Learning Relations from Social Tagging Data

Hang Dong[1,2], Wei Wang✉[2], and Frans Coenen[1]

[1] Department of Computer Science, University of Liverpool, Liverpool, UK
{HangDong,Coenen}@liverpool.ac.uk
[2] Department of Computer Science and Software Engineering,
Xi'an Jiaotong-Liverpool University, Suzhou, China
Wei.Wang03@xjtlu.edu.cn

Abstract. An interesting research direction is to discover structured knowledge from user generated data. Our work aims to find relations among social tags and organise them into hierarchies so as to better support discovery and search for online users. We cast relation discovery in this context to a binary classification problem in supervised learning. This approach takes as input features of two tags extracted using probabilistic topic modelling, and predicts whether a broader-narrower relation holds between them. Experiments were conducted using two large, real-world datasets, the Bibsonomy dataset which is used to extract tags and their features, and the DBpedia dataset which is used as the ground truth. Three sets of features were designed and extracted based on topic distributions, similarity and probabilistic associations. Evaluation results with respect to the ground truth demonstrate that our method outperforms existing ones based on various features and heuristics. Future studies are suggested to study the Knowledge Base Enrichment from folksonomies and deep neural network approaches to process tagging data.

1 Introduction

Many social media platforms allow users to annotate online data and resources with tags. The accumulated social tags, contributed by millions of online users (folks) collaboratively, are referred to as folksonomies [26]. The original idea was that such folksonomies can provide efficient content organisation mechanisms to support searching of online resources. However, over the years these folksonomies have become a dormant collection of unstructured, noisy and often ambiguous "keywords", which has shown little usefulness.

To address this issue an interesting line of research is to extract "useful" tags and organise them into some forms of structured knowledge, e.g., concept hierarchies or lightweight terminological ontologies [14, 19, 28]. This task is quite different from ontology learning from textual corpora [7] in which it is usually assumed that enough textual data covering specific domains is available. There are several reasons that make learning from tagging data challenging: (i) the difficulties in capturing the intrinsic semantic relations among tags, (ii) sparsity

of the tagging data and (iii) the significant amount of noise (e.g. syntactical variations, typos and spam) and ambiguity (e.g. polysemy and synonymy).

Current methods rely heavily on exploiting co-occurrences of data or external lexical resources to infer tag pair relations [9]. For example, by using heuristics based on set inclusion [18, 19] or graph centrality analysis [14], it is possible to derive the relations, but it is difficult to interpret their meanings explicitly. The co-occurrence based methods typically need to re-compute the whole model when new data is available and therefore do not scale well. In the case of methods based on external lexical resources [8, 12], relations can be explicitly defined; however, a limitation of this class is the low coverage of social tags and their senses typically found in such lexical resources.

We chose academic research as the domain of study because the structured knowledge that can be derived from academic resources is of particular interests to the research community. Learning relations from tagging data in academic domains is also more challenging than learning from general domains, as in the former case many tags are phrases which have complex meanings. To address the limitations of existing work, we propose a new approach to automatically learn relations between tag pairs. The objective is to create knowledge hierarchies by organising tags according to subsumption relations; more specifically, the "broader" and "narrower" relations in the SKOS vocabulary [17] were adopted.

The main contributions of the work include:

- A method to extract domain independent feature sets for articulating the meaning of tags based on probabilistic association analysis, which addresses the aforementioned challenges associated with learning from tagging data;
- A supervised learning method to detect subsumption relations between tags based on their features; the idea being that model trained in one domain can be used in other domains; and
- Extensive experiments and evaluation using two large real world datasets (Bibsonomy[3] and DBpedia[4]) to demonstrate the effectiveness of the proposed approach.

The rest of the paper is organised as follows. Related work on learning subsumption relations and probabilistic topic analysis are presented in Sect. 2. The proposed approach to learn subsumption relations from social tagging data is described in Sect. 3. Experiment and evaluation results are demonstrated in Sect. 4. Finally, conclusion and future works are presented in Sect. 5.

