Computational and Crowdsourcing Methods for Extracting Ontological Structure from Folksonomy

Huairen Lin and Joseph Davis

Knowledge Discovery and Management Research Group,
School of IT, The University of Sydney, NSW 2006, Australia
{lin,jdavis}@it.usyd.edu.au

Abstract. This paper investigates the unification of folksonomies and ontologies in such a way that the resulting structures can better support exploration and search on the World Wide Web. First, an integrated computational method is employed to extract the ontological structures from folksonomies. It exploits the power of low support association rule mining supplemented by an upper ontology such as WordNet. Promising results have been obtained from experiments using tag datasets from Flickr and Citeulike. Next, a crowdsourcing method is introduced to channel online users' search efforts to help evolve the extracted ontology.

1 Introduction

Social tagging systems, such as Flickr[1] , Delicious[2], have recently emerged as some of the rapidly growing web 2.0 applications. The product of this kind of informal social classification structure, also known as folksonomy, has provided a convenient way that allows online users to collectively annotate and categorize large number of distributed resources from their own perspectives. However, as the amount of resources annotated using folksonomy has increased significantly, exploration and retrieval of the annotated resources poses challenges due to its flat and non-hierarchical structure with unsupervised vocabularies.

At the same time, the development of semantic web is creating a cyberspace that contains resources with relations among each other and well-defined machine-readable meaning. In this vision, ontology is the enabling technology for most of the semantic applications, such as semantic search. However, there are significant challenges to be overcome before we can build tools for sophisticated semantic search. It is not easy to establish a single and unified ontology as a semantic backbone for a large number of distributed web resources, and manual annotation of resources requires skilled professionals or ontology engineers [1]. Furthermore, ontology needs to be constantly maintained to adapt the knowledge emerging from daily work of users [2].

[1] flickr.com Flickr is an online photo management and sharing application.

[2] delicious.com Delicious is a social bookmarking service.

L. Aroyo et al. (Eds.): ESWC 2010, Part II, LNCS 6089, pp. 472–477, 2010.

The goals of our research is to extract ontological structure from folksonomy and to facilitate its automatic evolution with changing usage patterns, in such a way that the resulting structure can better support semantics-based browsing and searching of online resources. With the unification of the seemingly exclusive features of folksonomy and ontology, they can complement each other by providing full advantage of colloquial terms from folksonomy and semantic relations from ontology. We can exploit the semantic relation in the ontological structure to satisfy users' query or navigation requests using terms that they are familiar with in order to access millions of annotated resources, and translating and integrating the resources from different sources.

2 State of the Art

Computational Approaches. There are several promising techniques for extracting knowledge from the existing resources, such as hierarchical clustering [3], statistical model [4], and association rules mining [5]. Most of hierarchical clustering algorithms are based on bottom-up methods. First it computes pairwise tag similarities, and then merges most similar tags into groups. After that, pairs of groups are merged as one until all tags are in the same group [3]. Association rule mining has also been adopted to analyse and structure folksonomies. The output of association rule mining on a folksonomy dataset are association rules like $A \rightarrow B$, which implies that users assigning the tag A to some resources often tend to also assign the tag B to them [5].

To further discover the relationships within tags in clusters, several existing upper ontology resources can be used as references, such as WordNet and other semantic web resources. Ontology mapping and matching techniques are commonly applied to identify relationships between individual tags, between tags and lexical resources, and between tags and elements in an existing ontology. For example, by mapping "apple and fruit" in a food ontology, we can find the relation that "apple" is a subclass of "fruit" [6,7].

Crowdsourcing Human Computation. While the most sophisticated computational techniques cannot substitute the participation of knowledge engineers, the recently proposed crowdsourcing method provides new ways to have users engage in ontology engineering and to aggregate their deep knowledge through a mass collaboration technique [8][9]. A computer program that can attract human's interest, fulfill their needs, and collect, interpret human's solution is important. Ontogame [10] proposed a game for ontology building. One of the game scenarios is to asks users to check the structure and abstraction from random wiki pages. Recently, experiments show that online service such as freeware, or a successful login procedure [11] can also be used to motivate public users to participate in a specific task. With a purpose-designed system, we will be also able to embed the task of building and maintaining ontologies into users' everyday work process and create the conditions for the ontology to continuously evolve without the help of knowledge engineers [2]. In [12], a semantically enriched bookmarks navigation system provides functionality that enables users to reject or accept the more general/narrow tags. These inputs were recorded for further ontology maintenance.

