Instanced-Based Mapping between Thesauri and Folksonomies

Christian Wartena and Rogier Brussee

Telematica Instituut
P.O. Box 589
7500 AN Enschede, The Netherlands
{Christian.Wartena,Rogier.Brussee}@telin.nl

Abstract. The emergence of web based systems in which users can annotate items, raises the question of the semantic interoperability between vocabularies originating from collaborative annotation processes, often called folksonomies, and keywords assigned in a more traditional way. If collections are annotated according to two systems, e.g. with tags and keywords, the annotated data can be used for instance based mapping between the vocabularies. The basis for this kind of matching is an appropriate similarity measure between concepts, based on their distribution as annotations. In this paper we propose a new similarity measure that can take advantage of some special properties of user generated metadata. We have evaluated this measure with a set of articles from Wikipedia which are both classified according to the topic structure of Wikipedia and annotated by users of the bookmarking service del.icio.us. The results using the new measure are significantly better than those obtained using standard similarity measures proposed for this task in the literature, i.e., it correlates better with human judgments. We argue that the measure also has benefits for instance based mapping of more traditionally developed vocabularies.

1 Introduction

Describing collections of books, articles, pictures or movies by assigning keywords to the objects in the collection has a long tradition. Traditionally this has been done by authors, publishers and librarians. Recently, keyword-like metadata are also provided by readers through collaborative tagging systems (1). The nature of these reader provided metadata, usually called tags, differs from the traditional keywords (see e.g. (2)). In particular, keywords are often taken from a restricted vocabulary, e.g. a thesaurus or ontology, while the vocabulary for tagging is always unrestricted. However, only a small part of all tags for a given collection is used frequently (2; 3; 4). The system of terms used in a tagging system, resources (e.g. documents), users and the relations between them is often called a folksonomy (5). More precisely, we will understand a folksonomy as a set of assignments of tags to resources by distinguishable users.

The fact that different collections are described with different vocabularies gives rise to interoperability problems. These problems have been acknowledged

A. Sheth et al. (Eds.): ISWC 2008, LNCS 5318, pp. 356–370, 2008.

as one of the most important obstacles for realizing a large scale semantic web and has led to a large research area on ontology matching (6). The emergence of folksonomies adds the problem of finding mappings between these vocabularies and traditional thesauri and ontologies as a new and interesting issue to this field. One of the main differences between folksonomies and ontologies is the fact that ontologies are usually designed carefully and subsequently might be used to annotate data, whereas folksonomy terms are in the first place used for annotation and the resulting system is only subsidiary. Together with the absence of structure and relations between the terms in a folksonomy this makes instance based mapping a natural choice for finding relations with concepts from a folksonomy.

This paper proposes a new method to map tags, to terms from thesauri or taxonomies (and vice versa), and gives an information theoretic measure for the quality of that mapping. We evaluate our method by mapping Wikipedia categories onto del.icio.us tags and comparing the found mappings to correspondences established by existing methods.

The organization of this paper is as follows: After an overview of related work, we introduce some of the basic concepts used in this paper (section 3). In section 4 we give an overview of dissimilarity measures and introduce a new measure that is especially suited for mapping terms of folksonomies. Section 5 describes an experiment carried out for evaluation and presents its results. We conclude the paper with a discussion for further applications of the mapping method proposed in this paper.

2 Related Work

Euzenar and Shvaiko (6) give an overview of ontology matching systems based on similarity of instances. Isaac e.a. (7) focus more specifically on instance based mapping between ontologies of keywords that uses annotated data to compute similarities between terms. As pointed out by (7) one of the crucial factors for this kind of mapping is the dissimilarity measure used to compare terms. They compare the effects of choosing different dissimilarity measures and find that in their case a slightly modified variant of the well known Jaccard coefficient gives best results. Our focus is also on the dissimilarity measure. We define a new dissimilarity measure that takes advantage of the property of tagged data, that we know the number of people that assigned a tag to an item. The results obtained using this dissimilarity measure are much better than using the other measures we tested.

The FCA-Merge algorithm (8), an approach to ontology merging based on formal concept analysis (FCA), is in fact also a good example of an instance based mapping technique. In FCA concepts are characterized by their instances. Concepts from two ontologies that are characterized by a similar set of instances are likely to be related. This observation is exploited in the FCA-Merge approach. In order to get enough data for merging Stumme e.a. (8) consider occurrences of concepts in documents instead of common instances. Our approach can be

regarded as a statistical version of FCA-merge, in that we do not consider a (binary) occurrence relation of concepts in documents but a probability that a concept occurs in a document. Another difference is that we consider collections of terms and neither use ontological relations between terms nor produce them.

