Learning Personalized Tag Ontology from User Tagging Information

Endang Djuana **Yue Xu** **Yuefeng Li**
School of Electrical Engineering and Computer Science
Queensland University of Technology
GPO Box 2434, Brisbane, QLD 4001
{e.djuanatjhwa, yue.xu, y2.li}@qut.edu.au

Abstract
The cross-sections of the Social Web and the Semantic Web has put *folksonomy* in the spot light for its potential in overcoming knowledge acquisition bottleneck and providing insight for "wisdom of the crowds". *Folksonomy* which comes as the results of collaborative tagging activities has provided insight into user's understanding about Web resources which might be useful for searching and organizing purposes. However, collaborative tagging vocabulary poses some challenges since tags are freely chosen by users and may exhibit *synonymy* and *polysemy* problem. In order to overcome these challenges and boost the potential of *folksonomy* as emergence semantics we propose to consolidate the diverse vocabulary into a consolidated entities and concepts. We propose to extract a tag ontology by ontology learning process to represent the semantics of a tagging community. This paper presents a novel approach to learn the ontology based on the widely used lexical database WordNet. We present personalization strategies to disambiguate the semantics of tags by combining the opinion of WordNet lexicographers and users' tagging behavior together. We provide empirical evaluations by using the semantic information contained in the ontology in a tag recommendation experiment. The results show that by using the semantic relationships on the ontology the accuracy of the tag recommender has been improved.

Keywords: collaborative tagging, *folksonomy*, ontology learning, personalization, tag recommendation

1 Introduction
The development of World Wide Web has leaded the research activities into cross-sections of two worlds: the Social Web and the Semantic Web. The Social Web is represented by a class of web sites and applications in which user participation is the primary driver of value which often referred by the phrase "collective intelligence" or "wisdom of crowds" to refer to the value created by the collective contributions of all these people (Gruber 2008). This trend was firstly mentioned in article by OReilly (2005) as Web 2.0.

The Semantic Web is an extension of the existing World Wide Web. It provides a standardized way of expressing the relationships between web pages, to allow machines to understand the meaning of hyperlinked information (Berners-Lee 2001). This may create the "web of data" in which metadata in the form of ontology, explicit specification of the conceptualization of a domain (Gruber 1993), plays important role in achieving this vision.

However, after several years on, this vision still has challenges due to knowledge acquisition bottleneck such as development and maintenance of ontologies. Ontology learning has been developed to overcome this barrier (Maedche and Staab 2001). Ontology learning or semi-automatic way of constructing ontology relies on machine learning and automated language-processing techniques to extract concepts and ontological relations from structured or unstructured data such as database and text (Navigli, Velardi and Gangemi 2003).

Folksonomy (Vander Wal 2005) which is emerging from collaborative tagging activities has been acknowledged as potential source for constructing ontology, as they capture the vocabulary of the users which may be aggregated to produce emergent semantics, from which people may develop lightweight ontologies (Mika 2007). The growing availability of *folksonomies* has motivated the work introduced in this paper for constructing lightweight ontology from collaborative tagging data.

User tagging or collaborative tagging describes the process by which many users add metadata in the form of keywords to Internet resources with a freely chosen set of keywords (tags) (Marlow et al 2006, Golder and Huberman 2006).

Research works have been conducted in utilizing tagging information to improve searching, clustering, and recommendation making. However, collaborative tagging vocabulary poses some challenges since tags are freely chosen by users and may exhibit *synonymy* and *polysemy* problem. Moreover, the relationships among tags haven't been maximally utilized, which could provide valuable information us to better understand users since there exists rich relationships among tags.

In this paper we present our approach to construct personalized tag ontology based on user tagging information and the widely used general knowledge ontology WordNet (Fellbaum 1998). We begin by introducing the background of user tagging collection and the main motivation for this work in Section 2. We then review related works in Section 3. In Section 4 we introduce our ontology learning approach including the ontology personalization approach. In Section 5 we present novel methods for improving tag recommendation based on the proposed tag ontology. In Section 6 we

present an experiment and the initial results. Section 7 concludes this paper and gives some ideas for further work.

2 Key Concept and Motivation

2.1 User Tagging

A user tagging collection involves three entities: items, tags, and users, which are described below:

- Users $U = \{u_1, u_2, .. u_{|U|}\}$ contains all users in an online community who have used tags to organize their items.
- Tags $T = \{t_1, t_2, .. t_{|T|}\}$ contains all tags used by the users in U. Tags are typically arbitrary strings which could be a single word or short phrase.

