Personalized Recommender Systems
Integrating Tags and Item Taxonomy

Huizhi Liang\textsuperscript{a,} *, Yue Xu\textsuperscript{b}, and Yuefeng Li\textsuperscript{b}

\textsuperscript{a} School of Information Technology, Queensland University of Technology, Brisbane, Australia
oklianghuizi@gmail.com
\textsuperscript{b} School of Information Technology, Queensland University of Technology, Brisbane, Australia
\{yue.xu, y2.li\}@qut.edu.au

Abstract. The social tags in Web 2.0 are becoming another important information source to profile users’ interests and preferences to make personalized recommendations. To solve the problem of low information sharing caused by the free-style vocabulary of tags and the long tails of the distribution of tags and items, this paper proposes an approach to integrate the social tags given by users and the item taxonomy with standard vocabulary and hierarchical structure provided by experts to make personalized recommendations. The experimental results show that the proposed approach can effectively improve the information sharing and recommendation accuracy.

Keywords: Recommender Systems, Tags, User Profiling, Personalization, Web 2.0, Taxonomy

1. Introduction

Recommender System is one effective tool to deal with information overload issue [7]. As explicit ratings are not always available in real life applications [1], the problem of how to make recommendations based on users’ implicit rating information has become an important research focus.

In Web 2.0, the user generated textural content information such as tags, blogs, reviews and comments is becoming more and more popular. Instead of using numerical information, people use one or more pieces of textural information to express their opinions and interest, collect and organize items, share experiences, and build up social networks etc. This kind of information, provided by users and processed by the “wisdom of crowds”, is becoming another important information resource, in addition to the information provided by the websites. Tags (i.e., social tags or folksonomy) are a typical type of Web 2.0 information. They are one or more keywords provided by users to label and organize items. Tags have been used in various web application areas, for example del.icio.us, CiteULike, Amazon.com, and LibraryThing. Folksonomy has the distinctive advantages of: being given by users explicitly and proactively, reflecting users’ topic preferences and personal viewpoints on item descriptions or classifications, and being relatively small in terms of data size. In addition to users’ click streams and browsing history, tags are becoming another important new implicit rating information source to profile users’ interests. They can be used to make personalized recommendations [29], improve personalized searching [2] and generate user and item clusters [16].

However, since there is no restriction or boundary on selecting words for tagging items, the tags used by users is free-formed, lack of standardization, and contain a lot of ambiguity. These problems resulted in low information sharing and inaccurate user profiling. Moreover, both the items and tags follow the power law distribution, which means only a small number of items or tags are being used by a large proportion of users while the majority of items or tags are only used by a small number of users [16]. The free-formed vocabulary and the power law dis-
tribution of items and tags bring challenges to the task of generating a proper neighborhood for making recommendations through calculating the tag and item overlaps, resulting in low recommendation performances.

On the other hand, each item itself is directly or indirectly associated with a set of controlled vocabulary such as item's keywords, taxonomy or ontology [8] [10] [14]. Item taxonomy is a set of controlled vocabulary terms designed to describe or classify items. Item taxonomy/ontology is widely available for various domains. Some typical item taxonomies include product classification taxonomy of Amazon.com (http://www.amazon.com), ACM Computing Reviews (http://www.reviews.com), Google Directory (http://directory.google.com), and Yahoo Directory (http://www.yahoo.com). Library of Congress Subject Headings [18] (http://www.loc.gov) and WordNet (http://wordnet.princeton.edu) are popularly used world knowledge ontology. Because item taxonomy is usually designed and developed by experts, it reflects the common viewpoints to the description and classification of items. Item taxonomy provides not only a standard vocabulary but also a hierarchical structure to represent the relationships among concepts or categories. Thus, item taxonomy can be used to reduce the influences caused by the free-formed vocabularies of tags.

In this paper, we propose to integrate tags that were contributed by users and item taxonomy that were developed by experts together to profile users' interests in order to make personalized recommendations. Section 2 will briefly review the related work. In Section 3, firstly some notations and definitions used in this paper will be given. After that, we will describe a novel method to find the semantic representation of each tag. Then, the approaches of user profiling, neighborhood formation, and recommendation making will be discussed. In Section 4, we will discuss the experimental results and evaluations. Finally, the conclusion will be given in Section 5.

2. Related Work

Recommender systems have been an active research area for more than a decade. Typically, users' explicit ratings are used to represent users' interests or preferences. Based on users' interests or preferences, similar users or items can be found to make recommendations. However, because users' explicit ratings are not always available, the recommendation techniques based on users' implicit ratings have drawn more and more attention recently [1] [30]. The tasks of recommender systems include rating prediction and top N recommendation. The former task is to predict the rating value a user will give to a rated item while the latter one is to recommend unrated/new items to the target user [1]. The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are widely used to measure the accuracy of the rating prediction task while precision and recall are commonly used for the top N recommendation task. For explicit ratings, both tasks are applicable while for implicit ratings, the top N recommendation is more applicable [1] [33]. Recommender systems can be broadly classified into three categories: content-based, collaborative filtering (CF), and hybrid approaches [6].

The content based approaches are mainly based on the contents of items such as keywords, taxonomic/ontology topics or categories/genes. Ziegler [36] proposed an approach to convert the implicit item ratings to the item taxonomic topic preferences. The term vector model and latent semantic topic model, such as latent Dirichlet allocation (LDA) and Probabilistic Latent Semantic Indexing (PLSI) are popularly used to process large textural corpus to recommend the most relevant items to users [1]. The collaborative filtering approach can be classified into memory based and model based approaches. The user and item based K nearest neighborhood (KNN) based approaches are two kinds of memory based CF approaches. More recently, the model based CF approaches such as matrix factorization techniques [15] get better performances for the rating prediction task. This approach works well on large scaled explicit rating dataset such as Netflix dataset. But how to use matrix factorization approaches to recommend top N unrated items to the target user and how to apply them to implicit ratings still remain open research questions [15]. Therefore, for implicit ratings, the memory based CF approaches are still popularly used. The hybrid approaches that combining the CF and the content based approaches have been applied in many applications [1] [3].