In summary, although several computational approaches have been proposed to bring structure to folksonomies, they do not come without limitations. These include the inability to decide the super/sub class relations of terms generated by association rule mining. Such problem can be partially solved by introducing an upper ontology such as WordNet. However, such an approach can only deal with the standard terms and has no effect on terms that do not appear in the upper ontology. Moreover, several attempts have shown that crowdsourcing human computation is promising method to bring non-experts together to tackle some difficult problems. These include problems like ontology refinement and evolution which normally need domain experts' participation.

3 Methodology

In this paper, our research concerns following specific research problems: (1) How to extract shared vocabularies from large folksonomy datasets? (2) How to find the semantic relations for these shared vocabularies? (3) How to handle the non-standard tags in the folksonomies? For instance, terms like 'folksonomy', 'ESWC' that cannot be found in traditional dictionary. (4) How can the resulting ontological structure be automatically evolved with the constant change of domain knowledge and patterns of usage?

We first propose an integrated computational approach to extract ontological structures. Our approach combines the knowledge extracted from folksonomies using data mining techniques with the relevant terms from an existing upper-level ontology. Specifically, low support association rule mining is used to analyze a large subset of a folksonomy. Knowledge is expressed in the form of new relationships and domain vocabularies. We further divided the tag word-formation into standard tag, compound tag and jargon tag and handle them respectively. Standard tags in the vocabulary are mapped to WordNet to get semantic relations. Jargon tags and user defined compounds are then incorporated into the hierarchy based on domain knowledge extracted from folksonomy. Thus, the hidden semantic knowledge embedded in the folksonomies is transformed into formalized knowledge in the form of ontological structures.

A semantic search assist is designed based on crowdsourcing model to update the extracted ontology for evolving folksnomy while it suggests helpful search terms and semantic relationships to help refine users' search. First, we elicit the inputs from users by providing terms semantically related with their query keywords and candidate semantic relationships such as 'is-a' after user conducts a normal search. With this assist, user can make explicit the semantic concept of what s/he is looking for by simply selecting the related term provided by ontology and assigning the relationship between the query keyword and related term. The semantic search engine will then return better result with a reasoning technology based on the disambiguated query. For example, by appointing 'apple' as 'is-a' kind of 'computer', the system will expand the results to more specific class such as 'Mac' or a individual model such as 'MacBook Air' and remove results belong to 'fruit'. We then collect and aggregate these terms and relationships from different search sessions. Every user-assigned relationship is recorded even it is a disagreement with existing knowledge. The long-term records will be split into several clusters to reflect knowledge from different domains. We assume that the user specified semantic relationship is correct based on some aggregation mechanisms such as the rule of majority. After that, we introduce a mechanism to periodically merge changes with old version of ontology and release improved version. In short, we show how users' search intent can be captured to help to evolve the ontology while helping to improve their desired search results.

We attempt to engage web users in evaluation tasks using a crowdsourcing medium such as Amazon Mechanical Turk (MTurk)[3]. Based on this service, we ask users to manually evaluate the collected term pairs and give monetary award to every complete task. We will also attempt to verify the quality of the extracted ontology against other manually built gold standard ontology. The measurement should reflect how well the extracted terms cover the target domain and the accuracy of the relationships among the terms, especially for standard terms. Furthermore, we take the task-based evaluation approach to measure how far the extracted ontological structure will help to influence and improve the search result. We investigate four potential application scenarios of the extracted ontological structures: multi-dimensional views, cataloguing and indexing, query expansion and tagging suggestion. We will use those widely used measures such as precision, recall, and F-measure to assess the quality and see how ontology would improve the search result.

