Distributional Semantics of Tags

Rogier Brussee*

University of Applied Sciences Utrecht
Faculty of Communication and Journalism
Crossmedialab
PO Box 8611, 3503 RP Utrecht, The Netherlands

Christian Wartena

Hochschule Hannover, University of Applied Sciences and Arts
Faculty of Information, Media and Design
Department of Information and Communication
Expo Plaza 12, 30539 Hannover, Germany

Abstract

Tags are a convenient way to label resources on the web. An interesting question is whether one can determine the semantic meaning of tags in the absence of some predefined formal structure like a thesaurus. Many authors have used the usage data for tags to find their emergent semantics. Here, we argue that the semantics of tags can be captured by comparing the contexts in which tags appear. We give an approach to operationalizing this idea by defining what we call paradigmatic similarity: computing co-occurrence distributions of tags with tags in the same context, and comparing tags using information theoretic similarity measures of these distributions, mostly the Jensen-Shannon divergence. In experiments with three different tagged data collections we study its behavior and compare it to other distance measures. For some tasks, like terminology mapping or clustering, the paradigmatic similarity seems to give better results than similarity measures based on the co-occurrence of the documents or other resources that the tags are associated to. We argue that paradigmatic similarity, is superior to other distance measures, if agreement on topics (as opposed to style, register or language etc.), is the most important criterion, and the main differences between the tagged ele-

*Corresponding author
Email addresses: rogier.brussee@hu.nl (Rogier Brussee), christian.wartena@hs-hannover.de (Christian Wartena)

ments in the data set correspond to different topics.

Keywords:
Distributional Similarity, Distributional Semantics, Social Tagging, Folksonomies

1. Introduction

Over the past decade, collaborative tagging has emerged as an important mechanism to annotate content on the World Wide Web. While users tag for different reasons, it is tempting to compare tagging with classification and annotation by professional editors and librarians. Professionals generally use restricted vocabularies and well defined meta-data schemes. Often, they also use terms in restricted vocabularies from a thesaurus or ontology where semantics, relations to other terms and restrictions on their use are explicitly captured. Tags, however, do not have such clearly defined properties. Several studies ([3], [4], [23]) have appeared that try to derive semantic properties from usage data, in particular to reconstruct their semantic similarity. Most of these studies use some form of co-occurrence to determine the similarity of tags, the underlying hypothesis being that tags that frequently co-occur, have a similar meaning. In other words, it is assumed that *syntagmatic* relations of terms are good indicators for semantic similarity.

In natural language, however, words in a syntagmatic relation are not usually synonyms but only have a dependence on each other. Consider, e.g., the synonyms *harbour* and *port*. They seldom occur together since people tend to either use the word harbour or the word port. In fact, in a corpus with 10 million newspaper sentences from the University of Leipzig ([25]) we find that *harbour* is not among the 25 most significant co-occurrences of *port*. Conversely, port is a much more significant co-occurrence for harbour, even though it still only has rank eight. For both words, the word ship is one of the most frequent co-occurring terms[1]. In this case, the semantic similarity of port and harbour is detected by their co-occurrence with the same words rather than simply the co-occurrence in the same documents. This is a phenomenon of second rather than first order co-occurrence, and we will consider it as an approximation to *paradigmatic* similarity, i.e. similarity based on associations between words. Such a second order approach will naturally detect semantic relations that are broader than synonymy because port and harbour will

[1] Data retrieved from `http://corpora.informatik.uni-leipzig.de/` on April 27th 2012

co-occur with words like ship, quay, or navigator that are related to, but not necessarily synonyms of harbour/port. To obtain a more precise, and statistically more stable, measure of similarity we will therefore have to keep track of the precise frequencies of co-occurrence rather than co-occurrence per se.

There are only a few studies that investigate paradigmatic similarity for tags. Beforehand, it is not evident that paradigmatic similarity is even useful for tags, since the nature of tags is rather different from words in natural language. Tags are isolated words without underlying syntactic rules restricting their selection, they are assigned by many users, and the same tag can be assigned many times to the same resource. In fact, tags are often presented as a "word cloud", that emphasizes both their lack of order and the importance of their frequency of use.

In the present paper, we will give an overview of paradigmatic approaches to distributional similarity for tags, and give a simple formalization. We then argue that for tags, just as for words in texts, paradigmatic similarity is a better indication for synonymy than syntagmatic similarity.

The remainder of this paper is organized as follows. In Section 2 we discuss a number of approaches to distribution-based similarity of words and tags. In Section 3 we introduce and formalize our approach to distributional similarity of tags. The resulting similarity measure between tags is studied and compared to alternative similarity measures in Section 4 using three different data sets.

2. Related Work

The relation between language and its meaning, or more generally semiotics, has a long history in philosophy, and linguistics. The computer science oriented literature usually sidesteps this discussion, either by defining semantics in terms of set theoretic interpretation of logics, making it the problem of a domain expert or by defining some machine learning problem. The latter approach usually involves some form of statistics. Here we also follow a statistical approach.

2.1. Distributional Similarity of Words in Texts

A central idea in structural linguistics is that essential properties of words can be derived from their distributional characteristics. De Saussure ([7]) makes the important distinction between *syntagmatic* relations between words defined by co-presence in a linguistic structure (e.g. a text, sentence, phrase, fixed window, words in a certain grammatical relation to the studied word and so on, see e.g. [10]) and more associative relations now usually called *paradigmatic* relations. In particular, words that appear in "similar semantic contexts" are said to be in a

paradigmatic relation. Harris suggests ([11]) that to a great extent, words can be described in terms of the contexts in which they appear, i.e., by their paradigmatic and syntagmatic relations. In linguistics, this statement has become known as the distributional hypothesis.

An unfortunate methodological problem of the distributional hypothesis, is that "semantically similar contexts" are hard to define in a way suitable for algorithmic detection. Harris was aware of this problem and suggested methods for detecting paradigmatic relations in a text through equivalence relations *generated* by intra-textual syntagmatic and grammatical relationships [12]). His approach can be understood as 1. doing (hard) linguistic analysis to track words occurring in syntagmatic contexts and 2. declaring words equivalent if they occur in the same role in an otherwise identical syntagmatic contexts (possibly after using some universal text independent grammatical equivalences such as plural to singular and passive to active sentence). This formally defines an equivalence between words that can be used to repeat step 1 and 2. Harris claims that this relation reflects paradigmatic similarity. In any case, while quite different from our methods, his method is, in spirit, using the second order relation from words to syntagmatic relations and then back to words again.

To find the distribution of words conditional on the usage of a given word, Schütze and Pederson [26] suggest constructing a vector of co-occurrence probabilities from a complete word co-occurrence matrix. To avoid a combinatorial explosion, co-occurrences are counted in a fixed size window. The cosine similarity of the vectors associated to words then provides a similarity measure for the words themselves. Other authors, e.g. Niwa and Nita [24], or Lund and Burgess [20], use different ways to define co-occurrence vectors. Curran [6] evaluates a number of different measures for semantic similarity, including some version similar to our first order similarity. He also gives a discussion of the difficulties evaluating semantic similarity measures. The approach closest to ours for tags, is that of Lindèn and Piitulainen [19] who take all words in a dependency relation to the word under consideration, and compute the probability distribution over all words in this dependency context. As in our approach, they then use the Jensen-Shannon divergence to compare the distributions for different words.

The method of representing the paradigmatic relations of words by a (context dependent) vector is quite similar to the query language models used in pseudo-relevance feedback methods in information retrieval (see e.g. Lafferty and Zhai [17], [32]). In these approaches, finding a set of relevant documents for a query term is a two stage process. First, all documents containing the query term are retrieved. Then the average distribution of words in the documents is computed,

which in their terminology is called "the query language model". Finally, documents are ranked according to the similarity between the document distribution and the query language model computed using the Kulback-Leibler divergence. If the query consists of a single word, the language model is very similar to the co-occurrence distribution of that query term as we will define below in section 3.1. Both the language model and our co-occurrence model give the frequency of terms that occur in the contexts in which the query term was found. As is to be expected, details of the computation are somewhat different.