The recommendation techniques based on user's implicit ratings mainly focused on the web log analysis of the users' usage information or navigated web content information [1]. For example, web mining on navigation patterns, click streams, browse history and purchase record etc. In Web 2.0, the user created contents such as tags, blogs, reviews becomes important implicit rating information to profile users' interests and preferences [34][28]. Compared with
web logs, these kinds of new user information have the advantages of lightweight, small sized, and being explicitly and proactively provided by users [22].

Recently, social tags are becoming an important research focus. The research of tag based recommender systems mainly focuses on how to recommend tags to users. The problem of tag recommendation can be described as given a target user and a set of items, how to recommend tags to a set of items for the user [13]. Some approaches proposed to use the co-occurrence of tags [16], association rules [12], folkrank [13] and tensor [20] to make tag recommendations. Since the task of recommending a tag to a user with the purpose of labeling an item is different with the task of recommending an item to a user, the tag recommendation approaches usually cannot be directly used to recommend items [23].

Currently, not so much work has been done on the item recommendations based on social tags. Since tagging is a kind of implicit rating behavior [23] [27] and tags are pieces of textural information describing the content of items, the memory based CF and content based approaches are mainly used. Diederich [9] proposed an exploratory user based CF approach that adopted the tag based user profiles. The tf-idf weighting approach that is similar to the tf-idf approach in text mining was used for each user’s tags. The work of Tso-shuter [29] extended the binary user-item matrix to binary user-item-tag matrix and used the Jaccard similarity measure approach to generate user neighborhood in order to make item recommendations. It was claimed that tag information failed to significantly improve the accuracy of memory based CF approaches due to the tag quality problem [29]. In our previous work [17], we proposed a recommendation approach based on the derived user-item, user-tag and tag-item sub matrices. But the performances still need to be improved.

More recently, the noise of tags or the quality [24] and usefulness [4] of tags arouses attentions. Some content based approaches that deal with the noise of textural contents were proposed. In the work of Niwa [19] and Shepitsen [25], the clustering approaches were used to find the item and tag clusters based on the tag-based tf-idf content representations. The tag content mapping between each user and item clusters was used to recommend web pages. The Latent Semantic Analysis such as PLSI [11] [32] and LDA [26] based approaches have been proposed to reduce the noise of tags and build latent semantic topic models to recommend items to users. Besides these memory based CF approaches and content filtering models, in the work of Sen [23], a special tag rating function was used to find users’ preferences for tags. Moreover, along with the tag preferences, the click streams, tag search history of each user were also used to get users’ preferences for items through the inferred tag preferences. Since some extra information and special functions were needed, it makes Sen’s work incomparable and gives restrictions to the applications of this work. Zhen [35] proposed to integrate tag information and explicit ratings to improve the accuracy of rating predictions of a model based CF approach.

In this paper, we propose an approach to integrate social tags and item taxonomy to make better recommendations. Although, some work such as Ziegler’s work [36] and Weng’s work [31] has discussed how to combine item taxonomy with users’ implicit ratings to make better recommendation, they didn’t consider tag information.

3. The Proposed Approach

3.1. Notations

For easy describing the proposed approach, we define some concepts and entities as below. In this paper, topics, concepts, categories and nodes are interchangeably used.

- **Users:** $U = \{u_1, u_2, \ldots, u_{|U|}\}$ contains all users in an online community who have used tags to organize items.
- **Items (or Products, Resources):** $P = \{p_1, p_2, \ldots, p_{|P|}\}$ contains all items tagged by users in $U$. Items could be any type of online information resources or products in an online community such as web pages, videos, music tracks, photos, academic papers, books etc. We assume that each item can be described by a set of tags given by users and a set of item taxonomic topics given by experts.
- **Tags:** $T = \{t_1, t_2, \ldots, t_{|T|}\}$ contains all tags used by the users in $U$. A tag is a piece of textural information given by users to label or collect items.
- **Item Taxonomy:** $\mathcal{O} = \langle C, R \rangle$. $C$ is a set of item taxonomic topics and $R$ is a set of relations between any $c_x \in C$ and $c_y \in C$. If $R$ is empty, then $\mathcal{O}$ consists of only a set of taxonomic topics and no relationships exist in any pair of topics. In this work, only the typical “is-a” relationship is considered. The “is-a” relationship in this paper
is the “sub topic of” relationship between two topics, denoted as $<$. For any two taxonomic topics $c_x, c_y \in C$, if $c_x < c_y$, then $c_x$ is a sub topic of $c_y$. The taxonomic topic $c_y$ expresses general or broad concepts, whereas $c_x$ expresses specific and narrow concepts. The taxonomy tree has exactly one root topic that represents the most general topic. The leaf topics that do not have any direct sub topics represent the most specific topics.

- **Item taxonomic descriptors**: Every item $p_k \in P$ is associated with a set of taxonomic descriptors $D_{p_k} = \{d_1, d_2, ..., d_g\}$. A taxonomic descriptor is a set of ordered topics, denoted by $d_i = \{c_{o_i}, c_{y_i}, ..., c_{y_i}\}$, where $d_i \in D_{p_k}$, $p_k \in P$, $c_{o_i}, c_{y_i}, ..., c_{y_i} \in C$, $c_{o_i}$ is the root topic, $c_{y_i}$ is a leaf topic with no sub topics, and $c_{o_i} < ... < c_{x_i} < c_{y_i} < c_{o_i}$. An item can have multiple descriptors because the item might possess a broad range of concepts. Strictly categorising the item under one single concept might be imprecise.