2 Related Work

There are three broad categories of method that are used to learn relations from social tagging data: (i) heuristic/rules, (ii) external lexical resource and (iii)

[3] `https://www.bibsonomy.org/`
[4] `http://dbpedia.org/`

machine learning. Heuristic/rules based methods make use of various heuristics or rules to define and consequently infer relations. Some well-known examples include the use of generality measures based on set inclusion [18, 19] and popularity-generality measures using graph centrality [1, 4, 14]. However, it is known that this category of method cannot formally define the semantic relations among social tags [11, 22]. Another problem is that it is difficult to establish meaningful relations if tagging data are sparse. In other words, two tags may co-relate to each other even though they do not co-occur. Furthermore, this category of methods needs to re-compute the whole model whenever new data becomes available and is thus not likely to scale well.

The second category of method is to ground social tags to external lexical resources to find relations, for example, using WordNet[5] [8], DBpedia and other resources in the Linked Open Data Cloud[6] [12]. However, the methods suffer from the limited coverage of the external resources. The relatively static (or slow-evolving) lexical resources or domain ontologies in general cannot effectively capture data evolution in social media data. It has been found in [2, 3] that WordNet can only represent less than half (48.7%) of the tags in the popular general social tagging dataset del.icio.us [29]; moreover, for many of those that are actually present in WordNet, no intended senses can be found.

The third category of method is to use either unsupervised or supervised machine learning techniques to discover desired hierarchical patterns. The study in [30] proposed an unsupervised divisive clustering algorithm based on Deterministic Annealing to generate a reasonable tag hierarchy; however, it could not discriminate among subordinate, related and parallel relations. By casting relation learning as a supervised classification problem, the work in [22] proposed to detect subsumption relations using association rule mining, set-based tag inclusion measures and graph searching measures. One advantage is that various heuristics or metrics can be used to learn relations. It is also shown in [22] that when using supervised learning with a combination of feature sets, higher F-measures can be achieved compared to any individual approach in the heuristic/rule based category. However, these methods only extract features based on co-occurrence and therefore have similar disadvantages as heuristic/rule based methods. We adapt the idea of supervise learning, but with distinct feature sets to detect semantic relations between tags.

To address the problem of data sparsity, it is necessary to reduce the dimensionality of tagging data. A more effective method is also required to capture the intrinsic semantic meanings of social tags, in different contexts, to disambiguate their meanings. The study in [28] applied probabilistic topic analysis, e.g. Latent Dirichlet Allocation (LDA) [5, 23], to a collection of abstracts of scientific publications from which subsumption relations could be derived. The study in [25] also defined several metrics based on the distribution of topics for concepts to learn ontologies from folksonomies. However, they only suggest how different two tags are, not how they are associated or co-related. Our work addresses this

[5] `http://wordnet.princeton.edu/`
[6] `http://lod-cloud.net/`

problem by extracting domain independent features from tag pairs according to similarity, topic distributions and probabilistic associations. These features are subsequently used for supervised learning.

Similar to probabilistic topic analysis, word embedding approaches can also be used to represent tags in the form of a low dimensional space and consequently better capture the similarity between tags than co-occurrence representation [20]. However, a key disadvantage is that dimensions in word embeddings are not probabilistically and semantically interpretable as probabilistic topic representations (*cf.* [6]). Therefore in this study we chose probabilistic topic analysis as the data representation technique and leave word embedding approaches for a further study.

3 A Supervised Model for Learning Tag Relations

In social tagging platforms, *users* create *tags* to annotate *resources*. Thus, a folksonomy can be formally represented using tuples of the form $\mathbb{F} := \langle U, T, R, Y \rangle$ where U, T and R are finite sets representing *users*, *tags* and *resources* respectively; Y is a ternary relation between them, $Y \subseteq U \times T \times R$ [15]. Due to the noisy nature of tagging data (e.g. special characters, typos, and spam), data cleaning is necessary. Variants of tags (e.g. ontology/ontologies, machine_leaning/machine-leaning) also need to be handled using morphological analysis. Once cleaned the folksonomy is transformed to $\mathbb{F}^{clean} := \langle U, C, R, Y \rangle$, where T is replaced by the new finite set C, whose elements are *tag concepts* or *tag groups*.

