An Integrated Approach to Extracting Ontological Structures from Folksonomies

Huairen Lin, Joseph Davis, and Ying Zhou

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

Abstract. Collaborative tagging systems have recently emerged as one of the rapidly growing web 2.0 applications. The informal social classification structure in these systems, also known as folksonomy, provides a convenient way to annotate resources by allowing users to use any keyword or tag that they find relevant. In turn, the flat and non-hierarchical structure with unsupervised vocabularies leads to low search precision and poor resource navigation and retrieval. This drawback has created the need for ontological structures which provide shared vocabularies and semantic relations for translating and integrating the different sources. In this paper, we propose an integrated approach for extracting ontological structure from folksonomies that exploits the power of low support association rule mining supplemented by an upper ontology such as WordNet.

Keywords: Ontological structure, Folksonomy, Collaborative tagging system.

1 Introduction

Organization of digital resources has been a major challenge on the World Wide Web. Recently, collaborative tagging systems (CTS) have emerged as a mechanism for users to organize and share such resources. CTS allow users to use any keywords or tags relevant to the content to annotate their favorite resources on the web. For example, Flickr[1], an online photo management and sharing application launched in Feb. 2004, counts more than 40 million monthly visitors and is ranked as one of the world's top 30 websites. As of November 2007, 2 billion photos are stored on the site [1]. Another example is Citeulike[2], a free online bibliography manager allowing users to gather, organize, and share scholarly papers. It has become popular especially among researchers and other academic users.

This kind of informal social classification in CTS where users use their own language or terminology to describe and classify the content was first recognized as folksonomy by Vander [2]. When a user tags an online resource, s/he is creating an informal taxonomy. These tags are aggregated to help find the information they represent. With bottom-up, user-driven and freely chosen vocabularies, folksonomies stand

[1] http://www.flickr.com
[2] http://www.citeulike.org

L. Aroyo et al. (Eds.): ESWC 2009, LNCS 5554, pp. 654–668, 2009.

in contrast to taxonomies which use controlled terms. The relationships among the terms are typically contributed by domain experts in order to classify resources.

As the amount of resources annotated using folksonomies has increased significantly, exploration and retrieval of the annotated resources pose challenges. The major problem with folksonomies is that the tags used to describe the content can be idiosyncratic and not understood by many users. Most tags are chosen based on individual users' own experience and linguistic styles and preferences. Furthermore, 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 created them [3]. Folksonomies tend to include all kinds of tags ranging from standard dictionary words and compound expressions created by individual users such as "evolutionary- genomics" to jargon and nonsense words.

Various solutions have been proposed to improve the quality of queries based on folksonomy. One stream of research has attempted to refine the query result using meaningful knowledge derived from the folksonomy itself. Clustering and tag clouds are widely used approaches. Clustering techniques group the search result into several subsets and recommend related resources based on selected tags. However, clustering techniques rely heavily on statistical association or co-occurrence of tags. The effectiveness of this approach can be limited as the relations derived are not based on meaning. Tag cloud is a somewhat rough approach to organizing tags. It shows a subset of frequently used tags in sizes relative to their frequencies [4]. It is easy for the user to see the "hot" keywords. However, tag cloud normally contains very general terms such as "computer" or "picture" and do not show any semantic relation between the tags. Another stream of research takes an existing upper ontology as the base structure and uses it to facilitate organizing query results. Although the need for relevant ontological structures to support CTS systems is well understood, the upper ontology may not be well matched with the tags in the folksonomy[5]. For example, WordNet is widely used as an upper ontology because it describes very general concepts across all domains. However, methods heavily dependent on WordNet frequently get poor results for accuracy due to the fact that many of the tags from collaborative tagging system do not exist in WordNet.

In this paper, we propose an integrated approach to extract ontological structures from folksonomies. 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. 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.

In particular, this paper answers the following questions:

- How to extract shared vocabularies from large data sets?
- How to find the semantic relations for these shared vocabularies?
- How to handle the non-standard tags in the folksonomies?

The rest of the paper is organized as follows: In section 2, related research is reviewed. In section 3, we present our integrated approach and discuss the detailed steps. Then in section 4, experimental results are presented and evaluated. Finally, we discuss our conclusion and future work.

2 Related Research

One of the main strengths of folksonomies and CTS is that they directly reflect the users' vocabularies, their choices in terminology, and subjective meaning [6]. Despite the fact that people assign one or more freely chosen tags to each of the resources and these tags are based on their own knowledge or professional background, there is a common basis of understanding of the tags used to communicate with each other [7]. Thus, a folksonomy has its uses and has the potential to be a weak knowledge base. Since users add new contextual dimension during collaborative tagging, most of the tags or keywords annotated by users tend to be more highly correlated than keywords automatically extracted by machine, such as Term Extraction Web Service from Yahoo[3]. Thus, folksonomies and CTS can be seen as potential sources of semantic information to support ontology evolution [8].