The above approaches, suffer from the problem that words are represented by vectors in a vector space of very high dimension. Moreover, as more words are tracked, the vectors for each word get sparser. This makes the paradigmatic approach computationally very expensive. A more efficient way to compute paradigmatic similarity is random indexing ([14], [15]). In random indexing, we assign a random vector to each word. The dimension of the random vector space can be chosen freely . The syntagmetic context of each word is then characterized by the weighted sum of all random vectors assigned to the words in this syntagmatic context, i.e. we make a random projection of the vectors in the original vector space model to a lower dimensional subspace. Random indexing thereby allows for computing paradigmatic similarities between words in a much lower dimensional space, with a loss in quality that depends on the dimension of the vector space, but that tends to be small.

2.2. Tag Similarity

To define paradigmatic similarity for social tags we have various possibilities for the "textual context" and the similarity measure to compare them. A systematic overview is given in [23]. Like most other approaches we are aware of, all methods described there use syntagmatic relations to compute similarity.

Approaches using paradigmatic similarity are proposed by [3] and ourselves ([30]). The former authors use the set of tags assigned in one assignment by one user to one resource as the context of a tag. However, it is expected, and observed, that the tags in a single such context are unlikely to be synonyms, even though they can be closely related in other ways. Clements et al. [4] note that the most closely related tags are seldom used by the same user. They exemplify this phenomenon by the tag pair *color/colour* which will be used by British respectively American users, and propose to detect synonymous tags, by looking for tags with similar distributions over documents which simultaneously have a low similarity of distributions over users.

3. Distributional Similarity

In this paper we follow a statistically oriented second order approach with a syntagmatic flavour. We study, the statistics of words in larger corpora that co-occur in the same "syntagmatic context" We thus go from the *first order* correlation between words and syntagmatic contexts, to the *second order* correlation among words that are syntagmatically correlated to the same contexts. In fact this statistical setup is quite general, and applies to any set-up with an association of words to recognizable "contexts", and therefore with a notion of co-occurrence. In particular this setup applies to social tags where people associate tags (i.e. words) to "contexts" like web pages or online pictures. Using different kinds of "context" is also possible, in particular the "context" of being tagged by the same person.

In a statistical approach, it is natural to use a quantitative version of co-occurrence, and weigh contexts depending on the number of times a word occurs. For example, two web pages with the tag "Politics" both give co-occurrence relations with other tags. However, web pages that are given the tag "Politics" often, should arguably be given more weight. In such an approach it is natural to define similarity itself based on the statistical properties of co-occurrences to capture the notion of semantic similarity. Finding paradigmatically related tags is then operationalized to finding words that are close in a comparison measure of the distribution of co-occurring tags.

3.1. Formalization

In a formalization of the statistical approach to paradigmatic similarity we adopt a language suitable for the application to social tagging. Therefore, we speak of "items" rather than syntagmatic contexts, "users" rather than authors and "tags" instead of words. We also say that a tag is associated to an item rather than contained in a syntagmatic context.

Given a corpus there is a collection of items (web resources, syntagmatic contexts) $\mathscr{I} = \{i_1, \ldots, i_k\}$, a collection of tags $\mathscr{T} = \{t_1, \ldots t_l\}$ and a collection of users $\mathscr{U} = \{u_1, \ldots u_m\}$. Given an item $i \in \mathscr{I}$, let $p(t|i)$ be the probability distribution on $\mathscr{T}$, that gives the likelihood of the occurrence of tag t when randomly selecting a tag associated to the item i. More generally, given a subset of items $\mathscr{J} \subset \mathscr{I}$ there is a probability distribution $p(t|\mathscr{J})$, which is the probability that a random tag selected from an item in $\mathscr{J}$ is t. In particular, for a tag w we can consider the subset of items $\mathscr{I}_w \subseteq \mathscr{I}$ that are (also) associated to w. By the above we then have a probability distribution $p(t|\mathscr{I}_w)$ which we abbreviate to $p(t|w)$, suppressing $\mathscr{I}$ in the notation.

As an example consider the set of items $\mathscr{I}$, the set of English sentences in some text corpus. Then $p(t|w)$ is the likelihood for a word (a.k.a. tag) t to occur in an English sentence in that corpus that contains w. Clearly we can construct similar probability distributions with paragraphs, or documents in the corpus. Note that this changes both the interpretation and the values of $p(t|w)$.

To compute the co-occurrence distribution, let $n(i,t,u)$ be the number of times user u associates tag t to item i. Furthermore let $n(i,t) = \sum_u n(i,t,u)$ be the number of times tag t was associated to item i, $n(u,t) = \sum_i n(i,t,u)$ the number of times tag t was associated by user u, $n(t) = \sum_i n(i,t)$ the number of associations ("occurrences") of tag t, $n(i) = \sum_t n(i,t)$ the number of tag associations to item i, and $n(u) = \sum_t n(u,t)$ the number of tag association by user u. Finally, let $n = \sum_t n(t) = \sum_i n(i) = \sum_u n(u)$ be the total number of tag associations. We then define

$$\begin{aligned} p(t|i) &= n(i,t)/n(i) \text{ the tag distribution (i.e. on } \mathscr{T}) \text{ of item } i \text{ ,} \\ P(i|w) &= n(i,w)/n(w) \text{ the item distribution (i.e. on } \mathscr{I}) \text{ of tag } w \text{ ,} \\ p(t) &= n(t)/n \text{ the background tag distribution (i.e. on } \mathscr{T}) \text{ .} \end{aligned}$$

Similar distributions can be defined for the combination tags and users, or users and items. The probability distributions $p(t|i)$ on the set of tags $\mathscr{T}$, and the distribution $P(i|z)$ on the corpus of items $\mathscr{I}$ describe how tag associations to a given item i are distributed over different tags, respectively how the associations of a given tag w are distributed over different items.

As in [17], we use Markov chains to compute the *co-occurrence distribution* as the second order statistical relation between tags. Consider a Markov chain on $\mathscr{T} \cup \mathscr{I}$ having transitions $\mathscr{I} \to \mathscr{T}$ with transition probabilities $p(t|i)$ and transitions $\mathscr{T} \to \mathscr{I}$ with transition probabilities $P(i|t)$. The chain allows us to propagate probability distributions from tags to items and vice versa. Given a distribution $\pi(t)$ on the set of tags $\mathscr{T}$, the one step Markov chain evolution gives us an item distribution $\pi^{(1)}(i)$, the probability to find a tag occurrence on a item i given that the tag distribution of the occurrences is π:

$$\pi^{(1)}(i) = \sum_t P(i|t)\pi(t).$$

The item distribution $\pi^{(1)}$ gives the *first order* statistical correlations between tags and items coming from the tag associations. Likewise, given an item distribution $\Pi(i)$, the one step Markov chain evolution gives us a tag distribution

$$\Pi^{(1)}(t) = \sum_i p(t|i)\Pi(i).$$

Since $\Pi(i)$ is the likelihood to find a tag on item i, $\Pi^{(1)}$ is the Π-weighted average of the tag distributions of the items. It is also defined by a *first order* statistical correlations between tags and items. Combining these, i.e. running the Markov chain twice, every tag distribution $\pi(t)$ gives rise to a new tag distribution

$$\pi^{(2)}(t) = \sum_i p(t|i)\pi^{(1)}(i) = \sum_{t',i} p(t|i)P(i|t')\pi(t')$$

In particular starting from the degenerate "known to be w" tag distribution $\delta_w(t) = 1$ if $t = w$, and 0 otherwise, we have the first order co-occurrence distribution over items i in $\mathscr{I}$

$$\delta_w^{(1)}(i) = \sum_{t'} P(i|t')\delta_w(t') = P(i|w)$$

which we recognize as the item distribution of the tag. The more interesting second order *co-occurrence distribution* (also called *t*he co-occurrence distribution) $p(t|w)$ on the set of tags $\mathscr{T}$

$$p(t|w) = \delta_w^{(2)}(t) = \sum_{i,t'} p(t|i)P(i|t')\delta_w(t') = \sum_i p(t|i)P(i|w). \tag{1}$$

It is thus the average of the tag distributions $p(t|i)$ of each item i weighted with the the probability $P(i|w)$ that tag w was assigned to item i.

