On Learning Knowledge Bases for Collabularies

Krisztian Buza, Dipl.-Ing.[1 3]
Leandro Balby Marinho, MSc.[1 3]
Lars Schmidt-Thieme, Prof. Dr. Dr.[2 3]
[1]*Research Assistant*
[2]*Professor*
[3] Information Systems and Machine Learning Lab (ISMLL)
University of Hildesheim, Hildesheim, Germany
{buza,marinho,schmidt-thieme}@ismll.uni-hildesheim.de

Összefoglalás: Nemrég javasoltunk egy új algoritmust felhasználói és szakértői tudást egyaránt leíró ontológiák (collabulary) automatikus tanulására és bemutattuk hogy ezek gyakorlati alkalmazásokban előnyösebbek a kizárólag szakértői tudást ábrázoló ontológiáknal. Ebben az cikkben az objektumpéldányok kategóriákhoz való hozzárendelésével foglalkozunk: összehasonlítunk egy adaptív és egy küszöbszám alapú megoldást, megvizsgáljuk a küszöbszám hatását.

Kulcsszavak: ontológia (taxonómia) tanulás, ontológia populáció, felhasználói és szakértői tudást egyaránt leiró ontológiák alkalmazásai, felhasználói és szakértői tudást egyaránt leiró ontológiák kiértékelése

Abstract: A collabulary is an ontology representing user knowledge and domain-expert knowledge in an integrated structure. Recently Marinho at al. proposed a new approach for automatically learning collabularies by means of semantic mapping and highly efficient frequent itemset mining techniques. They showed that learned collabularies may outperform domain-expert ontologies in practical applications. Now we focus on the problem of populating collabularies: we compare an adaptive and a threshold-based approach and investigate the influence of that threshold.

Keywords: ontology (taxonomy) learning, ontology population, knowledge base learning, collabulary applications, collabulary evaluation

1. Introduction

Due to the concrete advances towards the Semantic Web vision [1], ontologies are growing in use, specially in areas concerning information finding and organization. However, their massive adoption usually needs huge human effort: task of assembling ontologies is usually assigned to domain experts and knowledge engineers. Although ontology learning can help to some extent, the participation of the expert is still usually required since the learned representations are not free of inconsistences (in a semantic level at least) and therefore require manual validation and fine tuning.

A more promising solution to this problem lies in the rapid spread of the Web 2.0 paradigm: it has the potential to motivate ordinary users towards voluntary semantic annotation. The increasing popularity of Web 2.0 applications can be partly explained by the fact that no specific skills are needed for participating, where anyone is free to add and categorize resources at will in the form of free keywords called *tags*. Tags do not need to conform to a closed vocabulary and therefore reflect the latest terminology in the domain under which the system operates.

Although this freedom can cause a selfish behavior, the exposure to each other tags and resources creates a fundamental trigger for communication and sharing, thus lowering the barriers to cooperation and contributing to the creation of collaborative lightweight knowledge structures known as *folksonomies.* Despite the compelling idea of folksonomies, its uncontrolled nature can bring problems, such as: synonymy, homonymy, and polysemy, which lowers the efficiency of content indexing and searching. Another problem is that folksonomies usually disregard relations between their tags, what restricts the support for content retrieval. If tags are informally defined and continually changing, then it becomes difficult to automate workflow and business processes. In this sense, it is necessary to find a compromise between the flexibility and dynamics of folksonomies and the rigid structure of controlled vocabularies. This compromise is usually known as *collabulary*[2], which corresponds to a portmanteau of the words *collaborative* and *vocabulary*.

2. Review of Previous and Related Work

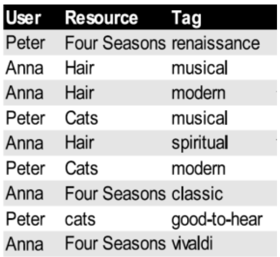
In [3] we deal with collabulary learning. Now we review this work shortly. In [3] we (i) defined the the problem of *collabulary learning* formally, (ii) we proposed a method for automatically enriching folksonomies with domain-expert knowledge, (iii) we proposed a new, fast and flexible algorithm based on efficient frequent itemsets mining techniques for collabulary learning and (iv) we introduced a new benchmark for task-based ontology evaluation in folksonomies. For (ii) the idea is to take a folksonomy and a domain-expert ontology as input and project them into an enriched folksonomy through semantic mapping. For (iii) we proposed the learning of a special taxonomy from the enriched folksonomy. This taxonomy represents both expert knowledge and user knowledge integrated. Thus it is called *collabulary*.

