This is the author's version of a work that was submitted/accepted for publication in the following source:


This file was downloaded from: http://eprints.qut.edu.au/41882/

© Copyright © 2010 by the Association for Computing Machinery, Inc. (ACM).

Permission to make digital or hard copies of portions of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyright for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permission to republish from: permissions@acm.org or contact the Publications Dept., ACM, Inc. Fax +1 (212) 869-0481. For other copying of articles that carry a code at the bottom of the first or last page, copying is permitted provided that the per-copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923 (USA).

Notice: Changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published source:
Personalized Recommender System Based on Item Taxonomy and Folksonomy

Huizhi Liang, Yue Xu, Yuefeng Li, Richi Nayak
Faculty of Science and Technology
Queensland University of Technology, Brisbane, Australia
oklianghuizi@gmail.com, {yue.xu, y2.li, r.nayak}@qut.edu.au

ABSTRACT
Item folksonomy or tag information is popularly available on the web now. However, since tags are arbitrary words given by users, they contain a lot of noise such as tag synonyms, semantic ambiguities and personal tags. Such noise brings difficulties to improve the accuracy of item recommendations. In this paper, we propose to combine item taxonomy and folksonomy to reduce the noise of tags and make personalized item recommendations. The experiments conducted on the dataset collected from Amazon.com demonstrated the effectiveness of the proposed approaches. The results suggested that the recommendation accuracy can be further improved if we consider the viewpoints and the vocabularies of both experts and users.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval-Information Filtering; H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces-Collaborative computing

General Terms
Algorithms, Experimentation

Keywords
Recommender Systems, Folksonomy, Tags, Taxonomy

1. INTRODUCTION
Recommender system is an effective tool to deal with the issue of information overload. Besides users’ item preferences that profiled with explicit or implicit ratings, how to profile users’ topic preferences is very important [1]. Traditionally, the taxonomy or ontology of items or the content information of items is used to find users’ topic preferences [1]. Item taxonomy is a set of controlled vocabulary terms or topics designed to describe or classify items. The advantages of taxonomy include: vocabulary are standard and controlled, having relationship information among concepts, well recognized as common knowledge, and independent with user communities. One limitation is that it does not reflect users’ personal viewpoints or preferences information.

Different with item taxonomy, folksonomy is contributed by users. Folksonomy has distinctive advantages which include being given by users explicitly and proactively, reflecting users’ topic preferences and personal viewpoints on item descriptions or classifications, having multiple functions such as organizing and sharing items, building networks, and expressing explicit opinions [6]. It becomes another important information source to find users’ topic preferences [6]. However, since there is no restriction or boundary on selecting words for tagging items, the tags used by users are free-formed and contain a lot of noise such as semantic ambiguity, tag synonyms and personal tags [7] [9]. The tag quality [7] [8] [9] problem generates difficulties in improving the accuracy of item recommendations based on tags.

An important research question is that can we integrate item folksonomy and taxonomy to overcome the tag quality problem and benefit from both. Some pioneer work discussed how to use hybrid taxonomy and folksonomy for knowledge organization [2] and navigation [3]. Very few work discussed how to use both information sources to find more accurate user topic preferences and make personalized item recommendations so far. In this paper, we propose to use both item folksonomy and taxonomy to make personalized item recommendations.

2. RELATED WORK
Item taxonomy is one important traditional information source to find users’ topic preferences [1]. The important recommendation approaches based on item taxonomy include the work of Ziegler [4]. However, the taxonomic topic weighting approach did not consider the popularity of each taxonomic topic. Currently, the existing recommender systems only used one kind of the two information sources. For example, the recommendation approaches based on item taxonomy [4] or tags [6]. Very few work discussed how to use both information sources to make personalized item recommendations. Our previous work [5] proposed to convert users’ preferences to tags into users’ preferences to taxonomic topics. However the folksonomy vocabulary and users’ personal viewpoints were not considered to profile users’ topic preferences and items’ topic descriptions.

3. Notations
In this paper, we focus on the top $N$ item recommendation task. Some key concepts and entities are defined as below.