The similarity between tags can now be expressed in different ways using the associated probability measures by choosing different similarity measures for probability measures. It is natural to use information theoretic divergences as they have a natural interpretation as the amount of information gained by using particularities of a distribution rather than an estimate like the background distribution. In particular, we will use the Jensen-Shannon divergence which has the nice properties of being symmetric, and finite unlike the Kullback-Leibler divergence D (see e.g. [5, Section 2.3]) Recall that the Jensen-Shannon divergence of two probability distributions p and q can be computed in either of the equivalent ways below

$$\begin{aligned} \mathrm{JSD}(p,q) &= \tfrac{1}{2}D(p||\tfrac{1}{2}p+\tfrac{1}{2}q) + \tfrac{1}{2}D(q||\tfrac{1}{2}p+\tfrac{1}{2}q) && (2)\\ &= \log 2 + \frac{1}{2}\sum_{t:p(t)\neq 0 \wedge q(t)\neq 0} p(t)\log\left(\frac{p(t)}{p(t)+q(t)}\right) + q(t)\log\left(\frac{q(t)}{p(t)+q(t)}\right) && (3)\end{aligned}$$

The Kullback-Leibler divergence is non negative, so combining the two expressions (2) and (3), we see that $0 \leq \mathrm{JSD}(p,q) \leq \log(2)$. The latter expression (3) is

computationally efficient for sparse distributions. The Jensen-Shannon divergence can be interpreted as the number of bits per symbol that is saved by optimally compressing two equally long streams of symbols with probability distribution p respectively q separately, rather than compressing them both with the optimal compression for the mixed stream. In the following we will refer to (dis)similarity of tags based on the divergence of co-occurrence distribution as second order co-occurrence similarity. The Jensen-Shannon divergence is not a proper metric itself, but its square root does have this property [8]. Thus, if we want to go beyond ranking and real distances are needed, e.g. for clustering, we will use the square root of the Jensen-Shannon divergence.

We can then define zeroth, first and second order co-occurrence Jensen-Shannon similarity of tags as

$$\mathrm{jsd\,sim}_0(w_1,w_2) = \mathrm{JSD}(\delta_{w_1},\delta_{w_2}) = 0 \text{ if } w_1 = w_2; \text{and } \log(2) \text{ otherwise} \tag{4}$$

$$\mathrm{jsd\,sim}_1(w_1,w_2) = \mathrm{JSD}(\delta^{(1)}_{w_1},\delta^{(1)}_{w_2}) = \mathrm{JSD}(P(-|w_1),P(-|w_2)) \tag{5}$$

$$\mathrm{jsd\,sim}_2(w_1,w_2) = \mathrm{JSD}(\delta^{(2)}_{w_1},\delta^{(2)}_{w_2}) = \mathrm{JSD}(p(*|w_1,p(*|w_2)). \tag{6}$$

Clearly $\mathrm{jsd\,sim}_0$ is just a convoluted way to write (in)equality, and is only shown for expository purposes. Note that for $\mathrm{jsd\,sim}_0$ and $\mathrm{jsd\,sim}_2$ the sum implied in the definition of the Jensen-Shannon divergence JSD runs over tags t, whereas in $\mathrm{jsd\,sim}_1$ it runs over items i. We can likewise use the cosine similarity or a χ^2 test to get similarity measures $\cos\mathrm{sim}_1$, $\cos\mathrm{sim}_2$ respectively $\chi^2\,\mathrm{sim}_1$ or $\chi^2\,\mathrm{sim}_2$.

Both $\mathrm{jsd\,sim}_1$ and $\mathrm{jsd\,sim}_2$ are *broad* second order measures because for two tags w_1 and w_2 we compare the full probability distributions $p(t|w_1)$ and $p(t|w_2)$ where t runs over all of $\mathscr{T}$. Alternatively, however, we can use the the co-occurrence probability $p(w_1|w_2)$ or a symmetrized version like

$$(p(w_1|w_2)p(w_2|w_1))^{1/2} = p(w_1 \text{ and } w_2)p(w1)^{-1/2}p(w_2)^{-1/2} \tag{7}$$

as a similarity measure. We will call this a *narrow* measure because it only uses a very small part of the co-occurrence distribution.

For reference purposes, we will note one more popular first order broad co-occurrence similarity, the Jaccard coefficient, that can be defined using our previous definitions as

$$\mathrm{jaccsim}_1(w_1,w_2) = \frac{|\{t_i \mid \delta^{(1)}_{w_1}(t_i) > 0 \text{ and } \delta^{(1)}_{w_2}(t_i) > 0\}|}{|\{t_i \mid \delta^{(1)}_{w_1}(t_i) > 0 \text{ or } \delta^{(1)}_{w_2}(t_i) > 0\}|} \tag{8}$$

3.2. Example

Consider the following examples taken from a sample of LibraryThing data (see subsection 4.1) to get a better feeling for the differences between first and second order co-occurrence.

Table 1 shows the 20 most similar terms for the tag Socrates according to 3 different similarity measures. Besides the broad second order similarity $\mathrm{jsd\,sim}_2$ and the narrow co-occurrence probability (1), we show results for first order co-occurrence similarity using the more common $\cos\mathrm{sim}_1$ rather than $\mathrm{jsd\,sim}_1$. The tag *Socrates* was used 49 times in our data set to tag 25 different books. The narrow co-occurrence probability favors very frequent terms (for this corpus) like *philosophy* or *non-fiction*. The first order cosine similarity and second order Jensen-Shannon similarity computed using document co-occurrence give more or less the same results for the 5 most similar tags. After that the second order similarity seems to contain more terms related to ancient Greece in general, without special relation to Socrates.

A similar picture arises if we compare tags close to *Aristotle*, given in Table 2. Here we see in both cases terms related to Aristotle. The cosine distance results in a high rank for *ethics* and *virtue* , concepts that Aristotle has written two influential works about. In contrast, the second order co-occurrence similarity results in a high rank for *Socrates* and *Plato*, two philosophers living at the same time as Artistotle. This is exactly the behavior we hoped for: both measures come up with strongly related terms, but the second order co-occurrence similarity favors similar and synonymous terms.

4. Experiments

To test whether the second order co-occurrence similarity really corresponds to semantic similarity, we have to make a comparison to some independent measure of semantic similarity, preferably based on human judgment. E.g.[3] use Wordnet as a ground truth. This implicitly assumes that the information in Wordnet (or another available lexicon with synonymy information) is correct and complete. Moreover it assumes that the semantics of words in text can be transferred directly to words used as tags. Since the coverage of tag vocabularies by dictionaries is low as tags are often abbreviations or small phrases consisting of more than one word, this is doubtful. We will therefore evaluate the proposed similarity measure indirectly by studying its behavior in a number of different scenarios rather than using a external reference source.

