Linking Multimedia Items to Semantic Knowledge Base with User-Generated Tags

Shuangyong Song, Chao Wang, Haiqing Chen

Alibaba Group, Beijing 100102, China.
{shuangyong.ssy; chaowang.wc; haiqing.chenhq}@alibaba-inc.com

Abstract. Multimedia items account for an important part of Linked Open Data (*LOD*), but currently most of the semantic relations between multimedia items and semantic knowledge base (*KB*) are based on manual semantic annotation. With the popularity of multimedia hosting websites, plentiful tagging information makes it possible to automatically generate semantic relations between multimedia items and *KB*. In this paper, we propose a mechanism for linking multimedia items to *KB* with user-generated tags, while taking into account topical semantic similarity between tags. Experimental results show the effect of our work on this task.

Keywords. multimedia *LOD*; knowledge base; user-generated tags; topic model

1 Introduction

Multimedia *LOD* has attracted considerable attention, but how to share and search multimedia items on semantic web remains a significant yet challenging research issue. Most of the semantic relations between multimedia items and semantic ontologies are based on manual semantic annotation, which is very time-consuming. Some researches try to automatically tag multimedia items based on their web page text [1], however, the complexity of web text generates much noise data and makes it difficult to detect the exact text that are really related to a target multimedia item.

Online multimedia hosting systems, such as Flickr, YouTube and Last.fm, have attracted great attention in that they enable an effective way for users to organize, tag and share multimedia items. Researches on understanding multimedia items based on their tags have been attached importance [2]. In this paper, we aim to create links between multimedia items and semantic *KB* by using their tagging information, in this way to realize semantic multimedia retrieval and detection of multimedia relations.

We firstly detect tags which can be unambiguously linked to ontologies in high-quality public semantic *KB* and process some ambiguous tags with simple tags' Co-occurrence relation, and then use the multimedia "item-tag" relation matrix to train topic models, through which to calculate topical semantic relations between tags, and then detect implicit semantic links between tags and semantic ontologies. Finally, we link multimedia items to semantic *KB* through tag-based mediation. In the following parts of this article, we will illustrate the details of the problem definition and proposed model, and report the experimental results.

2 Problem Definition & Approach

2.1 Problem Definition

Referring to [2], we give a definition of *multimedia ontology* as "ontology of a multimedia item with a unique URI and available links to public recognized semantic *KB*". We use *DBpedia* as referred semantic *KB*, where identifiers of category-level ontologies begin with "dbo", such as "dbo:movie", and identifiers of instance-level ontologies begin with "dbr", such as "dbr:creditcard". We define the set of multimedia items as $M = \{m_1, m_2, \ldots, m_i, \ldots, m_I\}$, and for m_i, a series of user-defined tags are given as $T(m_i)$. Our goal is to create a set of ontologies $O(m_i)$ by linking m_i to *KB*, with considering both explicit and implicit semantic links between $T(m_i)$ and *KB*.

2.2 Approach

1) Linking Tags to Ontologies: We firstly link all unambiguous tags to *KB*, which means that a tag has only one matched ontology definition in *KB*. For example, when $T(m_i)$ contains a tag "credit_card", we can detect an unambiguous matched ontology as "dbr:creditcard", then we add it to $O(m_i)$. However, when $T(m_i)$ contains a tag with multiple matched ontologies in *KB*, such as the tag "Apple", we design a simple tags' Co-occurrence relation based method to determine linking an ambiguous tag to which ontology or not linking it to any ontology. The formula of co-occurrence relation $R(t_i^a, t_j^u)$ between an ambiguous tag t_i^a and an unambiguous tag t_j^u is given below:

$$R(t_i^a, t_j^u) = \sum_{k=1}^{U} R(t_i^a, t_k^u, t_j^u) = \sum_{k=1}^{U} \frac{C(t_i^a, t_k^u) * C(t_k^u, t_j^u)}{F(t_k^u)} \quad (1)$$

