User Preference Mining and Privacy Policy Recommendation for Social Networks

*Haoran Xu and Yuqing Sun**
School of Computer Science and Technology,
Engineering Research Center of Digital Media Technology, Ministry of Education of PRC
Shandong University
China
hr_xu1990@163.com, sun_yuqing@sdu.edu.cn

Abstract

Users now high rely on social services not only for entertainment but also for work, and a lot of user data such as profiles and actions are stored on social service platform. Privacy setting is an important means to protect these private data. To help users better manage their privacy information, we propose a user preference based privacy policy recommendation approach for the current privacy setting modes. We investigate user preferences from their own privacy policies and recommend similar settings when a new friend is added or a new item is uploaded. To evaluate our methods, we propose several criteria and perform a lot of experiments on some practical datasets. The experimental results show that our algorithms are applicable for both person assignments and item management.

Keywords: user preference, privacy policy, social networks

1 Introduction

Nowadays, more and more people rely on web-based social network services, such as Facebook, Twitter and Google+. They communicate with each other, make new friends, share video or music, discuss in groups, play games with others, and etc. With the convenience of social networks, many organizations and companies even adopt different kinds of social services as their business and work platform [14, 22]. Social network services have been integrated into people's daily lives not only for entertainment and leisure but also for communication and work. Therefore, quite a lot of user profiles are stored on the social service platform, as well as a large amount of user data such as social actions and uploaded files. For example, 350 million photos are uploaded every day in Facebook [24]. These data are sometime sensitive to the owners, who want to control the access to these data as their expectation. That is to say, on one side users want to share the ideas or photos only with target visitors for social or business purpose; on the other side they want to limit unexpected people accessing these data. So, user privacy setting is an important issue in social networks.

Currently, social service sites mainly utilize access control policies to help users manage their data or profiles. There are two levels of policies, a default level policy is applicable for all user data and a local policy is applicable for some specified person or file. To avoid policy conflicts, a local policy dominates a default policy. Most social networks allow a user to define *permit* or *deny* as a default policy to every friend or visitor on accessing user data. There are two popular ways for a user to specify local policies. In the appointed person mode, a user authorizes the access rights only to the appointed persons. The group mode allows a user to classify friends into several groups according to their relationships and authorize different access rights to each group. The persons in a group have the same rights. The group mode is widely adopted by the influential social service platforms, such as the 'circles' in Google+, and the 'lists' in Facebook and Twitter [10], which are called 'social groups' in this paper.

Since a user may often upload data and add new visitors as friends, it is not convenient to manually configure privacy settings using the above methods. For example, people consider the privacy setting interface in Facebook very complex and tedious to configure privacy setting as new friends added [8]. To tackle this problem, some visualization tools are proposed to help users understand their privacy settings, such as Pviz [11], Privacy Mirrors [1]. However, in practice, most users do not know how to set appropriate privacy policies and they need privacy recommendation. These tools cannot help them on this point. Some recommendation approaches and policy managers are proposed to suggest users some privacy policies [17], such as *policy wizard* [7]. It allows a user to classify his/her friends into several groups according to friend profiles and relationships, and ask user to select some friends from each group to make privacy settings. Then it learns user preferences from

* Corresponding author: Yuqing Sun; E-mail: sun_yuqing@sdu.edu.cn

these settings and recommends similar privacy policies for other persons in the same group [5]. Although such approaches are effective, the learning process needs much user interaction, which is time-consuming and error prone. Shehab et al. propose a privacy policy recommendation approach, which first computes user similarity according to social graph propagation properties and then recommends privacy settings to other similar users in the social network who are not familiar with privacy settings [18]. Since users may have different preferences on grouping friends, this approach is not suitable for the group based authorization mode.

To help users better manage their private data, we propose a user preference based privacy policy recommendation approach for the popular privacy setting modes. It consists of two phases: user preference mining and privacy setting recommendation. For the group mode, user privacy preferences include grouping friends and authorizations on items. So we analyze user group memberships against friend profiles and investigate access right assignments against the item tags of user data so as to find user privacy preferences. In the recommendation phase, when a new friend is added, we compute an appropriate group as a recommendation based on user preferences so as to assign the access rights of this group. For a new item such as a photo, the recommendation algorithms compute the appropriate privacy settings according to user authorizations on accessing items. For the appointed person mode, we directly compute the right assignment similarity by collaborative filtering, and recommend similar privacy settings for a new friend or item. To evaluate our methods, we propose several criteria and perform a lot of experiments on some practical datasets. The experimental results show that our algorithms are applicable for both person assignments and item managements.