Table 1: *LibraryThing tags most similar to the tag* Socrates *according to second order co-occurrence similarity, (first order) co-occurrence probability and cosine similarity of the document distribution.*

$\mathrm{jsd\,sim}_2$		$p(*\|\text{Socrates})$		$\mathrm{cos\,sim}_1$	
socrates	0,00	philosophy	0,21	**socrates**	1,00
plato	0,12	**socrates**	0,04	plato	0,62
ancient philosophy	0,14	non-fiction	0,04	greek philosophy	0,60
greek philosophy	0,14	classics	0,04	ancient philosophy	0,51
classical philosophy	0,17	greek	0,03	classical philosophy	0,47
platonism	0,19	fiction	0,03	oligarchy	0,43
aristotle	0,19	greece	0,03	republic	0,30
dialogues	0,21	historical fiction	0,03	peloponnesian war	0,29
western philosophy	0,22	history	0,03	political philosophy	0,26
philosophy	0,24	ancient greece	0,02	platonism	0,25
ancient greece	0,25	plato	0,02	classes	0,23
history of philosophy	0,26	read	0,02	political thought	0,21
stoicism	0,26	ancient philosophy	0,02	philosophy	0,20
ancient	0,27	politics	0,01	ancient	0,20
political thought	0,27	ancient history	0,01	ancient greece	0,20
greece	0,27	christianity	0,01	dialogues	0,20
classical greece	0,27	ancient	0,01	ancient civilization	0,19
greeks	0,27	classic	0,01	greeks	0,19
greek	0,27	literature	0,01	dialogue	0,18
political philosophy	0,27	existentialism	0,01	pdf	0,18

4.1. Datasets

We use tags from three data sets: a set of tags crawled from *Delicious*, a publicly available data set crawled from LibraryThing and a publicly available dump from Bibsonomy.

LibraryThing[2] is an interactive web service that allows users to maintain personal lists of books, write reviews, rate books and add tags. The librarything data set we use was collected by Maarten Clements (Technical University Delft) and can be downloaded from the websites of the Delft Multimedia Information Retrieval Lab[3]. After initial retrieval, the data set was pruned so as to create a collection of annotations by users that have supplied both ratings and tags to at least 20 books and with books that were annotated by at least 5 people. The pruned data set was made available. Details on the data set are given in Table 3. For our purposes, pruning the data set does not pose a real problem because no co-occurrence similarity measure can be expected to give sensible results with very infrequent tags, and items with only a few tags provide hardly any information on tag co-occurrence.

Delicious[4] is a bookmarking service for URLs. A sample of the data containing, among other things, tag, userid and URL was crawled by Matthias Lux from Klagenfurth University [21]. In the experiment in subsection 4.2, we only used a small subset of this dataset with tags assigned to pages from the English Wikipedia. The number of tagged URLs, tags and users in this subset are given in Table 3.

Bibsonomy[5] is a system allowing users to store and tag references to (scientific) papers [2]. In the tag clustering experiment described in 4.3 below, we used a dump from July 2010 [16]. Characteristics of this data set are also given in Table 3.

In a number of experiments we only consider tags used at least 5 times, since almost no interesting and useful distributional statistics can be derived for tags with less usage. In Table 3 we additionally give the number of tags, users and items occurring at least 5 times in a user-item-tag relation. Note, however, that this restriction is not the same one as the pruning condition used to construct the LibraryThing dataset. While the smaller data set contains only triples of user, item

[2] `http://www.librarything.org`

[3] http://homepage.tudelft.nl/5q88p/LT/

[4] `http://www.delicious.com`

[5] `http://www.bibsonomy.org`

Table 2: *Related tags according to second order co-occurrence similarity and cosine similarity of tag distributions. Bold terms are the tags for which similar terms were computed. Tags are ordered according to their similarity to this term.*

JSD	Cosine
aristotle	**aristotle**
ancient philosophy	ancient philosophy
greek philosophy	ethics
plato	virtue
classical philosophy	greek philosophy
western philosophy	loeb
socrates	political thought
history of philosophy	western philosophy
political thought	antiquity
platonism	generosity
philosophy	desk

Table 3: Characteristics of 3 datasets with tags

	LibraryThing sample		Delicious sample		Bibsonomy	
Users	7,279	(7,279)	50.097	(18,214)	6,659	(4555)
(users that tagged at least 5 times)						
Items	37,232	(37,232)	53,345	(20,670)	468,265	(72,330)[6]
(items with at least 5 tags)						
Unique tags	10,559	(10,559)	49,603	(7624)	217,948	(46,028)
(tags used at least 5 times)						
Tag assignments	2,056,487		278,693		2,622,423	

and tag, where each tag occurs at least 5 times, we still consider assignments from users or resources that occur less than 5 times in the data set for computation of the co-occurrence distribution. The effect of the different way of sampling can clearly be observed in Figures 1, 2 and 3 where we clearly see that distribution of users, tags and items follow a typical power law in the Bibsonomy and Delicious data sets. The LibraryThing data sets has a different behavior for the low frequency occurrences. The figures should be read in a such away that the upper left most data point in Fig. 1 means that there are 11966 users in the Delicious data sample that have associated exactly 1 tag. The next point right of this point indicates that there are 8858 users that have made 2 tag associations, and so on.

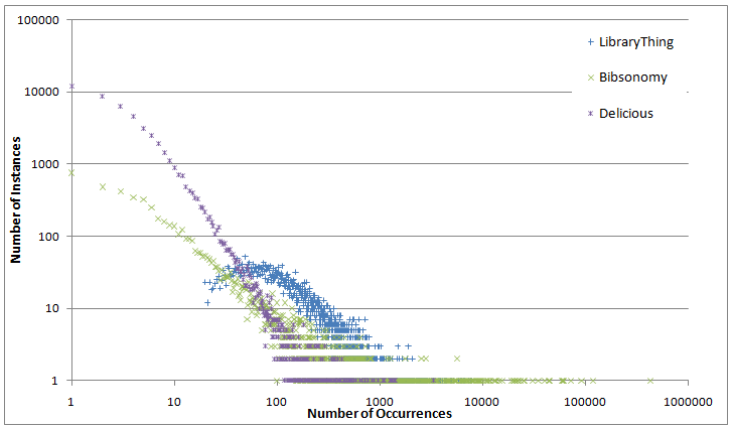

Figure 1: Number of users (y-axis) that have assigned a given number of tags (x-axis) for three datasets.

4.2. Experiment 1: Ontology Mapping

The first scenario in which we use second order co-occurrence similarity is in finding a mapping between two vocabularies. The experiment models mapping category assignments from a formal classification system to free-style tags and vice versa. Unfortunately, we do not have at our disposal a sufficiently large data sets that are both formally classified and have enough user generated tags. We therefore use Wikipedia articles which, while not formally classified, are organized in a system of categories that is kept in check by the Wikipedia maintainers. For the tagging data we use the tags from the Delicious data set assigned to these Wikipedia articles. The experiment is described in more detail in [30].

For the computation of the tag co-occurrence distributions we consider both user generated tags and category labels as one collection of formal "tags" in the

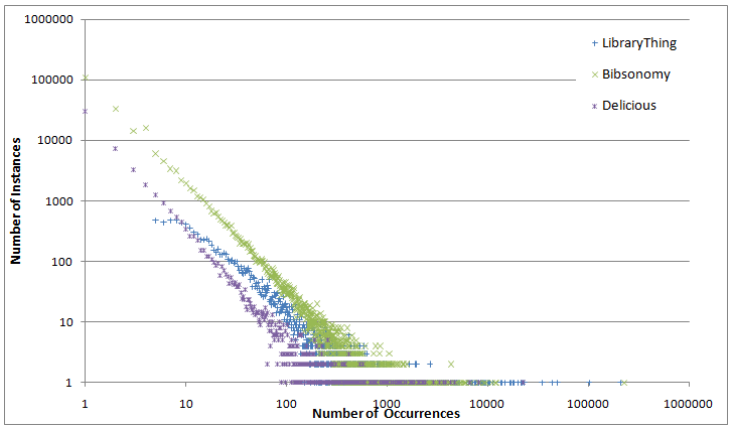

Figure 2: Number of tags (y-axis) that have be assigned a given number of times (x-axis) for three datasets.

