Integration of Computational and Crowd-Sourcing Methods for Ontology Extraction

Huairen Lin, Joseph Davis, Ying Zhou

School of Information Technologies, The University of Sydney
Sydney, New South Wales, Australia
{lin,jdavis,zhouy}@it.usyd.edu.au

Abstract— Ontological approach is one of those promising solutions to improve search precision, resource navigation and retrieval by providing shared vocabularies and semantic relations for translating and integrating the different sources. But the progress has slowed down due to the knowledge acquisition bottleneck. In this paper, we present an integrated conceptual framework for extracting ontological structures from folksonomies that exploits the combined power of traditional computation and pooled, human knowledge through crowd-sourcing.

I. INTRODUCTION

Ontology is an important field that has considerable potential to improve the organization, retrieval, and management of information with the shared understanding of specific domains [1]. However, the development of ontology-driven applications has slowed down due to the knowledge acquisition bottleneck [2]. The processes of ontology building, alignment and merging are usually handled manually and often needs the involvement of domain experts and ontology engineering professionals [3].

Recently, folksonomies have emerged as an informal social classification mechanism. While ontological approach tends to build a top-down organization using controlled terms pre-defined by domain experts, folksonomy permits a community of users to annotate and classify resources with freely chosen tags based on their own terminology/language and aggregate the tags into a bottom-up organization. With the use of folksonomy, collaborative tagging system (CTS) has emerged as a popular approach for organizing and sharing web resources and has been integrated into most web 2.0 applications. However, as the amount of resources annotated using CTS has increased exponentially, exploration and retrieval of the annotated resources pose major challenges. The key problem with folksonomies is that the concept and internal structure are not explicit to the machine or to other systems even though the tags may be meaningful and coherent to the user who assigned them [4].

Considering the pros and cons of ontologies and folksonomies, we contend that significant benefit can be gained by integrating these two approaches. Using folksonomy as the resource for building and extracting knowledge from it, we attempt to create an ontology that reflects the terminology of the users. The integrated approach will preserve strengths of folksonomy and ontology, that is: terms are drawn from the user’s usage that they are familiar with the possibility of high precision and accuracy in search.

We argue that, innovative solutions that are socio-technical in nature are needed to tackle these tough problems in ontology engineering. Recent development in Web 2.0 enables us to harness collective human intelligence among a population in ways traditional face-to-face meeting cannot [5].

While the most sophisticated computational techniques cannot substitute the participation of domain experts, the recently proposed crowd-sourcing method provides new ways to solve some of the awkward problems in ontology refinement and evolution. The crowd-sourcing method is able to aggregate the intelligence of large number of online users though a mass collaboration technique. It harnesses the collective intelligence of a vast number of individuals to offer solutions to the problem, and the winning ideas are typically awarded some form of a reward [5].

In this paper, we develop and demonstrate a systematic approach to achieving an integrated working model of computational and crowd-sourcing methods.

In the following section, related research is reviewed. Then, the general idea and the conceptual framework are explained followed by our conclusions.

II. LITERATURE REVIEW

A. Computational Approaches

Ontology generation has been considered as a hierarchical clustering problem by many researchers. Most of hierarchical clustering algorithms are based on bottom-up methods. First it computes pair-wise 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 [6]. On the other hand, top-down methods starts from the top/root level and then move to its subclass level tags [7].

Statistical model is used for finding concept hierarchy which usually arranges terms hierarchically using a subsumption relation. [8]. In order to induce ontology from Flickr, Schmitz added additional filters to the subsumption model, such as a threshold of the number of authors using a tag, to control for idiosyncratic vocabulary and selected candidate term pairs. The experimental results show that the model can generally reflect distinct facets [9].

Data mining techniques such as association rule mining have also been adopted to analyse and structure
folksonomies. The output of association rule mining on a folksonomy data set are association rules like \( A \to B \), which shows that users assigning the tag \( A \) to some resources often tend to also assign the tag \( B \) to them [10]. Association rule based approach has been extended in [11] to mine structural features of taxonomies by pruning the less important relations between tags.

By representing folksonomy as a tripartite network of users, tags and objects, semantics such as relation between broader/narrower tags has been unveiled through a process of graph transformation using social network analysis [12].