Approaches such as clustering and related tags do not make the hierarchical relations explicit between tags. As a result, it is difficult for a user to find related resources with broader or narrower tags during navigation which may better represent the user's current interests and help a user who has limited subjective knowledge. 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 [9]. Aiming to converting a large corpus of tags from folksonomy into a navigable hierarchical taxonomy of tags, an algorithm using graph centrality has been proposed. Cosine similarity between tags has been used to measure the distance from one tag to another and organize them into a hierarchical tree by starting with a single "root" node representing the top of the tree, and adding other tags to the tree in decreasing order of distance [10].

Association rule mining has also been adopted to analyze and structure folksonomies. Since folksonomies provide a three-dimensional dataset (user, tag, and resources), Schmitz proposed a conceptual level notation to reduce the three-dimensional folksonomy to a two-dimensional formal context and apply association rule mining. The output of association rule mining on a folksonomy data set are association rules like $A \rightarrow B$, which shows that users assigning the tag A to some resources often tend to also assign the tag B to them [11]. Association rule based approach has been extended in [12] to mine structural features of taxonomies by pruning the less important relations between tags.

Significant research progress in the field of semantic techniques has offered promising prospects for extracting semantic structures and relations from the folksonomies. 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

[3] http://developer.yahoo.com/search/content/V1/termExtraction.html

mapping "apple and fruit" in a food ontology, we can find the relation that "apple" is a subclass of "fruit" [13] [14]. WordNet has been successfully applied in many applications as a reliable upper ontology. An et al [15] presented an approach to automatically build a domain ontology by interweaving sub-taxonomies of WordNet with information extracted from deep web pages. In this research, concept and relations from WordNet have been used to bridge the concept gap and tie together ontology fragments into a single ontology. Laniado et al [16] illustrated an approach to integrating WordNet noun hierarchy in the *related tags* panel of del.icio.us (a collaborative tagging system which allows users to tag, manage and share Web pages). By mapping *related tags* to WordNet and getting the related terms, the tags and terms will be organized into a navigation tree according to a semantic criterion.

The ontological structure extracted from folksonomies can be useful in many areas of CTS, such as providing multi-dimensional views, cataloguing and indexing, query translation and tagging suggestion. It can be used to organize the search result into different dimensions like topic, date, location, etc. For each dimension, relevant resources can be organized in a hierarchical structure. Second, ontological structures provide an expressive way to catalog and index large digital resources. While a query needs a pre-specified keyword for information retrieval, the ontological structures give users a quick understanding of the subjective knowledge and let them directly browse for further information. Third, with query translation based on an ontological structure, we can enhance the precision and recall by matching the query keywords and the potential results at the level of semantics. For example, the keyword in the query can be replaced by their approximations and related instances will be returned to user. Finally, suggesting relevant ontological classes to the user will not only improve the tagging experience but increase classification quality [7, 17, 18].

In summary, folksonomies have their own shared vocabularies and relations, which can be extracted as an ontological structure and used to improve the exploration and retrieval of digital resources. Although several approaches have been proposed to bring structure to folksonomies, they do not come without limitations. These include the inability to decide the rules generated by association rule mining as to which term is more general or narrow, and tags that cannot be found in the upper ontologies.

3 System Architecture

In folksonomies, natural language has been used to annotate resources and to recall resources. As the result of non-controlled human language, the vocabularies used in folksonomy are shaped into following four word-formations:

- Standard tags, which can be found in traditional dictionaries, e.g. "genomics".
- Compound tags, a non-standard expression, part of them can be found in dictionary, e.g. "evolutionary-genomics".
- Jargon tags, another non-standard expression frequently used to quickly express user's ideas, e.g. "scientometrics", "folksonomy", "CSCW".
- Other nonsense tags, such as misspelling tags.