3 Annotated Data

Tags are terms that users give to items, like photos, movies or articles, usually on the internet. Users have different motivations to tag items, the most important being (1) organizing and finding back their favorite items and (2) describing non-textual items. A typical example of the first usage is provided by the bookmarking service del.icio.us. Examples for the latter usage are given by websites for sharing photos or videos. On these websites people tag the items they add to the site to make them findable for other people. In both cases the tags are very similar to keywords in that they provide one word descriptions for (part of) the content of the tagged object. Keywords assigned in a more traditional way differ from tags in that they are often taken from a predetermined list of terms and that they are chosen carefully to reflect the content of an item. Thus, tags contain more noise. Moreover, not all tags describe the content, e.g. opinionating tags ('interesting'), tags like 'to_read' or tags describing a personal context ('thesis') are found (see (1) for an overview of tag types). However, many tagging systems keep track of the number of times a tag was assigned to an item. It is likely that only the relevant descriptional tags reach high frequencies. Halpin e.a. (4) found that the distribution of tags for frequently tagged items tends to become stable over time.

In Wikipedia articles are classified according to categories by the article's authors. These categories are organized hierarchically. Since moreover the category system of Wikipedia is rather stable and the result of many debates on the correct structure, this system and its usage is more similar to a classical taxonomy and its typical usage than to a folksonomy (9), (10).

3.1 Formal Setup

For the following we consider a collection of tagged items (or documents) $\mathcal{C} = \{d_1, \ldots d_M\}$. Furthermore, we consider two collections of n, respectively n' annotations or tag occurrences $\mathcal{W}$ and $\mathcal{W}'$. Each tag occurrence is an instance of a tag t in $\mathcal{T} = \{t_1, \ldots t_m\}$ and $\mathcal{T}' = \{t'_1, \ldots t'_{m'}\}$, respectively. In the following we will assume that $\mathcal{T}$ and $\mathcal{T}'$ (and hence $\mathcal{W}$ and $\mathcal{W}'$) are disjoint. Each occurrence occurs on a tagged item (e.g. document) d in $\mathcal{C}$. Let $n(d,t)$ be the number of occurrences of tag t on d, $n(t) = \sum_d n(d,t)$ be the number of occurrences of tag t, $N(d) = \sum_t n(d,t)$ the number of tag occurrences in d and $D(t) = \{d \mid n(d,t) > 0\}$ the set of documents tagged by T. The size of this set $df(t) = |D(t)|$ is called the document frequency of t.

4 Similarity of Terms

Instance based ontology mapping relies on the presence of a similarity concept for terms based on their instances, or in our case, on their usage as annotations. One of the most obvious things to do is to look at the co-occurrence of annotations from different vocabularies on items in a collection that is annotated according to both systems. In the discussion (section 6) we will also sketch another possibility.

4.1 Co-occurrence Coefficients

A well known family of measures for the degree in which terms co-occur is provided by the co-occurrence coefficients, like the Dice coefficient, the overlap coefficient or the Jaccard coefficient (see e.g. (11) for an overview). In (7) the Jaccard coefficient was used for instance based mapping. We will also use this coefficient to make results comparable. The Jaccard coefficient is given by:

$$JC(t,t') = \frac{|D(t) \cap D(t')|}{|D(t) \cup D(t')|} \tag{1}$$

Isaac e.a. give a slight variation of the Jaccard coefficient that gives smaller scores to low frequency co-occurring annotations (7). They got slightly better results using this coefficient that is defined by

$$JC_{corr}(t,t') = \frac{\sqrt{|D(t) \cap D(t')| \cdot (|D(t) \cap D(t')| - 0.8)}}{|D(t) \cup D(t')|} \tag{2}$$

Both coefficients give values between 0 and 1, where 1 indicates perfect similarity. As a measure for dissimilarity we therefore use $1 - JC(t,t')$ and $1 - JC_{corr}(t,t')$.

4.2 Co-occurrence Distributions

There are two important types of information on the annotations that are not used by the co-occurrence coefficients discussed above. In the first place the number of occurrences of an annotation for an object is not taken into account. This type of information is usually not available for collections annotated with keywords, but is a very important source of information for user tagged data collections, since it allows to suppress "noise" that is always present in these data. In the second place the co-occurrence coefficients look only at the co-occurrence of two annotations but not at other annotations that co-occur with the annotations that are compared: if two terms co-occur often with the same terms, they are likely to be similar, even if their mutual co-occurrence is not very high.

