Tag Relatedness Using Laplacian Score Feature Selection and Adapted Jensen-Shannon Divergence

Hatem Mousselly-Sergieh[1,2], Mario Döller[3], Elöd Egyed-Zsigmond[2], Gabriele Gianini[4], Harald Kosch[1], and Jean-Marie Pinon[2]

[1] Universität Passau, Innstr. 43, 94032 Passau, Germany
[2] Université de Lyon, 20 Av. Albert Einstein, 69621 Villeurbanne, France
[3] FH Kufstein, Andreas Hoferstr. 7, 6330 Kufstein, Austria
[4] Università degli Studi di Milano, via Bramante 65, 26013 Crema, Italy
firstname.lastname@uni-passau.de, firstname.lastname@insa-lyon.fr, mario.doeller@fh-kufstein.ac.at, gabriele.gianini@unimi.it

Abstract. Folksonomies - networks of users, resources, and tags allow users to easily retrieve, organize and browse web contents. However, their advantages are still limited according to the noisiness of user provided tags. To overcome this problem, we propose an approach for identifying related tags in folksonomies. The approach uses tag co-occurrence statistics and Laplacian score feature selection to create probability distribution for each tag. Consequently, related tags are determined according to the distance between their distributions. In this regards, we propose a distance metric based on Jensen-Shannon Divergence. The new metric named AJSD deals with the noise in the measurements due to statistical fluctuations in tag co-occurrences. We experimentally evaluated our approach using WordNet and compared it to a common tag relatedness approach based on the cosine similarity. The results show the effectiveness of our approach and its advantage over the adversary method.

Keywords: Folksonomies, Tag Relatedness, Laplacian Score, JSD.

1 Introduction

In the current internet era, collaborative tagging systems become ubiquitous tools which allow users to add contents to the web, annotate them using keywords called tags, and share them with each other. This results in a complex network of users, resources and tags which is commonly referred to as a folksonomy. 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 variety of terms. In narrow folksonomies, the tagging activity is mainly performed by the content creators. Image folksonomies like Flickr[2] belong to this category.

[1] www.delicious.com
[2] www.flickr.com

C. Gurrin et al. (Eds.): MMM 2014, Part I, LNCS 8325, pp. 159–171, 2014.

Tags simplify resource retrieval and browsing. Additionally, tagging allow users to annotate the same resources with several terms, thus, they can organize their resources in multiple categories. However, the unsupervised way of tag creation makes them suffer from noise, such as redundancy (different tags vs. same meaning) and ambiguity (same tag vs. different meanings) [2]. To overcome these problems, researches worked on techniques for identifying related tags in folksonomies (e.g. [3–5]). The proposed solutions help to identify redundant tags and to resolve tag ambiguity by providing the needed context through groups of related tags. Generally, the main research directions of most contributions are on investigating and proposing efficient clustering algorithms to determine similar tags. However, little research has focused on the (dis)similarity measure which is used to create the tag dissimilarity matrix (the main input for clustering algorithms). Most approaches follow a simple procedure for creating the tag dissimilarity matrix based on the cosine similarity of tag co-occurrence vectors. Although the cosine method seems to be efficient, we believe that more sophisticated measures can help to boost the performance of the tag clustering algorithms.

In this paper we propose a tag relatedness measure which deal with tags as probability distributions. Initially, a probability distribution is generated for each tag in the folksonomy. This is done based on the co-occurrence of each tag with a subset of tags called the feature set. To determine the feature set, we present a solution based on the idea of Laplacian score for feature selection [6]. Next, related tags are identified by calculating the distance between the corresponding probability distributions. For this purpose, we propose a new distance measure based on the well-known Jensen-Shannon Divergence (JSD). The new measure, called Adapted Jensen-Shannon Divergence (AJSD), can efficiently deal with fluctuations in the samples from which the probability distributions are created. We experimentally evaluated the proposed approach and compared it to a common method for tag relatedness based on the cosine similarity. The results are promising and show the advantage of our approach.

Section 2 surveys related work. In section 3, a definition for folksonomies is provided. In section 4 the proposed approach is presented. Section 5 shows experimental results. Section 6 provides a conclusion and discusses future work.

