Mining Users’ Opinions based on Item Folksonomy and Taxonomy for Personalized Recommender Systems

Huizhi Liang  
Faculty of Science and Technology  
Queensland University of Technology  
Brisbane, Queensland, Australia  
oklianghuizi@gmail.com

Yue Xu  
Faculty of Science and Technology  
Queensland University of Technology  
Brisbane, Queensland, Australia  
yue.xu@qut.edu.au

Yuefeng Li  
Faculty of Science and Technology  
Queensland University of Technology  
Brisbane, Queensland, Australia  
y2.li@qut.edu.au

Abstract—Item folksonomy or tag information is a kind of typical and prevalent web 2.0 information. Item folksonomy contains rich opinion information of users on item classifications and descriptions. It can be used as another important information source to conduct opinion mining. On the other hand, each item is associated with taxonomy information that reflects the viewpoints of experts. In this paper, we propose to mine for users’ opinions on items based on item taxonomy developed by experts and folksonomy contributed by users. In addition, we explore how to make personalized item recommendations based on users’ opinions. The experiments conducted on real word datasets collected from Amazon.com and CiteULike demonstrated the effectiveness of the proposed approaches.

Keywords—Recommender Systems; Folksonomy; Tags; Opinion Mining; Personalization; Taxonomy

I. INTRODUCTION

Recommender system is an effective tool to deal with the issue of information overload. Besides the typical explicit ratings, in Web 2.0, users’ explicit textural information such as tags, blogs, reviews and comments becomes popular. Instead of using numeric numbers, people use one or more pieces of textual information to express their opinions and interest, collect and organize items, share experiences, and build up social networks. How to mine users’ opinions based on these kinds of user created textural information and then recommend items to users becomes very important. Currently, mining users’ sentimental orientations to items based on reviews, blogs and comments are the major focus of opinion mining [3]. Since tags are mainly keywords instead of sentences, usually it is difficult to conduct opinion mining using the traditional sentimental analysis approaches.

However, we argue that tags can be used as another important information source to profile users’ opinions. Different from reviews, blogs and comments that expressing users’ sentimental orientation to items, tags express users’ opinions on item classifications and descriptions. For example, for the book “The World is Flat”, assume user $u_1$ labeled it with a tag “globalization” while user $u_2$ described it with a tag “outsourcing”. The two tags not only reflect their topic interests or preferences [6], but also reflect these people’s different opinions for the classifications and descriptions of this item [4]. Another kind of opinion information that contained in tags is that if a set of items are put together by one user, then, these items are similar or closely related in the opinion of that user. Therefore, tags contain rich opinion information.

Besides folksonomy that expresses users’ opinions on item classifications, each item is associated with item taxonomy that reflects experts’ viewpoint on item classification and descriptions, for example, the product classification taxonomy of Amazon.com, and world knowledge ontology such as Library of Congress Subject Headings [12]. Because item taxonomy has the advantages of having standard and controlled vocabulary, independent with user communities and being well recognized as common knowledge, it can be used to reduce the noise of tags [7] and profile users’ opinions on items.

Currently, the rich opinion information in folksonomy is ignored. Some pioneer work discussed how to hybrid taxonomy and folksonomy for knowledge organization [2], and navigation [15]. The existing recommender systems only used one of the two information sources. For example, the recommendation approaches based on item taxonomy [8] or tags [6]. In this paper, we propose to use the rich opinion information on item classification and description based on both item folksonomy and taxonomy to make item recommendations.

Tags and taxonomic topics are categories which represent various conceptual aspects of the items so that users can classify items into these categories according to their opinions or understanding to the items. Thus the tags and taxonomic topics can be considered representing different features of items. In this paper, we propose to use tags and taxonomic topics as the features of items. Different from the typical “negative”, “neutral” and “positive” values, we use the numeric value ranging from 0 to 1 to express users’ opinions on the features. The higher the value, the more the user agrees that the item can be described with the feature.

This paper is organized as follows. Firstly, the related work is briefly reviewed in Section II. Then, some important notations are given in Section III. The proposed approaches are discussed in details in Section IV. Firstly, the approaches of mining users’ opinion on items are presented and then, the collaborative filtering recommendation approaches based on users’ opinions are discussed. In Section V and VI, the design of the experiments, experimental results and discussions are presented. The conclusions and future work are discussed in Section VII.
II. RELATED WORK

Opinion mining is an important research area. The techniques of text mining, natural language processing and sentiment analysis are employed to find users’ opinions [3]. The major tasks of opinion mining include mining the features of items and finding the users’ sentimental orientations to the features or items [3]. Some opinion mining approaches based on users’ reviews [13], blogs [11] and forum posts are proposed. Traditionally, recommender systems operate based on user-behavior and rating data [14]. How to incorporate users’ opinions in recommender system arouses attentions [14]. For example, the work [10] discussed how to improve the recommendation accuracy through combining the opinion information contained in users’ reviews and the explicit ratings. However, how to use the rich opinion information of tags to make recommendation still remains an open research question.

Currently, the recommendation approaches based on tags are mainly focus on how to recommend tags to users [16]. Not so much work has been done on the item recommendations based on tags. Since recommending a tag to a user to label an item is different with recommending an item to a user, the tag recommendation approaches usually cannot be used to recommend items directly [16]. The current approaches ignored the rich opinion information in tags. How to make use of the rich opinion information of tags to improve the accuracy of item recommendations still need to be further investigated.