- Users: $U = \{u_1, u_2, ..., u|U|\}$ 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, photos, documents and books 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.

• Item Taxonomy: $O = \subseteq C, R \supset C = \{c_0, c_1, ..., c_{|C|-1}\}$ is a set of topics or categories given by experts. We define $R = \{\prec\}$, $\prec$ as a “sub topic” of relationship, for any two topics $c_x, c_y \in C$, if $c_x \prec c_y$, then $c_y$ is a sub topic of $c_x$. The taxonomy tree has exactly one root topic. It represents the most general topic. The leaf topics represent the most specific topics.

• Item taxonomic descriptors: Each item $p_k$ is associated with a set of item taxonomic descriptors $D(p_k) = \{d_1, d_2, ..., d_j\}$. A taxonomic descriptor is a sequence of ordered topics, denoted by $d_i = \{c_0, c_{y_1}, c_{y_2}, ..., c_{y_n}\}$, $c_0$ is the root topic, $c_a$ is a leaf topic, and $c_{y_1} \prec \cdots \prec c_{y_m} \prec c_{y_n} < c_0$.

Figure 1 (a) illustrates an example of tagging. For example, user $u_4$ has used the tag $t_5$ and tagged item $p_5$ and $p_6$. Figure 1 (b) shows an example of item taxonomy.

4. THE PROPOSED APPROACHES

4.1 Tag Representation

By nature tags are given by users to describe their own items. The process of finding the personalized semantic meanings of each tag for each individual user is called tag representation.

Definition 1 (Tag Representation): represents the relevance of each tag $t_x \in T$ to each taxonomic topic $c_y \in C$ and each tag $t_x \in T$ with respect to user $u_i$. Let $s_{u_i,t_x}(c_y)$ denote how strong $t_x$ is related to $c_y$ with respect to $u_i$. The relationship between a tag and a set of taxonomic topics with respect to a user can be defined as a mapping $C^t: U \times T \rightarrow \mathbb{R}^{C \times [0,1]}$, such that $C^t(u_i, t_x) = \{[c_y, s_{u_i,t_x}(c_y)] | c_y \notin \mathcal{C}\}$. Let $r_{u_i,t_x}(t_z)$ denote how strong $t_z$ is related to $t_x$ with respect to $u_i$, the relationship between a tag and a set of tags with respect to a user can be defined as the mapping $\mathcal{T}^t: U \times T \rightarrow \mathbb{R}^{C \times [0,1]}$, such that $\mathcal{T}^t(u_i, t_x) = \{(t_z, r_{u_i,t_x}(t_z)) | t_z \in T\}$. The tag representation of $t_x$ with respect to user $u_i$ is defined as $C^t(u_i, t_x) = \{C^t(u_i, t_x), \mathcal{T}^t(u_i, t_x)\}$.

For a given user $u_i$ and a tag $t_x$, the strength of $c_y$ being related to a tag $t_x$ for the user $u_i$ can be estimated based on the relevance weight of $c_y$ to the items collected in the tag $t_x$ of the user $u_i$. Let $h_{k,y}^p$ denote the relevance weight of $c_y$ to item $p_k$. $s_{u_i,t_x}(c_y)$ can be calculated as:

$$s_{u_i,t_x}(c_y) = \sum_{p_k \in P_{u_i|t_x}} \frac{h_{k,y}^p}{|P_{u_i|t_x}|}$$

Where $P_{u_i|t_x}$ is the item set collected by $u_i$ with tag $t_x$. How to calculate $h_{k,y}^p$ is very important. Each item $p_k$ is associated with a set of item taxonomic descriptors $D(p_k)$ given by experts. Let $f(c_y, d_j)$ denote the weight of topic $c_y$ in descriptor $d_j \in D(p_k)$ of item $p_k$. Suppose a descriptor $d_j = \{c_0, c_{y_1}, c_{y_2}, ..., c_{y_n}\}$, inspired by Ziegler’s approach [4], we take the structural information of taxonomy into consideration to calculate the weight $f(c_y, d_j)$ for $c_y$ in $d_j$. For the non-leaf topic $c_z$ in the example descriptor $d_j$ given above, $f(c_z, d_j)$ can be calculated as:

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

After resolving Equation 3, we can get the value of $x$ (i.e., $f(c_z, d_j)$). Based on the leaf node weight $f(c_z, d_j)$ and Equation 2, we can get the weight of each non-leaf topic in $d_j$.