where $R(t_i^a, t_k^u, t_j^u)$ means partial co-occurrence relation between t_i^a and t_j^u created through t_k^u. U means number of all unambiguous tags. $C(t_i^a, t_k^u)$ means co-occurrence frequency of t_i^a and t_k^u, and $F(t_k^u)$ means frequency of t_k^u. Finally, a t_j^u with the maximum $R(t_i^a, t_j^u)$ of t_i^a will be detected as a vicarious tag of t_i^a. In particular, we also detect some tag-combined ontologies for expanding the links' range. For example, if an item has both "DigitalCamera" and "Conon" as its tags, we will check if there is a semantic ontology "DigitalCamera_Conon" or "Conon_DigitalCamera" in *KB*.

2) Detecting Semantic Relations between Tags: Probabilistic topic models, such as Latent Dirichlet Allocation model (LDA), have been proved to be powerful tools for identifying latent topical information. We utilize *JGibbLDA* [4] to detect topical information from item-tag matrix. Since the item-tag matrix is sparse and difficult to be well analyzed, topical information can help to discover implicit semantic relations between tags, by calculating similarity between tags' topical vectors. Besides, this step can be regarded as a dimension reduction of tags' vector space, which can greatly reduce computational consumption of semantic similarity between tags. If we set the number of topics as K, then each tag can be represented as a K-dimension vector.

3) Extending Ontology Linked Tags: The aim of this step is to expand the scope of "ontology linked tags" by considering topical semantic similarities between tags, as well as some lexical analysis. Lexical analyses include synonyms analysis, plurals analysis and gerund analysis, while topical semantic similarity based method focuses

on topical vectors of tags. For checking the possibility of linking an unlinked tag t_1 to semantic *KB*, we calculate the topical semantic similarity between t_1 and all *KB* linked tags t_* with cosine-similarity $C(t_1,t_*)$ of their topical vectors, and choose those with similarity greater than threshold σ to be synonymous tags, where $\sigma = 1\text{-}10^{-d}$, and d is a positive integer, while a larger d means a stricter threshold. Besides, for two tags t_1 and t_2, supplemented with *Inclusion Relation* and *Levenshtein Distance* between them, we design some rules for judging if they are similar tags:

a) If t_1 and t_2 have *Inclusion Relation*, such as "motor" and "motorcycle", and $C(t_1,t_2)$ is bigger than σ, we judge them as similar tags;

b) If *Levenshtein Distance* value between t_1 and t_2 is equal to or smaller than a threshold β, which is a positive integer, and $C(t_1,t_2)$ is bigger than σ, we judge them as similar tags;

c) If t_1 and t_2 don't have above relations, we judge them as dissimilar tags.

d) If t_1 and t_2 are judged as similar tags, and just one of them has semantic link to *KB*, we link the other one to the same ontology with a probability $C(t_1,t_2)$.

4) Linking Multimedia items to Ontologies: Based on mapping relations between tags and *KB*, we link multimedia items to *KB* by taking tags as medium. For selecting predicates in <*subject*, *predicate*, *object*> triples, we use 77 million triples collected from a semantic database *LOD4ALL* [1] as criteria. We create links between multimedia items and *KB* ontologies with unambiguous predicates. For example, we use predicate "dbo:locationCountry" to create links from image items to object ontologies such as "China" or "Denmark". For objects with multiple predicates, links will not be created while for other unknown objects, we tentatively use predicate as "rdfs:seeAlso" which means "further information about the subject resource".

3 Experiments

3.1 Dataset and Parameter Settings

Three datasets of images, videos and music are respectively prepared: 1) mirFlickr dataset [2] was utilized, which contains 0.7M images and 0.65M tags; 2) YouTube-8M dataset [3] consists of about 8M videos and 4716 unique tags; 3) Last.fm dataset [2] consists of about 0.5M music tracks and 0.52M unique tags. We remove multimedia items without any tag or just with "stopword" tags, and remaining datasets are used to train domain-sensitive topic models respectively with an empirical K value of 200.