The aim of the collabulary is to enhance the ability of users for structuring and finding information. The obvious question one can ask is how and to which extent this *collabulary* really helps both users and experts. Looking at the literature on ontology learning from folksonomies (eg. [4,5,6,7,8]), we see that most of the proposed approaches are motivated by facilitating navigation and information finding, even though they do not quantify to which extent ontologies really help on this task. Instead, the quality of the learned ontologies is measured based on how good they match people's common sense or how similar they are to a reference ontology. We argue that in this context, an ontology is as good as it helps users finding useful information. Therefore, for the evaluation of ontologies and collabularies (note, that a collabulary is a special ontology) the idea is to plug the investigated knowledge structures in collaborative filtering algorithms for recommender systems and evaluate the outcome as an indicator of the ontologies' usefulness, since collaborative filtering [9] is one of the most successful and prominent approaches for personalized information finding. [3] was the first effort towards thorough empirical investigation of the trade-off between folksonomies and controlled vocabularies. We conducted experiments on a real-life dataset and demonstrated the effectiveness of our approach.

Given the novelty of the problem, there are still very few related works. In [10] e.g., the authors rely on external authority sources or on Semantic Web ontologies to make sense of tag semantics. Even though this can help finding more interesting relations than co-occurrence models, it can somewhat restrict the relation discovery, since if a relation is not defined in these external sources, it is assumed that the tags are not related, even if they frequently co-occur in the dataset. We instead, infer the relations directly from the data and thus are not dependant on external sources.

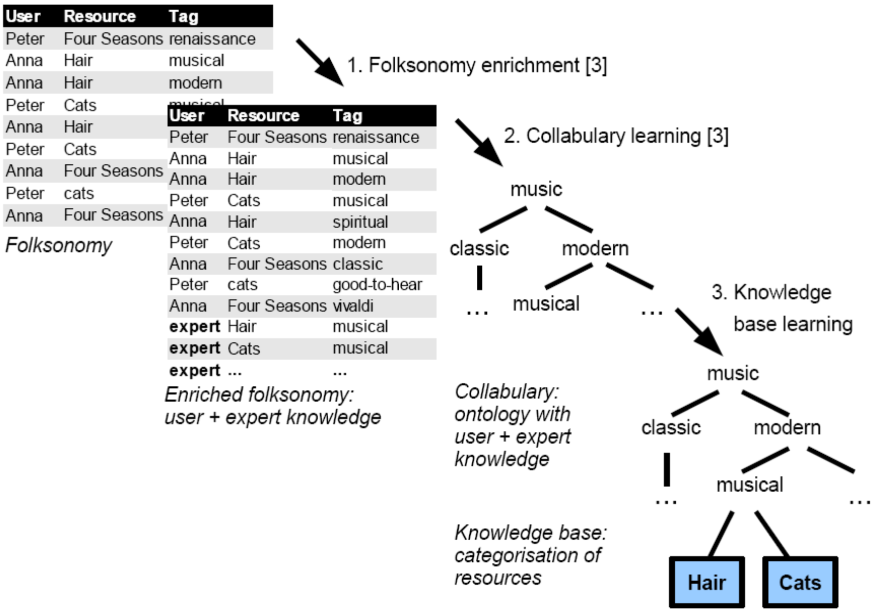
3. Learning Knowledge Base for a Collabulary

In our case a *folksonomy* is a set of user-resource-tag triples. Suppose (*u*,*r*,*t*) is one of these triples. It means that user *u* has labeled resource *r* with tag *t.* See Fig. 1 for an example.



User	Resource	Tag
Peter	Four Seasons	renaissance
Anna	Hair	musical
Anna	Hair	modern
Peter	Cats	musical
Anna	Hair	spiritual
Peter	Cats	modern
Anna	Four Seasons	classic
Peter	cats	good-to-hear
Anna	Four Seasons	vivaldi

Figure 1.
An example folksonomy with its user-resource-tag triple. The first triple means that user „Anna" has labeled the resource „Hair" with tag „modern", i.e. according to Anna's opinion Hair is modern.



User	Resource	Tag
Peter	Four Seasons	renaissance
Anna	Hair	musical
Anna	Hair	modern
Peter	Cats	musical
Anna	Hair	
Peter	Cats	
Anna	Four Seasons	
Peter	cats	
Anna	Four Seasons	

Folksonomy

User	Resource	Tag
Peter	Four Seasons	renaissance
Anna	Hair	musical
Anna	Hair	modern
Peter	Cats	musical
Anna	Hair	spiritual
Peter	Cats	modern
Anna	Four Seasons	classic
Peter	cats	good-to-hear
Anna	Four Seasons	vivaldi
expert	Hair	musical
expert	Cats	musical
expert	...	...