To further discover the relationships within tags in clusters, several existing ontology resources can be used as references, such as WordNet (despite its limitations) and other semantic web resources. Ontology mapping and matching techniques are commonly applied to identify relationships between 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” [13] [14].

B. Crowd Sourcing Human Computation

Crowd-sourcing model describes a process of organizing a large number of online users and tapping their knowledge/intelligence to complete assigned tasks. The contributing users could be rewarded for their efforts. The central aspect of crowd-sourcing model is human computation. Human computation method has proved its strength in distributed problem solving and it often performs better than traditional approaches due to the typically human capabilities such as acute visual perception and aesthetic judgment[15].

Crowd-sourcing human computation systems are typically web-based. Various kinds of web applications, such as community, wiki, search engine, collaborative tagging systems, and online games can realize and aggregate the crowd wisdom and organize a mass of users into productive workers [5]. Ontology maturing [16] introduces a process allows the emergence of ideas from each individual and consolidation in communities for a common terminology which is also expected to overcome the problem of the time lag between emergence of topics and their inclusion in an ontology. Wiki technique based web ontology editor are discussed to overcome the current difficulties in ontology construction and facilitating collaborative ontology editing. With wiki-based ontology editor, we are able to provide collaborative editing functions such as versioning, user roles and ranks, mapping to discover similarities, supporting for community consensus, voting, and ontology visualization such as tag clouds [17, 18].

Although humans have an innate ability to gather and analyse data, they tend to make decisions motivated by self-preservation and on their selfish needs [19]. A computer program that can attract human’s interest, fulfil their needs, and collect, interpret human’s solution is also important.

The concept of “game with a purpose” was proposed by Luis von Ahn to enable humans to solve problems within a gaming context. Applications designed under this consideration are expected to collect energy from people around the world when they spend billions of hours playing computer games, such as Peekapoon [20], a game asking characters to locate objects in images. The idea of “games with a purpose” by von Ahn has been applied to ontology-related problems. Ontogame [21] proposed a game for ontology building that asks users to check the structure and abstraction from random wiki pages.

Playing games can be a good incentive for some people, but not for the majority of the online users. Focus has shifted to free services such as free email, downloading a useful resource, and monetary rewards [22], or a successful login procedure [23], which are more commonly employed with online users. reCAPTCHA [23] improves the process of digitizing old printed material by challenging users to decipher scanned words from books that could not be read by OCR software. These segmented words are presented as a part of CAPTCHA test. Users are required to transcript the words appears at the bottom of Web registration forms as a test solution to complete a login procedure. This text transcription method has achieved high word accuracy almost at the level of professional human transcribers, and illustrated it as a fast transcription technique.

A review of the research literature reveals that: even candidate concepts and relationships may be generated by ontology learning toolsets, human labour is still needed to verify the suggestions and complete the ontology; On the other hand, several attempts have shown that crowd-sourcing human computation is promising method to bring non-experts together to tackle some tough problems.

III. A CONCEPTUAL FRAMEWORK: INTEGRATION OF COMPUTATIONAL AND CROWD-SOURCING APPROACHES

Based on a review of the literature, this section presents a conceptual framework of the procedure to extract ontological structures from folksonomies. It systematizes and integrates the two methods discussed so far: (1) utilize the computational method for deriving ontological structures (2) utilize crowd-sourcing method to facilitate collaborative ontology refinement and evolvement.

The proposed framework is divided into three phases as illustrated in Fig.1, First an integrated bottom-up and top-down extraction approach is introduced to derive draft ontological structures from folksonomies; then a semantic search engine based on the candidate ontology is introduced to improve and involve the ontology by analysing users’ search behaviours. It not only provides a semantic search function, but also allows users to refine the search result by modifying the related ontology; finally, a wiki-based community is setup to facilitate collaborative ontology construction and refinement.
A. The Use of Computational Intelligence

The first extraction stage combines the knowledge extracted from folksonomies using data mining techniques with the relevant terms from an existing upper-level ontology such as WordNet. 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. 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 knowledge embedded in the folksonomies is transformed into formalized knowledge in the form of ontological structures.