The rest of this paper is organized as follows. Section 2 surveys the related works. Then we present the user preference based privacy policy recommendation framework in Section 3. The personalized privacy policy recommendation algorithms are given in Section 4, followed by experiments and results analysis in Section 5. Finally, Section 6 summarizes the contributions of our work and discusses future research directions.

2 Related Work

2.1 Personalized Policy Recommendation

The most related work with ours is the personalized policy recommendation. In social networks, the existing privacy policy specification tools are often complex and difficult to understand. Some assistant tools on privacy setting are proposed to recommend users how to control the access rights to their data.

The user similarity based policy recommendation approach explores user relationships and recommends privacy policy for similar users, such as *PolicyMgr* [17] and *Privacy Wizard* [7]. These methods classify friends into several clusters, and require users provide example policy settings as training sets for each cluster. Then they adopt supervised model automatically configure privacy settings for other friends in the same cluster. However, the learning process involves much interaction with users, which is time-consuming and error prone. Shehab et al. propose a fine-grained policy recommendation system, which first computes user similarity and then recommends existing privacy policy for similar users [18]. Yet this approach do not consider user preference difference on grouping users.

The *xAccess* method adopts a role-based access control model to capture the privacy preference of social users. By analyzing both social network structure data and historical activity data, it extracts some *social roles* and recommends a set of privacy settings based on these roles [24]. Since extraction of social roles is complicated, this approach is not appreciate for practical social networks with a large number of dynamic users. Besides, the extraction of social roles takes into account the common points of similar users without consideration of user privacy preferences.

Tags are widely adopted in social service platforms for users to efficiently manage their data. Some approaches use tags to manage privacy policies by tag based setting, such as *APPGen* [23] and *A3P* [21]. However, these approaches only analyze the similarity between tags without considering the association between items and tags. Besides, they discuss little about the impact of subjects in privacy policies, which also reflect the user privacy preference. Different from these method, our privacy policy recommendation is based on the mining of user preference, which considers the features of both subject and object in a policy.

2.2 Circle Detection

Our work also seems related with circle detection, which studies the problem of automatically discovering users' social circles. Most current social networks allow a user to categorize his/her friends into social circles by manually assigning labels to them [2, 15]. The friends with the same labels are considered in the same circles. McAuley et al. propose an unsupervised method to learn which features of user profile lead to different social circles [12]. However, their work mainly focus on clustering current friends and do not provide a solution to recommend a circle

for a new added friend. Squicciarini et al. group a user's friends into social circles by extracting common interests from users' profiles and predict policies based on the group [19, 20]. However, their recommendation is based on the similarity of users or items, which is not suitable in group authorization mode.

Some works consider personalized recommendation among friends in the same circle [9, 16, 25], which are based on the facts that friends in the same circle often share similar interests. From this point, the detected circle structures are better than just using network topology information and the circle based recommendation gets a better accuracy. However, these approaches are not appropriate for privacy settings since privacy preferences are quite different from other social activities, such as buying goods or reading books. Even good friends may have totally different privacy preferences due to their characteristics such that these approaches are not applicable. Our purpose is to mine such preferences from groups and recommend similar privacy settings to users for new requirements.

3 The User Preference Based Privacy Policy Recommendation Model

In this section, we propose the user preference based privacy policy recommendation model and formally specify the basic notions. The basic idea of this model is to recommend a privacy policy for a user based on one's previous privacy settings. A privacy policy specifies who can access what private data, which contains three elements: subjects, objects and operations. In social network, a subject refers to a person or a friend group who are authorized to access items and an object refers to an item in a user data set, such as an uploaded photo or a blog. An operation means one of the accessing modes provided by a social service platform, such as "view" or "comment". So, our model focuses on two sides, one is to recommend appropriate accessing rights to a new added friend and the other is to recommend appropriate subjects accessing an uploaded item. All these recommendations are based on the analysis of user privacy preferences.

3.1 The Model

We consider two representative authorization modes in current social service platforms: the group mode and the appointed person mode. Our proposed privacy policy recommendation model consists of two parts: User Preference Mining and Privacy Policy Recommendation, shown as Fig.1.