  In this paper, a tag is defined as a sequence of terms. For $t \in T$, $t =< term_1, term_2, \ldots\ldots, term_m >$. A function is defined to return the terms in a tag:

  $tagset(t) = \{term_1, term_2, .. term_m\}$
- Items $I = \{i_1, i_2, .. i_{|I|}\}$ contains all domain-relevant items or resources. What is considered by an item depends on the type of user tagging collection, for instance, in Amazon.com the items are mainly books.

Based on the three entities, a user tagging collection or a collaborative tagging system is formulated as 4-tuple: $F = (U, T, I, Y)$ (Jaschke et al 2008) where U, T, I are finite sets, whose elements are the users, tags and items, respectively. Y is a ternary relation between them, i.e., $Y \subseteq U \times T \times I$, whose elements are called tag assignments or taggings. An element $(u, t, i) \in Y$ represents that user u collected item i using tag t.

Tags in a tag collection may exhibit many variations such as *synonymy* where different tags may have the same or closely related meanings. Different users may tag an item using different tags which have similar meaning. The other variation is *polysemy* where one tag has multiple meanings. A tag may be used by different users to tag different items that are not related to each other at all. Moreover, one tag may have semantic relationship to other tags, e.g. “inn” is a kind of “hotel” which shows the two tags are related with each other and “inn” has “more specific” meaning. This condition may not be utilized to relate items collected under these two tags because they are simply treated as two different tags.

2.2 Motivation

Many methods have been proposed to deal with the problems of synonymy and polysemy (Bischoff et al 2008, Suchanek and Vojnovic and Gunawardena 2008, Liang et al 2010). There are several works which try to infer relationship between tags (Tang et al 2009, Liu, Fang and Zhang 2010). However, these works mostly didn't base the inference on semantic measure but on statistical measure which may fail to capture the semantic relationships among tags. Also, the semantic relationships between tags need to be exploited more by existing tagging based applications including tag based recommenders.

In order to tackle these problems, it becomes desirable to find a way to consolidate the multiple facets and the relationships of tags into a consolidated entity which will help better understand the tags used by users. There are several possible solutions include using classification systems such as taxonomy or using conceptualization systems such as ontology. In this work we consider to use ontology to represent the semantics in tags collection because of the flexibility of an ontology and possibility of emerging semantics from the ontology learning process (Mika 2007, Lin, Davis and Zhou 2009).

3 Related Works

Work by Garcia-Silva et al (2012) compares most relevant approaches for associating tags with semantics in order to make explicit the meaning of those tags. They have identified three group of approaches which are based on 1) clustering techniques i.e. to cluster tags according to some relations among them (statistical techniques); 2) ontologies i.e. aiming at associating semantic entities e.g. WordNet, Wikipedia, to tags as a way to formally define their meaning; 3) hybrid approach i.e. mixing clustering techniques and ontologies. Our work falls into the second group which is based on ontologies.

Beside our work there are several works which tried to extract ontological structures from user tagging systems. Lin, Davis and Zhou (2009) extracted ontological structures by exploiting low support association rule mining supplemented by WordNet. Trabelsi, Jrad and Yahia (2010) focused more on extracting non-taxonomic relationships from folksonomies using triadic concepts with external resources: WordNet, Wikipedia and Google.

Tang et al (2009) and Liu, Fang and Zhang (2010) represents state of the art work for generating ontology from folksonomy based on generative probabilistic models i.e. tag-topic model and set-theoritical approach i.e. to produce tag subsumption graph respectively. Most of this works did not provide applications for the ontology such as tag recommendation.

As for the work in collaborative tag recommendation there are several notable works such as work by Sigurbjornsson, van Zwol and D’Silva (2008) which is based on tag co-occurrences. Although this work has achieved good result, it didn't rely on the actual meaning of tags which may miss the semantic relationships among tags.

Beside our work there are several works which utilize some format of ontology to assist in tag recommendation task. Baruzzo et al (2009) used existing domain ontology to recommend new tags by analyzing textual content of a resource needed to be tagged. They relied on existing domain ontology which is not always available for a particular domain and also they didn’t provide quantitative evaluation.

Tag recommendation approach by Tatu, Srikanth, D’Silva (2008) by mapping textual contents in Bibsonomy bookmarks, not just the tags to form conflated tags to normalized concepts in WordNet and similar approach by Lipczak et al (2009) which explored resource content as well as resource and user profiles are comprehensive. There is a drawback that they relied on extended textual contents provided by Bibsonomy which are not always available in other user tagging systems.

4 Ontology Learning from User Tagging

One stream of approach to the ontology construction relies on machine learning and automated language-processing techniques to extract concepts and ontological relations from structured or unstructured data such as database and text (Navigli, Velardi and Gangemi 2003).