Figure 1 shows an example of an item taxonomy tree. The taxonomy tree has nine taxonomic topics $c_0, c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8$. Within the item taxonomy depicted in this figure, “book” is the root topic covering the broadest concept while “programming” and “flowers” are two of the leaf topics expressing the most specific concepts. There are six unique item taxonomic descriptors in the taxonomy tree: $d_1 = \{c_0, c_1, c_4\}$, $d_2 = \{c_0, c_2, c_3\}$, $d_3 = \{c_0, c_3, c_4\}$, $d_4 = \{c_0, c_3, c_7\}$, $d_5 = \{c_0, c_3, c_8\}$, $d_6 = \{c_0, c_2\}$.

In this paper, we focus on the top $N$ item recommendation task. Let $u_i \in U$ be a target user, $P_{u_i}$ be the set of items tagged by the user $u_i$ already has, $p_k \in P - P_{u_i}$ be a candidate item, $\mathcal{A}(u_i, p_k)$ be the prediction score of how much user $u_i$ would be interested in the item $p_k$. The problem of item recommendation is defined as generating a set of rank-ordered items $p_1, ..., p_m \in P - P_{u_i}$ to the use $u_i$, where $\mathcal{A}(u_i, p_j) \geq ... \geq \mathcal{A}(u_i, p_m)$.

**Example 1** Suppose there are three users $u_1, u_2, u_3$ who used two tags $t_1, t_2$ and tagged six items $p_1, p_2, p_3, p_4, p_5, p_6$. Assume that user $u_1$ tagged item $p_1$ and $p_2$ with tag $t_1$ “apple”. With the same tag $t_1$, user $u_2$ has tagged item $p_3$ and $p_4$. User $u_3$ tagged item $p_5, p_6$ with tag $t_2$ “0403”. The taxonomic topic descriptors of the items $p_1, p_2, p_3, p_4, p_5, p_6$ are defined as: $D_{p_1} = \{d_1\}, D_{p_2} = \{d_2\}, D_{p_3} = \{d_3, d_4\}, D_{p_4} = \{d_3\}, D_{p_5} = \{d_3\}, D_{p_6} = \{d_3, d_5\}$, respectively. For example, item $p_3$ is associated with topic descriptor $D_{p_3} = \{d_3, d_4\}$, where $d_3 = \{c_0, c_3, c_7\}$, $d_4 = \{c_0, c_3, c_7\}$. Then, item $p_3$ is described by taxonomic topics (“book”, “computers”, “programming”) and (“book”, “computers”, “networks”).

### 3.2. The Semantic Representations of Tags

As mentioned in Introduction, user defined tags are free-formed and lack of standardization. As uncontrolled vocabulary, social tags suffer from many problems such as ambiguity in the meanings of terms, a proliferation of synonyms, varying levels of specificity, and lack of guidance on syntax and slight variations of spelling and phrasing [4]. The tag quality problem resulted in inaccurate user profiling and low information sharing.

To solve this problem, in this section, we propose an approach to extract the semantic meaning of a tag based on the taxonomic descriptors of the items of that tag. In a tag, a set of items are gathered together according to the user’s viewpoint. We believe that there must be some correlation between the user’s tag and the categories/topics of the items in that tag. Otherwise the user would not label the items with that tag. Thus, by combining the tags and the taxonomy of the items in the tags, we can derive a set of item taxonomic topics along with the structural relationship among them to represent the semantic meanings of each tag.
For each user $u_i \in U$, with tag $t_j$, a set of items has been collected by the user $u_i$, which is denoted as $P_{ui,t_j}$. Let $D_{pk}$ be the set of item taxonomic descriptors associated with item $p_k$. From $\bigcup_{p_k \in P_{ui,t_j}} D_{pk}$, we can get a set of item taxonomy descriptors for tag $t_j$ for user $u_i$. Apparently, the taxonomic topics in $\bigcup_{p_k \in P_{ui,t_j}} D_{pk}$ reflect the experts’ viewpoint to the classification of the items in $P_{ui,t_j}$, while the tag $t_j$ represents user $u_i$’s opinion about the items’ classification. The taxonomic topics are more standard. In order to reduce the noise of tags, we could use the standard taxonomic topics of the items to represent the semantic meanings of each tag for each user individually.

Before we discuss how to measure the relevance strength of a tag to a taxonomic topic in terms of an individual user, we firstly discuss how to measure the relevance of an item to a taxonomic topic.

### 3.2.1. Item representation

As mentioned above, each item can be described with a set of item taxonomic descriptors. An item taxonomic descriptor is a set of ordered taxonomic topics. It is a full path from the root topic node to a leaf topic node. With the “sub topic of” relationship, the taxonomic topics are structured hierarchically. The hierarchical structure of taxonomy imposes important information for finding the relevance between an item and a taxonomic topic. The process of determining the relevance between taxonomic topics and each item is to generate a representation for each item that can represent the semantic meaning of that item with taxonomic topics. This representation is called taxonomy based item representation. It is defined below.

[Definition 1] (Item Representation): For each item $p_k \in P$, $p_k$ can be represented by taxonomic topics together with the relevance of each taxonomic topic $c_y \in C$ to $p_k$. Let $w_{k,y}^P$ denote the weight of how much the taxonomic topic $c_y$ is relevant to the item $p_k$, the relationship between the item and taxonomic topics can be defined as the mapping $C^p: P \rightarrow 2^{C \times [0,1]}$, such that $C^p(p_k) = \{(c_y, w_{k,y}^P)|c_y \in C\}$. $C^p(p_k)$ is called the representation of item $p_k$.