2 Related Work

Tag relatedness is essential component for applications that depend on mining knowledge from collective user annotations. Conventionally, a tag relatedness measure 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 in [3] 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 using the computed tag similarity matrix as input to a spectral bisection clustering algorithm.

Gemmell et al. [4, 7] propose an agglomerative approach for tag clustering. For that purpose, they presented a tag relatedness measure based on the idea of term frequency-inverse document frequency (TF.IDF). Correspondingly, resources are considered as documents while the tags are considered as terms. Consequently, each tag is represented as a vector of tag-frequency-inversed resource frequency. Finally, the similarity between two tags is determined by the cosine similarity between the tag vectors. For their tag clustering approach, the authors of [8] 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 et al. [9] propose a tag relatedness approach which uses Jaccard measure to normalize tag co-occurrences. After that, the tags are 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 [5] is based on the notion of $(\mu, \epsilon) - cores$. Thereby, tags are organized in a graph with the edges weighted according to the structural similarity between the nodes. That means, tags that have a large number of common neighbors are considered related.

The presented works exploit tag co-occurrence counts to derive their tag relatedness measures. Additionally, either a simple threshold for tag co-occurrences [3, 9] or the cosine measure are used to identify similar tags [4, 7, 8]. The work in this paper, aims at addressing two important aspects which are less investigated in literature on tag relatedness. First, tags are dealt with as probability distributions and a new distance measure is proposed based on the well-known Jensen-Shannon Divergence. Second, to best of our knowledge, this work is the first to deal with the problem of feature selection for building tag co-occurrence vectors. In this regard, 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 $F = \{T, U, R, A\}$ [10]. T is the set of tags that are contributed by a set of users U to annotate a set of resources R. Two tags $t_1, t_2 \in T$ occur together if they are used by one or more user to describe a resource $r \in R$. This is captured by the assignment relation, $A \in U \times T \times R$. According to this definition a tag can be described in three different kinds of vector space representations: $\mathbb{R}^T$, $\mathbb{R}^U$ and $\mathbb{R}^R$ with respect to each of the three dimensions of the folksonomy T, U and R [11]. In $\mathbb{R}^T$ representation, which is called *tag-context*, a tag is represented as a vector, $v(t) \in \mathbb{R}^T$. The entries of the vector correspond to the co-occurrences of t with the other tags $t' \in T$. In the *user-context* representation each entry of $v(t) \in \mathbb{R}^U$ corresponds to a specific user and represents the number of times in which t was used by that user to annotate some resource. Finally, in the *resource-context* representation an entry

of $v(t) \in \mathbb{R}^R$, corresponding to a specific resource, represents the number of times t was used to annotate that resource.

Approaches for tag relatedness use one (or more) of the presented vector space representations to identify related tags. This is done by generating the corresponding vectors and calculating the cosine similarity between them. The importance of each of the mentioned vector space representations differs according to the nature of the folksonomy (narrow vs. broad) and the goal of the tag relatedness task (e.g. retrieval, recommendation, etc.). In narrow folksonomies, e.g. image folksonomies, where a limited user collaboration is observed, tag relatedness approaches are mainly based on the tag-context representation.

4 Our Approach

We propose a tag relatedness approach based on the tag-context representation. The approach consists of two-steps. 1) For each tag $t \in T$ a probability distribution is built based on the co-occurrence of t with a set of tags $T_f \subseteq T$. We call T_f the feature set. 2) For two tags $t_1, t_2 \in T$ their relatedness is determined according to the distance between their corresponding probability distributions. To compute the feature set T_f, we propose a feature selection approach for tag relatedness based on the Lapalcian score (LS) method [6]. To calculate the distance between two tag probability distributions, we apply the well-known Jensen-Shannon Divergence (JSD) [12] and propose an extension thereof called AJSD. The main characteristic of AJSD is its ability to deal with statistical fluctuations in the generated probability distributions.