Item taxonomy is one important traditional information source to profile users [1]. The important recommendation approaches based on item taxonomy include the work [8] and [17]. The work [8] proposed an approach to take the structural information of item taxonomy to make personalized item recommendations. The work [17] proposed to combine the implicit and explicit item preferences, with the topic preferences that generated based on the taxonomic topic weighting approach in [8] to make item recommendations. However, the weighting approach in [8] did not consider the popularity of each taxonomic topic.

Our previous work [7] proposed a recommendation approach based on item taxonomy and folksonomy. It converted users’ preferences on tags into users’ preferences on taxonomic topics. Although the input information sources of this approach included both item taxonomy and folksonomy, it only used taxonomic topics to profile users’ topic interests and did not use the folksonomy vocabulary and users’ viewpoints on item classifications and descriptions.

This paper extends the existing work through exploring how to combine the two information sources that reflecting the opinions of users and experts on item classifications and descriptions to mine users’ opinions on items and make personalized item recommendations.

III. NOTATIONS

In this paper, we focus on the top \(N\) item recommendation task. To describe the proposed approach, we define some key concepts used in this paper as below.

- **Users**: \(U = \{u_1, u_2, ..., u_M\}\) contains all users in an online community who have used tags to label and organize items.
- **Items (i.e., Products, Resources)**: \(P = \{p_1, p_2, ..., p_P\}\) contains all items tagged by users in \(U\). Items could be any type of resources or products in an online community such as web pages, videos, books, photos and papers etc.
- **Tags (i.e., Folksonomy)**: \(T = \{t_1, t_2, ..., t_T\}\) contains all tags used by users in \(U\). A tag is a piece of textual information given by one or more users to label or collect a set of items. Tags reflect the opinions of users on item classifications and descriptions.
- **Item Taxonomy**: \(O = \langle C, R \rangle, C = \{c_0, c_1, ..., c_{|C| - 1}\}\) is a set of taxonomic topics or categories given by experts to describe or classify items and \(R\) is a set of relations between any \(c_x \in C\) and \(c_y \in C\). Item taxonomic topics reflect the opinions of experts on item classifications and descriptions. In this paper, we only use the typical hierarchical relationship. We define \(R = \{<\}\) is a “sub topic of” relationship, for any two topics \(c_x, c_y \in C\), if \(c_x \subset c_y\), then \(c_x\) is a sub topic of \(c_y\). The taxonomy tree has exactly one root topic.
- **Item taxonomic descriptors**: Each item \(p_i\) is associated with a set of item taxonomic descriptors \(D(p_i) = \{d_1, d_2, ..., d_d\}\). A taxonomic descriptor is a sequence of ordered taxonomic topics, denoted by \(d_i = \{c_{o_i}, c_{y_i}, ..., c_{a_i}\}\), \(c_{o_i}, c_{y_i}, ..., c_{a_i} \in C\), \(c_0\) is the root topic, \(c_a\) is a leaf topic, \(c_o < ... < c_x < c_y < c_0\).
- **Features** \(F = \{\{c_{o_1}, c_{y_1}, ..., c_{a_1}\}, \{t_1, t_2, ..., t_T\}\}\), contains all the features that are used to describe items in \(P\). In this paper, tags and taxonomic topics are used as features describing the items, which are called tag based features and taxonomic topic based features respectively.

Figure 1 (a) illustrates an example of tagging. For example, user \(u_4\) has used the tag \(t_1\) and tagged item \(p_5\) and \(p_6\). Figure 1 (b) shows an example of item taxonomy. Suppose \(p_1\) is a book which is described by descriptor \(d_1\), \(D(p_1) = \{d_1\}\), where \(d_1 = \{c_{o_1}, c_{y_1}, c_{a_1}\}\). The book \(p_1\) is described by taxonomic topics based features {“book”, “garden”, “flowers”} in the viewpoint of experts and tag based features {“garden”, “apple”} in the opinion of user \(u_4\).

IV. THE PROPOSED APPROACHES

The proposed approaches include two major tasks. The first task is to mine users’ opinions on items. Then, based on users’ opinions on items, a set of items will be recommended.

A. Opining mining

The opinion mining process includes the following three sub tasks. 1) finding the relationships between features with respect to each individual user’s opinion and representing each tag based on the feature relationships for each user; 2) mining users opinions on the features for each item, based on all users’ tagging and taxonomy information; 3) determining each user’s opinion to each item feature.
1) Tag representation

How to find the relationship of features is a very important task for opinion mining [3]. Our purpose is to recommend those untagged or unrated items to each target user. Since the candidate items may be described by those very different features that are contributed by other users, if we can find the relationships among these features, we can estimate the target user’s opinions to the candidate items. For example, in Figure 1, the user u₄ described the item with tag "0403". Assume p₃ is the candidate item. If we know the relevance of "0403" and the tag based and taxonomic topic based features of p₃, then we can estimate how much u₄ is interested in the item p₃.

Basically, the taxonomic topic based features contain structural relationship information. Besides the “sub” topic relationship, the “related” topic relationship also can be inferred. For example, although two taxonomic topics are different, if they have the same common ancestors, these topics are related. Tag based features don’t have structural information. Usually, if two tags are used to describe the same item, then, these tags are closely related [5].