The key part of generating item representation of item $p_k$ is to calculate the value of $w_{k,y}^P$ that measures the relevance weight of a taxonomic topic $c_y$ and an item $p_k$. A commonly used approach is to use the frequency of each taxonomic topic in item taxonomic descriptors to measure the weight of a taxonomic topic. However, taxonomic topic nodes at higher levels in the taxonomy tree reflecting general concepts usually appear more frequently in item taxonomic descriptors than those at lower levels reflecting specific concepts. Therefore, the structural information of the whole taxonomy tree should be taken into consideration when calculating the weight value of a taxonomic topic. The following factors should be considered to determine the weight of a taxonomic topic $c_y \in C$ for the representation of an item $p_k \in P$:

- The frequency of a taxonomic topic. If taxonomic topic $c_y$ appears more frequently in the descriptors $D_{pk}$ of item $p_k$ than other taxonomic topics, then $c_y$ should have a higher weight value than the taxonomic topics which occur less frequently in $D_{pk}$.
- The semantic coverage of a taxonomic topic. The taxonomic topics that express more specific concept are more useful for identifying the feature of an item. If taxonomic topic $c_y$ expresses specific concepts compared to other taxonomic topics, then $c_y$ should have a higher weight value than taxonomic topics expressing general concepts.
- The structural information between one
taxonomic topic and another. Based on the specified direct "sub topic of" relationship among taxonomic topics, the inferred relationship among taxonomic topic nodes includes "child", "parent", "sibling", "grandparent", "grandchild", "ancestor" and others. The number of children, the number of siblings and the number of ancestors of $c_y$ in the taxonomy tree can affect the importance of taxonomic topic $c_y$ for the feature representation of item $p_k$ [31].

By taking these three factors into consideration, Ziegler et al. [36] proposed to decay the weight of a taxonomic topic node based on the number of children of a taxonomic topic node in the item taxonomy tree and the length of an item taxonomic descriptor. Inspired by the approach of Ziegler et al. [36], this paper proposes an approach that takes the structural information of item taxonomy into consideration when calculating the weight of a taxonomic topic for an item. The weight computation is conducted in a bottom up way. It is discussed as below.

Let $d_j \in D_{p_k}$ be an item taxonomic descriptor of item $p_k \in P$. $f(c_y, d_j)$ denote the weight of taxonomic topic $c_y \in C$ to item taxonomic descriptor $d_j$. Suppose the item descriptor $d_j = \{c_0, c_y, c_x, c_a\}$ and $c_a < c_x < c_y < c_0$. The calculations of non-leaf and leaf taxonomic topics are as below:

- **The calculation of non-leaf taxonomic topics.** For the non-leaf taxonomic topic $c_y \in C$ in the example descriptor $d_j$ given above, $f(c_y, d_j)$ can be calculated as:

$$f(c_y, d_j) = \frac{f(c_x, d_j)}{ch(c_y)}$$  

(1)

Where taxonomic topic $c_y \in C$ is the parent node of taxonomic topic $c_x \in C$ in item descriptor $d_j$. $ch(c_y)$ is the number of child nodes of taxonomic topic $c_y$. If $c_y \in C$ is not a taxonomic topic in item taxonomic descriptor $d_j, c_x \notin d_j$, then $f(c_y, d_j) = 0$.

- **The calculation of leaf taxonomic topics.** The total weight of all topics in one item taxonomic descriptor can be set to a positive number [36]. To facilitate comparison, the total weight of all the topics in $d_j$ is set to 1. Let $x$ be the weight of the leaf node $c_a$ of the example descriptor $d_j$. the following equation can be obtained:

$$x + \frac{x}{ch(c_2)} + \frac{x}{ch(c_2)ch(c_y)} + \frac{x}{ch(c_2)ch(c_y)ch(c_0)} = 1$$  

(2)

After solving Eq. (2), the value of $x$ (i.e., $f(c_a, d_j)$) can be calculated. Based on the leaf node weight $f(c_a, d_j)$ and Eq. (1), the weight value of each non-leaf topic in $d_j$ can be calculated. Leaf nodes have a higher weight value than those of non-leaf nodes calculated using Eq. (1) and Eq. (2).

However, if a taxonomic topic is popularly used to describe items, it is not a distinctive topic that represents the item. Thus, similar to the idf weighting approach in text mining, we propose to take the popularity of a topic for all items into consideration to measure the importance of a topic to a specific item. Let $c_y$ be a topic, $|P|$ be the total number of items, $idf(c_y)$ is defined as the inverse item frequency of tag $c_y$. Usually, $idf(c_y) = |P|/\log(|P_y|)$, where $|P_y|$ is the number of items that have been described with $c_y$ in the item set $P$. To get a value between 0 and 1 to facilitate comparison, we set $idf(c_y) = 1/\log(e + |P_y|)$, where $e$ is an irrational constant approximately equal to 2.72 and $0 < idf(c_y) \leq 1$.

Assuming each descriptor is equally important for the topic classification and description of item $p_k$, this paper uses the average value of $f(c_y, d_j)$ of all item descriptors in $D_{p_k}$ to measure the overall relevance weight of item $p_k$ to taxonomic topic $c_y$. Let $|D_{p_k}|$ denotes the number of descriptors of item $p_k$. The relevance weight $w_{k,y}$ can be calculated as:

$$w_{k,y} = \frac{1}{|D_{p_k}|} \sum_{d_j \in D_{p_k}} f(c_y, d_j) \cdot idf(c_y)$$  

(3)

Since the mapping $C^p(p_k)$ can be viewed as a vector: $C^p(p_k) = <w_{k, p_1}, \ldots, w_{k, |C| - 1}>$ for taxonomic topics $<c_{y_1}, \ldots, c_{|C| - 1}>$, each item $p_k$ can be described by a $|C|$-sized taxonomic topic vector $C^p(p_k)$. The values $w_{k,y}$ can be calculated by Eq. (3).