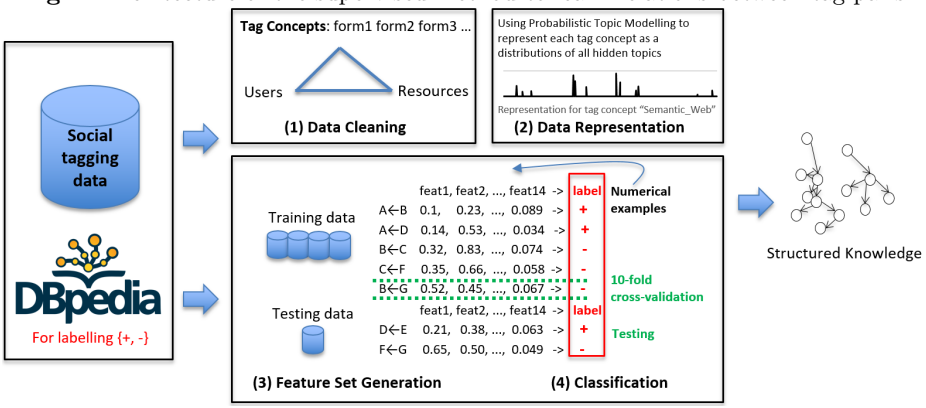
We cast the problem of deriving relations among social tags into a supervised learning problem. The input is a pair of tag concepts, C_a and C_b, represented as two probabilistic distributions in the latent space. The output is the relationship between them: a positive value means that C_a is a narrow concept of C_b. We aim to find these broader-narrower relations (including both direct and indirect ones) and optimise the sensitivity or recall [24] of the classification model.

As shown in Figure 1, the architecture for the proposed method consists of four main components: (i) **Data Cleaning**: cleaning of noisy tagging dataset and transforming it into a cleaned Folksonomy $\mathbb{F}^{clean}$; (ii) **Data Representation**: representing each tag as a distribution of topics in a low dimensional semantic space based on probabilistic topic analysis; (iii) **Feature Set Generation**: generation of feature sets based on topic distributions, similarity and probabilistic associations; and (iv) **Classification**: training and testing of classification models by optimising sensitivity to detect subsumption relations.

3.1 Probabilistic Topic Analysis of Tagging Data

Analogous to "bag-of-words", each resource from a tagging dataset can be represented as a "bag-of-tags". Using probabilistic topic analysis, we can infer the topic structure of tags in an unsupervised manner. One advantage of doing this is that it allows us to obtain the topic distributions in a low dimensional space

Fig. 1. Architecture of the supervised method to learn relations between tag pairs



which captures the meanings of tags under different contexts. This representation is also semantically interpretable: the value of an entry in the latent topic vector reflects its relatedness to that particular topic. With probabilistic topic analysis we can obtain a clear view on the semantic structure of the underlying data, with tag-topic distributions $p(C|\mathbf{z})$ and topic-resource distributions $p(\mathbf{z}|R)$.

Based on $p(C|\mathbf{z})$ and the the Bayes' rule, we represent each tag concept as a probability distribution, $p(z|C_a)$, computed as $p(z|C_a) \propto p(C_a|z) * p(z)$. The prior probability $p(z)$ is usually treated as uniform in the literature [23]. However, the prior distributions of the latent topics are certainly not uniform. We use a non-uniform prior $p(z)$ which respects the underlying dataset, computed as the ratio of the number of tokens sampled to a topic z, N_z, to the number of tokens in the whole dataset, N, $p(z) = \frac{N_z}{N}$. This can be obtained after approximation in probabilistic topic analysis [23]. Usually a tag concept is only closely related to few topics. As such, we introduce the notion of a *significant topic set* $\mathbf{z}_a^{sig}$ for a tag concept, which is specified as $\mathbf{z}_a^{sig} = \{z \mid z \in \mathbf{z} \text{ and } p(z|C_a) \geqslant p\}$, where p is a pre-defined threshold (0.1 in this work).

3.2 Assumptions for Feature Set Generation

We define three assumptions for extracting features in order to discover broader or narrower relations. They are proposed based on the human understanding of a subsumption relation and the cognitive processing of such a relation with respect to three aspects: (i) similarity (the two concepts should be similar), (ii) topic distribution (a more general concept should relate to more topics than a more specific one), and (iii) probabilistic association (given a concept in a certain context, one would be able to derive associated concepts).