Enriched folksonomy: user + expert knowledge

1. Folksonomy enrichment [3]
2. Collabulary learning [3]
music
classic
modern
…
musical
…
3. Knowledge base learning
Collabulary: ontology with user + expert knowledge
music
classic
modern
…
musical
…
Knowledge base: categorisation of resources
Hair
Cats

Figure 2.
The process of learning collabularies with their knowledge bases

In [3] we proposed to enrich a folksonomy with domain-expert knowledge by: doing a semantic mapping between an expert ontology and a folksonomy, and including extra triples representing the expert's tag assignments to resources. A collabulary, in our case, is a taxonomic relation containing both user and domain-expert tags. The knowledge base of a collabulary is the mapping between resources and taxonomical concepts. Fig. 2. illustrates the process of learning collabularies with their knowledge bases. This process consists of 3 steps:

1. *folksonomy enrichment*: domain-expert knowledge will be integrated in the folksonomy by adding new triples,
2. *collabulary learning*: build a taxonomy of tags,
3. *knowledge base learning*: resources will be mapped to categories.

In our previous work [3], we paid attention to the first two steps, we applied a simple solution for the 3rd step. Now we study this step in more detail, we compare two approaches for knowledge base learning.

After the categorization in the 3rd step, a resource may belong to several concepts. This models our intuitions well. Suppose, for example, that two of the taxonomical concepts are "musical" and "good to hear". Then "Hair" may be categorized both as "musical" and as "good to hear" as well. Note, that non-expert users often use not fully exact tags like "good to hear", "coooooll!!!", "I like it" or "makes me happy"… Even though these are very subjective, the different tags are typically used by different groups of users. For example, one of such tags may be characteristic for fans of musicals, the other one for users, who like rock. Thus it makes sense to include such subjective tags in the taxonomy as well; moreover, as shown in our previous work [3], this improves the quality of taxonomy in a practical point of view.

The pseudocodes of the both approaches for knowledge base learning are depicted in Fig. 3. The simple approach maps resources to taxonomical concepts (tags) based on a correlation threshold. This threshold is denoted by *a*. For each resource *r* the most correlating taxonomical concept (tag) is selected first. This most correlating tag is denoted by *t*. Suppose *t* co-occurs *c*-times together with *r*. Then *r* will be mapped to *all* the taxonomical concepts (tags) which co-occur more than (*a*c)*-times together with *r*. In [3] we applied the simple approach with *a*=0.3

The adaptive approach clusters tags into two groups based on how frequently they co-occur with a resource. A resource is than mapped to all tags in the cluster of often co-occurring tags. In this work, we use 1-dimensional *k*-means clustering with *k*=2.

```
a) learn_knowledge_base_simple(Folksonomy f, real number a) {
     for all resources r in f {
       t = tag which co-occurs most often together with r
       c = count how often t and r co-occur
       map r to all tags which co-occur with r more than (a*c)-times
     }
   }

b) learn_knowledge_base_adaptive(Folksonomy f, real number a) {
     for all resources r in f {
       t = tag which co-occurs most often together with r
       c[] = counts how often each tag co-occurs with r
       perform 1-dim. k-means clustering with k=2 on the values in c[]
          // this leads to a clustering of tags
       map r to all tags which are in the cluster with higher mean
     }
   }
```

Figure 3.
The pseudocode of different approaches (simple and adaptive) for knowledge base learning

4. Experimental evaluation

To quantitatively evaluate a collabulary together with its knowledge base we use it in a recommender system and evaluate the output of the recommender system. This leads to an indirect evaluation of a collabulary.

Based on some pieces of information on which user likes (purchases) which resource (product, piece of music in our case), a recommender system recommends some new resources to the users. If one wants to evaluate a recommender system, the data set is usually split into two subsets: into a train and test set. The recommender system calculates the recommendations based on the training data set. Then these recommendations are compared to the test set: if the system recommended a resource *r'* for the user *u'* and the test set contains that user *u'* liked (purchased) resource *r'* than this is a hit. The more hits, the better the recommender system. In our case the recommender system was the same during all the experiments, and the learned collabulary was an input of this system. This was the only input of the recommender system we changed during the experiments. Thus the quality of the recommendation shows the quality of the learned collabulary. Note, that a recommender system is an application which exploits the collabulary, in fact any other applications that exploits a collabulary (an ontology) is suitable for such an indirect evaluation.