In this section, we present our integrated bottom-up and top-down architecture that aims to extract ontological structures from folksonomies based on the four word-formations. A visual representation of the entire extraction architecture has been presented in Fig. 1. The system proceeds as follows:

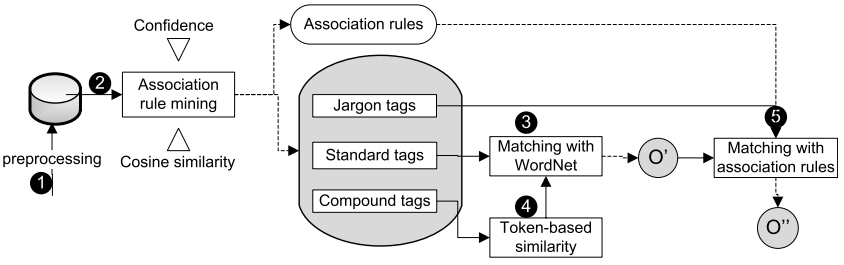

Fig. 1. The extraction process

1) In the data pre-processing phase, resources with only one tag or tagged by languages other than English are excluded. However, we should be very careful in this step not to delete jargon and compound tags. Thus methods like traditional dictionary filtering are not appropriate in our approach.
2) Low support association rule mining is applied to generate association rules representing relations between correlated tags. In brief, it has three sub tasks:
   - Discovering shared vocabularies or essential tags, where a tag should have certain relation with other tags, which is the basis for ontological structure.
   - Extracting the association rules between jargon and standard tags. Those association rules are treated as ontology matchers to incorporate jargon into ontological structure.
   - Retrieving associated terms and excluding non-relevant terms during the mapping and matching with WordNet.
3) WordNet is implemented as an upper ontology providing the semantic is-a relation, which is called as hypernyms in WordNet. After the standard tags have been connected to each other via semantic relations from WordNet, they were organized into a hierarchical structure.
4) A series of similarity filters are employed to interpret the compound tags before matching them with WordNet.
5) Jargon tags are incorporated into the previously built ontological structure by matching tags using association rules and similarity coefficient.

In following subsections, we discuss in detail each of the steps.

3.1 Low Support Association Rules Mining

Association rule mining technique is applied to the dataset to discover possible pair wise associations between tags. A priori association rule mining algorithm has been proposed to solve the "supermarket basket" problem and to discover interesting relations between items. If 90% of the transactions that include *butter* and *bread* also include *milk,* the relation is showed *as* $\{butter, bread\} \rightarrow milk$ with a confidence value of 0.9 [19, 20]. Such analysis is based on the past transaction data consisting of a

set of transaction D = ($d_1, d_2 \ldots d_k$) and a set of items, I = ($i_1, i_2 \ldots i_k$). In our approach, given a dataset from CTS, where every resource was annotated with a set of tags by several online users, the resources set corresponds to transactions D and the set of tags correspond to items I.

The aim of association rule mining in CTS is to generate associations in the form $t_a \rightarrow t_c$ between tags t_a and t_c that have support and confidence above certain thresholds, called minimum support and minimum confidence. Support of a rule is simply computed as the percent of the resources containing the tag pair. Confidence is computed as the ratio of the number of resources containing both t_a and t_c and the number of resources containing t_a only. While the confidence threshold reflects the strength of the rule, support threshold measures the coverage.

As folksonomy is collectively built by various users, the tags in folksonomies usually follow a Zipf distribution. Except for a few general tags, majority of the tags do not occur very frequently in the dataset. Traditional association rule mining algorithms normally set a relatively high support and confidence threshold to find common and strong rules. However, this is not the case for folksonomies. Setting a relatively high support threshold is likely to miss important associations among tags in the long tail of the Zipf distribution[21]. Hence we adopt a very low support threshold to include tags that do not occur very frequently in our analysis. Lower support may inadvertently bring lots of noise in the rule set. To offset this effect, we introduce cosine similarity[3, 22] to filter out possible noise.

To apply this measure, we first convert datasets from folksonomies into a metric space V. Given a pair of tag (x, y), tag x is expressed as a vector $\vec{x}$ in this space, where each dimension corresponds to a resource and value indicating whether or not a tag appears in a resource [23]. This tag-resources model can be converted into a 0/1 matrix because the times of a tag appears in a resource should be 0 or 1.

Table 1 shows such 0/1 matrix for tag(x, y), where each column represents a resource and each row represents a tag x or y, intersection (row, column) = 1 if a specific tag appear in the resource. If not, the value is 0. The traditional cosine similarity between (x, y) can be measured as Eq. (1). Considering the occurrence value is only 1 and 0 in folksonomies, the Eq. (1) can be simplified as Eq. (2) where the capital letter X and Y correspond to the set of resources having tags x or y.