In this work we propose to construct the tag ontology based on some existing ontology, which we call backbone ontology. The basic idea is to take advantage of hierarchies of concepts in the backbone ontology and to form the tag ontology by mapping the tags in the tag collection to the concepts on the backbone ontology and extracting the available relationships among concepts in the backbone ontology.

The lexical knowledge base WordNet (Fellbaum 1998) was chosen in this paper as the backbone ontology as it has wide coverage of concepts (over 200,000) and richness of relationships such as semantic relationships "is-a", "part-of", lexical relationships "synonymy" and "antonymy" as well as availability of accompanying corpus and other facility for disambiguation process. The backbone ontology is defined below.

Definition 1 (Backbone ontology): The backbone ontology is defined as a 2-tuple $BackboneONTO = (C, R)$ where $C = \{c_1, c_2, .., c_{|C|}\}$ is a set of concepts; $R = \{r_1, r_2, .., r_{|R|}\}$ is a set of relations representing the relationships between concepts.

A concept c in C is a 3-tuple $c = (id, synset, category)$ where id is a unique identification assigned by WordNet system to the concept c; $synset$ is a synonym set containing synonymic terms which represent the meaning of the concept c; and $category$ is a lexical category assigned by WordNet lexicographers to classify this concept c into a general category. A relation r in the relation set R is a 3-tuple $r = (type, x, y)$, where $type \in \{is\_a, part\_of, ...\}$; $x, y \in C$ are the concepts that hold the relation r.

For easy to describe the work, we denote the set of synonyms representing c by $synset(c)$ and the category of c by $category(c)$. For each term w in $synset$ (c), w is represented as a 2-tuple $(w, freq_c(w))$ where w is a synonym term of the concept c; $freq_c(w)$ is the frequency assigned by WordNet lexicographers to the term as an indication of how frequently this term has been used to represent the meaning of the concept c based on the accompanying WordNet corpus. For a term w, the set of concepts for which w is a synonymic term is defined as $con(w) = \{c|(w, f) \in synset(c)\}$.

4.1 Mapping Tags to Concepts

One tag may contain one or more terms. It is possible that a tag can be mapped directly to one or more concepts in the backbone ontology. It is also possible that only part of a tag may map to one or more concepts. We propose the following mappings to deal with different cases.

There are 3 different cases for finding possible mappings for a given tag, which are: (1) mapping the full tag to one or more concepts; (2) mapping part of the tag to one or more concepts; and (3) splitting the tag into a list of single words, then mapping each of the words to concepts separately. Readers are referred for a more detailed discussion for each case from previous publications in Djuana, Xu and Li (2011).

1. Direct Mapping

We define the following function to represent the whole mapping from a tag to concepts:
$Tag\_Concept_{whole}: T \rightarrow 2^C$

$$\forall t \in T, Tag\_Concept_{whole}(t) = \{c | \forall c \in C, \exists (w, f) \in synset(c), t == w\}$$

$Tag\_Concept_{whole}(t)$ is a set of concepts for each of which *t* is synset term.

2. Partial Mapping

The following function represents the partial mapping from a tag to concepts: $Tag\_Concept_{partial}: T \rightarrow 2^C$

$$\forall t \in T, Tag\_Concept_{partial}(t) = \{c | \forall c \in C, \exists (w, f) \in synset(c), MaxPostfix(t) == w\}$$

$MaxPostfix(t)$ stands for the largest postfix of *t*.

3. Term Mapping

The following function represents the term mapping from a tag to concepts: $Tag\_Concept_{term}: T \rightarrow 2^C$

$$\forall t \in T, Tag\_Concept_{term}(t) = \sum_{a \in tagset(t)} Tag\_Concept_{whole}(a)$$

Overall, $\forall t \in T$, the tag to concept mapping is defined as follows:

$$Tag\_Concept(t) = \begin{cases} Tab\_Concept_{whole}(t) & if\ t\ is\ directly\ mapped \\ Tag\_Concept_{partial}(t) & partically\ mapped \\ Tag\_Concept_{term}(t) & term\ mapped \end{cases} \quad (1)$$

4.2 Mapping Disambiguation

A tag can be mapped to multiple concepts. After all the possible mappings are found, we need to choose the most appropriate concept from the mapped concepts to represent the meaning of the tag for this particular tag collection.