Rather than the clustering technique, association rule mining is an unsupervised data mining method to find interesting association between data sets. In this stage, we applied the association rules to find semantically related tags, which is the basis for further ontology building. Furthermore, we simplified the a priori [24] algorithm to find 2-item set rules and introduced a new cosine coefficient, which significantly improved the efficiency in low support mining. We only calculate the relationship between tag pairs; both antecedent and consequent can only have one tag; additional cosine similarity threshold is set to offset the noise caused by low support and to compare the relevance between tags.

We use WordNet as the upper ontology and compute each semantic relation between tags in terms of hyponym relation from WordNet. A term that is more generic or more abstract than a given term is considered as a hyponym. Though WordNet as an upper ontology resource contains a sufficiently wide range of common words, it does not cover special domain vocabulary and cannot reflect usage change. In CTS, many of the tags are in the form of jargon and compound terms. Mapping terms with WordNet ontology is obviously not enough to find the relationships among them.

Thus, additional consideration was given to incorporate there terms into ontological structures by matching tags using association rule mining and token-based similarity.

We have implemented a prototype system for the described extraction strategy. The implementation and evaluation are reported in [25].

B. Crowd-sourcing Based on The Medium of The Web

In order to outsource the ontology refinement and evolution tasks to non-experts or online users, a semantic search engine and a wiki-based environment are designed to collect simple ideas, complex or theoretical opinions, and to facilitating collaborative ontology editing. Both the semantic search engine and wiki based environment are backed by the ontology extracted using computational techniques. However, they provide different mechanisms to collect user inputs.

The ontology evolving function integrated in search page only provides some simple editing features for easily updating of terms and relations. The wiki-based collaborative construction/refinement environment is setup to facilitate people who would like to further contribute for advanced editing. A link is provided at the search page for entering into the advanced editing functions showing information relevant to the keyword.

The proposed search engine will be a mashup of several resources and techniques, including Yahoo! BOSS\(^1\) search platform, Yahoo key term extractor\(^2\), ontology library built with our computational approach, ontology query language like SPARQL\(^3\), and an ontology reasoner such as Pellet\(^4\). It not only provides a semantic search function using ontology from users, but also allows users to refine the search result by modifying the related ontology.

Since the search engine is designed for easy and quick ontology evolving, all the adjustment is made in one page. Therefore, it does not provide advanced interface for editing of ontology. A wiki-based ontology refinement environment is setup to allow users to collaborate much more frequently, transparently and directly. Users can share their intelligence to improve the results from machine based approaches. In this step, users are introduced to reconstruct, approve, or merge the faceted ontology/partial structure.

The proposed wiki-based ontology construction environment aims to build a community providing support

---

1 http://developer.yahoo.com/search/boss/
2 http://developer.yahoo.com/search/content/v1/termExtraction.html
3 http://www.w3.org/TR/rdf-sparql-query/
4 http://clarkparsia.com/pellet
for collaborative knowledge engineering. It will not only simply integrate an existing Wiki system, but also provide functions for knowledge engineering, such as ontology review, visualization, alignment, merging, and community support that enables discussion and voting about the changes. It will be constructed based on OntoWiki [26], an open-source wiki-based ontology editor. The versioning control will keep the major change introduced by the users and thus do not need experts to evaluate the results created by other users. Voting and comments are other frames to control the accuracy.

The updating-while-search feature at the search page and the advanced editing feature at the wiki-based community provide two ways to enable a large number of users to collaborate in ontology evolution and refinement. Note that these are not isolated systems.

IV. CONCLUSION

Recent research indicates that the problem of ontological structure extraction and collaborative tagging systems constitutes a critical challenge with much potential for future research. The integrated framework proposed in this paper allows a systematic approach to this emerging area. We have not only identified computational methods for ontology extraction, but also introduced the proposal for crowdsourcing model which is capable of aggregating the human intelligence while reducing the costs and need for ontology experts. Search engines and wikis are recommended as the medium for this integration.

The application of this conceptual framework might assist collaborative tagging system to improve the query performance and enhance the organization of resources. Indeed, the framework presented here is also a useful guide for ontology construction in any domain. It also presents an illustration of how to leverage mass collaboration with Web 2.0 technologies.

REFERENCES