Assumption 1. *(Similarity) For two tag concepts C_a and C_b to have a broader-narrower relation, they must be similar to each other to some extent or they must not diverge greatly.*

Table 1. Feature sets **S1**, **S2** and **S3** corresponding to the three assumptions

Features	Description
	S1: Similarity Measure Features
Cos_sim	The cosine similarity of two topic distribution vectors
KL_Div1	The Kullback-Leibler Divergence from C_a to C_b
KL_Div2	The Kullback-Leibler Divergence from C_b to C_a
Gen_Jaccard	The generalised Jaccard Index of two topic distribution vectors
	S2: Topic Distribution Related Features
overlapping	Number of overlapping significant topics
diff_num_sig	Difference of the number of significant topics
diff_max	Difference of the maximum elements in two tag vectors
diff_aver_sig	Difference of the average probability of significant topics
	S3: Probabilistic Association Features
$p(C_a\|C_b)$	The probabilistic association of C_a given C_b
$p(C_b\|C_a)$	The probabilistic association of C_b given C_a
$p(C_a\|C_b, R_{a,b})$	The local probabilistic association of C_a given C_b and a common root concept $R_{a,b}$
$p(C_b\|C_a, R_{a,b})$	The local probabilistic association of C_b given C_a and $R_{a,b}$
$p(C_a, C_b)$	The joint probabilistic association of C_a and C_b
$p(C_a, C_b\|R_{a,b})$	The local joint probabilistic association of C_a and C_b given $R_{a,b}$

Assumption 2. *(Topic distribution) A broader concept should have a topic distribution spanning over more dimensions; while the narrower concept should span over less dimensions within those of the broader concept. This reflects that the narrower concept tends to have a focus on less topics but with higher probabilities than the broader one.*

Assumption 3. *(Probabilistic association) For two tag concepts C_a and C_b to have a broader-narrower relation, they should have a strong association with each other. In a certain context, given one concept, one should be able to associate the other. This can be modelled using the conditional and joint probability of latent topics in a probabilistic framework.*

3.3 Feature Set Generation

The three assumptions are translated into three feature sets, as listed in Table 1. For Assumption 1, we extract features based on a number of similarity/divergence measures, i.e., Cosine similarity, Kullback-Leibler (KL) Divergence and Generalised Jaccard Index, together denoted as the feature set **S1**. (KL) Divergence is an asymmetric measure of the divergence of two probability distributions, which is also the relative entropy of one distribution with respect to another. Since it is asymmetric, we generate two features, denoted as *KL_Div1* and *KL_Div2* as in Table 1. In [28], the difference between KL Divergences was used to discover relations; however, we found that it is difficult to determine a suitable noise threshold.

Topic distribution Based Features The intuition behind Assumption 2 is that the significant topic sets $\mathbf{z}^{sig}$ for two tag concepts C_a and C_b that have a broader-narrower relation tend to be similar or significantly overlapped. While the probability distribution for $\mathbf{z}^{sig}$ of C_a tends to be more uniform, the distribution for C_b tends to be more imbalanced. This reflects the fact that the meaning

of a narrower tag concept is more specific and is concentrated on fewer topics. This is translated into the features on number of overlapped significant topics, difference of the number of significant topics, difference of maximum probability, and difference of the average probability of significant topics. They are referred to as feature set **S2** (see Table 1).

Probabilistic Association Based Features The idea of probabilistic association between two words has its root in cognitive psychology and was first introduced in [13]. It measures the associative relations between words, which can be computed as a conditional probability over a response word given a cue word. The probabilistic association between two tag concepts can also be computed based on this idea.

We model the associations using both conditional and joint probabilities in the latent semantic space. While the conditional probability measures how a tag concept would be associated given another one as a cue, the joint probability measures how two tag concepts would be associated together.

We further propose to compute these two types of probability associations with reference to a specific context. This is done by computing the probabilistic associations conditioned on a third tag concept, which is usually the root concept of a specific domain or sub-domain under consideration. This allows us to learn relations and build a tag concept hierarchy in a progressive, top-down manner. As an example, if machine learning is the domain of consideration, then the concept "Machine Learning" is used as the root concept or context. As the features are extracted by considering a particular context, they are referred to as local associations. The relevant features are together denoted as **S3** (see Table 1) and explained below.