4.1 Tag Probability Distribution

In folksonomy F, an empirical probability distribution for a tag $t \in T$ can be created by quantifying the co-occurrences with each of the tags of the feature set $f \in T_f$ by counting the number of times $\#(t, f)$ in which t was used together with f to annotate the same resources. We can use this set of counts to create a histogram in the variable f. Then, by normalizing this histogram with the total number of co-occurrences of t with the elements of the set T_f, we obtain the empirical co-occurrence probability distribution $P_t(f)$ for the tag t with the elements $f \in T_f$: $P_t(f) = \frac{\#(t,f)}{\sum_{f \in T_f} \#(t,f)}$.

In this equation $P_t(f)$ represents the value of the distribution at the histogram channel which corresponds to the feature f. The empirical probability distribution of the tag t over the complete set of features T_f is denoted as $P_t(T_f)$.

4.2 Feature Selection for Tag Similarity

Identifying similar tags in a folksonomy is an all pairs similarity search problem (APSS). Given the set of |T| tags and considering that each tag is represented by a d dimensional vector, the naive approach will compute the similarity between

all tag pairs in $O(|T|^2.|d|)$ time. In the case of tag-context approach where $d = |T|$ the algorithm will have a complexity of $O(|T|^3)$.

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, $\mathbb{R}^{T_f}$ where $T_f \subset T$ and $|T_f| \ll |T|$. A key requirement of the set T_f is that it should have no (or minimal) impact on the quality of the tag relatedness measure. This gives rise to the problem of feature selection for tag relatedness.

Laplacian Score Feature Selection. A simple approach to build the feature set T_f, is to select a subset of the most occurring (frequent) tags in the folksonomy (e.g. [11, 13]). This technique seems to be effective; however, most frequent tags can have uniform co-occurrence patterns with all other tags in the folksonomy. In that case, all tags would be considered related to each other since they will have very similar patterns of distribution over the set of most frequent tags. Therefore, a more sophisticated approach for identifying T_f is required. A possible solution for this problem is provided by the Laplacian score (LS) feature selection method [6]. LS is an unsupervised process for identifying good features for clustering problems. Therefore, it is also suitable for tag relatedness approaches, since the focus is also on finding clusters, i.e., groups of related tags.

The basic idea of LS is to evaluate the features according to their locality preserving power. To achieve that, the data points are organized in a weighted indirect graph in which the nodes correspond to the data points. An edge is drawn between two nodes if they are mutually "close" to each other. Furthermore, the edges are weighted according to the similarity between the connected data points. Now, the importance of a feature can be determined according to which extent it respects the graph structure. Specifically, a feature is considered "good" if and only if for every two data points, which are close based on this feature, there is an edge between these points. This can be formulated as a minimization problem with the following objective function:

$$L_f = \frac{\sum_{ij}(f_i - f_j)^2 S_{ij}}{Var(f)} \tag{1}$$

f_i and f_j correspond to the values of a feature f at the data points i and j respectively, while S_{ij} is the corresponding similarity. $Var(f)$ is the variance of the of feature f. The minimization of the objective function (equation 1) implies preferring features of larger variances. This conforms to the intuition that features with higher variance are expected to have more expressive power.

The feature selection algorithm and estimation for the solution of the objective function are summarized in the following steps (a mathematical justification can be found in [6]):

1) For the set of n data points a 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. That is, x_i belongs to the set of k nearest neighbors of x_j and vice versa.

2) The edges between close nodes are weighted according to a similarity functions. A widely used function is the Gaussian similarity $S_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}}$, where t is free parameter that can be determined experimentally. Pairwise similarities are then combined in a similarity matrix S.
3) For a feature f that is defined as a vector over the data points let:

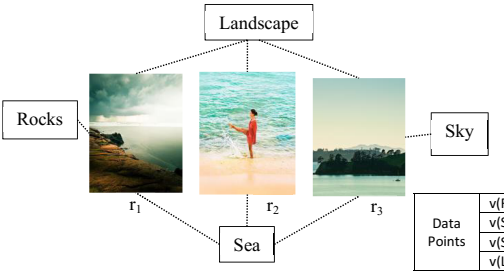
$$\bar{f} = f - \frac{f^T D\mathbb{1}}{\mathbb{1}^T D\mathbb{1}}\mathbb{1} \tag{2}$$

$\mathbb{1} = [1...1]^T$ is the identity matrix. $D = diag(S\mathbb{1})$ is a diagonal matrix, the entries of which d_{ii} correspond to the sum of the entries of the column i in S.
4) Let $L = D - S$ be the Laplacian matrix of the similarity graph [14]. The Laplacian score of the feature f is then calculated as:

$$L_f = \frac{\bar{f}^T L \bar{f}}{\bar{f}^T D \bar{f}} \tag{3}$$

Accordingly, the final feature set will contain features with the top scores.