However, if a topic is popularly used to describe items, it is not a distinctive topic to represent an item. Let $\text{iff}(c_y)$ denote the inverse item frequency of topic $c_y$, we set $\text{iff}(c_y) = 1/\log(e + |P_{c_y}|)$, where $|P_{c_y}|$ is the number of items that have been described with $c_y$ in the item set $P$, $e$ is an irrational constant approximately equal to 2.72 and $0 < \text{iff}(c_y) \leq 1$. Let $|D_{p_k}|$ denotes the number of descriptors of item $p_k$, the weight $h_{k,y}^p$ can be calculated as:

$$h_{k,y}^p = \frac{1}{|D_{p_k}|} \sum_{d_j \in D_{p_k}} f(c_y, d_j) \cdot \text{iff}(c_y)$$

For a given user $u_i$ and a tag $t_x$, the strength of a tag $t_x$ being related to the tag $t_x$ for the user $u_i$ 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 $u_i$ [8]. Let $|U_{p_k|t_x}|$ be the number of users 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 $Pr(t_x | p_k) = \frac{|U_{p_k|t_x}|}{|U_{p_k}|}$. Thus, $r_{u_i,t_x}(t_z)$ can be calculated as:
\[ r_{ux,t_2}(t_2) = \sum_{p_k \in P_{ux,t_2}} \frac{\Pr(t_2 | p_k)}{|P_{ux,t_2}|} \]  

**Example 1 (Tag Representation)** The descriptors of the items in Figure 1 (a) are defined as: \(D_P = \{d_1\}, D_p = \{d_2\}, D_{p_k} = \{d_1, d_4\} \), \(D_{p_k} = \{d_2\}\), \(D_p = \{d_3\} \), \(D_{p_k} = \{d_3\} \), \(D_{p_k} = \{d_4, d_3\} \), where \(d_1 = \{c_0, c_1, c_4\} \), \(d_2 = \{c_0, c_2, c_3\} \), \(d_3 = \{c_0, c_2, c_6\} \), \(d_4 = \{c_0, c_2, c_7\} \). Figure 1 (c) shows an example of the tag representations of tag \(t_2\) for \(u_1\) and \(u_2\). Personalized semantic meanings of \(t_2\) are generated for different users and the semantic ambiguity can be removed. Similarly, we can get the related taxonomic topics and tags of each personal tag (i.e., \(t_5\) “0403”). Moreover, the tag synonyms can be found through comparing the tag representations. The noise of tags can be reduced.

### 4.2 Item representation

With tag weighting, each item is not only associated with a set of taxonomic topics, but also is described by a set of tags contributed by users. The item representation can be defined as:

**Definition 2** (Item Representation): represents the relevance of each item \(p_k \in P\) to each taxonomic topic \(c_y \in C\) and each tag \(t_x \in T\). Let \(h_{k,y}^x\) 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 by the mapping \(C^P: P \rightarrow 2^{C^{[0,1]}}, \) such that \(C^P(p_k) = \{c_y, h_{k,y}^x\} \in C\). Let \(w_{k,x}^y\) 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 \(T^P: P \rightarrow 2^{T^{[0,1]}}, \) such that, \(T^P(p_k) = \{(t_x, w_{k,x}^y) | t_x \in T\}\). The item representation of \(p_k\) is defined as:

\[ C^T(p_k) = \{C^P(p_k), T^P(p_k)\} \]