Table 4: Characteristics of the Wikipedia category mapping experiment

Articles	53,345
Wikipedia Categories	42,445
Category assignments	222,640
Unique Delicious tags	49,603
Delicious tag assignments	278,693

sense of section 3.1. However, we do not unify tags and category labels, even if they are represented by the same strings. Thus, we leave open the possibility that the same word has a different meaning when used as a tag in Delicious and as a category label in Wikipedia. The mapping is then computed by finding the closest tag for each category label and vice versa. To determine closeness we use the second order $\mathrm{jsd}\,\mathrm{sim}_2$ measure and $\mathrm{jaccsim}_1$, the first order Jaccard coefficient . The latter was reported to be superior to other co-occurrence similarities in similar experiments [13]. Other measures were not taken into account, because the computation of the mapping is computationally expensive and more importantly, the manual evaluation of the resulting mapping is very time consuming. Table 4 summarizes the number of users, tags, and Wikipedia articles (items) for this experiment.

We only computed mappings for all categories and tags that occur on at least 10 Wikipedia articles, since we cannot expect to get reasonable results for very infrequent tags. For the targets we also restricted the set of candidates to that same subset. Thus, we found 2355 mappings from tags onto a Wikipedia category

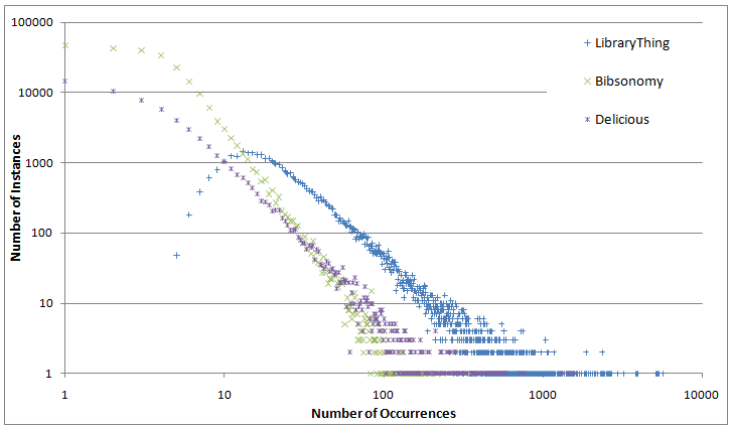

Figure 3: Number of resources (y-axis) that have been assigned a given number of tags (x-axis) for three datasets.

and 1827 mappings from a category onto a tag for both dissimilarity measures.

For the evaluation, we manually classified 10% of the mappings. We selected the mapping with the 100 most and least frequent source terms with the remaining mappings chosen at random. Initially, we classified each of the mappings into one of the following 8 categories: identical term, synonym, broader term, narrower term, related term, unrelated term, unclassifiable term, unknown term. The classes were subsequently grouped together into 4 classes: synonym, related term, unrelated term, unclassified. Here unclassified means that either the meaning of the tag was unclear, or the source label was an organizational tag or category (like *to-read* or *important*) for which no corresponding wiki category label exists. The numbers of examples for each class are given in Figure 4. Mappings in both di-

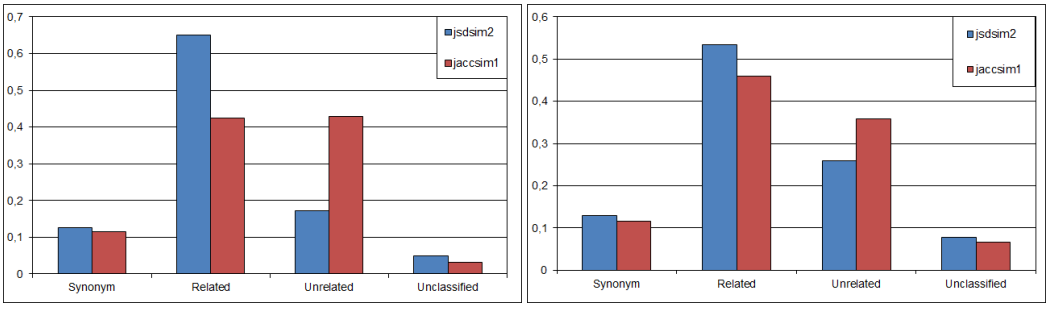

Figure 4: Fraction of mappings from Wikipedia categories onto tags (left) and vice versa (right) using different dissimilarity measures for each evaluation category

rections show similar percentages of synonyms and unclassified mappings for the second order co-occurrence similarity $\mathrm{jsd\,sim}_2$ and the first order Jaccard coefficient. However, for both directions, the second order co-occurrence similarity gives many more related terms than the Jaccard coefficient.

A more detailed analysis shows in particular, that the second order co-occurrence gives better results for infrequent tags and categories. This becomes especially clear for the mapping from (Delicious) tags to (Wikipedia) categories, where the Jaccard coefficient does not perform too badly. Figure 5 gives the evaluation for the mapping of the 100 most frequent tags (413 to 22 267 occurrences; if we would exclude the three most frequent tags, *Wikipedia*, *Wiki* and *Reference*, the range goes up to 4630 for *history*) and the 100 least frequent tags (13 to 15 occurrences). For the frequent tags both similarity measures perform almost equally well. For the infrequent tags the second order co-occurrence clearly outperforms the Jaccard coefficient.

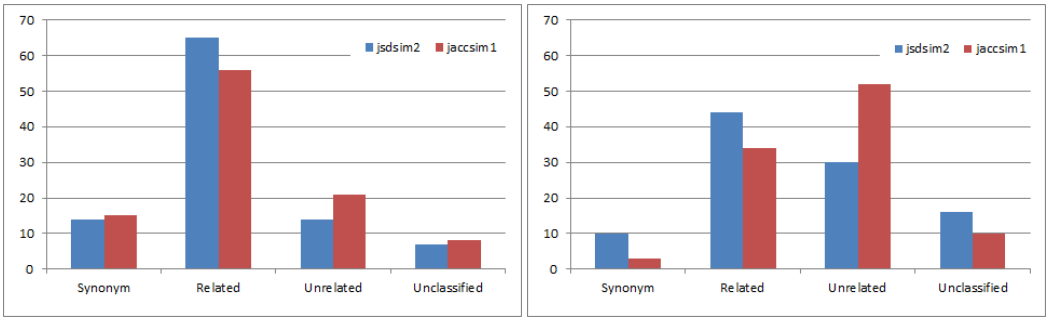

Figure 5: Fraction of mappings from tags to Wikipedia categories for 100 most frequent tags (left) and 100 least frequent tags (right) using different dissimilarity measures for each evaluation category

4.3. Experiment 2: Tag Clustering

Clustering is an interesting way to create more structure in a large set of tags. It can help users getting a better insight in the tagged collection [28] and was also shown to improve the accuracy of automatic recommendation [31]. The topical coherence of the clustering is strongly dependent on the distance measure used by the clustering algorithm.Thus, the effectiveness of clustering can be used to evaluate distance measures for semantic relevance. Here we compare the second order co-occurrence similarity $\mathrm{jsd\,sim}_2$ with the first order cosine similarity $\cos\mathrm{sim}_1$ and Jaccard coefficient $\mathrm{jaccsim}_1$ in a clustering experiment with tags from Bibsonomy.

Before computing distances between tags, we first normalized the tags by unifying obvious spelling variants. For each tag t, we collected all less frequent tags with an edit distance less than $1+0.5\text{length}(t)$ as potentially equivalent. We then assigned penalties of 0.5 and 0.1 for insertion or deletion of a final 's' and non-alphanumeric characters, respectively. After this procedure, tags t on this list that either have an edit distance of less than 0.5 (i.e. differ only in non alphanumeric characters) or that have $\text{jsd}\,\text{sim}_2(t,t') < 0.45\log(2)$ were considered spelling variants of t. The combination of the edit distance condition with the JSD condition ensured that only those tags are identified that have both a similar spelling and are used in similar contexts. This was setup to identify spelling variants with the same meaning only. Before normalization 28,129 tags were occurring more than twice with the papers selected below. After normalization only 23,252 tags were left.