In our case, the data points as well as the features correspond to the tags of the folksonomy. In other words, we consider each element $t \in T$ as a data point, i.e., a multi-dimensional vector $\boldsymbol{v}(t)$. The components of the vector correspond to the complete set of tags and the values correspond to the co-occurrence counts. On the other hand, the features (corresponding also to the tags) are represented as vectors over the data points.



		Feature Vectors			
		f(Rocks)	f(Sea)	f(Sky)	f(Landscape)
Data Points	v(Rocks)	0	1	0	1
	v(Sea)	1	0	1	3
	v(Sky)	0	1	0	1
	v(Landscape)	1	3	1	0

Fig. 1. Simple folksonomy with the corresponding data points and feature vectors

For better understanding consider the simplified folksonomy (user links omitted) shown in Fig. 1. In that example, three resources (images) $R = \{r_1, r_2, r_3\}$ are annotated with tags form the set $T = \{Rocks, Sea, Sky, Landscape\}$. The table on the right shows the co-occurrence counts of the tags. The data points correspond to the row of the table while each column of the same table corresponds to a single feature vector. In the next step, the generated data points and features vectors can be processed according to the LS method to identify the final feature set.

4.3 Distance of Tag Probability Distributions

To determine if two tags are related, the distance between their corresponding empirical co-occurrence probability distributions must be calculated. In the literature, Jensen-Shannon Divergence (JSD) [12] is a widely used measure which has shown to outperform other measures [15]. It is based on Kullback-Leibler Divergence (D_{KL}); however, it is symmetric and has always a finite value. In our previous work, we proposed an extension of JSD called AJSD [13] which deals with the statistical fluctuations due to the finiteness of the sample.

Before discussing the new metric, first, we explain how to calculate JSD between two tag probability distributions. Given two tags $t_1, t_2 \in T$ and the corresponding empirical co-occurrence probability distributions $P(T_f)$ and $Q(T_f)$ respectively, over the feature set $T_f = \{f_1, ..., f_m\}$ (To avoid mathematical cluttering, we will omit the feature set from the notation). The values of P and Q at a specific feature $f_i \in T_f$ are given by $P(f_i)$ and $Q(f_i)$, respectively. Now, the JSD between P and Q is given by:

$$\begin{aligned} D_{JSD}(P||Q) &= \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M) \\ &= \frac{1}{2}\sum_{f\in T_f}\left(P(f)\log\frac{2P(f)}{P(f)+Q(f)} + Q(f)\log\frac{2Q(f)}{P(f)+Q(f)}\right) \end{aligned} \tag{4}$$

Adapted Jensen-Shannon Divergence (AJSD). If, as in our case, the probabilities P and Q are not available, rather we have an estimate of them through a finite sample represented in the form of a histogram for P and a histogram for Q, then 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 counts k_f and h_f respectively. We define the following measured frequencies $x_f \equiv k_f/n$ and $y_f \equiv h_f/m$, where $n = \sum_f k_f$ 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. As a consequence 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 equations 4 we get:

$$d = \frac{1}{2}\sum_f\left(x_f\log\frac{2x_f}{x_f+y_f} + y_f\log\frac{2y_f}{x_f+y_f}\right) \tag{5}$$

The value of this expression is not in general the maximum likelihood estimate of JSD. That is, due to the unequal variances of the terms in the sum. 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.
a) Thanks to the normality condition stated above, the ML estimate of $P(f)$ correspond to $x_f = k_f/n$ with the variance given in 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 $\sigma^2_{Q(f)} = h_f/m^2$
b) Consider the individual addendum term in the sum expression of equation 5:

$$z_f \equiv x_f \log \frac{2x_f}{x_f + y_f} + y_f \log \frac{2y_f}{x_f + y_f} \tag{6}$$