The first type of information could be used by considering annotations as vectors in a document space and computing some geometrical distance between the vectors or by taking the angle between two vectors as a dissimilarity measure. In our experiments it turned out that almost all annotations are completely orthogonal to each other and the mapping based on these dissimilarity measures does

not produce any useful results. Nevertheless, in other experiments useful results were obtained using the cosine similarity (4). For other tasks, like clustering of keywords this measure also gives decent results (12). Taking into account the co-occurrence of other annotations is typical for latent semantic indexing (13). In the following we will introduce a more direct approach that takes both types of information into account.

For a term (tag or keyword) t we compute the co-occurrence probabilities with all other terms. More precisely, for each term t' we compute the probability that an annotation for an item annotated with t is an instance of t', weighted with the importance of t for that item. Arranged in the right way, this gives us for each term a probability distribution over all terms. This approach is very similar to the setup in (14) (section 3). The difference is that we keep track of the density of a term in an item rather than just the mere occurrence or non occurrence of a term. Finally, we can take a standard information theoretic dissimilarity measure between probability distributions in order to compare terms.

To make things more precise we consider (conditional) probability distributions Q on $\mathcal{C}$ and q on $\mathcal{T}$.

$$Q_t(d) = n(d,t)/n(t) \text{ on } \mathcal{C}$$
$$q_d(t) = n(d,t)/N(d) \text{ on } \mathcal{T}$$

The distribution $Q_t(d)$ is called the *source distribution of* t and can be interpreted as the probability that a randomly selected occurrence of term t has source d. Similarly, $q_d(t)$, the *term distribution of* d is the probability that a randomly selected term occurrence from item d is an instance of term t. Now we define the average co-occurrence distribution as

$$\bar{p}_z(t) = \sum_d q_d(t) Q_z(d). \tag{3}$$

We use the notation $\bar{p}_z$ since this distribution is just the weighted average (hence the bar) of the tag distributions of documents containing z where the weight is the probability to find (an instance of) z on item d. We can also interpret this distribution as the transformation of the simple distribution p_z, that is defined by

$$p_z(t) = \begin{cases} 1 & \text{if } t = z\ , \\ 0 & \text{otherwise.} \end{cases}$$

The transformation is given by

$$\sum_{d,t'} q_d(t) Q_{t'}(d) p_z(t') = \sum_d q_d(t) Q_z(d) = \bar{p}_z(t) \tag{4}$$

which is a two step evolution in a Markov chain that connects terms to documents and document to terms.

4.3 Similarity of Distributions

We will use the distributions of co-occurring terms as a base for the definition of the dissimilarity between terms. A standard measure for this is the Jensen-Shannon divergence. The Jensen-Shannon divergence or information radius (11; 15) between two distributions p and q is defined as

$$\mathrm{JSD}(p||q) = \frac{1}{2}D(p||m) + \frac{1}{2}D(q||m)$$

where $m = 1/2(p+q)$ is the mean distribution and $D(p||q)$ is the relative entropy or Kullback-Leibler divergence between p and q which is defined by

$$D(p||q) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right)$$

The dissimilarity between two terms based on the average co-occurrence distributions defined above, is thus given by

$$\mathrm{JSD\,dis}(s,t) = \mathrm{JSD}(\bar{p}_s, \bar{p}_t).$$

This distribution provides a way to express the similarity of the contexts in which two terms occur.

5 Evaluation

To evaluate the quality of instance based ontology mapping using the tag similarity defined in the previous section we have performed two experiments. In the first small scale experiment we have mapped tags assigned to a small set of video fragments by high-school students onto the thesaurus based keywords provided by archive of the Dutch public broadcasting companies, and vice versa. Since the results from this experiment were very encouraging, we performed a second experiment with a much larger data set. In this larger scale experiment we compared the categories of English Wikipedia articles with the tags assigned by del.icio.us users, and evaluated the relation between the dissimilarity of the term mapping and the quality of the mapping. We also evaluated the influence of tag frequency on the quality of the mapping, and compared the dissimilarity measure proposed here with other measures proposed for this purpose.