For disambiguating the concepts, we propose to measure the strength of the mapping by using the word frequency provided by WordNet. A matrix $T\_C[t_i, c_j]_{m \times n}$ is defined to represent the strength of the mapping between tags and concepts, where m=|T| and n=|C|. In order to make the frequency comparable between different concepts, we normalize the frequency value to a scale of [0, 1]. The mapping strength based on frequency is defined below:

$$T\_C_{frequence}[t_i, c_j] = \begin{cases} \frac{f_{c_j}(t_i)}{\sum_{c_k \in Tag\_Concept(t_i)} f_{c_k}(t_i)} & c_j \in Tag\_Concept(t_i) \\ 0 & otherwise \end{cases} \quad (2)$$

For a tag t_i, the concept c_j should be chosen as t_i's concept if $T\_C[t_i, c_j]$ is the highest value for all $c_j \in Tag\_Concept(t_i)$. After the disambiguation, each tag *t* will be mapped to one and only one concept. This can be defined by a one to one disambiguation mapping $M_{frequence}: T \rightarrow C$

$$M_{frequence}(t) = \underset{c \in Tag-Concept(t)}{\operatorname{argmax}} (T\_C_{frequence}[t, c]) \quad (3)$$

On the other hand, multiple tags may be mapped to one concept. The following function defines the mapping from a concept to tags: $Concept\_Tag: C \rightarrow 2^T$
$Concept\_Tag(c) = \{t | \forall t \in \mathrm{T}, M_{frequence}(t) == c\}$

4.3 Relationship Extraction Process

After the mapping and disambiguation processes, each tag will be mapped to a concept on the backbone ontology. Based on the mappings, we retrieve the available relationships ("is-a" relations) from the mapped concept c consecutively until we reach the top of the hierarchy. This operation is the same operation as finding an ancestor in a tree-based structure. The top of the hierarchy in the backbone ontology is a general category defined by WordNet.

We can then extract the mapped concepts together with the relationships in the backbone ontology to form the tag ontology. As the result of the tag to concept mapping and the relationships extraction, we can construct the tag ontology which is defined as below:

Definition 2 (Tag Ontology): The tag ontology is defined as 2-tuple $TagOnto = (TC, TR)$ where $TC = \{tc_1, tc_2, .., tc_{|TC|}\}$ is a set of tag-concepts, i.e., $\subseteq C \times 2^T$, and $TR = \{tr_1, tr_2, .., tr_{|TR|}\}$ is a set of tag relations. Each element in TC is a pair of a concept c and a set of tags $\{t_1, t_2, .. t_n\}$, i.e., $tc = (c, \{t_1, t_2, .. t_n\}) \in TC$, which represents that each tag in $\{t_1, t_2, .. t_n\}$ can be mapped to concept c. TR is defined as:

$$TR = \left\{ r = (type, c_1, c_2) \middle| \begin{matrix} r \in R, \\ Concept\_Tag(c_1) \neq \emptyset, \\ Concept\_Tag(c_2) \neq \emptyset \end{matrix} \right\}$$

4.4 Personalization in Mapping Disambiguation

The tag ontology constructed using the approach described in previous sections mainly utilizes the structural information between concepts and the frequencies of synset terms provided by WordNet. The tag-to-concept mapping is mainly determined based on the synset term frequencies which are derived based on WordNet corpus.

However, for a given tagging collection, the synset term frequencies may not adequately reflect the interests of the users in this particular collection. To reduce the bias caused by solely using the synset term frequency, we propose to take user tagging information into consideration in disambiguating the mapping from tags to concepts.

Let (U, T, I, Y) be a tagging system, the following strategy is proposed to generate personalized tag ontology for users in *U*. The personalization in the context of this paper is for a tagging community rather than for individual users. The idea here is to find tag relevance based on the tagging information of users in a tagging community and then map tags onto the backbone ontology based on the tag relevance.

In WordNet, each concept is assigned into one and only one category. Let *CA* denote the set of categories in WordNet ontology, for a concept $c \in C$, $\varepsilon(c) \in CA$ is defined as the only category assigned to the concept *c*. Different concepts can be categorized into one category. On the other hand, for a category *Ca*, it may have multiple concepts. A function $concept(C_a) = \{c | \forall c \in C, \varepsilon(c) == C_a\}$, is defined to return all the concepts that belong to category *Ca*.

Moreover, the categories of a tag *t* can be obtained from the category of *t*'s concepts (i.e., $Tag\_Concept(t)$). The set of categories of a given tag *t* is defined as: $category(t) = \{\varepsilon(c) | c \in Tag\_Concept(t)\}$. A category can have multiple concepts. Similarly, a category *Ca* can have multiple tags which belong to *Ca*. A function $tag(Ca) = \{t | \forall t \in T, C_a \in category(t)\}$ is defined to return all the tags that belong to category *Ca*.

For an item, different users may collect it using different tags and these tags must have something in common which reflects some characteristic of the item. Therefore, by looking at the tags that have been used by users in *U* to tag the same items, we can find related tags with respect to the users in *U*. For a given tag $t \in T$, the related tags of *t* is defined by the following equation:

$$t\_related(t) = \{t_j | \forall i \in I_t, \exists t_j \in T_i, \exists u \in U, (u, t_j i) \in Y\} \quad (4)$$

where I_t is a set of items that are collected by users with tag *t*, T_i is a set of tags that are used by users to tag item *i*.