- **Probabilistic Association** The probabilistic association between two tag concepts is computed as the probability of one tag concept given another as a cue in a global context. By the global context we mean that the conditional probability is computed not conditioned on any other concepts. As this measure is asymmetric, we generate two features, $p(C_a|C_b)$, and $p(C_b|C_a)$; the higher the probability, the stronger the association and the more likely that a tag concept can be associated by another. We adopt the method proposed in [13] to compute the two features.
- **Joint Probabilistic Association** The joint probabilistic association captures the likelihood of associating two tag concepts together without references to any specific context. It is a symmetric measure, denoted as $p(C_a, C_b)$, which is computed as $p(C_a, C_b) = p(C_a|C_b) \sum_{z \in \mathbf{z}} p(C_b|z)p(z)$.
- **Local Probabilistic Association** To better capture the association between two tag concepts under a particular context, we propose the idea of local probabilistic association, conditioned on the common root, $R_{a,b}$, of both tag C_a and tag C_b. Since the association is asymmetric, we generate two features denoted as $p(C_a|C_b, R_{a,b})$ and $p(C_b|C_a, R_{a,b})$, respectively. The feature is computed as $p(C_a|C_b, R_{a,b}) = \sum_{z \in \mathbf{z}} p(C_a|z)p(z|C_b, R_{a,b}) =$

$\sum_{z \in \mathbf{z}} \frac{p(C_a|z)p(C_b|z)p(R_{a,b}|z)p(z)}{p(C_b,R_{a,b})}$, where $p(C_a|z)$, $p(C_b|z)$, and $p(R_{a,b}|z)$ can be obtained from the LDA analysis, $p(C_b, R_{a,b})$ can be computed using the joint probabilistic association.

– **Local Joint Probabilistic Association** The local joint probabilistic association is calculated conditioned on the root concept $R_{a,b}$ for both tag C_a and tag C_b. It measures how the two tags are jointly generated within a particular context. It is also a symmetric measure, denoted as $p(C_a, C_b|R_{a,b})$. Similarly, it is computed as $p(C_a, C_b|R_{a,b}) = p(C_a|C_b, R_{a,b})p(C_b|R_{a,b})$, where $p(C_a|C_b, R_{a,b})$ can be obtained using local probabilistic association.

4 Experimental Results and Evaluation

To evaluate the proposed mechanism for learning the relations from social tagging data, a series of experiments were conducted using two large, real-world datasets: Bibsonomy and DBpedia. The tagging data from Bibsonomy was cleaned and only the quality, frequently occurred tags were kept and matched to the terms in DBpedia, which had been organised in a hierarchy. The features were extracted by using the proposed method and used for training and testing different classification models. We also re-implemented and compared to (i) the features proposed in [22] denoted as **S4** with different feature sets and (ii) the "Information Theory Principle for Concept Relationship" in [28] related to our feature set **S1**. The evaluation results demonstrated that our method achieved the highest recall, precision and F_1.

Dataset and Feature Extraction We used the open dataset from Bibsonomy[7], which contains 3,794,882 annotations, 868,015 distinct resources and 283,858 distinct tags contributed by 11,103 users, accumulated from 2005 to July 2015. We cleaned the dataset using morphological and statistical methods, following the four steps in [10]: (i) specific character handling, (ii) multiword and single tag group extraction, (iii) tag selection using selected metrics and (iv) tag selection by language. After these, We selected the tag groups and annotations only for academic publication resources. Each resource was represented as a "bag-of-tags", including all tags used by different users to annotate the resource. We further removed the resources which have less than 3 tag tokens. Finally, we obtained a cleaned, potentially high quality dataset, comprising 7,458 tag concepts and 128,782 publication resources.

To infer latent topics from social tags, we ran the LDA and Gibbs sampling based on the MALLET Machine Learning Library[8]. The topic-word hyperparameter α was set to $50/|\mathbf{z}|$, where $|\mathbf{z}|$ is the number of latent topics, and the document-topic hyperparameter β was set to 0.01. We held out 10% of the data to optimise the perplexity of the LDA model and set $|\mathbf{z}|$ as 600.

[7] `https://www.kde.cs.uni-kassel.de/bibsonomy/dumps`, the "2015-07-01" version.
[8] `http://mallet.cs.umass.edu/`

For tag grounding and instance labelling, we used DBpedia[9] through querying two ontological relations, *skos:broader* and *dct:subject*. Six categories with which we are familiar were chosen (i.e., Machine learning, Semantic Web, Data mining, Natural language processing, Social information processing and Internet of Things). The extracted concepts were matched to the tag concepts in Bibsonomy. In total, we extracted 355 tag pairs with direct broader relations grounded to all the six DBpedia categories, which were used as positive instances. It should be noted that an instance in our method represents features extracted with respect to a pair of tag concepts and a common root of the two tags. Negative instances were created by reversing the broader relations in the positive instances, and generating some random negative relations under each category. We finally obtained 1,065 instances for both training and testing. For each of the instances, we extracted all the 14 features proposed in Section 3.3.