5.1 The Data Sets

For the first experiment we used tags that were assigned to a set of 115 video fragments by high-school students from different schools in an experiment on tagging (16). 244 students participated in this experiment. They assigned 4,359 different tags to the fragments with a total of 12,414 assignments (tag occurrences). The video fragments were also provided with keywords by the Dutch

Institute of Sound en Vision, the archive of the Dutch public broadcasting companies. The keywords are taken from the Gemeenschappelijke Thesaurus voor Audiovisuele Archieven (GTAA, Common Thesaurus for Audiovisual Archives), containing about 9,000 subject terms and extensive lists of person names, company names and geographical names (17). For the annotation of the selected 115 fragments 269 different keywords were used, with a total of 638 assignments.

For the second experiment we used articles from the English Wikipedia that were also bookmarked by users in a sample of del.icio.us data. To access the category information for the Wikipedia pages we used an SQL dump of Wikipedia from January 3th, 2008 (`http://download.wikimedia.org/enwiki/20080103/`). Besides a large number of categories that are used to classify the content of an article, Wikipedia also has a small number of categories that keep track of the status of an article, e.g. that it needs references, violates copyrights etc. Most of these categories can easily be identified by unique prefixes. We have left out these categories from our data by filtering on the following prefixes: *Wikipedia*, *All_*, *cleanup*, *Unprintworthy*, *Articles*, *Redirects*. Moreover, we have restricted the dataset to article pages, and did not consider previous versions, discussion or history pages etc. From the cleaned up set of pages we selected the subset for which we have at least one tag from a sample of del.icio.us bookmarks, obtained by continuous aggregation at Klagenfurt University and kindly provided to us by Mathias Lux. This gives us 58,345 pages (i.e. about a quarter of all English Wikipedia articles), 42,445 different Wikipedia categories for these pages and 222,640 category assignments together with 49,603 different tags for the selected articles and 278,693 tag assignments.

5.2 Experimental Setup

In the experiments we computed for each tag from a vocabulary $\mathcal{T}$ the nearest tag from vocabulary $\mathcal{T}'$ and vice versa. Thus we have produced two mappings for each experiment and each dissimilarity measure. Since we cannot expect to find useful results for very low frequency terms we only computed the mapping for terms t for which $df(t) > 3$ in the first experiment and $df(t) > 10$ in the second experiment. In order to reduce computation time we also restricted the set of possible candidates to tags with document frequencies higher than 3 and 10, respectively. This restriction has an influence on a very small part of the results only, since these very low frequency tags are unlikely to match the more frequent ones. Thus, we have computed 33 mappings from user tags to GTAA terms and 97 mappings the other way around in the first experiment. In the second experiment 2355 tags were mapped onto a Wikipedia category and 1827 categories onto tags for each evaluated dissimilarity measure.

5.3 Evaluation Criteria

Since there exist no reference mappings for the vocabularies we used, any evaluation will always be somewhat subjective. Moreover, rather than classifying mappings as good or bad, we wanted to have a more fine grained evaluation. We

have therefore defined a number of categories for the quality of a mapping and manually classified a sample of the mappings. We used the following classes:

i Identical. Since the same term might have different meanings in different ontologies or folksonomies, mapping of a term to a literally identical term might not be correct per se. Nevertheless, in the absence of more detailed knowledge of the vocabularies we will consider these mappings as good. Terms with variations in capitalization, usage of blanks, underscores and hyphens and singular/plural variations are classified also classified as identical. Note that we have assumed that the vocabularies $\mathcal{T}$ and $\mathcal{T}'$ are disjoint. Since we keep track of the source of the annotations this is satisfied, even if both vocabularies contain terms with identical string values.

s Synonym. This categories contains synonyms and abbreviations. Examples are pairs like *vista – Windows Vista* or *Human-Computer interaction – hci*.

b Broader term. A mapping is classified as 'broader' if the source term is mapped onto a broader term. Broader term has to be understood in an informal and intuitive way, and not according to some formal ontology. Examples are pairs like *Windows software – Windows* or *War correspondents – journalist*.

n Narrower term. The opposite from the previous category.

r Related term. The term is clearly related but does not fall in any of the previous categories. Examples are pairs like *Pharmacology – drug* or *Digital typography – font*. Note that related terms are not necessarily worse than broader or narrower terms. E.g. *presidential elections* is only a related term of *presidential candidates*, while *people* is a broader term.

u Unrelated. Mappings between terms that are not, or only very loosely related. In this category we find many pairs the relatedness of which terms can only be understood in a specific collection, like *Vermont culture – poetry* or *People from Texas – presidents*

x The source term does not classify the content of an article. Thus it cannot be expected that a meaningful mapping can be found. Examples are tags like *important*, *to_read* or *Wikipedia*. Also some Wikipedia categories that escaped from our filtering, fall into this class.

q We did not know the exact meaning of one of the terms.