Another kind of important information that can be used to find the relationship of features is the personal tagging information. Tags are given by users to organize or describe their own items. It forms a three dimensional relationship User-Tag-Item among the three entities [6]. The User-Tag-Item relationship records the personal tagging information of each individual user [5]. In opinion of user u₄, the collected items under one tag are similar or closely related in some way, otherwise the user won’t put them together and labeled them with the same tag. For example, in Figure 1 (a), with tag t₁, user u₁ collected p₁ and p₂ while user u₂ collected the p₁ and p₄. In u₁’s opinion, p₁ and p₂ are more similar than other items, while in u₂’s opinion, p₁ and p₄ are similar. Thus, the features of these items can be used to represent the conceptual categories covered by or related to the tag in terms of the user’s opinion. The process of finding the relationships of features based on each individual user’s opinion and representing each tag with the features is called tag representation, which is defined as below.

\[
|s_{u_{i,t_x}}(c_y)| \leq |t_x|
\]

**Definition 1** (Tag Representation): represents the relationships of features based on user u₄’s opinion. Specifically, it represents tag tₓ ∈ T’s relevance to each taxonomic topic based feature cᵧ ∈ C and each tag based feature tₓ ∈ T in user u₄’s opinion. Let \( s_{u_{i,t_x}}(c_y) \) denote how strong tₓ is related to cᵧ with respect to user u₄, the relationship between a tag and a set of taxonomic topics in the opinion of a user can be defined as the mapping \( C^T: U \times T \rightarrow 2^{C \times [0,1]} \), such that \( C^T(u_i,t_x) = \{(c_y, s_{u_{i,t_x}}(c_y)) | c_y \in C\} \). Let \( r_{u_{i,t_x}}(c_y) \) denote how strong tₓ is related to tₓ with respect to user u₄, the relationship between a tag and a set of tags in the opinion of a user can be defined as the mapping \( T^T: U \times T \rightarrow 2^{T \times [0,1]} \), such that \( T^T(u_i,t_x) = \{(t_y, r_{u_{i,t_x}}(t_y)) | t_x \in T\} \).

To calculate the relevance weight of a tag to a taxonomic topic (i.e., \( s_{u_{i,t_x}}(c_y) \)) and the relevance weight between two tags (i.e., \( r_{u_{i,t_x}}(c_y) \)) based on the opinion of each individual user uᵢ is very important. For a given user uᵢ and a tag tₓ, the strength of a taxonomic topic cᵧ being related to a tag tₓ for the user uᵢ can be estimated based on the relevance weight of cᵧ to the items collected in the tag tₓ of the user uᵢ. Let \( h^p_{k,y} \) denote the relevance weight of taxonomic topic cᵧ to item pₓ. The items in tₓ of uᵢ is denoted as \( P_{u_{i,t_x}} \). We could use any of \( h^p_{1,y}, \ldots, h^p_{in,y} \) to estimate the relevance of cᵧ to tₓ for user uᵢ. Assuming that \( h^p_{1,y}, \ldots, h^p_{in,y} \) are equally important to the user uᵢ to calculate the relevance of cᵧ to tₓ, we use the average value of \( h^p_{1,y}, \ldots, h^p_{in,y} \) to estimate the relevance of cᵧ to tₓ. Let \( s_{u_{i,t_x}}(c_y) \) denote the relevance weight of cᵧ and tₓ in terms of uᵢ, it can be calculated as:

\[
s_{u_{i,t_x}}(c_y) = \sum_{p_k \in P_{u_{i,t_x}}} \frac{h^p_{k,y}}{|P_{u_{i,t_x}}|}
\]

Figure 1. An example of tagging and representations.
Each item $p_k$ is associated with a set of item taxonomic descriptors $D_{pk}$ given by experts. Ziegler [8] proposed to decay the weight of the taxonomic topic node based on the number of children of the taxonomic topic node in the item taxonomy tree and the length of the descriptor. Let $f(c_y, d_j)$ denote the weight of topic $c_y$ in descriptor $d_j \in D_{pk}$ of item $p_k$. Suppose a descriptor $d_j = \{c_0, c_2, c_3, c_4\}$ and $c_0 < c_2 < c_3 < c_4$, inspired by Ziegler’s approach, we take the structural information of item taxonomy into consideration to calculate the weight $f(c_y, d_j)$ for $c_y$ in $d_j$. For the non-leaf topic $c_2$ in the example descriptor $d_j$ given above, $f(c_y, d_j)$ can be calculated as:

$$f(c_y, d_j) = \frac{f(c_y, d_j)}{\text{child}(c_y)}$$ (2)

Where $c_2$ is the parent node of topic $c_y$ in $d_j$, $\text{child}(c_y)$ is the number of child nodes of topic $c_y$. To facilitate comparison, the total weight of all the topics in $d_j$ is equal to 1. Let $x$ be the weight of the leaf node $c_0$ of the example descriptor $d_j$, we get the following equation:

$$x + \frac{x}{\text{child}(c_2)} + \frac{x}{\text{child}(c_2) - \text{child}(c_3)} + \frac{x}{\text{child}(c_2) - \text{child}(c_3) - \text{child}(c_4)} = 1$$ (3)

After resolving (3), we can get the value of $x$ (i.e., $f(c_y, d_j)$). Based on the leaf node weight $f(c_0, d_j)$ and (2), we can get the weight of each non-leaf topic in $d_j$. Apparently, the leaf nodes have higher weight values than those of non-leaf nodes now. However, if a topic is popularly used to describe items, it is not a distinctive topic to represent this item. Similar to the idf weighting approach in text mining, we should consider the popularity of a topic for all items. Let $idf(c_y)$ denote the inverse item frequency of topic $c_y$, usually, $idf(c_y) = \log(|P|/|P_{c_y}|)$, where $|P_{c_y}|$ is the number of items that have been described with $c_y$ in the set $P$. To get a value between 0 and 1 to facilitate comparison, we set $idf(c_y) = 1/\log(e + |P_{c_y}|)$, 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 of item $p_k$, we use the average value of $f(c_y, d_j)$ in $D_{pk}$ to measure the relevance weight of item $p_k$ to the topic $c_y$. Let $|D_{pk}|$ denotes the number of descriptors of item $p_k$, the relevance weight $h_{k,y}^P$ can be calculated as:

$$h_{k,y}^P = \frac{1}{|D_{pk}|} \sum_{d \in D_{pk}} f(c_y, d_j) \cdot idf(c_y)$$ (4)