Classification evaluation 80% of the data was randomly selected for training and 20% for testing. In both the training and testing data, the ratio of the number of positive to negative instances was around 1:2. For the current work we aimed to train classification models with high sensitivity. The evaluated metrics used include precision, recall, F_1 score, accuracy and the Area Under the receiver operating characteristic Curve (AUC). For imbalanced data, as in the case of our experiments, precision, recall, F_1 score and AUC are more suitable evaluation metrics than accuracy [24].

We trained a number of classifiers, Logistic Regression (LR) and Support Vector Machine (SVM), on the data described above. For parameter tuning, 10-fold cross validation was used. For the SVM model, we used the standard radial basis (RBF) kernel and tuned two parameters C and γ [16] to optimise the sensitivity. In addition, the weighted-SVM [21, 27] was used to boost the recall. Weighted-SVM specifies two different misclassification cost parameters for the two classes: C^+ for positive observations and C^- for negative observations. The ratio of the misclassification cost parameters was set to 2, i.e., $\frac{C^+}{C^-} = 2$.

The evaluation and comparison results are presented in Table 2. From the table, it can be seen that in all experiments, SVM performed better than LR in terms of recall, precision, F_1 score, accuracy and AUROC. A sensitivity of 73.2% was obtained using the standard SVM with RBF kernel. This setting also produced the highest precision, F_1, accuracy and AUROC values. By heavily penalising the misclassification cost on positive instances, the weighted-SVM achieved 100% recall. However, this setting produced a very high false positive rate and therefore, the precision and accuracy were lower than the best results obtained using the standard SVM.

Compared to the feature set **S4**, proposed in [22], which was mainly based on tag co-occurrences, our proposed mechanism performed significantly better. This is attributed to the well-founded assumptions based on the semantically interpretable latent topics. Also, the recall and precision were not improved when

[9] `http://downloads.dbpedia.org/2015-10/core/`, the "2015-10" version

Table 2. Classification results using different feature set combinations

		Recall	Precision	F_1 Score	Accuracy	AUC
S1+S2+S3 (Full features in our approach)	LR	54.9%	60.0%	57.4%	72.8%	0.808
	SVM	**73.2%**	65.0%	68.9%	77.9%	0.814
	weighed-SVM	**100.0%**	42.0%	59.2%	54.0%	0.792
Wang et. al [28] (**S1**)	LR	12.7%	47.4%	20.0%	66.2%	0.585
	SVM	38.0%	58.7%	46.2%	70.4%	0.648
Rêgo et. al [22] (**S4**)	LR	16.9%	63.2%	26.7%	69.0%	0.657
	SVM	22.5%	57.1%	32.3%	68.6%	0.563
S1+S2+S3+S4	LR	56.3%	62.5%	59.3%	74.2%	0.808
	SVM	71.8%	64.6%	68.0%	77.5%	0.818
S2	LR	22.5%	59.3%	32.7%	69.0%	0.752
	SVM	59.2%	55.3%	57.1%	70.4%	0.688
S3	LR	4.2%	37.5%	7.6%	65.7%	0.769
	SVM	5.6%	50.0%	10.1%	66.7%	0.794
S1+S2	LR	42.3%	61.2%	50.0%	71.8%	0.761
	SVM	63.4%	57.0%	60.0%	71.8%	0.699
S1+S3	LR	32.4%	54.8%	40.7%	68.5%	0.700
	SVM	62.0%	64.7%	63.3%	76.1%	0.776
S2+S3	LR	33.8%	60.0%	43.2%	70.4%	0.787
	SVM	59.2%	60.9%	60.0%	73.7%	0.743

* **S1** denotes Similarity and Divergence Based Features; **S2**, Topic distribution Based Features; **S3**, Probabilistic Association Features; **S4**, the baseline feature set in [22] including support, confidence, cosine similarity, inclusion and generalisation degree, mutual overlapping and taxonomy search.

we combined our features with the baseline features (**S1**+**S2**+**S3**+**S4**), 71.8% recall, 55.0% precision compared to 73.2% and 55.0% when only **S1**+**S2**+**S3** was used, showing that the co-occurrence based features do not provide any further contribution to the results.