In the second experiment we did not evaluate all mappings but evaluated every tenth mapping, the mappings being sorted by the frequency of the source term. Since we are interested in the relation between the quality of the mapping and the frequency of the source terms and between the quality and the dissimilarity of the mapped terms we also evaluated the mappings for the 100 most and 100 least frequent terms, the 100 mappings with the largest dissimilarity between the found terms and the 100 with the smallest dissimilarity. This resulted in the numbers of evaluated mappings as given in Table 1.

5.4 Results

Results for mappings using divergence of average co-occurrence distributions. Fig. 1 shows the fraction of mappings that can be classified according

Table 1. Number of evaluated mappings for two different mapping directions and three different dissimilarity measures

	JSD dis	Jaccard	Jaccard corr.
Categories to tags	522	498	511
Tags to categories	584	568	587

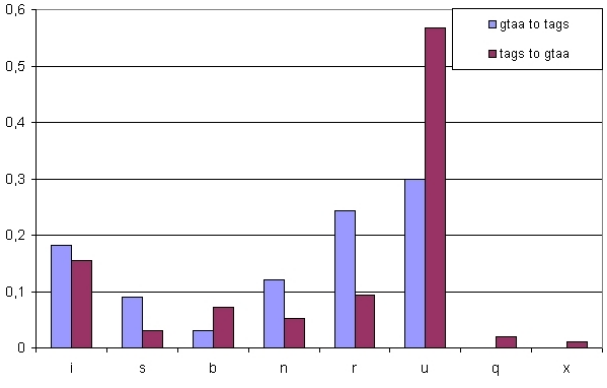

Fig. 1. Fraction of mappings from GTAA terms onto tags and vice versa using JSD dis for each evaluation category

to each of the evaluation classes discussed above[1]. For the thesaurus terms, in about 70% of the cases a synonym or related term could be found. In the opposite direction, for more than half of the tags no related thesaurus term was found. These results are largely due to the very small data set. Recall that we computed matching terms for terms with only more than tree occurrences.

The corresponding results for the experiment with del.icio.us tags and Wikipedia categories is given in Fig. 2, again using JSD dis of average co-occurrence distributions to compute similarities. Again, we see that the mapping from keywords onto tags is much better than the mapping the other way around. However, the overall quality is clearly better. Furthermore, we observe a strong tendency to map the Wikipedia categories to more general tags, whereas the tags tend to be mapped to more specific categories. This suggests that the Wikipedia categories in general are more specific than the user tags. This can also be observed by inspecting the data more closely. The category names are often rather long and specific, whereas the corresponding tags tend to be short and hence in many cases more general, e.g. *20th century classical composers* is mapped onto the tag *composers* or *Software development process* onto *softwaredevelopment*. We should also note that del.icio.us does not support tags consisting of more than one word, but uses a blank as a tag separator. Many tags suggest moreover

[1] The complete set of data from the experiment is available at `https://doc.telin.nl/dsweb/View/Collection-19536`

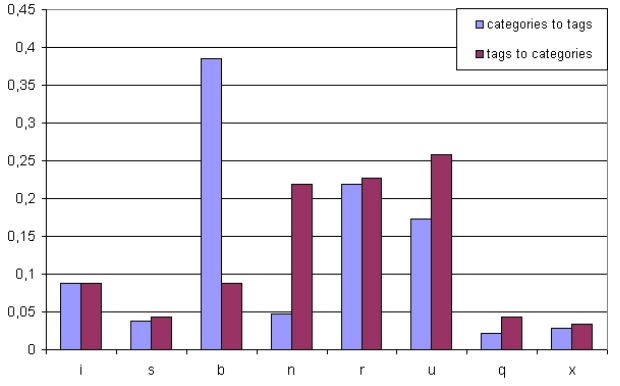

Fig. 2. Fraction of mappings from Wikipedia categories onto tags and vice versa using JSD dis for each evaluation category