To evaluate clustering of tags, we manually created 12 reference clusters of tags. We started by choosing 12 scientific disciplines for which a substantial number of tags is available in Bibsonomy. For each of these disciplines, we selected two or three important and typical journals (Table A.6 in Appendix Appendix A). and collected the most common spelling variants and abbreviations as used in Bibsonomy. We then collected all tags assigned to papers published in these journals, and filtered out all tags that are either too frequent in the whole collection (e.g. "imported") or that matched regular expressions describing highly personal tags (e.g. "my-.*"). Finally, for each discipline we select tags that occur at least 5 times for a selected journal within that discipline and for which at least 80% of the occurrences was on an article within its discipline. This ensured that we have at least one unambiguously way to assign a tag a reference cluster although other other reasonable classifications might of course exist. This resulted in a set of 215 distinct (normalized) tags (listed in Appendix A) with a total of 6,449 occurrences in the selected, and 38,258 occurrences in the whole data set.

We clustered the 215 tags using the first order $\arccos(\cos\text{sim}_1)$ distance, the first order Jaccard coefficient jaccsim_1, and the second order $\sqrt{\text{jsd}\,\text{sim}_2}$ distance between tags (see section 3.1). For each distance measure we use the same set of (journal, tag) pairs, and the same k-means clustering algorithm with random cluster initialization described in [1]. Results were averaged over 20 runs to estimate the effect of the random choice. The co-occurrence distributions were computed over all 23,252 normalized forms of tags and used all 38,258 occurrences in the whole data set. Since the clustering results for the two first order measures turned out to be indistinguishable well within the standard deviation from the variation resulting from random initialization, in the following we only report on the cosine

similarity.

Following [18] we evaluated the computed clustering $C = \{c_1, \cdots, c_n\}$ against the reference clustering $C^* = \{c_1^*, \cdots, c_{12}^*\}$ corresponding to the 12 scientific disciplines. For each cluster $c \in C$ and reference cluster $c^* \in C^*$ we define a recall measure $\mathrm{rec}(c, c^*) = |c \cap c^*|/|c^*|$, a precision measure $\mathrm{prec}(c, c^*) = |c \cap c^*|/|c|$ and an F value

$$F(c, c^*) = \frac{\mathrm{rec}(c, c^*)\mathrm{prec}(c, c^*)}{\frac{1}{2}(\mathrm{rec}(c, c^*) + \mathrm{prec}(c, c^*))}.$$

Let $F_{\max}(c^*) = \max_{c \in C} F(c, c^*)$ be the F-value of the best fitting found cluster. Finally define $\|C^*\| = \sum_{c^*} |c^*|$ and the mean maximal F-value by

$$F = \frac{1}{\|C^*\|} \sum_{c^* \in C^*} |c^*| F_{\max}(c^*)$$

A value of $F = 1$ means that the set of selected tags was clustered exactly according to the reference clustering by subject. The overall F-values for clustering with the different similarities are given in Figure 6. We also determined the purity of the clusters (figure 7). Purity is the average highest precision of the clustering and is defined as follows ([22, chap 16]). As above, let $\|C\| = \sum_{c \in C} |c|$, and $\mathrm{prec}_{\max}(c) = \max_{c^* \in C^*} \mathrm{prec}(c, c^*)$

$$\begin{aligned} \mathrm{Purity}(C) &= \frac{1}{\|C\|} \sum_{c \in C} |c| \mathrm{prec}_{\max}(c) \\ &= \frac{1}{\|C\|} \sum_{c \in C} \max_{c^* \in C^*} |c \cap c^*|. \end{aligned}$$

Note that the purity measure does not "punish" a breakup of clusters and becomes 1 for clusters of size 1.

4.4. Relative Entropy of Co-occurrence Distributions

Tags may express many things. The most common intention seems to be classifying the topic of the item. However, a tag may also express something about the form of an item, the medium, the relation between the tagger and the tagged item, the context in which the tagger has found the item, and so on. For example the tagger may express where he found an item ("You Tube", "Bull. ACM") state an opinion ("nice", "OMG cuuuute"), or the purpose of the tag ("PhD", "Holiday2010"). It is even possible that a tag is ambiguous between two types of usage. Assuming that most tags are related to a topic, the co-occurrence distribution of

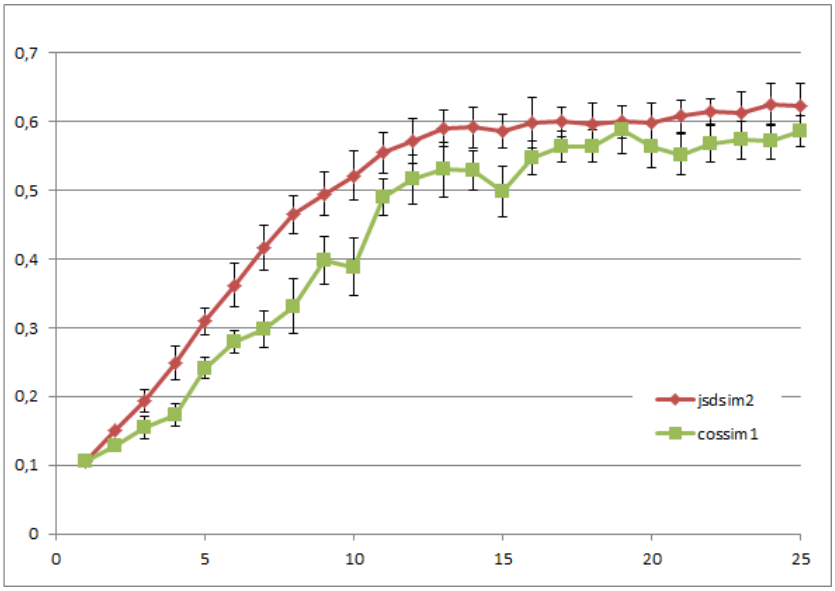

Figure 6: Overall F-Measure for clustering of 214 Bibsonomy tags matched against 12 reference clusters. Error bars give ± 1 standard deviation for 20 runs with random initial points. The red line gives results using $\mathrm{jsd\,sim}_2$, the green line the results using $\cos\mathrm{sim}_1$.

a topic-related tag should roughly correspond to a corpus dependent distribution of topics, weighted by a corpus independent distribution of tags given a topic. Hence, we expect the co-occurrence distribution $p(-|t) = \delta_t^{(2)}$ of a topic related tag to have a relatively high divergence $D(p(-|t)||p_0)$ with the mean tag distribution p_0. On the other hand a topic independent tag should have a low divergence with the mean distribution because they can occur with most items. Indeed, this exactly the behavior we observe. As an example, consider the tags *borrowed from library* and *Kansas City* we found in the LibraryThing dataset. The tags have similar frequencies (125 and 118 occurrences resp.) and occur on a similar number of items (119 and 118 items resp.). The first tag (*borrowed from library*) is clearly not about a topic while the second tag is. This is reflected by the relative entropy of the co-occurrence distribution, which is 0.740 for *borrowed from library* and 2.39 for *Kansas City*.

In the construction of the co-occurrence distribution in section 3.1 we can use users just as well as items, so we can compute a co-occurrence distribution for users instead of items. We then expect to find that personal tags tend to have higher divergence because by the same reasoning, personal tags should correspond to people or groups of like minded people.