Table 1. 0/1 matrix view

	t_1	t_2	t_3	t_4	...	...	...	t_i
x	1	0	1	0	...	0	0	0
y	0	1	1	1	...	0	0	0

$$\cos(x, y) = \frac{\sum_{i \in |v|} \vec{x}_i * \vec{y}_i}{\sqrt{\sum_{i \in |v|} \vec{x}_i^2} * \sqrt{\sum_{i \in |v|} \vec{y}_i^2}} \tag{1}$$

$$\cos(x, y) = \frac{|X \cap Y|}{\sqrt{|X| * |Y|}} \tag{2}$$

Compared to the support measure, cosine similarity measure not only provides a correlation value between two tags, but it also enable us to prune the rule set because it does not calculate the resources that contain neither x nor y. The cosine similarity measure also helps to exclude high confidence but poorly correlated rules.

Considering the above mentioned specialty in our approach, Apriori, the earliest and highly efficient algorithm to mine association rules, does not fit our purpose well. We modify it and develop a simplified version of Apriori algorithm, LApriori: 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.

Algorithm 1. LApriori to discover association rules mining in folksonomies

```
1)  L1 = count frequent 1-item sets;
2)  for every resource r do
3)          for each pair of tags {t_a , t_c} in r do
4)                  if t_a ⊂ L1 and t_c ⊂ L1 then
5)                          increase support of {t_a , t_c} by 1;
6)                  end if
7)          end for
8)  end for
9)
10) for every frequent 2-item set {x,y} do
11)         cos(x,y)= Support(x,y)/sqrt(support(x)*support(y));
12)         if cos(x,y)> min_sim  then return x->y;
13)         end if
14) end for
```

3.2 Standard Tags: Ontology Matching with WordNet

We use WordNet as the upper ontology and compute each semantic relation between tags in terms of hypernym relation from WordNet. A term that is more generic or more abstract than a given term is considered as a hypernym. For example, in table 2, the term wine has the following upper hypernyms: alcohol, beverage, drink, red, etc.

Table 2. A hypernym example of "wine" from WordNet 3.0

Sense 1:	Sense 2:
wine, vino	wine
=> alcohol, alcoholic beverage, intoxicant,	=>red
=> beverage, drink, drinkable, potable	=>color
=> food, nutrient	

Possible semantic relations between them are described as *more general* ($\supseteq$), *less general* ($\subseteq$), *equivalence* (=) [24]. $x \supseteq y$, if x is a hypernym of y. For example alcohol is a hypernym of wine, then we can say that alcohol is *more general* than wine, or wine *is-a* kind of alcohol, alcohol $\supseteq$ wine.

In folksonomies, we added another two definitions: *essential tags*, and *candidate hypernyms*.

Essential tags are all distinct tags existing in association rules filtered by predefined thresholds.

Candidate hypernyms: Given a tag, only the hypernyms that exist in its related tags are regarded as *candidate hypernyms*. For example, if *beverage* and *food* are two hypernyms for *wine* and also related to *wine* through association rules, then *beverage* and *food* are *candidate hypernyms* for wine. On the other hand, although *alcohol* and *red* are also hypernyms for wine, we do not consider them as *candidate hypernyms* because they have no relationship with *wine* in the generated association rules.

We only use hypernyms both existing in WordNet and association rules, because those hypernym terms not related to certain tags in folksonomies do not reflect the subjective knowledge well.

Based on the above mentioned considerations, we design the following Folk2Onto algorithm to find *more general* term for each *essential tag*.

Algorithm 2. Folk2Onto to find *more general* term for each *essential tag*

```
for each tag t_k in essential tags do
U_k = the more general term for t_k , set U_k = null;
S_k = get all tags related to t_k from association rules;
W_k = get all hypernyms for t_k from WordNet;
candidate hypernyms {h_1 ... h_n ...} = S_k ∩ W_k ;
    for each h_n do
        if U_k is null
            U_k = h_n ;
        else if U_k is not null and h_n is a hyperhym of U_k
            break ;
        else if U_k is not null and U_k is a hyperhym of h_n
            U_k = h_n ;
        end if
    end for
end for
```

For example, given a set of tags {food, beverage, wine, milk}, following semantic relations or ontological structure were generated as Fig 2:

beverage $\supseteq$ wine,
beverage $\supseteq$ milk,
food $\supseteq$ beverage,

Fig. 2. An ontological structure for "wine"

Beside *hypernyms*, WordNet also provides semantic relations such as *meronyms, synonyms, and antonyms which* can potentially be helpful in our approach.

3.3 Compound Tags: Token-Based Similarity Matching

Compound tags are non-standard terms and thus cannot be processed by WordNet without transformation. Here we adopt a series of filters provided by Jawbone[25] to analyze the compound tags. If they match certain defined criteria, the compound tags will be reserved and represented by its base term for more general parent finding.