In this paper, we propose to estimate the relevance between a tag t_i and a concept c_j by exploiting the relevance between the tag and its t-related tags that belong to the same category of c_j to measure the strength from t_i to the concept c_j. Let $p(t_i|t_k)$ represent the probability of using t_i to tag some items given that t_k has been used to tag the items. If $p(t_i|t_k)$ is high, it can be considered that t_i is highly relevant to t_k.

We propose the following equation to measure the relevance of a tag to a concept based on the relevance of the tag to its related tags that belong to the same category of this concept:

$$t\_relevance(t_i, c_j) = \sum_{t_k \in t\_related(t_i) \cap tag(category(c_j))} p(t_i|\mathrm{t_k}) \quad (5)$$

Given tags t_i and t_k, the probability of using t_i and t_k to tag an item *a* can be calculated by the equation: $p(a | t_i, t_k) = \frac{p(t_i | a, t_k) p(a | t_k)}{p(t_i | t_k)}$, from which, we can get the following equation to calculate $p(t_i|t_k)$:

$$p(t_i|t_k) = \sum_{a \in I} p(t_i|a, t_k) p(a|t_k) \quad (6)$$

Let $UI_{t_j} = \{(u_i, i_k) | \forall u_i \in U, \forall i_k \in I, (u_i, t_j, i_k) \in Y\}$ be a set of user-item pairs each of which represents that a user tags an item using tag t_j (i.e., the tag assignments using t_j); $U_{t_j, i_k} = \{u_i | \forall u_i \in U, (u_i, t_j, i_k) \in Y\}$ be a set of users who have used tag t_j to tag item i_k.

For a given tag *t*, the probability of using *t* by any user to tag any item, denoted as $p(t)$, can be defined as the ratio between the number of tag assignments using *t* and the total number of tag assignments, i.e., $p(t) = \frac{|UI_t|}{|Y|}$.

The probability of using tag *t* to tag item *a* by any users can be defined as the ratio between the number of users

who used t to tag a and the total number of tag assignments, i.e., $p(t,a) = \frac{|U_{t,a}|}{|Y|}$.

Similarly, $p(t_1,t_2,a) = \frac{|U_{t_1,a} \cap U_{t_2,a}|}{|Y|}$, it is the ratio between the number of users who have used both t_1 and t_2 to tag item a and the total number of tag assignments.

Based on these probabilities, we can calculate the two probabilities, $p(a|t)$ and $p(t_1|a,t_2)$, as:

$$p(a \mid t) = \frac{p(t,a)}{p(t)} = \frac{|U_{t,a}|}{|UI_t|}$$

$$p(t_1 \mid a,t_2) = \frac{p(t_1,t_2,a)}{p(t_2,a)} = \frac{|U_{t_1,a} \bigcap U_{t_2,a}|}{|U_{t_2,a}|}$$

Thus, equation (6) becomes:

$$p(t_i|t_k) = \sum_{a \in I} \frac{|U_{t_i,a} \cap U_{t_k,a}|}{|UI_{t_k}|} \qquad (7)$$

With Equation (7), we can calculate the relevance between a tag and a concept using Equation (5). The normalized tag relevance is used to measure the relevancy from a tag to a concept . $T\_C_{relevance}[t_i, c_j]_{m \times n}$ is defined as below:

$$T\_C_{relevance}[t_i, c_j] = \frac{t\_relevance(t_i, c_j)}{\sum_{c \in Tag\_Concept(t_i)} t\_relevance(t_i, c)} \qquad (8)$$

For different sets of users, $T\_C_{relevance}[t_i, c_j]$ can be different because they are based on user tagging information, while $T\_C_{frequency}[t_i, c_j]$ will be the same for all user sets because it is based on the term frequency provided by WordNet .

The mapping disambiguation based on tag relevancy can be defined as $M_{relevance} : T \rightarrow C$

$$M_{relevance}(t) = \underset{c \in Tag\_concept(t)}{\operatorname{argmax}} (T\_C_{relevance}[t, c]) \qquad (9)$$

5 Tag Recommendation based on Tag Ontology

A tag recommender is a specific kind of recommender systems in which the goal is to suggest a set of tags for a user to use for tagging a particular item. One of our goals in this paper is to investigate whether the semantic information captured in the constructed tog ontology can be utilized to improve the accuracy of tag recommendation.