**Example 2** (Item Representation) Figure 2 shows an example of the taxonomy based item representations of $p_3$. For example, the relevance weight of item $p_3$ to taxonomic topic $c_6$, $w_{3,6}$ is shown as follows.

As defined in Example 1, there are two descriptors $d_3$ and $d_4$ for item $p_3$. $D_{p_3} = \{d_3, d_4\}$. $w_{3,6} = \frac{1}{|D_{p_3}|} \sum_{d_j \in D_{p_3}} f(c_6, d_j) \cdot idf(c_6)$ = $\frac{1}{2} \cdot f(c_6, d_3) + f(c_6, d_4) \cdot idf(c_6)$. As shown in the taxonomy tree in Figure 1, $ch(c_3) = 3$, $ch(c_6) = 3$. If $f(c_6, d_3) = x$, then based on Eq. (2), $x + \frac{x}{3} + \frac{x}{3} = 1$. After solving this equation, $x = 0.69$. Since descriptor $d_4$ does not contain taxonomic topic $c_6$, $f(c_6, d_4) = 0$. As $c_6$ has
been used to describe items $p_3$, $p_5$ and $p_6$, $i.e., c_6 = \frac{1}{\log(e+3)} = 0.57$. Thus, $w^{p}_{3,6} = \frac{1}{2} (0.69 + 0) \cdot 0.57 = 0.197$.

The taxonomy based item representation of $p_3$ is: $\mathcal{E}^p(p_3) = \{ (c_0, 0.035), (c_3, 0.12), (c_6, 0.197), (c_7, 0.26) \}$. Item $p_3$ is related to taxonomic topics $c_3$ “computers”, $c_6$ “programming” and $c_7$ “networks”. 

### 3.2.2. Tag representation

For a given user $u_i$ and a tag $t_x$, the strength of a topic $c_y$ being related to the tag $t_x$ for the user $u_i$ could be estimated based on the relevance weight of $c_y$ to the items collected in the tag $t_x$ of the user $u_i$. Based on Eq. (3), we can measure the relevance of $c_y$ to an item $p_k$. Let $P_{u_i,t_x} = \{ p_{1, i}, p_{2, i}, \ldots, p_{m, i} \}$, we could use any of $w^P_{1, y}, w^P_{2, y}, \ldots, w^P_{m, y}$ to estimate the relevance of $c_y$ to the tag $t_x$ for user $u_i$. That means, we can estimate the relevance of any taxonomic topic to the tag $t_x$. Thus, each tag can be represented by taxonomic topics together with the relevance of each topic to the tag. Similar to item representation, tag representation can be defined as follows:

**[Definition 2] (Tag Representation):** For each tag $t_x \in \mathcal{T}$ and a user $u_i$, $t_x$ can be represented by taxonomic topics together with the relevance of each taxonomic topic $c_y \in \mathcal{C}$ to the tag $t_x$ with respect to the user $u_i$. Let $s_{u_i,t_x}(c_y)$ denote how strong $t_x$ is related to $c_y$ with respect to user $u_i$, the relationship between a tag and a set of taxonomic topics for the user $u_i$ can be defined as the mapping $\mathcal{E}^t; U \times \mathcal{T} \rightarrow 2^{\mathcal{C} \times [0,1]}$, such that $\mathcal{E}^t(u_i, t_x) = \{ (c_y, s_{u_i,t_x}(c_y)) | c_y \in \mathcal{C} \}$. $\mathcal{E}^t(u_i, t_x)$ is called the representation of tag $t_x$ with respect to the user $u_i$.

Assuming that $w^P_{1, y}, w^P_{2, y}, \ldots, w^P_{m, y}$ are equally important to the user $u_i$ to calculate the relevance of $c_y$ to $t_x$, in this paper, we propose to use the average value of $w^P_{1, y}, w^P_{2, y}, \ldots, w^P_{m, y}$ to estimate the relevance of $c_y$ to $t_x$. Let $s_{u_i,t_x}(c_i)$ denote the relevance weight of the taxonomic topic $c_i$ and the tag $t_j$ in terms of the user $u_j$, it can be calculated as:

$$s_{u_i,t_x}(c_i) = \sum_{p_k \in P_{u_i,t_x}} \frac{w^P_{i, k}}{|P_{u_i,t_x}|}$$

Thus, the free-formed social tags that given by users can be converted to a set of standard and relatively small sized item taxonomic topics given by experts, which can reduce the differences of user tag vocabularies, incorrect syntax and spelling and semantic ambiguity.

**[Example 3] (Tag Representation)*** Figure 3 shows an example of the taxonomy based representations of tag $t_1$ “apple” for user $u_1$ and user $u_2$. For example, the calculation of the relevance of tag $t_1$ and taxonomic topic $c_6$ in terms of user $u_2$ can be calculated as: $s_{u_2,t_1}(c_6) = \sum_{p_k \in P_{u_2,t_1}} \frac{w^P_{k, 6}}{|P_{u_2,t_1}|} = \frac{w^P_{1, 6} + w^P_{2, 6}}{2}$. Based on Eq. (3), $w^P_{1, 6} = 0.197$, $w^P_{2, 6} = 0$. As a result, $s_{u_2,t_1}(c_6) = \frac{0.197 + 0}{2} = 0.099$. The relevance weight of tag $t_1$ and taxonomic topic $c_6$ in terms of user $u_1$ can be calculated as: $s_{u_1,t_1}(c_6) = \sum_{p_k \in P_{u_1,t_1}} \frac{w^P_{k, 6}}{|P_{u_1,t_1}|} = \frac{w^P_{1, 6} + w^P_{2, 6}}{2} = 0$.