In detail, the following term filters are applied to check whether the compound tag has a particular relationship to another term existing in WordNet:

1) EndWithFilter operates by splitting the compound into independent token of standard terms. The last one is used to represent the whole compound. For instance, "collaborative_tagging" is represented by tagging.
2) StartsWithFilter operates in a similar way as EndWithFilter except that the first token is used to represent the whole word. We apply this filter after the EndWithFilter because the first part of a compound is usually a definitive term while the last part is usually a subject which reflects the main meaning of the compound tag.

Note that we do not replace or transform the compound to the standard term, but only use them as interpreters for semantic relation discovery.

3.4 Jargon Tags: Combination of Association Rules and Similarity Ranking

In this step, jargon tags are incorporated to the previously built ontological structure with a matcher using graph centrality in a similarity graph of tags[10]. Although jargon tags are also non-standard and cannot be recognized, the association rules show their relations with other common tags. Considering each jargon and its related standard tags as separate subset in vector spaces, the tag similarity graph for each subset is a subgraph where each tag is represented by a vertex and the cosine similarity measures the distance between them.

The incorporation considers each jargon tag as the central node of a subgraph. Then it adds each related standard tags in the subgraph. Based on the matcher between this jargon and its related standard tag, the jargon tag is incorporated to the structure. If there is more than one standard tag associated with the jargon tag, the tag with the highest cosine similarity index will have the priority. Association rules involving jargons usually have the jargon as the antecedent. Thus, the jargon tag will be considered as a child of its consequent in the rule. This incorporation repeats until all jargon tags have been connected with their related standard tags in the structure.

For example, a jargon tag "folksonomy" is associated with four standard tags, "tagging, plurality, social, ontology". Ranking by cosine similarity, the rule "folksonomy $\rightarrow$ tagging" was selected. Based on this match, folksonomy was incorporated to the ontological structure as a child of tagging.

4 Experimental Evaluation

The experiment was based on collections from two separate CTS, citeulike.org and flickr.com. The collection from Citeulike was crawled from using several keywords, including "science", "philosophy", "research". We got 30,769 rows of data, where

each row represents a paper and a set of tags described by online users. Another dataset from Flickr was collected with flickr API, which consists of a set of callable methods for users, photos, photosets and other uniquely identifiable objects. We crawled the data using a narrow keyword "fruit" and collected 18,555 rows of data.

Pre-processing operations were performed to clean up the datasets. For dataset from Flickr, we only kept one record for each user because many users batch upload multiple photos with same tags. These repetitive tags will give us biased support count in association rules mining step and thus were excluded. Other common cleanup methods were applied to remove the tags called "no-tag" (a system generated tag for empty tag). We also removed objects with only one tag.

The descriptive statistics on the datasets after pre-processing is shown in Table 3:

Table 3. Statistics of collections used in experiment

	Collection	
	Citeulike	**Flickr**
Resources	30,769	18,555
After cleaning	25,937	6,462
Distinct tags	26,709	16,832
Users	4,068	6,462
Seed keywords	science, philosophy, research	fruit

4.1 Low Support Association Rules

There are three parameters, minimum support, confidence, and cosine similarity (denoted as *minsup*, *minconf*, and *mincos*) need to determine in our approach. We counted the number of *essential tags* with different *minsup* threshold and observe that most of the *essential tags* do not occur frequently (see Table 4). Moreover, the investigation of the initial association rule set reveals some interesting patterns of the cosine similarity. The value of similarity between pair of synonyms or sub-classes under same upper class tends to be high, sometimes close to 1. On the other hand the similarity value between a sub-class tag and its parent tag or upper class tag tends to be low. For instance, *food* is the parent of *beverage* in WordNet, the cosine similarity between *food* and *beverage* is low because *food* is a general term and it is associated with many other tags in the data set.

Thus the *mincos* was set to a relatively low value 0.2 to preserve the relations between upper and sub class tags and the relations between subclass tags. The *minsup* was set a very low value, 0.02% to include low-occurrence tags and reflect their relations. As usual, minconf was set to a 0.8, a relatively high value.

We observe that in total 152,372 rules are generated from citeulike at 0.02% minsup. These rules are significantly reduced to 24,025 by 0.2 cosine similarity and 0.8 confidence thresholds. Approximately 4,000 essential tags have been found. Table 5 also demonstrates the necessity of very low support threshold. In both these experiments, support 0.02% retains relations between around 4,000 essential tags. It only keeps relations between 300 essential tags if we increase the support threshold to 0.18%, a low support in traditional associational rule mining.