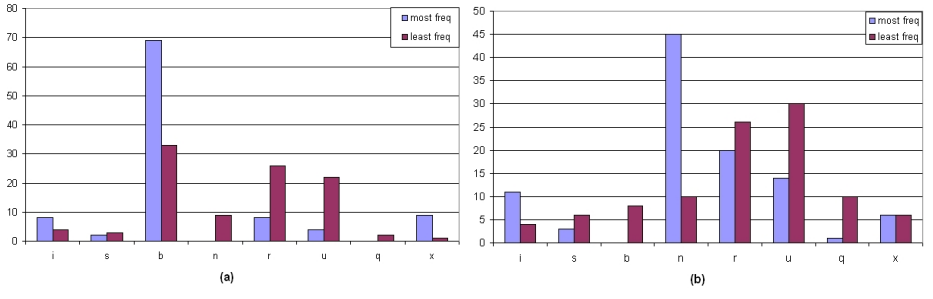

Fig. 3. Evaluation of mappings from the 100 most and least frequent Wikipedia categories onto tags (a) and tags onto categories (b) using JSD dis for each evaluation category

that many users are not aware of this feature. On the other hand in the Wikipedia category system more general terms are available. However, the most specific terms are used to annotate articles. E.g. the term *20th century classical composers* is used to annotate 1,706 articles, the more general terms *classical composers* and *composers* only for 14 and 313 articles, respectively.

Next we inspect the influence of the frequency of terms (tags or categories) on the quality of the found mappings. The results are presented in Fig. 3. Clearly, the results for the high frequency terms are much better than for the least frequent ones. Nevertheless, for both directions the results for the low frequency terms still show substantially more mappings to related terms (including synonyms and broader and narrower terms) than to unrelated terms.

We also investigated whether, for a mapping from t onto t', the divergence of the average co-occurrence distributions, JSD dis(t, t'), can serve as an indication for the quality of the mapping. This is an important feature in practical applications, since this gives the possibility to automatically decide whether a mapping

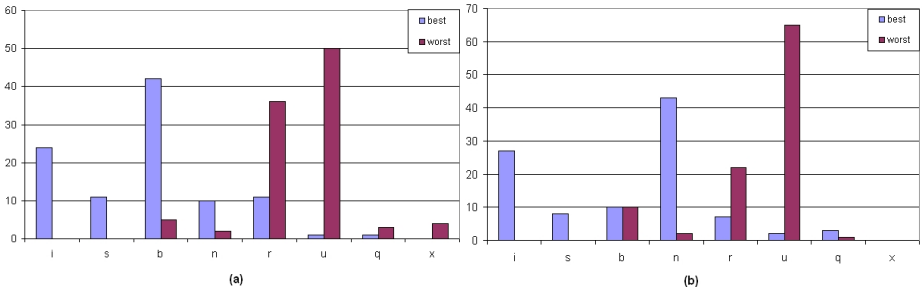

Fig. 4. Evaluation of mappings of the 100 mappings from Wikipedia categories onto tags (a) and tags onto categories (b) with smallest (best) and largest (worst) dissimilarity using JSD dis for each evaluation category

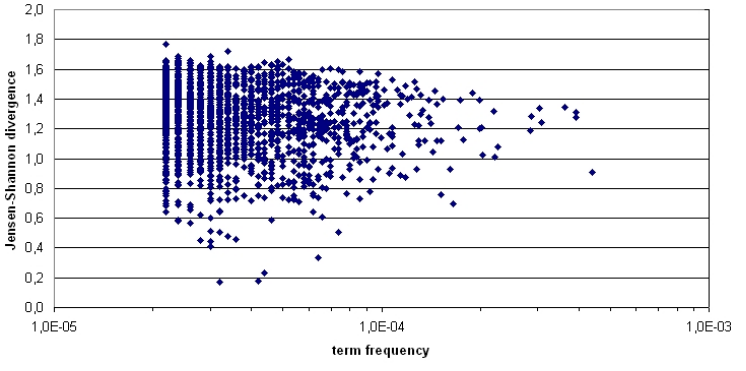

Fig. 5. Frequency of a Wikipedia category (logarithmic scale) vs. the Jensen-Shannon divergence of the category and the best fitting tag

is good enough to be used or not. The results for the evaluation of the mappings with the smallest and largest dissimilarity are given in Fig. 4. The tendency is rather clear and suggest that we can indeed use the dissimilarity as an indication of mapping quality.

Finally, we did not find a strong correlation between the frequency of a Wikipedia category and the dissimilarity of the category and the best fitting tag (see Fig. 5 for the direction from Wikipedia categories to del.icio.us tags).