For a given user $u_t$ and a tag $t_x$, the strength of a tag $t_x$ being related to the tag $t_x$ for the user $u_t$ can be estimated based on the probabilities of $t_x$ being used to tag the items collected in the tag $t_x$ of the user $u_t$ [5]. Let $|U_{p_k,t_x}|$ be the number of users that tagged $p_k$ with $t_x$, $|U_{p_k}|$ is the number of users that have tagged item $p_k$, the conditional probability of $t_x$ being used to tag item $p_k$ given the item $p_k$ denoted as $Pr(t_x | p_k)$ can be calculated as:

$$Pr(t_x | p_k) = \frac{|U_{p_k,t_x}|}{|U_{p_k}|}$$ (5)

Let $r_{u_i,t_x}(t_y)$ denote the relevance of a tag $t_x$ to a tag $t_y$ for user $u_i$, it can be calculated as:

$$r_{u_i,t_x}(t_y) = \frac{\sum_{p_k \in P_{u_i,t_x}} Pr(t_y | p_k) h_{k,y}^P}{|P_{u_i,t_x}|}$$ (6)

**Example 1 (Tag Representation)** The descriptors of the items in Figure 1 (a) are defined as: $D_{p_1} = \{d_1\}$, $D_{p_2} = \{d_2\}$, $D_{p_3} = \{d_3\}$, $D_{p_5} = \{d_4\}$, $D_{p_6} = \{d_5\}$, where $d_1 = \{c_0, c_1, c_4\}$, $d_2 = \{c_0, c_1, c_5\}$, $d_3 = \{c_0, c_2, c_5\}$, $d_4 = \{c_0, c_2, c_7\}$, $d_5 = \{c_0, c_5, c_7\}$. Figure 1 (c) shows an example of the tag representations of tag $t_2$ for $u_1$ and $u_2$.

The calculation of the relevance of tag $t_2$ and $c_6$ for $u_2$ is shown as follows: since $d_2$ collected item $p_3$ and $p_4$ with tag $t_2$, $s_{u_2,t_2}(c_6) = \sum_{p_k \in P_{u_2,t_2}} h_{k,6}^P + h_{p,6}^P$. There are two descriptors $d_3$ and $d_4$ for item $p_3$, $h_{p,6}^P = \sum_{d \in D_{p_3}} f(c_6, d) \cdot idf(c_6) = \frac{1}{2} \cdot f(c_6, d_3) + f(c_6, d_4) \cdot idf(c_6)$. Show in Figure 1 (b), $child(d_3) = 3$, $child(d_4) = 0$. Suppose $f(c_6, d_3) = x$, $x + \frac{x}{3} + \frac{x}{3} = 1$, $x = 0.69$. Similarly, $f(c_6, d_4) = 0$. $c_6$ has described $p_3$, $p_5$ and $p_6$, $idf(c_6) = \frac{1}{\log(4+3)} = 0.57$. Thus, $h_{p,6}^P = \frac{1}{2} \cdot 0.69 + 0 = 0.57$, $h_{k,6}^P = 0.397$, $s_{u_2,t_2}(c_6) = 0.599$. The relevance of $t_2$ and $t_3$ for $u_2$ can be calculated as:

$$r_{u_2,t_2}(t_3) = \sum_{p_k \in P_{u_2,t_2}} Pr(t_3 | p_k) h_{k,3}^P$$

For user $u_1$, the tag $t_2$ “apple” is mainly related to $c_1$ “garden”, $c_4$ “flowers” and $c_5$ “fruit”. While for user $u_2$, mainly related to $c_3$ “computers”, $c_7$ “programming”, $c_7$ “networks” and $c_8$ “databases”. In the folksonomy, $t_2$ is related to $t_1$ “garden” for user $u_1$, while $t_2$ is related to $t_3$ “globalization” and $t_4$ “internet” for $u_2$.

Therefore, the relationships of features based on each individual user’s opinion are obtained. Based on the tag representations, we can find more accurate item descriptions or classifications and users’ opinion on the features, which will be discussed as follows.

2) **Item representation**

How to incorporate users’ opinions to rank, organize and classify items is another important task [3]. Typically, items are classified by experts and described with taxonomic topic based features. In Web 2.0, users use tags to express their own opinions on item classifications and descriptions. How to classify items and describe each item with related features based on all users’ opinions and experts’ viewpoint is the major focus of this subsection. The process of determining the related features of each item and represent each item with the features is called item representation.

**Definition 2 (Item Representation)**: represents the relationships between item $p_k$ and the features based on the opinions of users and experts. Specifically, for each item $p_k \in P$, it represents the item $p_k$’s relevance to each taxonomic topic based feature $c_y \in C$ and each tag based feature $t_x \in T$. $h_{k,y}^P$ denote the weight of how much the item $p_k$ is relevant to the taxonomic topic $c_y$, the relationship between an item and a set of taxonomic topics can be defined
as the mapping $C^p: P \rightarrow 2^{C \times [0,1]}$, such that $C^p(p_k) = \{(c_y, h_{k,y}^p)\}_{c_y \in C}$. Let $w_{k,x}^p$ denote the weight of how much the item $p_k$ is relevant to the tag $t_x$, the relationship between an item and a set of tags can be defined as the mapping $T^p: P \rightarrow 2^{T \times [0,1]}$, such that, $T^p(p_k) = \{(t_x, w_{k,x}^p)\}_{t_x \in T}$.