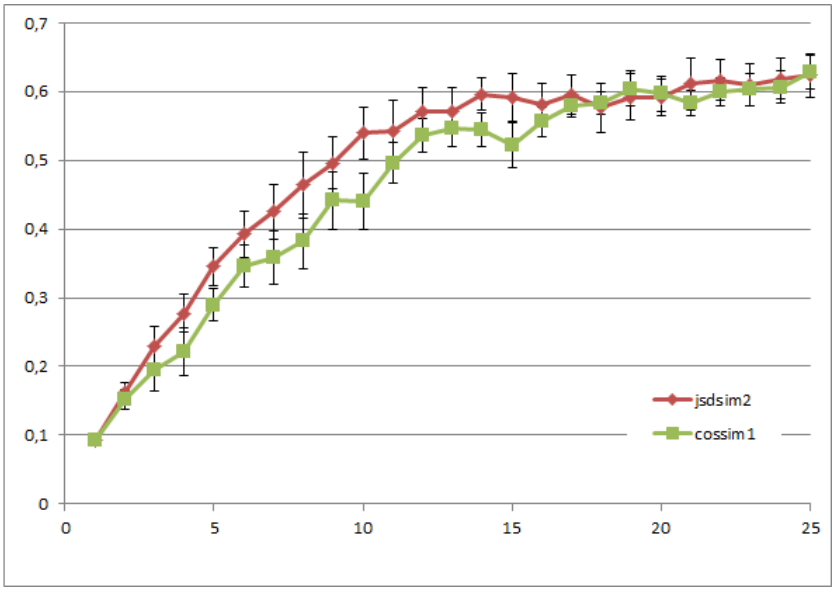

Figure 7: Purity for the clustering of 214 Bibsonomy tags matched against 12 reference clusters. Error bars give ± 1 standard deviation for 20 runs with random initial points. The red line gives results using $\mathrm{jsd\,sim}_2$, the green line the results using $\cos\mathrm{sim}_1$.

For a larger scale experiment, we first categorized tags in 5 different classes motivated by [29]. The class "Attributes" includes genre and some additional properties at the level of the work and at the level of expression, including the usage context. Self-referential tags are typically organizational tags, or tags related to the copy e.g. its physical appearance. Author tags refer to the author of the tagged book. We then normalized the tags in a way similar to the normalization described in Section 4.3 and manually labeled 675 tags with a class label. A random selection of about 500 tags was labeled using 5 different class labels. We stopped labeling after a certain number of examples for each class was found. We then searched for more examples of opinion and self referential tags to get balanced classes. Moreover, additional examples of attributes were selected to include enough examples of different types of attributes, like usage context. Finally, we manually assigned every tag a class label and determined the average relative entropy of the co-occurrence distribution of tags in each class. The result with co-occurrence distributions computed over both items and users is given in Table 5.

Here we clearly see that for the item based co-occurrence we find a higher relative entropy for the classes that are strongly related to topics (Topic and Author)

Table 5: Averaged relative entropy (standard deviation) of co-occurrence distributions of tags from given classes using Kullback-Leibler divergence with natural logarithms

Class name	Nr. of examples	Rel. Entropy of item based co-occurrence distr.	Rel. Entropy of user based co-occurrence distr.
Topic	200	2.90 (1.16)	1,53 (0,66)
Author	100	2.10 (0.674)	1,79 (0,59)
Attribute	200	1.67 (0.853)	1,58 (0,83)
Opinion	65	1.13 (0.502)	1,90 (0,87)
Self reference	110	0.928 (0.554)	1,91 (0,78)

and a low relative entropy for classes of tags that are independent from a topic (Opinion and Self reference) (difference between values for author and opinion are significant at the level of $\alpha = 0.1\%$ using the student's t-test), whereas for the user based co-occurrence distribution the relative entropy of the personal Opinion and Self reference tags are clearly higher (difference between author and opinion tags significant at the the level of $\alpha = 5\%$). In [29] we have shown that two similar eccentricity measures are a very useful feature for automatic classification of tags.

5. Discussion and Conclusion

We have shown how second order co-occurrence distributions of tags can be used to capture some aspects of the semantics of tags, in particular for an operationalization of paradigmatic tag similarity measure. We put forward the hypothesis that this second order operationalization of paradigmatic similarity corresponds more closely to the real semantic similarity of tags as interpreted by people, than first order syntagmatic similarity measures. We found evidence for this hypothesis in three different experiments: an ontology alignment task and a clustering experiment and an experiment to recognize the topicality of tags.

For the mapping between Wikipedia category labels and Delicious tags the second order co-occurrence similarity seems to be superior to first order co-occurrence measures. while we did not find more synonyms, we found significantly more related and less unrelated terms than with Jaccard coefficient.

For the clustering of tags on scientific articles from Bibsonomy we also see a clear improvement with the second order measure: in all cases, second order similarity gave both better F-measure and better purity than the syntagmatic similarity.

Finally, an investigation of tags on books from LibraryThing data shows that co-occurrence distributions give information on other semantic aspects of tags, in this case the degree of topicality of a tag. We argued that this is due to fact that the majority of tags is related to a topic in the data sets we consider. Therefore, the (co-occurrence) distribution over tags should behave approximately as if it were generated by a corpus dependent distribution over topics and a corpus independent distribution of tags given a topic. In the tasks we have used for evaluation, this emphasis on topics is well in line with the goal of the task, which might be the underlying reason for the good performance of the paradigmatic approach.

Van Vliet et al. ([27]) report on a small scale experiment also evaluating the use of the methods presented in this paper. They compare the semantic similarity of tags of pictures to the topical coherence of the tags as perceived by experts and also find a tendency in this direction. However, evidence is inconclusive for lack of sufficient number of tags and evaluations.

6. Further research

In this paper we intended to investigate the value of a paradigmatic similarity for social tags without caring (much) for efficient computation or scalability. Now that the value of the approach has been shown, it is a topic for further research to investigate to what extend the results carry over to more efficient forms of paradigmatic similarity like e.g. random indexing (see e.g. [9] for a discussion of this topic).

acknowledgements

This research has received funding from the European Community's Seventh Framework Program within the MyMedia project (grant agreement Nº 215006) and the PetaMedia network of excellence (grant agreement Nº 216444).

References

[1] David Arthur and Sergei Vassilvitskii, *k-means++: the advantages of careful seeding*, Proceedings of the eighteenth annual acm-siam symposium on discrete algorithms, 2007, pp. 1027–1035.

[2] Dominik Benz, Andreas Hotho, Robert Jäschke, Beate Krause, Folke Mitzlaff, Christoph Schmitz, and Gerd Stumme, *The social bookmark and publication management system bibsonomy*, The VLDB Journal **19** (2010), no. 6, 849–875.

[3] Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme, *Semantic grounding of tag relatedness in social bookmarking systems*, Iswc '08: Proceedings of the 7th international conference on the semantic web, 2008, pp. 615–631.

[4] Maarten Clements, Arjen P. de Vries, and Marcel J. T. Reinders, *Detecting synonyms in social tagging systems to improve content retrieval*, SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008, pp. 739–740.

[5] T.M. Cover and J.A. Thomas, *Elements of information theory*, Wiley-Interscience, 2006.

[6] James Richard Curran, *From distributional to semantic similarity*, Ph.D. Thesis, 2004.

[7] Ferdinand de Saussure, *Cours de linguistique générale*, V.C. Bally and A. Sechehaye (eds.), Paris/Lausanne, 1916. English translation: Course in General Linguistics. London: Peter Owen, 1960.

[8] B. Fuglede and F Topsoe, *Jensen-Shannon divergence and Hilbert space embedding*, Proc. of the internat. symposium on information theory, 2004, pp. 31–39.

[9] James Gorman and James R. Curran, *Scaling distributional similarity to large corpora*, Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 2006, pp. 361–368.

[10] G. Grefenstette, *Use of syntactic context to produce term association lists for text retrieval*, Sigir '92: Proceedings of the 15th annual international acm sigir conference on research and development in information retrieval, 1992, pp. 89–97.

[11] Z. S. Harris, *Distributional structure*, Word **10** (1954), no. 23, 146–162.

[12] Zellig S. Harris, *Discourse analysis*, Language **28** (1952), no. 1, 1–30.

[13] Antoine Isaac, Lourens van der Meij, Stefan Schlobach, and Shenghui Wang, *An empirical study of instance-based ontology matching*, ISWC'07/ASWC'07: Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference, 2007, pp. 253–266.

[14] Pentti Kanerva, *The Spatter Code for Encoding Concepts at Many Levels*, ICANN '94: Proceedings of the International Conference on Artificial Neural Networks, 1994, pp. 226–229.