The tag representations of tag $t_1$ for user $u_1$ and $u_2$ are: $\mathcal{E}^t(u_1, t_1) = \{(c_0, 0.035), (c_3, 0.147), (c_4, 0.175), (c_5, 0.175)\}$, $\mathcal{E}^t(u_2, t_1) = \{(c_0, 0.035), (c_3, 0.12), (c_6, 0.099), (c_7, 0.13), (c_8, 0.22)\}$. For user $u_1$, the tag $t_1$ “apple” is related to taxonomic topics $c_1$ “garden”, $c_4$ “flowers” and $c_6$ “fruit”. Whereas, for user $u_2$, it is related to $c_3$ “computers”, $c_6$ “programming”, $c_7$ “networks” and $c_8$ “databases.

We can see that the semantic meaning of a tag can be generated distinctively for different users and thus the semantic ambiguity can be reduced. The related taxonomic topics of each personal tag (e.g., $t_2$ “0403”) can also be determined.
3.3. User Profile Generation

User profile is used to describe user’s interests and preferences information. Usually, a user-item rating matrix is used in collaborative filtering based recommender systems to profile a user’s item interests and preferences. Due to the long tail of items, the size of user-item matrices is usually very big but the overlap of preferred items between users is very low, which makes it difficult to find similar users.

One effective way to improve information sharing is to find users’ common information topic interests besides the common item ratings or item preferences. Some approaches have been proposed to generate user’s taxonomic topics through converting the user-item rating vector into user-taxonomic topic vector [36]. However, these approaches did not consider the tag information. We believe that tag information well reflects user’s topic interests and preferences and should be used to profile users.

In this paper, we profile each user with his/her item preferences and topic preferences as well. Since there is no explicit rating information available for typical tagging communities [29], the binary implicit ratings are used to represent each user’s item preferences. How to obtain each user’s topic preferences is the major focus of this sub section. The topic preferences of each user are represented by a set of taxonomic topics with their weights, or called user representation, which is defined as below:

**Definition 3 (User representation):** For each user $u_i \in U$, $u_i$ can be represented by taxonomic topics together with the preference to each taxonomic topic $c_y \in C$ by the user $u_i$. Let $w_{iy}^c$ denote the weight of how much the user $u_i$ is interested in the taxonomic topic $c_y$, the relationship between a user and a set of taxonomic topics can be defined as the mapping $C^u_i : U \rightarrow 2^{C \times [0,1]}$, such that $C^u_i(u_i) = \{(c_y, w_{iy}^c) | c_y \in C\}$. $C^u_i(u_i)$ is called the user representation of user $u_i$.

To calculate how much $u_i$ will be interested in $c_i$, we firstly calculate how much the user is interested in a tag $t_x$ which has been used by the user. For each user $u_i$, we define the probability of user $u_i$ tagging items as the ratio between the number of items that are tagged by the user and the total number of items, that is $\Pr(u_i) = \frac{|P|}{|P|}$, where $|P|$ is the number of items that user $u_i$ has tagged and $|P|$ is the total number of items. $\Pr(u_i) = 0$ if the user never tagged any items, $\Pr(u_i) = 1$ if the user has tagged all the items. The probability of user $u_i$ tagging items using a specific tag $t_x$ denoted as $\Pr(u_i,t_x)$ is defined as the ratio between the number of items that are tagged by the user using tag $t_x$ and the total number of items. $\Pr(u_i,t_x) = \frac{|P_{t_x}|}{|P|}$, where $P_{t_x}$ denotes the items that are tagged by user $u_i$ with the tag $t_x$. Based on the two probabilities above, we can calculate how much user $u_i$ is interested in $t_x$, which is defined as below.

$$\Pr(t_x | u_i) = \frac{\Pr(u_i,t_x)}{\Pr(u_i)} = \frac{|P_{t_x}|}{|P|}$$

Based on Eq. (4), we can get the relevance weight $s_{u_i,t_x}(c_y)$ between the tag $t_x$ and the topic $c_y$ for user $u_i$. Similar with the item representation, we also take the occurrence of a topic (i.e., topic popularity) for all users into consideration to measure the general importance of a tag in the identification of the topic preference of a user. Let $c_y$ be a topic, $iuf(c_y)$ is defined as the inverse user frequency of topic $c_y$. Similar with $iuf(c_y)$, we set $iuf(c_y) = 1/\log(e + 0.035)$.
By taking the inverse item frequency into consideration, the weight of a user's preference to a taxonomic topic can be calculated with the equation below.

$$w_{i,j}^t = \sum_{t \in T} \Pr(t_x | u_i) \cdot s_{u_i}(c_x) \cdot iuf(c_x) \quad (6)$$

The mapping $C^u(u_i)$ can be viewed as a vector: $C^u(u_i) = \langle w_{u_i}[0], \ldots, w_{u_i}[|C|-1] \rangle$ for taxonomic topics $\langle c_0, \ldots, c_{|C|-1} \rangle$. Therefore, each user $u_i$ can be profiled by two vectors: $\varrho^u(u_i)$ and $C^u(u_i)$, where $\varrho^u(u_i)$ is a $|P|$-sized binary item vector representing $u_i$’s item preferences and $C^u(u_i)$ is a $|C|$-sized taxonomic topic vector representing $u_i$’s topic preferences.