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

$$\sigma^2(z_f) \simeq \left(\frac{\partial z_f}{\partial x_f}\right)^2 \sigma^2(x_f) + \left(\frac{\partial z_f}{\partial y_f}\right)^2 \sigma^2(y_f) \tag{7}$$

$$\simeq \log^2 \frac{2x_f}{x_f + y_f} \sigma^2(x_f) + \log^2 \frac{2y_f}{x_f + y_f} \sigma^2(y_f) \tag{8}$$

$\sigma^2(z_f)$ can be easily calculated by substituting the quantities of step (a) in equation 8.
c) 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 5 is given by the following weighted sum:

$$\hat{d} = \frac{\sum_f w_f z_f}{\sum_f w_f}; \text{with } \sigma^2(\hat{d}) = \frac{1}{\sum_f w_f} \tag{9}$$

We use $\hat{d}$ as adapted Jensen-Shannon Divergence (AJSD). Note that AJSD, due to the statistical fluctuations in the samples, 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 AJSD provides a ranking for the terms that are correlated with the true ranking in a stronger way than JSD.

5 Experimental Results

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.

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

Qualitative Insight. For each of the 4,776 unique tags in the dataset, we identified its most related tags. Table 1 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 their counterparts which are identified using JSD and the cosine approaches (e.g. Car, Garden in Table 1). That is, because AJSD generates non-zero similarity even two tags have different sample distributions (section 4.3).

Table 1. Sample tags with the corresponding most related tags

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

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. 2 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

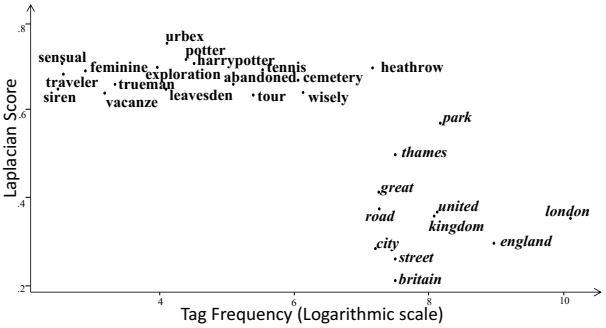


Fig. 2. Tags importance (Laplacian Score) vs. tag frequency

co-occurrence with all other tags in the folksonomy. Therefore, it is less discriminative (has a low LS) than a more specific tag like *potter* which expected to have non-uniform tag co-occurrence distribution.

Evaluation 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 [11, 16, 17]. 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 [18]. 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. Fig. 3 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 frequency based approach (FRQ). 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].

[4] `http://wordnet.princeton.edu/`

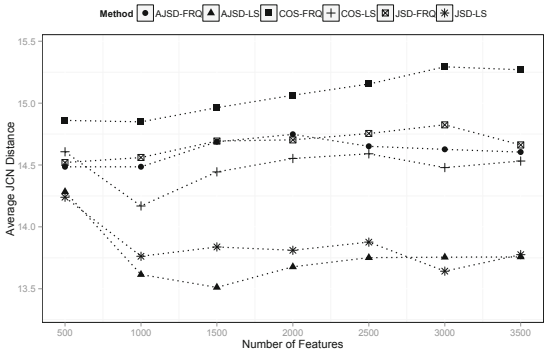


Fig. 3. Average JCN

LS leads to lower average JCN distance than FRQ for all similarity measures and disregarding the number of features (Fig. 3). 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 (Fig. 3).

Since the distributional properties of the investigated measures can be different, we followed the evaluation method described in [16]. In this approach, the performance of two tag relatedness measure can be compared according to the order of the ranks of the set of most similar tag pairs generated by each of them. This is done by calculating the correlation between the rankings of each tag relatedness approach and the corresponding rankings using WordNet. A suitable measure is provided by *Kendall* τ rank correlation coefficient.