Table 4. Distribution of essential tags

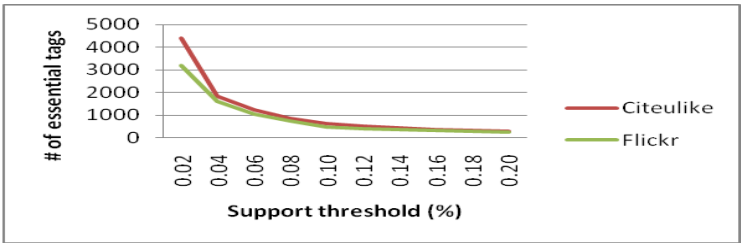

Table 5. Rules generated with 0.02% support, 80% confidence

Rules			Support	Confidence	Cosine	Accept?
folksonomy	→	tagging	1.59%	0.82	0.722	Y
macroeconomics	→	economics	0.09%	0.96	0.2671	Y
cyber-ethnography	→	ethnography	0.06%	1.00	0.2872	Y
asc	→	collaboration	0.03%	1.00	0.172	N
final	→	social	0.04%	0.90	0.1679	N
seeking	→	information	0.03%	0.85	0.1605	N

Table 5 shows the effect of the three thresholds. It contains 6 randomly selected low support rules generated at support threshold 0.02% and confidence threshold 0.8. Low support value helps to preserve rarely occurred pairs while cosine similarity acts as a guard to exclude rules consisting of tag pairs not highly related. For example, the relation between *macroeconomics* and *economics* was revealed under low support threshold. On the other hand, although the confidence for the rule *final* $\rightarrow$ *social* is higher than 0.8, it was excluded because cosine similarity is less than *mincos*. If we set *minsup* higher than 0.18% or *mincos* higher than 0.3, both second and third rules will not be discovered and will not be included in the final ontological structure.

4.2 Ontological Structures and Evaluation

In this section, we present and evaluate the results. Taking the task-based evaluation approaches [26], we measure how far the extracted ontological structure will help to influence and improve the results of certain tasks including multi-dimensional view and cataloguing and indexing.

Multi-dimensional View: The result retrieved with the "fruit" keyword was successfully organized into several dimensions in our approach. In these concept dimensions, "*produce*, *plant*, *food*, and *color*" contain most sub-classes. In Fig 3, it shows that detailed sub-classes are organized into multiple level ontological structures.

We provide a subjective evaluation of our ontological structure (Fig 3) by comparing it to an ontology (Fig 4) extracted from sei.cmu.edu and cluster results from flickr.com (Fig 5).

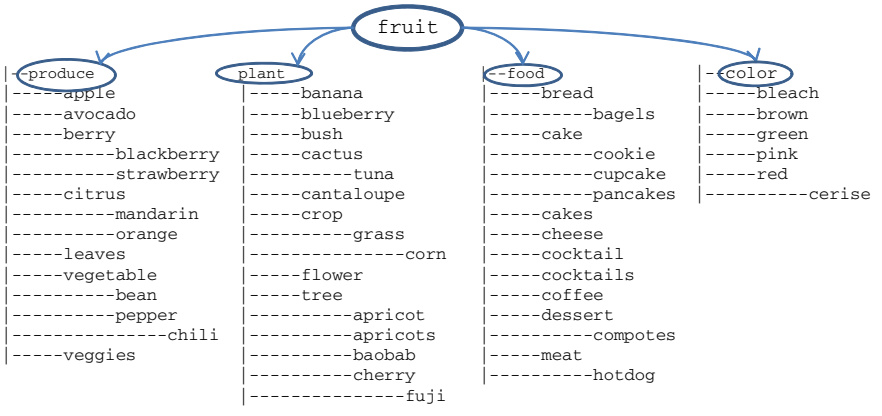

Fig. 3. A fragment output of "fruit" ontological structure, extracted from Flickr dataset

We observe that the terms from our result have a reasonable coverage and similar structure with the ontology in Fig 4. Such as the name and color of fruits like "berry, citrus, and red". The result also shows that our method produces more specific terms and additional levels than the ontology in Fig 4. For example, the "citrus" includes sub-classes, "orange" and "mandarin". However, our result does not provide enough property information for each term, such as flavor and seedless of strawberry in the ontology in Fig 4. The reason is that we currently only consider the hypernym relation from WordNet.

