Ramezani, T. Schimoler, L. Christiansen, and B. Mobasher, "The impact of ambiguity and redundancy on tag recommendation in folksonomies," in *Proceedings of the third ACM conference on Recommender systems*, RecSys '09, (New York, NY, USA), pp. 45–52, ACM, 2009.

[4] K. Q. Weinberger, M. Slaney, and R. Van Zwol, "Resolving tag ambiguity," in *Proceedings of the 16th ACM international conference on Multimedia*, MM '08, (New York, NY, USA), pp. 111–120, ACM, 2008.

[5] G. Begelman, P. Keller, F. Smadja, *et al.*, "Automated tag clustering: Improving search and exploration in the tag space," in *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, pp. 15–33, 2006.

[6] J. Gemmell, A. Shepitsen, B. Mobasher, and R. Burke, "Personalization in folksonomies based on tag clustering," *Intelligent techniques for web personalization & recommender systems*, vol. 12, 2008.

[7] S. Papadopoulos, Y. Kompatsiaris, and A. Vakali, "A graph-based clustering scheme for identifying related tags in folksonomies," in *Proceedings of the 12th international conference on Data warehousing and knowledge discovery*, DaWaK'10, (Berlin, Heidelberg), pp. 65–76, Springer-Verlag, 2010.

[8] A. Budanitsky and G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness," *Computational Linguistics*, vol. 32, pp. 13–47, 2006.

[9] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," *Advances in Neural Information Processing Systems*, vol. 18, p. 507, 2006.

[10] J. Gemmell, A. Shepitsen, B. Mobasher, and R. Burke, "Personalizing navigation in folksonomies using hierarchical tag clustering," in *Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery*, DaWaK '08, (Berlin, Heidelberg), pp. 196–205, Springer, 2008.

[11] L. Specia and E. Motta, "Integrating folksonomies with the semantic web," in *Proceedings of the 4th European conference on The Semantic Web: Research and Applications*, ESWC '07, (Berlin, Heidelberg), pp. 624–639, Springer-Verlag, 2007.

[12] E. Simpson, "Clustering Tags in Enterprise and Web Folksonomies," *HP Labs Techincal Reports*, 2008.

[13] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme, "Information retrieval in folksonomies: Search and ranking," in *The semantic web: research and applications*, pp. 411–426, Springer, 2006.

[14] C. Cattuto, D. Benz, A. Hotho, and G. Stumme, "Semantic grounding of tag relatedness in social bookmarking systems," in *The Semantic Web-ISWC 2008*, pp. 615–631, Springer, 2008.

[15] C. Manning and H. Schütze, *Foundations of statistical natural language processing*. MIT press, 1999.

[16] R. J. Bayardo, Y. Ma, and R. Srikant, "Scaling up all pairs similarity search.," *WWW*, vol. 7, pp. 131–140, 2007.

[17] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, pp. 1373–1396, June 2003.

[18] F. R. Chung, *Spectral Graph Teory*, vol. 92. Amer Mathematical Society, 1997.

[19] N. Ljubešić, D. Boras, N. Bakarić, and J. Njavro, "Comparing measures of semantic similarity," in *30th International Conference on Information Technology Interfaces, Cavtat*, 2008.

[20] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[21] G. Srinivas, N. Tandon, and V. Varma, "A weighted tag similarity measure based on a collaborative weight model," in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pp. 79–86, ACM, 2010.

[22] 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*, pp. 641–650, ACM, 2009.

[23] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *arXiv preprint cmp-lg/9709008*, 1997.