With Equation 4, we can calculate how much item \(p_k\) is relevant to taxonomic topic \(c_y\). As discussed in [8], the relevance weight \(w_{k,x}^y\) of item \(p_k\) to a tag \(t_x\) can be calculated as:

\[ w_{k,x}^y = \sum_{t_x \in P_{k,x}} \frac{1}{M} r_{u_k,t_x}(t_x) \cdot iif(t_x) \]

where \(iif(t_x)\) is the inverse item frequency of tag \(t_x\), \(r_{u_k,t_x}(t_x)\) is the tag set of \(p_k\), \(U_{u_k}\) is the user set of \(p_k\), and \(M\) is the number of unique user-tag \((u_k, t_x)\) pairs of item \(p_k\). Since the two mappings \(C^P(p_k)\) and \(T^P(p_k)\) can be viewed as two vectors: \(C^P(p_k) = <h_{k,0}^x, h_{k,1}^x, ..., h_{k,C^{[0,1]}}^x>\) for topics \(<c_0, ..., c_{C^{[0,1]}}\>\), \(T^P(p_k) = <w_{k,0}^y, w_{k,1}^y, ..., w_{k,T^{[0,1]}}^y>\) for tags \(<t_0, ..., t_{T^{[0,1]}}\>\), each item \(p_k\) can be described by two vectors \(C^P(p_k)\) and \(T^P(p_k)\). Figure 1 (d) shows an example of item representation of item \(p_3\).

### 4.3 User profiling

We propose to use item taxonomic topics and tags to profile users’ topic preferences. The user representation is defined as:

**Definition 3** (User representation): represents each user \(u_i \in U\)'s preferences to each taxonomic topic \(c_y \in C\) and each tag \(t_x \in T\). Let \(h_{i,y}^x\) 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 \(C^U: U \rightarrow 2^{C^{[0,1]}}, \) such that \(C^U(u_i) = \{c_y, h_{i,y}^x\} \in C\). Let \(w_{i,x}^y\) 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 \(T^U: U \rightarrow 2^{T^{[0,1]}}, \) such that \(T^U(u_i) = \{(t_x, w_{i,x}^y) | t_x \in T\}\). The user representation of \(u_i\) is defined as:

\[ C^T(u_i) = \{C^U(u_i), T^U(u_i)\} \]

As there is no explicit rating information available for typical tagging communities [6], the binary implicit ratings are used to represent each user’s item preferences [6]. To calculate how much \(u_i\) will be interested in taxonomic topic \(c_y\) and tag \(t_x\), we can firstly calculate how much the user is interested in the tag \(t_x\). As discussed in [8], the strength of \(u_i\) will be interested in tag \(t_x\) can be calculated as \(r(t_x | u_i) = \frac{|P_{u_i}|}{|P_u|} \), where \(|P_{u_i}|\) is the number of items that user \(u_i\) has tagged. For a given user \(u_i\) and a tag \(t_x\), based on Equation 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 \(Pr(t_x | u_i)\). Let \(iuf(c_y)\) denote as the inverse user frequency of topic \(c_y\), the weight \(h_{i,y}^x\) can be calculated as:

\[ h_{i,y}^x = \sum_{t_x \in T} Pr(t_x | u_i) \cdot s_{u_i,t_x}(c_y) \cdot iuf(c_y) \]

Similarly, the weight \(w_{i,x}^y\) can be calculated as:

\[ w_{i,x}^y = \sum_{t_x \in T} Pr(t_x | u_i) \cdot r_{u_i,t_x}(t_x) \cdot iuf(t_x) \]

We profile each user \(u_i\) with item and topic preferences. Thus, each user \(u_i\) can be profiled by three vectors: \(i_f^U(u_i)\) and \(T^U(u_i)\). \(i_f^U(u_i)\) is a binary vector representing \(u_i\)’s item preferences. Figure 1 (e) shows an example user representation of user \(u_4\).

### 4.4 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. Cosine similarity is used to calculate the similarity of any two numeric vectors. The similarity of item preferences of two users is:

\[ sim_u^i(u_i, u_j) = \frac{\sum_{p_k \in P_{u_i}} r_{u_i,p_k} \cdot r_{u_j,p_k} \cdot iuf(p_k)}{|P_u|} \]

Thus, the similarity of two users is defined as below:

\[ sim_u^i(u_i, u_j) = \lambda_1 \cdot \frac{\sum_{p_k \in P_{u_i}} r_{u_i,p_k} \cdot r_{u_j,p_k}}{|P_u|} + \lambda_2 \cdot \cos(C^U(u_i), C^U(u_j)) + \lambda_3 \cdot \cos(T^U(u_i), T^U(u_j)) \]