The task of a tag recommender system is to recommend, for a given user $u \in U$ and a given item $i \in I$ which has not been tagged by the user, a set $\tilde{T}(u,i) \subseteq T$ of tags. In many cases $\tilde{T}(u,i)$ is computed by first generating a ranking on the set of tags according to some criterion, from which then the top n tags are selected.

5.1 CF based Tag Recommendation

A tag recommender has been proposed in (Jaschke et al 2008) which is based on the user-based CF method. To recommend tags to a target user for tagging an item, it first finds the neighbor users of the target user, then generates a set of candidate tags which have been used by the neighbor users to tag the item and finally rank the candidate tags based on the similarity between the target user and neighbor users to decide the top n tags as the final recommendations.

Let $CT(u,i)$ be a set of candidate tags which have been used by u's neighbors to tag item i. For a candidate tag t in $CT(u,i)$, its ranking can be calculated by the following equation:

$$w(u,t,i) = \sum_{v \in N_u^k} sim(\vec{x}_u, \vec{x}_v) * \delta(v,t,i),$$

$$\delta(v,t,i) = \begin{cases} 1 & (v,t,i) \in Y \\ 0 & otherwise \end{cases} \qquad (10)$$

where $sim(\vec{x}_u, \vec{x}_v)$ is the similarity of users, N_u^k is user u's neighborhood containing k similar users, $\delta(v,t,i)$=1 indicates the user v has used this tag t to tag the item i. The top n tags, denoted as $T(u,i)$, can be determined based on the ranking:

$$T(u,i) = argmax_{t \in T}^{n} w(u,t,i) \qquad (11)$$

5.2 Tag Recommendation based on Tag Ontology

Having the tag ontology in place we can explore the concept representation of a tag, its placement in the hierarchy and its relationships to other concepts. This brought us an idea to improve the recommendations in $T(u,i)$ based on the semantic information in the extracted ontology to see if the ontology can directly improve tag recommendations.

In the proposed method, we generate candidate tags based on neighbour users' preference and the synset information captured in the tag ontology as well, and rank the candidate tags based on both user similarity and tag popularity.

5.2.1 Candidate tag expansion

Let $CT(u,i)$be the set of candidate tags generated based on neighbor users' preferences. For each candidate tag t in $CT(u,i)$, by using the disambiguation mapping methods given in Equation (3) or (9), t can be mapped to concepts $M_{frequence}(t)$ or $M_{relevance}(t)$ in the tag ontology, respectively. From the synset terms of the mapped concepts, two expanded sets of candidate tags can be generated based on the two methods:

$$CT_{frequence}(u,i) = \bigcup_{t \in CT(u,i)} synset(M_{frequence}(t))$$

$$CT_{relevance}(u,i) = \bigcup_{t \in CT(u,i)} synset(M_{relevance}(t))$$

5.2.2 Recommendation ranking

For each of the candidate tag t in $CT_{frequence}(u,i)$ or $CT_{relevance}(u,i)$, its ranking is calculated by the following equation:

$$w_\gamma(u,t,i) = \begin{cases} \sum_{v \in N_u^k} sim(\vec{x}_u, \vec{x}_v) * \delta(v,t,i) & t \in CT(u,i) \\ \sum_{v \in N_u^k} sim(\vec{x}_u, \vec{x}_v) * \delta(v,t,i) * \mathcal{P}(t) & t \notin CT(u,i)), t \in CT_\gamma(u,i) \end{cases} \qquad (12)$$

where $\gamma \in \{frequence, relevance\}$ and $\mathcal{P}(t)$ is the popularity of tag *t*, which is calculated as: $\mathcal{P}(t) = |UI_t|/max_{t_i \in T}|UI_{t_i}|$.

As defined in Section 4.4, UI_t contains (user, item) pairs representing the tag assignments using tag *t*. $|UI_t|$ is the number of times that *t* has been used to tag items. The higher the $|UI_t|$, the more popular the tag *t* is.

$\mathcal{P}(t)$ is the ratio between $|UI_t|$ and the maximum number of times that a tag has been used to tag items in this tagging community.

Based on the two disambiguation methods, we can generate two lists of tags ranked by using Equation (12). Thus, two lists of top *n* tags can be determined based on the ranking:

$$T_{frequence}(u,i) = argmax^n_{t \in T} w_{frequence}(u,t,i) \quad (13)$$

$$T_{relevance}(u,i) = argmax^n_{t \in T} w_{relevance}(u,t,i) \quad (13)$$

In our experiments to be discussed below, the accuracy of recommendations using the result in (13), (14), or the combination of the two has been compared.