Comparison between divergence of average co-occurrence distributions and Jaccard coefficient. For a quite similar task (7) found that the Jaccard coefficient and a modification introduced by them (above repeated as 2) gave best results. We compared the results using these two coefficients with the results already discussed above. The fraction of identical mappings is given in Table 2. We see that the Jaccard and the modified Jaccard coefficient give almost the same results, whereas these are rather different from the mapping produced using JSD dis.

Table 2. Fraction of identical assignments by using three different dissimilarity measures for mapping of Wikipedia categories onto tags (first number) and vice versa (second number)

	JSD dis	Jaccard	Jaccard corr.
JSD dis	1 / 1		
Jaccard	0.42 / 0.50	1 / 1	
Jaccard corr.	0.47 / 0.50	0.86 / 0.97	1 / 1

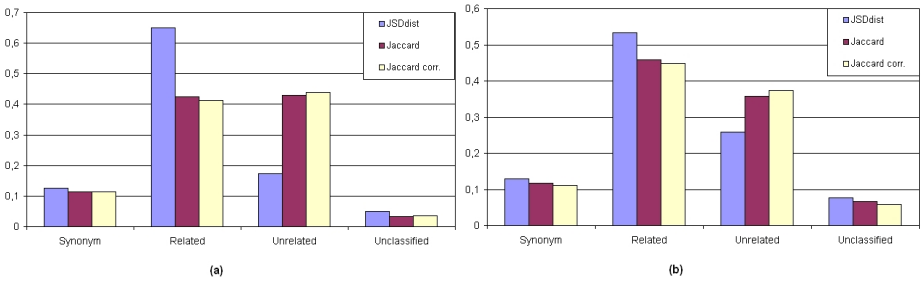

Fig. 6. Fraction of mappings from Wikipedia categories onto tags (a) and vice versa (b) using different dissimilarity measures for each evaluation category

Moreover, the mappings found with JSD dis are substantially better, as can be seen from Fig. 6. We did not find better results using the modified Jaccard coefficient as compared to the simple Jaccard coefficient.

Theoretically, the contexts of two terms t and t' can be very similar and the terms might be mapped onto each other using JSD dis even if t and t' never co-occur. In order to see whether this happens indeed, we computed the Jaccard coefficient for the pairs produced by the mapping from Wikipedia categories onto tags using JSD dis. We found that for 11 mappings the Jaccard coefficient was 0, indicating that there is no overlap. Moreover, there were many mappings with a Jaccard coefficient that was almost 0. Most of the 11 mappings were onto weakly related terms, e.g. *History of science* was mapped onto *philosophy-of-science*. These annotations never co-occur, but it is no surprise that they have similar contexts. In this case the tag found using the Jaccard coefficient was *sci*. Other examples are *1990 deaths – people* (*art-deco* using the Jaccard coefficient) or *science fiction critics – science-fiction* (*batman* using Jaccard coefficient).

6 Discussion

One of the main contributions of this paper is the introduction and usage of a novel similarity measure for terms, the Jensen-Shannon divergence of average co-occurrence distributions. In the experiments we found that this similarity measure gives better results that the Jaccard coefficient for finding corresponding terms in a taxonomy and a folksonomy. It is likely that this is, to a large extent,

the consequence of taking into account the frequency of tag assignments, while the Jaccard coefficient only uses the information whether an article is tagged with a term or not. However, we expect that our measure also gives better results in domains in which such frequency information is not available, since in contrast to simple co-occurrence coefficients like the Jaccard coefficient, we also make use of the context in which tags appear.

In another paper (12) we also obtained good results for clustering keywords using this measure. Together with the relative simplicity of this measure and its natural information theoretical interpretation, Jensen-Shannon divergence of co-occurrence distributions seems to be an interesting new way to compare terms. The theoretical time complexity of computing the underlying distributions $\bar{p}_z$ is a disadvantage for this approach. However, by coding distributions as efficient sparse vectors, the necessary computations are still practicable.

As we have seen above, contexts for two terms can be similar even if they never co-occur. This feature makes it possible to find mappings between the annotation systems of collections with only a small overlap. It should even be possible to find similarities between annotations in collections without overlap, if there is a number of annotations that is common (or for which the mappings are known from another source) to both collections. The context of terms can then be expressed in terms of distributions over these common terms. The divergence of these distributions again serves as a dissimilarity measure of terms. Whether this gives satisfying results, and how many common terms are needed, are questions that are subject for future work.