The item representation of $p_k$ is defined as $C^T(p_k) = \{C^p(p_k), T^p(p_k)\}$. $C^p(p_k)$ reflects the opinion of experts while $T^p(p_k)$ reflects the opinion of users.

Based on (4), we can calculate how much item $p_k$ is relevant to taxonomic topic $c_y$. Mainly, we discuss how to measure the relevance weight $w_{k,x}^p$ of item $p_k$ to a tag $t_x$ in this sub section. As discussed in the above sub section, the weight $r_{u_i,t_x}(t_x)$ estimates the relevance of a tag $t_x$ to a tag $t_x$ with respect to a user $u_i$. Since the items collected in $t_x$ must have something in common (otherwise the user will not put them together in one tag), the related tag $t_x$ should reflect some topics of the items in $t_x$. As discussed in [5], the relevance weight $w_{k,x}^p$ of item $p_k$ to a tag $t_x$ can be calculated as:

$$w_{k,x}^p = \sum_{u_i \in \{u \mid u \in [0,1]\}} \frac{1}{M} \cdot r_{u_i,t_x}(t_x) \cdot iif(t_x) \tag{7}$$

Where $iif(t_x)$ is the inverse item frequency of tag $t_x$, $T_{p_k}$ is the tag set of $p_k$, $U_{p_k}$ is the user set of $p_k$, and $M$ is the number of unique user-tag $(u_i, t_x)$ pairs of item $p_k$.

Since the two mappings $C^p(p_k)$ and $T^p(p_k)$ can be viewed as two value vectors: $C^p(p_k) = \langle h_{k,0}, \ldots, h_{k,|C|-1} \rangle$ for topics $\langle c_0, \ldots, c_{|C|-1} \rangle$, $T^p(p_k) = \langle w_{k,1}^p, \ldots, w_{k,T}^p \rangle$ for tags $\langle t_1, \ldots, t_T \rangle$, each item $p_k$ can be described by a $|C|$-sized taxonomic topic value vector $C^p(p_k)$ and a $|T|$-sized tag value vector $T^p(p_k)$. The values can be calculated by (4) and (7) respectively.

### Example 2 (Item Representation)

The item representation of $p_3$ is shown in Figure 1 (d). With the calculation process shown in Example 1, the relevance value of item $p_3$ with taxonomic topics can be calculated. The calculation of the relevance of item $p_3$ to tag $t_5$ is shown as follows. Shown in Figure 1 (a), the user-tag pairs of item $p_3$ include $(u_2, t_2), (u_2, t_3)$, and $(u_3, t_5)$. $w_{t_5}^p = \sum_{u_i \in \{u \mid u \in [0,1]\}} \frac{1}{3} \cdot r_{u_i,t_5}(t_5) \cdot iif(t_5) = \frac{1}{3} \cdot (r_{u_2,t_5}(t_5) + r_{u_3,t_5}(t_5)) = 0$. $w_{t_2}^p = \sum_{u_i \in \{u \mid u \in [0,1]\}} \frac{1}{3} \cdot r_{u_i,t_2}(t_2) = 0.52$. $w_{t_3}^p = \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{1}{2} = 0.52$. $w_{t_4}^p = 0.028$. Thus, item $p_3$ related to $c_3$ “computers”, $c_2$ “programming” and $c_4$ “networks” and tags $t_3$ “globalization” and $t_4$ “internet”.

### 3) User profiling

User profile is used to describe a user’s information such as interests, preferences, behavior and opinion [1]. Profiling each user’s opinions to items is crucial for the prediction of whether a user will be interested in a candidate item. Not only the items that a user has tagged or collected, but also the opinions of the user to the collected items should be profiled. Thus, we propose to use both the item set of each user and the user’s opinions on the classifications of these items to profile each user. The process of finding each user’s opinion on item classifications and represent each user with the features of items is called user representation.

**Definition 3 (User representation):** represents user $u_i$’s opinion on the features of items. Specifically, for each user $u_i \in U$, it represents the user $u_i$’s preferences to each taxonomic topic based feature $c_y \in C$ and each tag based feature $t_x \in T$. Let $h_{u_i,t}^p$ 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 $C^u: U \rightarrow 2^{C \times [0,1]}$, such that $C^u(u_i) = \{(c_y, h_{u_i,t}^p)\}_{c_y \in C}$. Let $iuf(t_x)$ denote the weight of how much the user $u_i$ is interested in the tag $t_x$, the relationship between a user and a set of tags can be defined as the mapping $T^u: U \rightarrow 2^T \times [0,1]$. Such that $T^u(u_i) = \{(t_x, h_{u_i,t}^p)\}_{t_x \in T}$. The user representation of $u_i$ is defined as $C^T(u_i) = \{C^u(u_i), T^u(u_i)\}$.