Table 2 also shows the results obtained when different combinations of the feature sets were used. With all three feature sets both LR and SVM produced the best recall, F_1, accuracy and AUC values. When only one feature set was considered, as can be seen from from the table, using the topic distribution related features (**S2** founded on Assumption 2) generated the best results when using SVM (59.2% recall and 57.1% F_1). Both topic distribution related features and similarity/divergence features significantly outperforms the baseline **S4**. When two sets of features were considered, the similarity/divergence (Assumption 1) and topic distribution related (Assumption 3) features produced the best results in all cases. If the similarity/divergence based feature set **S1** alone were used, then the method corresponds to the method on learning ontologies from publication abstracts as described in [28]. Surprisingly, the result produced from tagging data was not useful, which shows that the feature set **S1** alone is not sufficient for capturing the meaning of tag concepts. This can be attributed to the fact that publication abstracts contain much richer information than tags. Another notable finding is that, although probabilistic association based features **S3** alone led to low recall (5.6% by SVM and 4.2% by LR), it boosted the recall significantly (increasing it by 15.5% from 63.4% to 73.2%) when combined with the other two feature sets. Some example hierarchical relations predicted on the test set using SVM and the three feature sets are shown in Table 3.

Table 3. Examples of learned relations from Bibsonomy tags using the three feature sets **S1**+**S2**+**S3** with SVM

narrower → broader concept	**narrower → broader concept**
social_graphs → social_networks	semantic_analysis → machine_learning
mixture_model → data_mining	unsupervised_learning → machine_learning
folksonomy → collective_intelligence	latent_variables → bayesian_networks
semantic_search → semantic_web	sentiment_analysis → natural_language_processing
delicious → social_bookmarking	word_sense_disambiguation → natural_language_processing

5 Conclusion and Future Work

Social tagging data represent a potential source for extracting structured knowledge, which is of particular interests and importance to the research communities. In this on-going work, we have presented a novel method to derive domain independent features and learn broader/narrower relations among tag concepts in constructing such structured knowledge. Based on our understanding and perception of concept relations, three assumptions are proposed to determine if a broader-narrower relations holds between any two given tag concepts. These assumptions are then translated into a number of effective feature sets for the learning task, in particular, the probabilistic association based one, which helps capture tag relations based on human cognitive processing of information. The experiment and evaluation results confirmed the effectiveness of the method and showed that the proposed method significantly outperforms existing work in terms of recall, the primary focus, precision and the F_1 measure. The combination of feature sets is an effective strategy for detecting relations among tags. For future study, we plan to further fine-tune this approach to built knowledge hierarchies iteratively and progressively. We will extend our experiments, using DBpedia only, to several heterogeneous Knowledge Bases covering all domains for tag grounding and instance labelling. We also plan to evaluate the performance of our method to see if it can be used for the purpose of enriching existing Knowledge Bases with new concepts and relations. Deep neural network approaches, leveraging both probabilistic topic representation and word embeddings, are also to be explored.

Acknowledgment

This research is funded by the Research Development Fund at Xi'an Jiaotong-Liverpool University, contract number RDF-10-2015.

References

1. Almoqhim, F., Millard, D.E., Shadbolt, N.: Improving on popularity as a proxy for generality when building tag hierarchies from folksonomies. In: Social Informatics: 6th Int. Conf. pp. 95–111. Springer International Publishing, Cham (2014)
2. Andrews, P., Pane, J.: Sense induction in folksonomies: a review. Artificial Intelligence Review 40(2), 147–174 (2013)