4 Preliminary Results and Future Work

We have implemented a prototype system for the described computational extraction strategy. The implementation and evaluation are reported in [13]. Through the investigation into four kinds of word-formations (standard tags, jargon tags, compound tags, and nonsense tags) in folksonomies, our approach has produced promising initial results using two datasets from Flickr and Citeulike.

To explore potential application scenarios with the resulting ontological structure, we are building a semantic photo organizing system, SmartFolks, based on a subset of Flickr image collection. With the extracted ontology as background knowledge and Jena[4] tool kit as semantic web framework, the system enables user to find the resources through the navigation of ontological structure or to get better results with the technology of semantic web such as ontology based query expansion. This demo site is available at http://smartFolks.thetag.org

Our future work is focusing on the ontology evolution using crowdsourcing method. A semantic search assist component will be integrated in the SmartFolks system to channel users' search efforts for ontology evolution.

[3] mturk.com MTurk is a web based service that enable developers outsource certain task to human across the world. The unit work typically costs only few cents.

[4] jena.sourceforge.net/ Jena is a framework for building Semantic Web applications.

5 Conclusion

Recent research indicates that there are significant challenges in the area of ontological structure extraction from collaborative tagging systems. The integrated framework proposed in this thesis allows a systematic approach to this emerging area. We have not only identified computational methods for ontology extraction, but also presented a proposal for crowd-sourcing model which is capable of aggregating the human intelligence without the need for the involvement of ontology experts. A semantic search engine is recommended as the medium for this integration. The application of this conceptual framework might assist folksonomy based systems to improve the query performance and enhance the organization of resources. It is hoped that the crowdsourcing approach will complement the computational methods to help create robust ontological resources that can advance the state-of-the-art with respect to semantic search.

References

1. Wu, X., Zhang, L., Yu, Y.: Exploring social annotations for the semantic web. In: Proceeding of the 15th WWW Conference, Edinburgh, Scotland, pp. 417–426. ACM, New York (2006)
2. Braun, S., Schmidt, A., Walter, A.: Ontology maturing: a collaborative web 2.0 approach to ontology engineering, Banff, Canda (2007)
3. Wu, H., Zubair, M., Maly, K.: Harvesting social knowledge from folksonomies. In: The 7th Conference on Hypertext and Hypermedia, Odense, Denmark (2006)
4. Heymann, P., Garcia-Molina, H.: Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, InfoLab, Stanford (2006)
5. Schmitz, C., Hotho, A., Jaschke, R., Stumme, G.: Mining association rules in folksonomies. In: The 10th IFCS Conference, Studies in Classification, Data Analysis, and Knowledge Organization (2006)
6. Angeletou, S., Sabou, M., Motta, E.: Semantically enriching folksonomies with FLOR. In: CISWeb, p. 65 (2008)
7. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic grounding of tag relatedness in social bookmarking systems. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 615–631. Springer, Heidelberg (2008)
8. Brabham, D.C.: Crowdsourcing as a model for problem solving: An introduction and cases. Convergence 14(1), 75 (2008)
9. Niepert, M., Buckner, C., Allen, C.: Working the crowd: Design principles and early lessons from the Social-Semantic web (2009)
10. Siorpaes, K., Hepp, M.: Ontogame: Towards overcoming the incentive bottleneck in ontology building. In: 3rd International IFIP Workshop (2007)
11. von Ahn, L., Maurer, B., McMillen, C., Abraham, D., Blum, M.: reCAPTCHA: Human-Based character recognition via web security measures. Science 321(5895), 1465 (2008)

12. Limpens, F., Gandon, F., Buffa, M.: Collaborative semantic structuring of folksonomies. In: 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, pp. 132–135. IEEE Computer Society, Los Alamitos (2009)
13. Lin, H., Davis, J., Zhou, Y.: An integrated approach to extracting ontological structures from folksonomies. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E. (eds.) ESWC 2009. LNCS, vol. 5554, pp. 654–668. Springer, Heidelberg (2009)