How to obtain each user’s preferences on the features is the major focus of this sub section. To calculate how much $u_i$ will be interested in taxonomic topic $c_y$ and tag $t_x$, we firstly calculate how much the user is interested in the tag $t_x$. As discussed in [5], the strength of $u_i$ will be interested in tag $t_x$ can be calculated as $(t_x \mid u_i) = \frac{\sum_{p \in p_i} w_{t_x}^p}{|p_i|}$, where $|p_i|$ is the number of items that user $u_i$ has tagged. For a given user $u_i$ and a tag $t_x$, based on (1), we can get the relevance weight $s_{u_i,t_x}(c_y)$ between tag $t_x$ and taxonomic topic $c_y$ for user $u_i$. Thus, we can estimate each user $u_i$’s preferences to the taxonomic topic $c_y$ through calculating the product of $s_{u_i,t_x}(c_y)$ and $P_r(t_x \mid u_i)$. Let $iuf(t_x)$ denote as the inverse user frequency of topic $c_y$, the weight $h_{u_i,t}^p$ can be calculated as:

$$h_{u_i,t}^p = \sum_{t_x \in T} P_r(t_x \mid u_i) \cdot s_{u_i,t_x}(c_y) \cdot iuf(t_x) \tag{8}$$

Let $iuf(t_x)$ denote as the inverse user frequency of tag $t_x$, the weight $w_{t_x}^p$ can be calculated as:

$$w_{t_x}^p = \sum_{t_x \in T} P_r(t_x \mid u_i) \cdot r_{u_i,t_x}(t_x) \cdot iif(t_x) \tag{9}$$

The two mappings $C^u(u_i)$ and $T^u(u_i)$ can be viewed as two value vectors: $C^u(u_i) = \langle h_{u_i,0}, \ldots, h_{u_i,|C|-1} \rangle$ for topics $\langle c_0, \ldots, c_{|C|-1} \rangle$, $T^u(u_i) = \langle w_{u_i,1}^p, \ldots, w_{u_i,T}^p \rangle$ for tags $\langle t_1, \ldots, t_T \rangle$. We profile each user $u_i$ with item and feature preferences. Thus, each user $u_i$ can be profiled by three vectors: $h_{u_i}^p, C^u(u_i)$ and $T^u(u_i)$. $h_{u_i}^p$ is a $|P|$-sized binary item vector representing user $u_i$’s collected item set. If $u_i$ has item $p_k$, then the value of this item in vector $h_{u_i}^p$ is 1, otherwise is 0. $C^u(u_i)$ is a $|C|$-sized taxonomic topic value vector and $T^u(u_i)$ is a $|T|$-sized tag value vector.

### Example 3 (User Representation)

The user representation of $u_4$ is shown in Figure 1 (e). The calculation of $u_4$’s preferences to taxonomic topic $c_6$ “programming” is shown as follows:

$$h_{u_4,c}^p = \sum_{t_x \in T} P_r(t_x \mid u_4) \cdot s_{u_4,t_x}(c_6) \cdot iuf(t_x)$$

Based on the tagging graph in Figure 1 (a), we can get $P_r(t_5 \mid u_4) = 1$. $iuf(c_6) = 0.57$, $s_{u_4,t}(c_6) = 0.32$. Thus, $h_{u_4,c}^p = 0.57 \cdot 0.32 = 0.18$. The calculation of $u_4$’s preferences to tag $t_4$ “internet” is shown as follows.

$$w_{t_4}^p = \sum_{t_x \in T} P_r(t_x \mid u_4) \cdot r_{u_4,t_x}(t_x) \cdot iif(t_x)$$

This completes the calculation of $u_4$’s preferences to taxonomic topic $c_6$ “programming” and tag $t_4$ “internet”.


After user representations, $u_3, u_2$ and $u_4$ have preferences on $t_4$, $iuf(t_4) = 0.57$. Thus, $w_{t_4} = 1 \cdot \left( \frac{1}{2} \cdot 0.57 \right) = 0.14$.

Shown in Figure 1 (c), although $u_4$ used a personal tag $t_5$ "0403" to collect items, we can find that $u_4$ also interested in topics $c_6$ "programming" and $c_3$ "computers". In the folksonomy of this user community, $u_4$ is also interested in tag $t_3$ "globalization" and $t_4$ "internet".

Therefore, each user and item is represented with a set of taxonomic topics and tags. Since memory based CF approaches are more popularly used for implicit ratings and other user behaviors, based on user and item representations, the memory based CF and content mapping can be used to form neighborhood and recommend items.

B. Personalized Item Recommendation making

How to recommend items based on the users’ opinions is another important research question [6]. With users’ opinion information, not only the similarity of the objective content information of items, but also the similarity of users’ opinions to these items will affect whether an item will be considered as similar to another item. Moreover, both the similarity of the collected item sets and the similarity of users’ opinions to the features of items will be considered to determine whether a user is a peer neighbor user to another user. Thus, users’ opinion information will affect whether an item will be recommended to a target user. After incorporating users’ opinion information, different neighborhood will be formed and the rank of the recommended items will be different. We discuss the neighborhood formation and recommendation generation approaches in the following sub sections.

1) Neighborhood Forming

Neighborhood formation is to generate a set of like-minded peers for a target user $u_i \in U$ or a set of similar peer items for an item $p_i \in P$. The more accurate a user profile or item representation is, the more similar neighbor users or items will be found. We use cosine similarity to measure the similarity of any two taxonomic topic value vectors as well as any two tag value vectors.