Finally, we want to remark that in this paper we have focused on the choice for a dissimilarity measure, not on the design of a terminology matching system. In such a system, substantially better results could be obtained, e.g. by reducing the noise arising from different spellings and variations of tags (especially hyphens, underscores, etc.) and by also using lexical similarity as a matching cue.

7 Conclusions

In this paper we introduced a novel measure for the similarity of terms that are used for annotation of items in large collections, like books in libraries, movies in archives, URLs on the internet, etc. This measure takes into account the contexts in which annotations occur and is based on the distribution of co-occurring annotations. We used this measure for instance based mapping between Wikipedia categories and tags from the bookmarking service del.icio.us. We compared the results with mappings produced using the Jaccard coefficient, that is reported to give best results in similar experiments in the literature. In a human evaluation we found that our similarity measure gives substantially better results than the Jaccard coefficient.

A second contribution of this paper is that we have investigated the correspondence of terms from a folksonomy and a more traditionally structured thesaurus. For most frequently used terms a correspondence could be found between categories assigned by the authors of Wikipedia articles and tags used by readers

to bookmark these articles on del.icio.us. However, some advanced statistical methods are needed to detect these correspondences and distinguish them from noise present in folksonomies.

Acknowledgements

We would like to thank Mathias Lux (Klagenfurt University) for kindly making available his collection of tags from del.icio.us. This research was funded by MultimediaN (`http://www.multimedian.nl`), sponsored by the Dutch government under contract BSIK 03031 and by the European Commission FP7 project MyMedia (`http://www.mymediaproject.org`) under the grant agreement no. 215006.

References

1. Golder, S.A., Huberman, B.A.: The structure of collaborative tagging systems. CoRR abs/cs/0508082 (2005)
2. Noll, M.G., Meinel, C.: Authors vs. readers: a comparative study of document metadata and content in the www. In: King, P.R., Simske, S.J. (eds.) ACM Symposium on Document Engineering, pp. 177–186. ACM, New York (2007)
3. Lux, M., Granitzer, M., Kern, R.: Aspects of broad folksonomies. In: DEXA Workshops, pp. 283–287. IEEE Computer Society, Los Alamitos (2007)
4. Halpin, H., Robu, V., Shepherd, H.: The complex dynamics of collaborative tagging. In: WWW, pp. 211–220 (2007)
5. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: BibSonomy: A Social Bookmark and Publication Sharing System. In: Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures, pp. 87–102 (2006)
6. Euzenat, J., Shvaiko, P.: Ontology matching. Springer, Heidelberg (DE) (2007)
7. Isaac, A., van der Meij, L., Schlobach, S., Wang, S.: An empirical study of instance-based ontology matching. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ISWC 2007. LNCS, vol. 4825, pp. 253–266. Springer, Heidelberg (2007)
8. Stumme, G., Maedche, A.: FCA-Merge: Bottom-up merging of ontologies. In: 7th Intl. Conf. on Artificial Intelligence (IJCAI 2001), pp. 225–230 (2001)
9. Ponzetto, S.P., Strube, M.: Deriving a large-scale taxonomy from Wikipedia. In: AAAI, pp. 1440–1445. AAAI Press, Menlo Park (2007)
10. Huijsen, W.O., Wartena, C., Brussee, R.: Learning ontologies from wikipedia for semantic annotation of texts. In: Proceedings of the 13th Knowledge Management Forum, Milano (November 2008) (to appear)
11. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts (1999)
12. Wartena, C., Brussee, R.: Topic detection by clustering keywords. In: DEXA Workshops. IEEE Computer Society, Los Alamitos (to appear, 2008)
13. Landauer, T., Foltz, P., Laham, D.: Introduction to latent semantic analysis. Discourse Processes 25, 259–284 (1998)

14. Li, H., Yamanishi, K.: Topic analysis using a finite mixture model. Inf. Process. Manage. 39(4), 521–541 (2003)
15. Fuglede, B., Topsoe, F.: Jensen-shannon divergence and hilbert space embedding. In: Proc. of the Internat. Symposium on Information Theory, p. 31 (2004)
16. Melenhorst, M., Grootveld, M., Veenstra, M.: Tag-based information retrieval of educational videos. EBU Technical Review Q2 (2008), http://www.ebu.ch/en/technical/trev/trev_2008-Q2_social-tagging.pdf
17. Malaisé, V., Gazendam, L., Brugman, H.: Disambiguating automatic semantic annotation based on a thesaurus structure. In: Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles, pp. 197–206 (2007)