3. Andrews, P., Pane, J., Zaihrayeu, I.: Semantic disambiguation in folksonomy: A case study. In: Advanced Language Technologies for Digital Libraries: International Workshops on NLP4DL 2009 and AT4DL 2009. pp. 114–134. Springer Berlin Heidelberg (2011)
4. Benz, D., Hotho, A., Stumme, G., Stützer, S.: Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge. In: Proc. of the 2nd Web Science Conference (WebSci10) (2010)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)
6. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Advances in neural information processing systems. pp. 288–296 (2009)
7. Cimiano, P.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006)
8. Djuana, E., Xu, Y., Li, Y.: Learning personalized tag ontology from user tagging information. In: Proceedings of the Tenth Australasian Data Mining Conference - Volume 134 (AusDM '12). pp. 183–189. Australian Computer Society, Inc. (2012)
9. Dong, H., Wang, W., Liang, H.N.: Learning structured knowledge from social tagging data: A critical review of methods and techniques. In: 2015 IEEE Int. Conf. on Smart City/SocialCom/SustainCom (SmartCity). pp. 307–314 (Dec 2015)
10. Dong, H., Wang, W., Frans, C.: Deriving dynamic knowledge from academic social tagging data: a novel research direction. In: iConference 2017 Proceedings. iSchools (2017)
11. García-Silva, A., Corcho, O., Alani, H., Gómez-Pérez, A.: Review of the state of the art: discovering and associating semantics to tags in folksonomies. The Knowledge Engineering Review 27(1), 57–85 (2012)
12. García-Silva, A., García-Castro, L.J., García, A., Corcho, O.: Social tags and linked data for ontology development: A case study in the financial domain. In: The 4th Int. Conf. on Web Intelligence, Mining and Semantics. pp. 1–10. ACM (2014)
13. Griffiths, T.L., Steyvers, M., Tenenbaum, J.B.: Topics in semantic representation. Psychological Review 114(2), 211 (2007)
14. Heymann, P., Garcia-Molina, H.: Collaborative creation of communal hierarchical taxonomies in social tagging systems. Tech. rep., Stanford (2006)
15. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. In: The 3rd European Conf. on The Semantic Web: Research and Applications (ESWC'06). pp. 411–426. Springer-Verlag (2006)
16. Hsu, C.W., Chang, C.C., Lin, C.J.: A practical guide to support vector classification. Tech. rep., Dept. of Computer Science, National Taiwan University (2003)
17. Isaac, A., Summers, E.: Skos simple knowledge organization system. Primer, World Wide Web Consortium (W3C) (2009)
18. Meo, P.D., Quattrone, G., Ursino, D.: Exploitation of semantic relationships and hierarchical data structures to support a user in his annotation and browsing activities in folksonomies. Information Systems 34(6), 511–535 (2009)
19. Mika, P.: Ontologies are us: A unified model of social networks and semantics. Web Semantics: Science, Services and Agents on the World Wide Web 5(1), 5–15 (2007)
20. Niebler, T., Hahn, L., Hotho, A.: Learning word embeddings from tagging data: A methodological comparison. In: Lernen, Wissen, Daten, Analysen (LWDA) Conference Proceedings. pp. 229–240 (2017)
21. Osuna, E., Freund, R., Girosi, F.: Support vector machines: Training and applications. Tech. rep., AI Memo 1602, Massachusetts Institute of Technology (1997)

22. Rêgo, A.S.C., Marinho, L.B., Pires, C.E.S.: A supervised learning approach to detect subsumption relations between tags in folksonomies. In: Proc. of the 30th Annu. ACM Symp. on Applied Computing (SAC '15). pp. 409–415. ACM (2015)
23. Steyvers, M., Griffiths, T.: Probabilistic topic models. Handbook of Latent Semantic Analysis 427(7), 424–440 (2007)
24. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining, (First Edition). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (2005)
25. Tang, J., Leung, H.f., Luo, Q., Chen, D., Gong, J.: Towards ontology learning from folksonomies. In: Proc. of the IJCAI. vol. 9, pp. 2089–2094 (2009)
26. Vander Wal, T.: Folksonomy. `http://vanderwal.net/folksonomy.html` (2007), [Online; accessed 07-June-2018]
27. Veropoulos, K., Campbell, C., Cristianini, N., et al.: Controlling the sensitivity of support vector machines. In: Proc. of the IJCAI. pp. 55–60 (1999)
28. Wang, W., Barnaghi, P.M., Bargiela, A.: Probabilistic topic models for learning terminological ontologies. IEEE Trans. Knowl. Data Eng. 22(7), 1028–1040 (2010)
29. Wetzker, R., Zimmermann, C., Bauckhage, C.: Analyzing social bookmarking systems: A del. icio. us cookbook. In: Proc. of the ECAI 2008 Mining Social Data Workshop. pp. 26–30 (2008)
30. Zhou, M., Bao, S., Wu, X., Yu, Y.: An unsupervised model for exploring hierarchical semantics from social annotations. In: The 6th Int. Semantic Web Conf., 2nd Asian Semantic Web Conf. pp. 680–693. Springer Berlin Heidelberg (2007)