As the folksonomy representative we have chosen *Last.fm* (http://last.fm), a social tagging system that provides personalized radio stations where users can tag artists and tracks they listen to. Representing the domain-expert we have chosen the *Open Music Project* (http://musicmoz.org), which is based on the *Open Directory* (http://www.dmoz.org) philosophy and aims to be a comprehensive database about music. We extracted the *style* hierarchy representing a taxonomy of music genres from *musicmoz* to constitute the core domain-expert ontology. (This hierarchy contains cross references which were disregarded in order to guarantee the tree structure.) Since we consider the aforementioned databases to be defined over the same set of instances, we eliminated all the resources that are not present in both Last.fm and Open Music database.

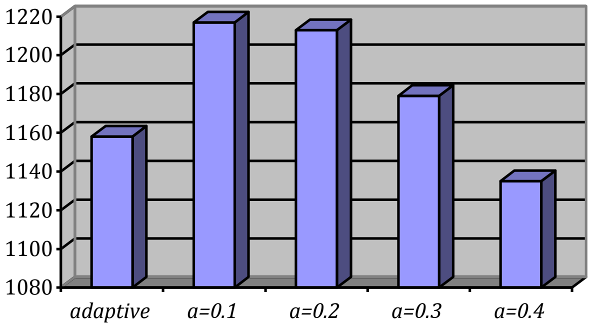


Figure 4.
Count of hits for the adaptive approach and different settings of parameter *a*

For the evaluation of collabularies together with their knowledge bases we use the benchmark introduced in [3], we use the same data. Now, in contrast to [3] we do not perform 5-fold-crossvalidation, because we observed previously, that the standard deviation in the count of correctly recommended items is very small w.r.t. the different folds. For more details on experimental settings we refer to [3]. Fig. 4. summarizes the experimental results. It shows the count of hits (the count of “good” recommendations) for the adaptive approach and for different settings of the parameter *a*.

5. Discussion and Conclusion

Our experiments show two interesting phenomena. First, the clustering-based adaptive approach (shown in Fig. 3.) does not necessarily outperform the simple one, if the parameter *a* is right chosen. Second, lowering the value of parameter *a* from 0.4 to 0.1 leads to more hits in an asymptotical fashion.

Both phenomena need further investigation. Concerning the first observation, it would be worth trying other clustering algorithms as well. Concerning the second phenomena, we plan to investigate the quality of the recommendation in a more fine-grained fashion for the case of $a < 0.1$. We are especially interested, which value of *a* maximizes the quality of recommendation.

References

[1] Berners-Lee T.; Hendler J.; Lassila O.: *The Semantic Web*. Scientific American, 2001;

[2] Folksonomy. Wikipedia Article. http://en.wikipedia.org/wiki/Folksonomy, accessed:May 2008;

[3] Marinho L. B.; Buza K.; Schmidt-Thieme L.: *Folksonomy-based collabulary learning*. “The Semantic Web - ISWC 2008, 7th International Semantic Web Conference” Springer Verlag, LNCS 5318, 2008;

[4] Zhou M.; Bao S., Wu X., Yu Y.: *An Unsupervised Model for Exploring Hierarchical Semantics from Social Annotations*. Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), Busan, South Korea, LNCS 4825, pp. 673 – 686, Springer Verlag, 2007;

[5] Heymann P.; Garcia-Molina H.: *Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems*. Stanford University, Stanford InfoLab Technical Report, 2006;

[6] Brooks C. H., Montanez N.: *Improved annotation of the blogosphere via autotagging and hierarchical clustering*. WWW '06: Proceedings of the 15th international conference on World Wide Web, Edinburgh, Scotland, ISBN 1-59593-323-9 pp.625 – 632, ACM, 2006;

[7] Schmitz P.: *Inducing Ontology from Flickr Tags*. Proceeding of the Workshop on Collaborative Tagging at WWW2006, Edinburgh, Scotland, 2006;

[8] Schmitz C.; Hotho A.; Jäschke R.; Stumme G.: *Mining Association Rules in Folksonomies*. Data Science and Classification: Proceedings of the 10th IFCS Conference, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 261 – 270, Springer, Berlin, Heidelberg, Germany, 2006;

[9] Resnick P., Iacovou N., Suchak M., Bergstorm P., Riedl J.: *GroupLens: An Open Architecture for Collaborative Filtering of Netnews*. Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, Chapel Hill, North Carolina, pp. 175 – 186, ACM, 1994;

[10] Specia L., Motta E.: *Integrating Folksonomies with the Semantic Web*. Proceedings of the European Semantic Web Conference (ESWC2007), volume 4519 of LNCS, pp. 624 – 639, Springer, Berlin, Heidelberg, Germany, 2007.