[15] J. Karlgren and M. Sahlgren, *From words to understanding*, Foundations of Real-World Intelligence, 2001, pp. 294–308.

[16] Knowledge and Data Engineering Group,University of Kassel, *Benchmark Folksonomy Data from BibSonomy, version of July 1st, 2010*.

[17] John D. Lafferty and ChengXiang Zhai, *Document language models, query models, and risk minimization for information retrieval.*, SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 2001, pp. 111–119.

[18] Bjornar Larsen and Chinatsu Aone, *Fast and effective text mining using linear-time document clustering*, KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999, pp. 16–22.

[19] K. Lindén and J. Piitulainen, *Discovering synonyms and other related words*, Computerm 2004: Proceedings of the 3rd international workshop on computational terminology, 2004, pp. 63–70.

[20] Kevin Lund and Curt Burgess, *Producing high-dimensional semantic spaces from lexical co-occurrence.*, Behaviour Research Methods, Instruments, & Computers **28** (1996), no. 2, 203–208.

[21] Mathias Lux, Michael Granitzer, and Roman Kern, *Aspects of broad folksonomies*, Dexa '07: Proceedings of the 18th international conference on database and expert systems applications, 2007, pp. 283–287.

[22] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze, *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA, 2008.

[23] Benjamin Markines, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho, and Gerd Stumme, *Evaluating similarity measures for emergent semantics of social tagging*, Www '09: Proceedings of the 18th international conference on world wide web, 2009, pp. 641–650.

[24] Yoshiki Niwa and Yoshihiko Nitta, *Co-occurrence vectors from corpora vs. distance vectors from dictionaries*, Coling '94: Proceedings of the 15th conference on computational linguistics - volume 1, 1994, pp. 304–309.

[25] U. Quasthoff, M. Richter, and C. Biemann, *Corpus portal for search in monolingual corpora*, Proceedings of the fifth international conference on language resources and evaluation, lrec 2006, 2006, pp. 1799–1802.

[26] H. Schütze and J.O. Pederson, *A cooccurrence-based thesaurus and two applications to information retrieval*, Proceedings of ria conference, 1994, pp. 266–274.

[27] H. Van Vliet, E. Hekman, N. Veldhoen, M. Rotte, and R. Brussee, *Publieks annotatie van cultureel erfgoed.*, Utrecht University of Applied Sciences. Crossmedia Lab., 2010.

[28] C. Wartena, R. Brussee, M. Wibbels, and Y. van Houten, *An Integrated Recommendation, Browsing and Search Interface Using Tags*, Nem summit, 2009.

[29] Christian Wartena, *Automatic classification of social tags*, Ecdl'10: Proceedings of the 14th european conference on research and advanced technology for digital libraries, 2010, pp. 176–183.

[30] Christian Wartena and Rogier Brussee, *Instance-based mapping between thesauri and folksonomies*, Iswc '08: Proceedings of the 7th international conference on the semantic web, 2008, pp. 356–370.

[31] Christian Wartena and Martin Wibbels, *Improving tag-based recommendation by topic diversification*, Ecir'11: Proceedings of the 33rd european conference on advances in information retrieval, 2011, pp. 43–54.

[32] Cheng Xiang Zhai and John D. Lafferty, *Model-based feedback in the language modeling approach to information retrieval.*, Cikm '01: Proceedings of the tenth international conference on information and knowledge management, January 17, 2005, pp. 403–410.

Appendix A. Details of Clusters of Bibsonomy tags

Table A.6: Selected disciplines and journals

Discipline	Journal
Artificial Intelligence	Artificial Intelligence
	Artificial Intelligence Review
	Genetic Programming and Evolvable Machines
	Machine Learning
Econometrics	Economic Modelling
	Journal of Econometrics
	Journal of Policy Modeling
Economics	Economics Letters
	European Economic Review
	Journal of International Economics
	Journal of Public Economics
Finance	Journal of Banking and Finance
	Journal of Financial Economics
	Journal of International Money and Finance
Heart	Am. J. Physiol. Heart Circ. Physiol.
	Circ. Res.
	Circulation
	Circulation research
Computer Science	Communications of the ACM
	IEEE Expert
	IEEE Intelligent Systems
	IEEE Software
Physics	Phys. Rev. B
	Phys. Rev. Lett.
	Physica A: Statistical Mechanics and its Applications
	Physical Review Letters
Neuro Science	Journal of Neuroscience
Physiology	Journal of General Physiology
	Physiological Review
	Physiology Reviews
	The Journal of General Psychology
Planetary Science	Earth and Planetary Science Letters
	Icarus
	Journal of Geophysical Research-Planets
	Journal Of Geophysical Research-Solid Earth And Planets
	Journal of Geophysical Research. E. Planets
Psychiatry	Archives of General Psychiatry
	Biol Psychiatry
	Santé mentale
Statistics	Computational Statistics and Data Analysis
	Statistics
	Statistics and Probability Letters
Stochastics	Chaos, Solitons & Fractals
	Stochastic Analysis and Applications
	Stochastic Processes and their Applications

Table A.7: Characteristic tags of the selected disciplines

Artificial Intelligence	artificial; boosting; Design; inaki; induction; inductive_programming; juergen; kiwi; LCS; Machine_Learning; program_synthesis; Selection; springer
Econometrics	Applied; CGE; Computable; Disequilibrium; Econometrics; EU; general; Globalization; Macroeconometric; Macroeconomics; Semiparametric; Specification; Test
Economics	aversion; Competition; cycles; Employment; Endogenous; Environmental; Fiscal; goods; Imperfect; Income; Inequality; Migration; Multinational; Optimal; Political; Tax; Trade; Unemployment; Wages; Welfare
Finance	Banks; CEO; directors; Dividends; Executive; fund; funds; IMF; Initial; Investor; investors; IPO; IPOs; Mergers; Microstructure; Mutual; offerings; Ownership; Trading; Venture
Heart	AMP-Dependent; Arrhythmia; beta-Agonists; Ca^{2+}-Transporting; Calcium-Binding; Cardiac; Cardiomegaly; Cardiovascular; Congestive; Diastole; Electrocardiography; Fusion; Ischemia; Isoproterenol; Knockout; Left; N.I.H; Rabbits; Recombinant; Ventricular
Computer Science	agile; cites.pclass; collaborative-filtering; device; ERP; kde; NLP; object-oriented; ontology; personalization; PIM; recommender_systems; research.cs.softeng; research.kr.ontologies; research.nlp; SemanticWeb; v1002; visual-information-seeking; wismasys0809; wwwbook
Physics	Boltzmann; Chaos; Complex; DNA; Econophysics; electron; experiment; frustrated-phase-separation; Granular; high-tc; htsce; htsct; Ising; Lattice; materials; Minority; Nonequilibrium; Nonextensive; Traffic; transitions
Neuro Science	Adult; attn; bg; bio; Brain; cereb; Cerebral; cond; devo; Dopamine; hip; ltp; Male; Nerve; Perception; Resonance; striatum; thal; vis; Visual
Physiology	Aniline; Cattle; Chelating; Cytosol; Dyes; Fluorescence; Fluorescent; In; Permeability; Triphosphate; Vitro; Xanthenes
Planetary Science	Atmosphere; Earth; Features; Galilean; Ganymede; Ice; Icy; Infrared; Jupiter; Meteorites; Moon; Origin; Reflectance; Satellites; Shell; Spectrometer; Spectroscopy; Thermal; Volcanism; Water
Psychiatry	accompagnement; Attention; contenance; écoute; émotions; entretien-informel; grant; neuroleptics; personnalité; personnes-âgées; santé-mentale; Schizophrenia; schizophrénie; TD; troubles-de-la-personnalité; vieillissement
Stochastics	Backward; Branching; deviations; Fleming-Viot; integral; integrals; Interacting; Large; logarithm; Markov; numbers; Poisson; queue; queues; Renewal; scenery; SDEs; sheet; times; Wiener