6 Evaluation

6.1 Experiment Setup

We have conducted experiments to evaluate the usefulness of the proposed tag ontology in making tag recommendations. Two datasets are used in the experiments:

(1). The Bibsonomy dataset used in ECML PKDD Discover Challenge 2009 (http://www.kde.cs.uni-kassel.de/ws/dc09/). The dataset contains public bookmarks and publication posts of Bibsonomy. The dataset that used in this experiment contains 1122 users, 19682 items and 6517 tags.

(2). The publicly available Delicious dataset (Wetzker, Zimmermann and Bauckhage 2008). The dataset contains all public bookmarks of users posted on delicious.com between September 2003 and December 2007. In this paper a portion of the data set is used which contains bookmarks from January to March 2004. This portion contains 1289 users, 863 items (URLs) and 215 tags.

Each of the datasets is split into a testing dataset and a training dataset based on posting date. The split percentage is 25% for testing dataset which is taken from newer posts and 75% for training dataset which is taken from older posts. This is to simulate the actual tag recommendation scenario in which users are normally given a recommendation list based on what tags previously stored in the system.

In the experiments we conducted 5 folds cross validation for all the users in the dataset. In each run of the experiment, we randomly take 20% portion as the target users while the remaining 80% is taken as the training users from whom we calculate similarities to the target users to find neighbors. The top *n* tags are recommended to each target user for each of the user's items in the testing set. The recommended tags are compared to the target user's actual tags of the items in the testing dataset. If a recommended tag matches with an actual tag, we calculate this as a hit. The standard precision and recall are used to evaluate the accuracy of tag recommendations.

6.2 Results

We have conducted the following runs to compare the performance between the baseline recommender, the user based CF method, and the proposed methods:

- User-CF: this is the user based CF tag recommender system proposed in (Jaschke et al 2008).
- Exp_Freq: this is the proposed method to expand candidate tags by using synset terms of the tag ontology mapped based on synset term frequency.
- Exp_Rel: this is the proposed method to expand candidate tags by using synset terms of the tag ontology mapped based on tag relevance.
- Freq&Rel: this method generates the tag recommendations by combining the results of Exp_Freq and Exp_Rel and selecting the top *n* tags.

The results of the experiments are presented in Table I to Table IV for Bibsonomy and Delicious datasets, respectively. As shown in these tables, the use of the ontology has improved the precision and recall for all the two datasets. From the results, we can see that, the Exp_Rel run achieved better results than that of Exp_Freq run, which means that the tag relevance generated based on user tagging behavior of the users in this tagging community is more useful than the term frequency given by WordNet lexicographers. The former reflects the specific perspective of the users in this particular community, while the latter reflects the general viewpoint of lexicographers. Especially, the combination of the two methods outperforms all the other methods. From the results of this experiment, we can say that the tag ontology can be used to improve the performance of recommendation.

N	5	10	15	20
User-CF	0.183	0.103	0.070	0.052
Exp_Freq	0.191	0.109	0.075	0.056
Exp_Rel	0.191	0.110	0.075	0.056
Freq&Rel	0.201	0.126	0.091	0.072

Table 1: Precision for Bibsonomy dataset

N	5	10	15	20
User-CF	0.435	0.474	0.479	0.479
Exp_Freq	0.445	0.489	0.491	0.50
Exp_Rel	0.445	0.491	0.50	0.52
Freq&Rel	0.481	0.513	0.531	0.561

Table 2: Recall for Bibsonomy dataset

N	5	10	15	20
User-CF	0.169	0.081	0.072	0.054
Exp_Freq	0.176	0.095	0.063	0.047
Exp_Rel	0.176	0.096	0.065	0.047
Freq&Rel	0.183	0.104	0.072	0.049

Table 3: Precision for Delicious dataset

N	5	10	15	20
User-CF	0.609	0.655	0.656	0.655
Exp_Freq	0.639	0.681	0.682	0.680
Exp_Rel	0.639	0.683	0.685	0.689
Freq&Rel	0.641	0.697	0.703	0.711

Table 4: Recall for Delicious dataset

7 Conclusion

Tagging is getting more and more popular in many Web sites. It provides useful data for better understanding users' information needs. The user self-defined tags not only reflect users' understanding to the content of the tagged items, but also provide rich information about item hierarchical classification.

In this paper, we proposed a novel approach to construct tag ontology from user tagging information to represent the semantic meaning and hierarchical relationship among tags. We believe the constructed tag ontology can be used in many applications such as item classification, item recommendation, and tag recommendation. In this paper, we presented a primary experiment to show the improvement to tag recommendation based on the tag ontology. There is room to further improve the recommendation by applying further the extracted ontology structural information in the process of generating recommendation.