The similarity of two users $u_i$ and $u_j$ includes two parts: the similarity of items and the similarity of features. The feature preferences are represented by taxonomic topic value vector and tag value vector, we linearly combine the similarities of them to measure the similarity of topic preferences. Since the approach of weighting each commonly rated item with its inverted user frequency or $iuf$ [9] performs better for binary ratings in many cases [9], we use this $iuf$ approach to calculate the similarity of items of two users denoted as $sim^p_{u}(u_i, u_j)$.

$$sim^p_{u}(u_i, u_j) = \frac{\sum_{p \in P_{u_i}} iuf(p_k) \cdot \delta_{p \in P_{u_i}}} {||P_{u_i}||}$$

Where $|P_{u_i}|$ is the number of items that $u_i$ has tagged. Thus, the similarity of two users is defined as below:

$$sim^p_{u}(u_i, u_j) = \lambda_1 \cdot sim^p_{u}(u_i, u_j) + (\lambda_2 \cdot \cos(C^u(u_i), C^u(u_j)) + \lambda_3 \cdot \cos(T^u(u_i), T^u(u_j)))$$ (11)

Where $0 \leq \lambda_1, \lambda_2, \lambda_3 \leq 1$ and $0 \leq \lambda_1 + \lambda_2 + \lambda_3 \leq 1$.

Similarly, the similarity of two items can be calculated as:

$$sim^p_{p}(p_i, p_j) = \eta \cdot \cos(C^p(p_i), C^p(p_j)) + (1 - \eta) \cdot \cos(T^p(p_i), T^p(p_j))$$ (12)

Where $0 \leq \eta \leq 1$. The $K$ nearest neighbor users who have similar user profiles with $u_i$ can be found, which is denoted as $N(u_i)$.

2) Recommendation Generation

A set of items that are most frequently tagged by the neighbors of the target user or most similar to the target user’s items will be recommended to the target user. The similarity of taxonomic topics and tags 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 tagged by the most similar users, but also have similar taxonomic topics and tags with the target user. We discuss both user and item based CF approaches that combine the topic mapping respectively.

4.5.1 User based approach

For each target user $u_i$, a set of candidate items will be generated from the items tagged by $u_i$’s neighborhood formed based on the similarity of user profiles. For each candidate item $p_k$, the prediction score of how much $u_i$ may be interested in $p_k$ is calculated in terms of the aspects of how similar those users who have the item $p_k$ and how similar the item’s features with $u_i$’s feature preferences. We use the simple linear combination to hybrid the two parts. Similarly, we linearly combine the feature match of both taxonomic topics and tags. For each candidate item $p_k$, the prediction score can be calculated as:

$$A_u(u_i, p_k) = \sum_{u_j \in N(u_i)} (\alpha_1 \cdot sim^p_{u}(u_i, u_j) + (\alpha_2 \cdot \cos(C^u(u_i), C^p(p_k)) + \alpha_3 \cdot \cos(T^u(u_i), T^p(p_k))))$$ (13)

Where $0 \leq \alpha_1, \alpha_2, \alpha_3 \leq 1$ and $0 \leq \alpha_1 + \alpha_2 + \alpha_3 \leq 1$.

4.5.2 Item based approach

For item based approach, the candidate item set can be the whole item set except those items that are already rated/tagged by the target user. To avoid unnecessary computation of item pairs, the top $K$ most similar items of each tagged item of the target user $u_i$ can be aggregated together as the candidate item set. For each candidate item $p_k$, we propose to calculate the prediction score of a candidate item based on the maximum score of the linear combination of the similarity with each tagged item and the similarity with the target user’s feature preferences. Similarly, we linearly combine the feature match of both taxonomic topics and tags. The prediction score for each candidate item $p_k$ can be calculated as:

$$A_p(u_i, p_k) = \max_{p_\in P_{u_i}} (\beta_1 \cdot sim^p_{p}(p_\in P_{u_i}, p_k) + (\beta_2 \cdot \cos(C^u(u_i), C^p(p_k)) + \beta_3 \cdot \cos(T^u(u_i), T^p(p_k))))$$ (14)

Where $0 \leq \beta_1, \beta_2, \beta_3 \leq 1$ and $0 \leq \beta_1 + \beta_2 + \beta_3 \leq 1$. 
V. EXPERIMENT DESIGN

A. Data preparation

We conducted the experiments with two real world datasets from Amazon.com and CiteULike.com. The former dataset has taxonomic and folksonomy information while the latter one only has folksonomy information.

1) Dataset D1: Amazon.com dataset. This dataset was crawled from Amazon.com on April, 2008. The items of the dataset are books. To avoid too sparse, we only select those users that have at least 5 items and those items that have been used by at least 3 users. The final dataset consists of 4112 users, 34201 tags, 30467 items. We also extracted the taxonomic descriptors [7] of each item from amazon.com. The taxonomy contains 9919 unique topics.

2) Dataset D2: CiteULike dataset. The “Who-posted-what” dataset (http://static.citeulike.org/data/current.bz2) is used. The items are research papers. We select those users that have at least 5 items and those items that have been used by at least 2 users. The final dataset comprises 7103 users, 78414 tags, 117279 items.

B. Experiments setup

To evaluate the proposed approaches, each dataset was 5 folded and split into 5 datasets. For each split dataset, 80% of users were used as the training users while 20% of users were randomly selected as the test users. For each test user, randomly, 20% of the items of this user were hidden as the test/answer set while 80% of each user’s items are used as his/her training set. The training set of each user contains user's items and corresponding tags information as well. For each test user, the recommender system will generate a list of ordered items that the test user didn’t collect. The top N items with high prediction scores will be recommended to the user. 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 one split dataset were recorded as one run of the results. The average precision and recall values of the 5 split datasets were used to measure the accuracy of the recommendations.

VI. RESULTS AND DISCUSSIONS

In this section, we firstly discuss the setting of the parameters. Then, we discuss the influences of folksonomy and taxonomic information to the accuracy of recommendations and the comparison of the proposed approaches with other related state-of-art work.