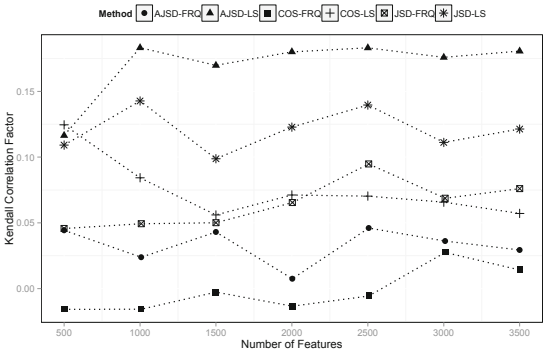


Fig. 4. Kendall Correlation

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 (Fig. 4). 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, a tag relatedness approach based on the Laplacian score feature selection (LS) and an adaptation of Jensen-Shannon Divergence (JSD) was presented. LS allows reducing the dimension of tag co-occurrence vectors without affecting the quality of the applied distance measure. The adapted JSD measure (AJSD) discovers tag pairs of smaller WordNet (JCN) distances and of higher correlation with WordNet than the original JSD measure. Furthermore, both AJSD and JSD performs better than cosine measure. In future work, we will work on improving the performance of our approach by determining the best parameter values for LS. Also, we aim at evaluating the performance of our approach by integrating it into a tag recommendation system.

References

1. Vander Wal, T.: Explaining and showing broad and narrow folksonomies (June 2005), `www.vanderwal.net/random/entrysel.php?blog=1635` (accessed July 30, 2013)
2. Bischoff, K., Firan, C.S., Nejdl, W., Paiu, R.: Can all tags be used for search? In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, pp. 193–202. ACM, New York (2008)
3. Begelman, G., Keller, P., Smadja, F., et al.: Automated tag clustering: Improving search and exploration in the tag space. In: Collaborative Web Tagging Workshop at WWW 2006, Edinburgh, Scotland, pp. 15–33 (2006)
4. Gemmell, J., Shepitsen, A., Mobasher, B., Burke, R.: Personalizing navigation in folksonomies using hierarchical tag clustering. In: Song, I.-Y., Eder, J., Nguyen, T.M. (eds.) DaWaK 2008. LNCS, vol. 5182, pp. 196–205. Springer, Heidelberg (2008)
5. Papadopoulos, S., Kompatsiaris, Y., Vakali, A.: A graph-based clustering scheme for identifying related tags in folksonomies. In: Bach Pedersen, T., Mohania, M.K., Tjoa, A.M. (eds.) DAWAK 2010. LNCS, vol. 6263, pp. 65–76. Springer, Heidelberg (2010)
6. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. Advances in Neural Information Processing Systems 18, 507 (2006)
7. Gemmell, J., Shepitsen, A., Mobasher, B., Burke, R.: Personalization in folksonomies based on tag clustering. Intelligent Techniques for Web Personalization & Recommender Systems 12 (2008)

8. Specia, L., Motta, E.: Integrating folksonomies with the semantic web. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 624–639. Springer, Heidelberg (2007)
9. Simpson, E.: Clustering Tags in Enterprise and Web Folksonomies. HP Labs Technical Reports (2008)
10. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 411–426. Springer, Heidelberg (2006)
11. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic grounding of tag relatedness in social bookmarking systems. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 615–631. Springer, Heidelberg (2008)
12. Manning, C., Schütze, H.: Foundations of statistical natural language processing. MIT press (1999)
13. Mousselly-Sergieh, H., Egyed-Zsigmond, E., Gianini, G., Döller, M., Kosch, H., Pinon, J.M.: Tag Similarity in Folksonomies. In: INFORSID 2013 (May 2013)
14. Chung, F.R.: Spectral Graph Teory, vol. 92. Amer Mathematical Society (1997)
15. Ljubešić, N., Boras, D., Bakarić, N., Njavro, J.: Comparing measures of semantic similarity. In: 30th International Conference on Information Technology Interfaces, Cavtat (2008)
16. Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., Stumme, G.: 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)
17. Srinivas, G., Tandon, N., Varma, V.: 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)
18. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008 (1997)