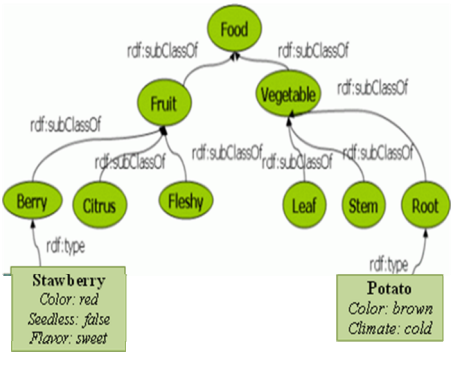

Fig. 4. An ontology of food[4]

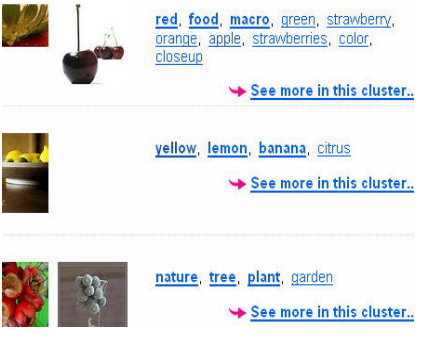

Fig. 5. Fruit clusters[5] from Flickr

After that, we compared our result with clusters of query result using keyword "fruit" from flickr.com (Fig 5). There are three main clusters in the screenshot. First is "red, food, etc." Second is "yellow, banana, etc." The third is "nature, tree, plant, etc." Although we can see that the third cluster are mainly about nature and is right to be separated from other clusters, there is no significant difference between first and

[4] http://www.sei.cmu.edu/isis/guide/technologies/owl-s.htm
[5] http://www.flickr.com/photos/tags/fruit/clusters/

second clusters since most of them are names of fruits. Furthermore, tags like "food, yellow, red" are not distinguished correctly, and are mixed in the two clusters.

On the other hand, the terms related to fruit are clearly classified into four dimensions in our results in Fig 3. Furthermore, our structure provides detailed sub-classes in each dimension, for example, the "berry" is placed under the "produce" dimension and could be further navigated into "blackberry and strawberry".

Although terms like "fleshy" are missing in our structure, it is actually rarely used by users. By checking the photos annotated with "fleshy" in Flickr website, we found that only 0.00003% photos were tagged by "fleshy.

In short, the extracted ontological structure reflects the fruit domain knowledge well and organizes the related resources into several navigable dimensions.

Cataloguing and Indexing: Fig 6 illustrates a fragment of the experiment result in science domain from Citeulike. The related terms are organized into a 5-level ontological structure, which gives users an overview of knowledge in science domain. It also provides a conceptual framework for cataloguing and indexing the resources. For example, from general to specific, anthropology and biology are organized under science catalogue. Then, biology is further divided into genetics and neurobiology. The number beside each term shows how many papers are contained in the corresponding catalogue.

```
|-science  (762)
|------anthropology  (111)
|------------ethnography (128)
|------biology  (256)
|------------genetics  (154)
|------------------evolutionary-genomics (41)
|------------------evolutionary-proteomics (22)
|------------------genomics (250)
|------------------proteomics (127)
|------------neurobiology (41)
|------------------neuroscience (199)
|------------------------neurophysiology (24)
|------------sociobiology (26)
|------system_biology (6)
|------------sysbio (74)
|------cryptography (25)
|------economics (259)
|------------macroeconomics (21)
|------informatics (141)
|------ip (54)
|------mathematics (163)
|------------geometry (78)
|------------statistics (456)
|------medicine (105)
|------------toxicology (12)
|------------biomedicine (11)
```

Fig. 6. A fragment of ontological structure in science domain

We evaluated the catalogues manually and observe that compound and jargon terms have been appropriately incorporated at the correct hierarchical level, such as, "evolutionary-genomics", "evolutionary-proteomics" and "sociobiology" (see fig 6). In total, 1540 terms were incorporated into the ontological structure. Among those terms, 35.65% of them were standard terms and more than 64% were non-standard terms, including 36.17% compound and 28.18% jargon terms.

5 Conclusion and Future Work

In this paper, we have proposed an integrated bottom-up and top-down approach to extract ontological structures from collaborative tagging systems. 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.

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. Rather than the clustering technique, association rule mining is a unsupervised data mining method to find interesting association between data sets. In this paper, we applied the association rules to find semantically related tags, which is the basis for further ontology building. Furthermore, we simplified the a priori algorithm to find 2-item set rules and introduced a new cosine coefficient, which significantly improved the efficiency in low support mining.

We observe that the ontological structures obtained could be enriched and deepened using larger tag datasets, other semantic relations provided by WordNet, and more specialized semantic lexical resources such as thesauri and subject-specific dictionaries. Additional work to represent the extracted ontologies in the web using RDF data format and SPARQL query language[6] will enable the integration with other web services, such as a collaborative ontology evolvement environment to reflect the knowledge and usage changes in CTS.