8 References

Baruzzo, A., Dattolo, A., Pudota, N. and Tasso, C. (2009), Recommending new tags using domain-ontologies, *Proc. of IEEE/WIC/ACM Web Intelligence and Intelligent Agent Technology-Workshops*, 409-412, IEEE.

Berners-Lee, T. (2001), The Semantic Web. Scientific American,http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21. Accessed 19 Oct 2012.

Bischoff, K., Firan, C.S., Nejdl, W., and Paiu, R. (2008), Can all tags be used for search? In *Proc. ACM Conference on Information and Knowledge Management*, 193-202, ACM Press.

Djuana, E., Xu, Y. and Li, Y. (2011), Constructing tag ontology from folksonomy based on WordNet, In *Proc. IADIS International Conference on Internet Technologies and Society*, IADIS. (In press)

Fellbaum, C. (ed.) (1998), *WordNet: An Electronic Lexical Database*, Cambridge, MA: MIT Press.

García-Silva, A., Corcho, O., Alani, H., Gómez-Pérez, A. (2012), Review of the state of the art: Discovering and associating semantics to tags in folksonomies, *The Knowledge Engineering Review*, Vol 27(01) 57-85.

Golder, S. and Huberman, B. (2006), The structure of collaborative tagging systems, HP Labs Tech. Report, http://www.hpl.hp.com/research/scl/papers/tags/tags.pdf, Accessed 19 Oct 2012.

Gruber, T.R. (1993), Towards principles for the design of ontologies used for knowledge sharing, In *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Guarino, N., Poli, R. (Eds.), Kluwer Academic Publishers.

Jaschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., and Stumme, G. (2008), Tag recommendations in social bookmarking systems, *AI Communications*, vol 21, 231-247, IOS Press.

Liang, H., Xu, Y., Li, Y., Nayak, R., Tao, X. (2010), Connecting users and items with weighted tags for personalized item recommendations. In *Proc. 21st ACM Conference on Hypertext and Hypermedia*, 51-60, ACM.

Lin, H., Davis, J., and Zhou, Y. (2009), An integrated approach to extracting ontological structures from folksonomies, *The Semantic Web: Research and Applications*, 654-668, Springer.

Lipczak, M., Hu, Y., Kollet, Y., Milios, E. (2009), Tag sources for recommendation in collaborative tagging systems, In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases Discovery Challenge.*

Liu, K., Fang, B., Zhang, W. (2010), Ontology emergence from folksonomies, In *Proc. ACM International Conference on Information and Knowledge Management*, 1109-1118, ACM Press.

Maedche, A. and Staab, S. (2001), Ontology learning for the Semantic Web. *IEEE Intelligent Systems* 16(2), 72-79, IEEE.

Marlow, C., Naaman, M., Boyd, D. and Davis, M. (2006), HT06, tagging paper, taxonomy, Flickr, academic article, to read, In *Proc. ACM Hypertext and Hypermedia*, 31-40, ACM Press.

Mika, P. (2007), Ontologies are us: A unified model of social networks and semantics. *Web Semantics.* 5(1), 5-15, Elsevier.

Navigli, R., Velardi, P., and Gangemi, A. (2003), Ontology learning and its Application to automated terminology translation", In *IEEE Intelligent Systems*, vol 18(1), 22-31, IEEE.

O'Reilly, T. (2005), What is Web2.0: Design patterns and business models for the next generation of software, http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html, Accessed 19 Oct 2012.

Sigurbjörnsson, B., van Zwol, R. (2008), Flickr tag tecommendation based on collective knowledge, In *Proc. World Wide Web Conference*, 327-336, W3C.

Suchanek, F. M., Vojnovic, M., and Gunawardena, D. (2008), Social tags: meaning and suggestions. In *Proc. ACM Conference on Information and Knowledge Management*, 223-232, ACM Press.

Tang, J., Leung, H., Luo, Q., Chen, D., Gong, J.(2009), Towards ontology learning from folksonomies, In *Proc. 21st International Joint Conference on Artificial Intelligence*, 2089-2094, AAAI Press.

Tatu, M., Srikanth, M. and D'Silva, T. (2008): Tag recommendations using bookmark content. In *Proceedings of RSDC'08*, 96-107.

Trabelsi, C., Jrad, A.B., and Yahia, S.B. (2010), Bridging folksonomies and domain ontologies: Getting out non-taxonomic relations, *Proc. IEEE International Conference on Data Mining Workshops*, 369-379 IEEE.

Vander Wal, T. (2005), Folksonomy coinage and definition, http://vanderwal.net/folksonomy.html. Accessed 19 Oct 2012.

Wetzker, R., Zimmermann, C. and Bauckhage, C. (2008), Analyzing social bookmarking systems: A del.icio.us cookbook", *Proc. of European Conference on Artificial Intelligence.*