**[Example 4]** (User Representation) Figure 4 shows an example of user representations of user $u_3$. For example, the calculation of user $u_3$’s preferences to taxonomic topic “programming” is:

$$w_{3,6}^t = \sum_{t \in T} \Pr(t_x | u_3) \cdot s_{u_3}(c_6) \cdot iuf(c_6) = \Pr(t_2 | u_3) \cdot s_{u_3}(c_6) \cdot iuf(c_6).$$

$$\Pr(t_2 | u_3)=1.$$ Based on Eq. (4), $s_{u_3}(c_6)=0.32$. According to the item taxonomy shown in Figure 1 and the item description defined in Example 1, the user set of $c_6$ contains $t_2$ and $t_3$. Therefore, $iuf(c_6)=\frac{1}{\log(\epsilon+2)} = 0.64$ and $w_{3,6}^t=1 \cdot 0.32 \cdot 0.64=0.2$.

The taxonomy based user representation of $u_3$ is: $C^u(u_3)=[(c_0, 0.021), (c_3, 0.07), (c_6, 0.2), (c_9, 0.06)]$. Although user $u_3$ used a personal tag $t_2$ “0403” to represent this user’s topic preferences, after the taxonomy based user representation, $u_3$ is actually interested in taxonomic topics $c_6$ “programming” and $c_3$ “computers”.

### 3.4. Neighborhood Formation

Neighborhood formation is to generate a set of like-minded peers for a target user $u_i$ or a set of similar peer items for an item $p_i$. The “$K$-Nearest-Neighbors” technique is used to select the top $K$ neighbors with shortest distances to $u_i$ or $p_i$, through computing the distances between $u_i$ and all other users or the distances between $p_i$ and all other items. The more accurate a user profile or item representation is, the more similar neighbor users or items will be found. The distance or similarity measure can be calculated through various kinds of proximity computing approaches such as cosine similarity and Pearson correlation. Cosine similarity is popularly used to calculate the similarity of two vectors. Since a vector of topics with their correspondent weights is used to represent each item and the topic preferences of each user in this paper, the topic similarity of each user pair, and the topic similarity between an item and a user can be measured through calculating the similarity of their topic vectors. For any two topic vectors $v_i$ and $v_j$, the cosine similarity is defined as:

$$\cos(v_i, v_j) = \frac{\sum_{y=1}^{|C|} v_{iy}v_{jy}}{\sqrt{\sum_{y=1}^{|C|} v_{iy}^2}(\sum_{y=1}^{|C|} v_{jy}^2)} \quad (7)$$

Since each user is profiled with item preference and topic preference as well, the similarity of two users $u_i$ and $u_j$ includes two parts: the similarity of topic preferences and the similarity of item preference. Cosine similarity is used to measure the similarity of topic preferences between two users.

$$\text{sim}^C(u_i, u_j) = \cos(C^u(u_i), C^u(u_j)) \quad (8)$$

To measure the similarity of item preferences with implicit binary ratings, a simple approach is to count the overlap of commonly rated items between two users [5]. Since the approach of weighting each
commonly rated item with inversed user frequency or \textit{iuf} [5] takes the user frequency of item into account, it performs better for binary ratings in many cases [5]. We use this \textit{iuf} approach to calculate the similarity of item preferences of two users, which is defined as below.

\[
sim^p(u_i, u_j) = \frac{\sum_{p_k \in \mathcal{P}_u \cap \mathcal{P}_v} \text{iuf}(p_k)}{|\mathcal{P}_u| |\mathcal{P}_v|}\] (9)

Thus, the similarity of two users is defined as:

\[
sim(u_i, u_j) = \eta \cdot \sim^p(u_i, u_j) + (1 - \eta) \cdot \sim^c(u_i, u_j) \] (10)

Where \(0 \leq \eta \leq 1\). Similar with the similarity measure of the users’ topic preferences, using the similarity measure approach, we can generate the neighborhood of the target user \(u_i\), which includes \(K\) nearest neighbor users who have similar user profiles with \(u_i\). The neighborhood of \(u_i\) is calculated as \(\mathcal{N}(u_i) = \{ u_j | u_j \in \arg \max_{u \in \mathcal{U}} \{ \text{sim}(u_i, u) \} \} \).

### 3.5. Recommendation Generation

Typically, from the generated neighborhood, a set of items that are not rated/tagged by the target user but most frequently rated/tagged by the neighbor users will be recommended to the target user. Since the topics of items and the topic preferences of users can be represented by item taxonomic topics, the topic similarity between the target user and the candidate item can be used to improve the accuracy of recommendations through selecting those items that are not only rated by the most similar users, but also have similar topics with the target user. With the topic matching approach, it makes the collaborative filtering approach take the benefits of content based recommendation approaches. It is a hybrid recommendation approach [3].

For each target user \(u_i\), a set of candidate items will be generated from the items tagged by \(u_i\)’s neighbourhood formed based on the similarity of user profiles, which is denoted as \(\varphi(u_i)\), \(\varphi(u_i) = \{ p_k | p_k \in \mathcal{P}_u, u_j \in \mathcal{N}(u_i), p_k \notin \mathcal{P}_u \} \), where \(\mathcal{P}_u\) is the rated/tagged item set of user \(u_j\). Let \(\mathcal{U}_{p_k}\) denote the user set of item \(p_k\), for each candidate item \(p_k \in \varphi(u_i)\), \(\mathcal{N}(u_i) \cap \mathcal{U}_{p_k}\) is the sub set of users in \(\mathcal{N}(u_i)\) who have tagged the item \(p_k\). The prediction score of how much \(u_i\) may be interested in \(p_k\) is calculated by considering two aspects: the similarity of users and the similarity of topic matching. We use the simple linear combination to hybrid the two parts. With Eq. (11), the similarity of two users can be measured. Let \(\text{sim}^m(u_i, p_k)\) denote the topic mapping between the target user \(u_i\) and the candidate item \(p_k\), the cosine similarity can be used to calculate the mapping, which is defined as below:

\[
\text{sim}^m(u_i, p_k) = \cos(c_u(u_i), c_p(p_k)) \] (11)

Thus, the prediction score \(\mathcal{A}(u_i, p_k)\) for each candidate item \(p_k \in \mathcal{P}(u_i)\) can be calculated as below:

\[
\mathcal{A}(u_i, p_k) = u_j \in \mathcal{N}(u_i) \text{sim}^m(u_i, p_k) + (1 - \eta) \cdot \text{sim}(u_i, u_j) \] (12)

Where \(0 \leq \alpha \leq 1\). The top \(N\) items with high prediction scores will be recommended to the target user \(u_i\).