References

1. Auchard, E.: Flickr to map the world's latest photo hotspots. Reuters.com (2007)
2. Vander, T.: Folksonomy Coinage and Definition (2007), http://www.vanderwal.net/folksonomy.html
3. Euzenat, J., Shvaiko, P.: Ontology matching. Springer, Heidelberg (2007)
4. Wikipedia: Tag cloud. Wikipedia, The Free Encyclopedia (2009)
5. Suchanek, F.M., Vojnovic, M., Gunawardena, D.: Social tags: Meaning and Suggestions. In: ACM Conference on Information and Knowledge Management (CIKM 2008), pp. 223–232. ACM, Napa (2008)
6. Mathes, A.: Folksonomies-Cooperative Classification and Communication Through Shared Metadata. Computer Mediated Communication, LIS590CMC (2004)
7. Stuckenschmidt, H., Harmelen, F.V.: Information Sharing on the Semantic Web. Springer, Heidelberg (2005)
8. A1-Khalifa, S., Davis, C.: Measuring the Semantic Value of Folksonomies. Innovations in Information Technology (2006)
9. Mika, P.: Ontologies are us: A unified model of social networks and semantics. Web Semantics 5, 5–15 (2007)

[6] http://www.w3.org/TR/rdf-sparql-query/

10. Heymann, P., Garcia-Molina, H.: Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Stanford InfoLab Technical Report (2006)
11. Schmitz, C., Hotho, A., Jäschke, R., Stumme, G.: Mining Association Rules in Folksonomies. In: Proceedings of the 10th IFCS Conference, Studies in Classification, Data Analysis, and Knowledge Organization (2006)
12. Schwarzkopf, E., Heckmann, D., Dengler, D., Kroner, A.: Mining the Structure of Tag Spaces for User Modeling. In: Workshop on Data Mining for User Modeling (ICUM 2007) (2007)
13. Specia, L., Motta, E.: Integrating Folksonomies with the Semantic Web. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 624–639. Springer, Heidelberg (2007)
14. Damme, C.V., Hepp, M., Siorpaes, K.: FolksOntology: An Integrated Approach for Turning Folksonomies into Ontologies. In: Proc. of the ESWC Workshop Bridging the Gap between Semantic Web and Web (2007)
15. An, Y.J., Geller, J., Wu, Y.-T., Chun, S.A.: Automatic Generation of Ontology from the Deep Web. Database and Expert Systems IEEE 2007 (2007)
16. Laniado, D., Eynard, D., Colombetti, M.: Using WordNet to turn a folksonomy into a hierarchy of concepts. In: Proc. of 4th Italian Semantic Web Workshop, Italy (2007)
17. Schreiber, A.T.G., Dubbeldam, B., Wielemaker, J., Wielinga, B.: Ontology-Based Photo Annotation. IEEE Intelligent Systems (2001)
18. Schmitz, P.: Inducing Ontology from Flickr Tags. In: Proceedings of the Collaborative Web Tagging Workshop (WWW 2006), Edinburgh, UK (2006)
19. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In: Proceedings of the 1993 ACM SIGMOD international conference on Management of data (1993)
20. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules in Large Databases. In: Proceedings of the 20th International Conference on Very Large Data Bases (1994)
21. Cohen, E., Datar, M., Fujiwara, S., Gionis, A., Indyk, P., Motwani, R., Ullman, J., Yang, C.: Finding interesting associations without support pruning. Transactions on Knowledge and Data Engineering 13, 64–78 (2001)
22. Spertus, E., Sahami, M., Buyukkokten, O.: Evaluating similarity measures: a large-scale study in the orkut social network. In: Proceedings of the 11th ACM SIGKDD intl. conference on Knowledge Discovery in Data mining, Chicago, USA (2005)
23. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York (1986)
24. Giunchiglia, F., Shvaiko, P., Yatskevich, M.: S-Match: an Algorithm and an Implementation of Semantic Matching. In: Bussler, C.J., Davies, J., Fensel, D., Studer, R. (eds.) ESWS 2004. LNCS, vol. 3053, pp. 61–75. Springer, Heidelberg (2004)
25. Wallace, M.: Jawbone (2007), `http://mfwallace.googlepages.com/jawbone.html`
26. Dellschaft, K., Staab, S.: Strategies for the evaluation of ontology learning. In: Buitelaar, P., Cimiano, P. (eds.) Ontology learning and population: Bridging the gap between text and knowledge. IOS Press, Amsterdam (2008)