Where \(0 \leq \lambda_1, \lambda_2, \lambda_3 \leq 1\) and \(0 \leq \lambda_1 + \lambda_2 + \lambda_3 \leq 1\). The similarity of two items can be calculated as:

\[ sim_t^i(p_i, p_j) = \eta \cdot \cos(C^P(p_i), C^P(p_j)) + (1 - \eta) \cdot \cos(T^P(p_i), T^P(p_j)) \]

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 \(N(u_i)\).

### 4.5 Recommendation Generation

For each target user \(u_i\), a set of candidate items will be generated from the items tagged by \(u_i\)'s neighbor users. For the user based collaborative filtering approach, the prediction score of each candidate item \(p_k\) can be calculated as:

\[ A_{p_k}(u_i) = \sum_{u_j \in N(u_i)} \alpha_1 \cdot sim_u^i(u_i, u_j) + \alpha_2 \cdot \cos(C^U(u_i), C^P(p_k)) + \alpha_3 \cdot \cos(T^U(u_i), T^P(p_k)) \]

Where \(0 \leq \alpha_1, \alpha_2, \alpha_3 \leq 1\) and \(0 \leq \alpha_1 + \alpha_2 + \alpha_3 \leq 1\). For the item based approach, the prediction score can be calculated as:
\[ A_p(u_i, p_k) = \max_{p_j \in P_u} \{ \beta_1 \cdot \text{sim}_p(p_j, p_k) + \beta_2 \cdot \cos(C^u(u_i), C^p(p_k)) + \beta_3 \cdot \cos(T^u(u_i), T^p(p_k)) \} \]

Where \( 0 \leq \beta_1, \beta_2, \beta_3 \leq 1 \) and \( 0 \leq \beta_1 + \beta_2 + \beta_3 \leq 1 \).

5. EXPERIMENT AND EVALUATIONS

5.1 Data preparation
The experiments were conducted on the dataset collected from Amazon.com. The items 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. The taxonomy formed by the descriptors is tree-structured and contains 9919 unique topics.

5.2 Experiments setup
To evaluate the proposed approaches, the 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. 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 values of the 5 split datasets were used to measure the accuracy.

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 had the best results. With \( \eta = 0.3, \beta_1 = 0.3, \beta_2 = 0.2, \beta_3 = 0.5 \), the proposed item based approach had the best results. The following discussions are given on the basis of the best settings of the parameters.

5.3 Taxonomy V.S. Folksonomy
We compared the top 3 recommendation precision values of the following approaches:

- **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 [4]. It used implicit ratings but not tag information nor item preferences. For a fair comparison, **TPR** combined item preferences and topic preferences generated based on Ziegler’s approach.

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**. Moreover, the proposed taxonomy model **CR-User** performed better than TPR that is based on Ziegler’s taxonomic topic weighing approach [4]. The improvement suggested that after considering both structural information of item taxonomy and the popularity of taxonomic topics, the accuracy of item recommendations based on item taxonomy can be improved. Another interesting finding is that the proposed folksonomy model **TR-User** performed much better than the proposed taxonomy model **CR-User**. It suggested that after removing the noise, folksonomy can be used as quality information to profile users and describe items.

6. CONCLUSIONS
In this paper, we proposed to integrate the item taxonomy developed by experts and the item folksonomy contributed by users to make personalized item recommendations. To reduce the noise of tags, we propose to find the related taxonomic topics and tags to represent the personalized semantic meaning of each tag for each individual user. Based on the tag representations, we proposed approaches to find related taxonomic topics and tags to represent the topics of each item and the topic preferences of each user. The experimental results suggest that after removing the noise of tags, the tag information can be used as quality user profiling and item content describing information source to boost the accuracy of item recommendations. Moreover, the results also suggest that integrating the standard item taxonomy vocabulary and users’ personal vocabularies as well as the viewpoints of both experts and users on item descriptions/classifications can further improve the accuracy of item recommendations.

7. REFERENCES