A. The setting of Parameters

The results indicated that with $\lambda_1=0.8, \lambda_2=0.1, \lambda_3=0.1, \alpha_1=0.3, \alpha_2=0.2, \alpha_3=0.5$, the proposed user based approach achieved the best performances for Dataset D1. With $\eta=0.3$, $\beta_1=0.3, \beta_2=0.2, \beta_3=0.5$, the proposed item based approach achieved the best performances for Dataset D1. Since there is no taxonomic information for Dataset D2, the values of $\lambda_2, \gamma, \eta$, and $\beta_2$ were set to 0. The results suggested that with $\lambda_1=0.9, \lambda_3=0.1, \alpha_1=0.4$, and $\alpha_3=0.6$, the proposed user based approach achieved the best performances for Dataset D2. With $\beta_1=0.4$ and $\beta_2=0.6$, the proposed item based approach achieved the best performances for Dataset D2. The following discussions are given on the basis of the best settings of the parameters.

B. Taxonomy V.S. Folksonomy

To measure the influences of taxonomic and folksonomy for the improvement of the recommendation accuracy, we compared the top 3 ($N=1,\ldots,3$) recommendation precision values of the proposed approaches that represent the topic preferences of each user and topics of each item using both taxonomic topics and tags with the approaches that represent users or items using taxonomic topics or tags only. We also compared the proposed approaches with the approach proposed by Ziegler [8]. The following are the 5 approaches compared in this part of experiments.

- **CTR-User and CTR-Item**: These are the proposed user and item based approaches that represent each user and item with both taxonomic topics and tags. For simplicity, they are called the combined models.
- **CR-User** and **TR-User**: CR-User is the proposed user based approach that only represents each user and item with taxonomic topics while tags are used for TR-User. CR-User is called taxonomy model. TR-User is called folksonomy model.
- **TPR**: Ziegler proposed an approach to acquire a user’s topic preferences based on item taxonomic topics [8]. For a fair comparison, TPR combined item preferences and topic preferences generated based on [8].

The top 3 precision values are shown in Figure 2.

![Figure 2. Top 3 Precision values of Dataset D1.](image)

**Discussions:**

As shown in Figure 2, the proposed user based approach **CTR-User** performed slightly better than the proposed item based approach **CTR-Item**. Both the combined models performed better than the proposed taxonomy model **CR-User** and folksonomy model **TR-User**. For example, the Top-1 ($N=1$) precision value of **CTR-User** was 0.41, while that of **TR-User** was 0.35 and that of **CR-User** was 0.29. It indicated that after combining the item taxonomy and folksonomy information, the accuracy of item recommendations can be further improved. Moreover, the proposed taxonomy model **CR-User** performed better than **TPR** that is based on the weighing approach proposed in [8]. The improvement suggested that after considering both structural information and the popularity of taxonomic topics, the recommendation accuracy based on item taxonomy can be improved.

Another important finding is that the proposed folksonomy model **TR-User** performed much better than the
proposed taxonomy model \( CR-User \). To further discuss the proposed folksonomy and taxonomy model, we select a set of tags whose popularity are larger than or equal to \( \theta \), and only retained those selected tags in the user and item representations. Then, we compared the Top-3 \((N = 3)\) precision values for the two models. The number of tags used by at least \( \theta \) users of both datasets and the Top-3 precision values of the proposed folksonomy model with different \( \theta \) values are plotted in Figure 3 where \( \theta \) was set from 1 to 10 incrementally.

![Figure 3. The number of tags used by at least \( \theta \) users and the Top-3 \((N = 3)\) Precision results with different \( \theta \) values.](image)

As shown in Figure 3, with \( \theta = 1 \), we retained all the tags (i.e., 34201) in the item and user representations and got the best precision value 0.31 for Dataset D1 with the proposed folksonomy model \( TR-User \). Similarly, the best precision value can be achieved for Dataset D2 with the proposed \( TR-User \) approach when we retain all the tags (i.e., 78414). The distributions of tags follow the power law distributions [3]. Figure 3 indicated that although keeping more tags not necessarily improved the precision values, the precision values decreased dramatically when a large number (i.e., 90\%) of tags with lower \( \theta \) values (i.e., \( \theta \leq 5 \)) was removed.

We compared the proposed folksonomy model \( TR-User \) with the proposed taxonomy model \( CR-User \) on Dataset D1. The Top-3 \((N = 3)\) precision value for the proposed taxonomy model was 0.24 and there were 9919 unique taxonomic topics in the dataset D1. As shown in Figure 3, only when we selected less than 4897 tags with \( \theta > 4 \), the proposed folksonomy model performed worse than the taxonomy model (i.e., \( \leq 0.24 \)). It suggested that after making use of the rich opinion information, folksonomy can be used as quality information source to provide more accurate personalized item recommendations than taxonomy.

**VII. CONCLUSIONS AND FUTURE WORK**

In this paper, we proposed to integrate the item taxonomy and folksonomy information to conduct opinion mining and make personalized item recommendations. The item and user based CF combining with the content filtering approaches are presented. The experimental results show that the proposed approaches are effective. The best performances of the proposed combined approaches suggest that integrating the standard item taxonomy vocabulary and experts’ viewpoint on item descriptions/classifications with users’ personal vocabularies and viewpoints can further improve the accuracy of item recommendations. The future work will explore how to integrate tags with other types of user opinion information such as reviews, blogs, and microblogs such as tweets to find users’ opinions on items and make better personalized item recommendations.

**REFERENCES**