### 4. Experiments and Discussions

#### 4.1. Data preparation

We conducted the experiments using the Amazon dataset. This dataset was crawled from Amazon.com on April, 2008. The items of the dataset are books. In pre-processing, we removed the books that are used by less than 3 users or whose taxonomy descriptors are not available. The final dataset comprises 4,112 users, 34,201 tags, 30,467 books. The book descriptors were also obtained from Amazon.com. The taxonomy formed by these descriptors is tree-structured and contains 9,919 unique topics.

#### 4.2. Experiments setup

We conducted the experiments using the dataset obtained from Amazon.com. The items of the dataset are books. In pre-processing, we removed the books that are only used by one user or whose taxonomy descriptors are not available. The final dataset comprises 5,177 users, 37,120 tags, 31,724 books and 242,496 records. The book descriptors are also obtained from Amazon.com. The taxonomy formed by these descriptors is tree-structured and contains 9,919 unique topics.

The precision and recall are used to evaluate the recommendation performance. The whole dataset is split into a training and test set. The split percentage is 50% each, and 50% of the items of each user were hidden as the test/answer set while 50% of each user’s items are used as his/her training set. The training set of each user contains this user’s items and
corresponding tags and taxonomy descriptors as well. For each test user, the recommender system will generate a list of ordered items that the test user did not collect. The top $N$ items with high prediction scores will be recommended. If an item in the recommendation list was in the test user's hidden item list, then the item was counted as a hit. The average precision and recall of the whole test users of the dataset were used to measure the accuracy performance of the recommendations.

4.3. Comparisons

To evaluate the effectiveness of the proposed approach, we compared the precision and recall of the recommended top $N$ items produced by the following approaches:

- **Item-Tag-Taxonomic approach**, which is the proposed approach that combines implicit item rating and topic preferences generated through integrating tags and item taxonomy.
- **Tag-Taxonomic approach**, which is the proposed approach that only uses topic preferences generated from tags and item taxonomy.
- **Taxonomic approach**, which is proposed by Ziegler’s [36] that uses topic preferences generated from item taxonomy only.
- **Item-tag approach**, which is our previous work that uses three derived matrixes user-item, user-tag and tag-item sub matrixes to make recommendations [17].
- **Standard CF approach**, which is the standard collaborative filtering (CF) approach that uses the implicit item ratings or item preferences only. We use the improved approach that takes the $iuf$ value of each item into consideration to measure the similarity of two users [5].

To evaluate the performances of the approaches in different situations, we conducted the comparison experiments with two datasets Dataset 1 and Dataset 2. Dataset 1 is the whole dataset covering all users' information, which is to evaluate the effectiveness in normal situation. The average number of books that a user has is 16.73. Dataset 2 is to evaluate the effectiveness when the dataset is very sparse. We selected 1,000 users that each user has no more than 20 books. It includes 1,000 users, 4,893 books and 5,228 tags. The average number of books that a user has is 6.84.

4.4. Results and Discussions

The precision and recall results of Dataset 1 are shown in Figure 5 and Figure 6 respectively, while the precision and recall results of Dataset 2 are shown in Figure 7 and Figure 8 respectively.

![Figure 5 Precision evaluation of Dataset 1.](image)

![Figure 6 Recall evaluation of Dataset 1.](image)

![Figure 7 Precision evaluation of Dataset 2.](image)
Figure 8 Recall evaluation of Dataset 2.

From the comparison of the proposed Tag-Taxonomic approach that uses both social tags and item taxonomy and Taxonomic approach proposed by Ziegler [36] that only uses item taxonomy, we can see that the proposed approach outperforms the latter one, which means that the social tags are helpful to mine user’s actual topic interests and preferences. More importantly, the experimental results show that the proposed approach is effective, especially in sparse situation. It integrates social tags and item taxonomy together to reduce the inaccuracy caused by the free-style vocabularies of social tags, and the problem of low information sharing caused by the long tails of items and tags.

Moreover, we can see that the proposed Item-Tag-Taxonomic approach that combine item preferences and topic preferences performs better than the other approaches in both relatively dense and sparse situations. From Figure 5 and Figure 6, we can see that item preferences played more important part when more books each user has on average. The results in Figure 7 and Figure 8 show that in very sparse situation, it becomes difficult to find similar users based on users’ item preferences or overlaps of items. In this case, the topic preferences played a major role to make recommendations.

5. Conclusion

In this paper, we propose an approach of combining social tags and item taxonomy to make personalized recommendations. Firstly, we propose an approach to extract tags’ semantic meanings and represent them with taxonomy topics. Then, we propose an approach to generate users’ topic preferences based on users’ interests to tags. The information sharing among users was improved after converting the user-tag vector into much smaller sized and standard user-taxonomic topics vector. Also, we propose to measure user similarity based on users’ topic preferences that were generated through integrating the user contributed tags and expert designed item taxonomy. Finally, a hybrid recommendation generation approach is proposed to recommend items that not only preferred by the target user’s peer neighbors, but also similar to the target user’s preferred topics. The experimental results show that the proposed approach outperforms the standard collaborative filtering approach and Ziegler’s approach.

This research contribute to improving the recommendation accuracy of traditional recommender systems through incorporating social tags in Web 2.0.

References