For evaluating the effects of different β, we set β as positive integer between 1 and 6 and compare the performance with different β, since we empirically judge that two tags with *Levenshtein Distance* more than 6 are with little probability to be similar tags. We randomly choose 500 tag-couple with *Levenshtein Distance* equals to or smaller than 6 as the dataset for evaluating different β, and we roughly set the recall as 1.0 when $\beta = 6$ considering we are unable to collect all valid tag-couples. We firstly get the *recall* and *precision* results and further the *F1*-value to evaluate β. Table 1 shows the results. As shown, we get the best performance when β is 2. Therefore, we set $\beta = 2$ in the following experiments.

[1] https://lod4all.net/zh/index.html

[2] https://labrosa.ee.columbia.edu/millionsong/lastfm

Table 1. Results with different β.

	$\beta=1$	$\beta=2$	$\beta=3$	$\beta=4$	$\beta=5$	$\beta=6$
Precision	0.901	0.892	0.692	0.439	0.295	0.248
Recall	0.555	0.601	0.688	0.755	0.835	1.000
***F*1-value**	0.687	**0.718**	0.690	0.555	0.436	0.397

Table 2. Results with different d (when $\beta = 2$)

	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$
Precision	0.322	0.545	0.921	0.877	1.000	1.000
Recall	1.000	0.887	0.728	0.411	0.169	0.055
***F*1-value**	0.487	0.675	**0.813**	0.560	0.289	0.104

Totally 144,318 tags are unambiguously mapped to *dbo* or *dbr*, which only account for 17.59% of all tags. To evaluate the effects of different d when $\beta = 2$, we manually label 600 multimedia items. We roughly set the recall as 1.0 when $d = 1$ considering we are unable to collect all possible related ontologies to an item. After getting both *precision* and *recall* with different d, we also use *F1*-value as evaluation criterion to evaluate d. Table 2 shows the results. As shown, the best performance shows when $d = 3$, which indicates that when $\sigma = 1\text{-}10^{-3}$ we can get best discriminant performance. Therefore, we set $d = 3$ in the following experiments.

3.2 Experimental Results

With our model, 591,850 tags are finally mapped to *KB*, which account for 72.18% of all tags, which makes tremendous growth compared to 17.59%. We compare our model with two baselines on the tag similarity calculation subtask:

• Word2Vec based model (*W2V*): Word2vec is a recently popular method for getting distributed representation of words, of which the output form is similar to *LDA*.

• Co-occurrence based model (*Co-occur*): *Co-occur* is another method for detecting relationship between words, which doesn't consider latent semantic information.

Table 3. Result comparison with F1-value.

Dataset	Flickr			YouTube			Last.fm		
Modes	*W2V*	*Co-occur*	***Our model***	*W2V*	*Co-occur*	***Our model***	*W2V*	*Co-occur*	***Our model***
***F*1-value**	0.356	0.654	**0.841**	0.401	0.686	**0.855**	0.388	0.640	**0.838**

We randomly choose 300 items from each dataset as test datasets. For each item, we manually check the validity of every created item-ontology link, which can help getting the precision easily for each model, and we take the union of all valid results of three different models as basis for getting recall of each model. Then the *F*1-value can be calculated and the results are shown in Table 3.

4 Future Works

This paper is only a preliminary work. Named entity disambiguation and named entity normalization will be considered in our next step work. Besides, *predicate* discriminate should be performed by considering adjacent tags as context.

5 References

1. Ding, G., Xu, N. Automatic semantic annotation of images based on Web data. In IAS'10.
2. Huiskes, M. J., Lew, M. S. The MIR Flickr Retrieval Evaluation. In SIGMM'08, pp.39-43.
3. Abu-El-Haija S, et al. YouTube-8M: A large-scale video classification benchmark. 2016.
4. Phan, X-H., Nguyen, L-M., Horiguchi, S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In WWW'08, pp.91-100.