User Preference Mining: User preference refers to user subjective tendency, which determines one's taken actions and choices under certain conditions. In social networks, user preferences consist of social links and actions (such as uploading pictures or grouping friends). For example, if a user establishes friendships with many people in a football club, he/she may be interested in football. Likewise, if a user posts, forwards or comments a certain topic, he/she may be interested in that topic. The contents, objects and time of user action together play a significant role in determining user preferences. Hence, we need to analyze user preferences according to their authorizations so as to recommend an appropriate privacy setting for each new uploaded item or new added person.

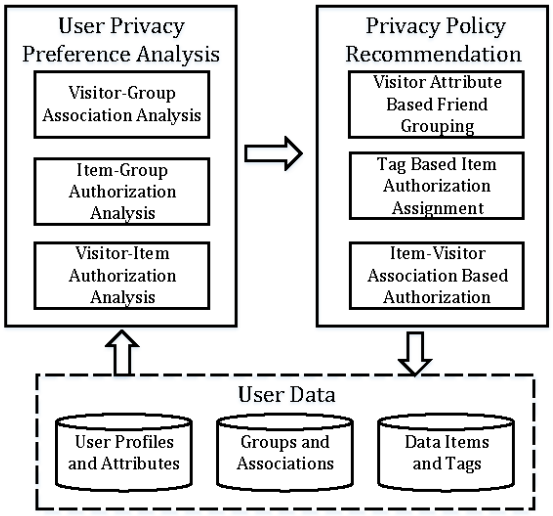


Fig.1 The user preference based privacy policy recommendation model

For the group based authorization mode, a group is regarded as the subject of authorization, where persons in one group get all the authorizations assigned to this group. We mine the hidden factors in friend-group associations against friend profiles, actions and groups in current settings, shown as the visitor-group associate analysis part in Fig.1. Likewise, we analyze the authorizations against groups and item tags so as to find the relationships between the access rights and items, shown as the item-group authorization analysis part. For the appointed person authorization mode, we directly analyze the relationship between visitors and items, shown as the item-group authorization analysis part in Fig.1.

Privacy Policy Recommendation: The privacy policy recommendation process considers the privacy settings on two cases. Considering a visitor requests a friendship for a user, some appropriate groups are recommended to this user based on user preferences and visitor attributes. After the user chooses some

groups, this visitor obtain all the access rights authorized to these groups. When a user uploads a new item, the recommendation algorithm computes some appropriate subjects who can access this item.

Consider the two popular authorization modes, the above process consists of three parts: the visitor attribute based friend grouping part, tag based item authorization assignment part and the item visitor association based authorization part. The first two parts are for the group authorization mode, while the last part is for the appointed person authorization mode.

Table 1 Notations

Symbol	Description
$u, \mathcal{U}$	a user and a user set
$v, \mathcal{V}_u$	a visitor and a set of u's visitors
$g_k, \mathcal{G}_u$	a group and a set of u's groups
$\mathcal{GV}_u$	the association matrix between u's groups and visitors
$e_j, \mathcal{E}_u$	a feature and a set of u's features
$\mathcal{VE}_u$	the association matrix between u's visitors and features
$\iota_t, \mathcal{I}_u$	an item and a set of u's items
$\mathcal{GI}_u$	the authorization matrix between u 's groups and items
$\mathcal{VI}_u$	the authorization matrix between u's visitors and items
$\tau_q, \mathcal{T}_u$	a tag and a set of tags associated with u's items
$\mathcal{IT}_u$	the authorization matrix between u 's items and tags

3.2 Basic Notions

In this subsection, we give the formal definitions of basic concepts in our method as follows. Some symbols and notations used in this paper are given in Table 1.

Definition 1 (*User*): A user u is a registrant in a social network, who uploads photos, videos, blogs, and other data objects. Let $\mathcal{U}$ be the set of all users in a social network.

Definition 2 (*Visitor*): For a given user $u \in \mathcal{U}$, a visitor $v \in \mathcal{U}$ refers to the user who requests to access u's data. All u's visitors are donated as a set $\mathcal{V}_u = \{v_1, v_2, \cdots, v_{|\mathcal{V}_u|}\}$, where $|\mathcal{V}_u| \in \mathbb{N}^+$ is the size of the visitor set.

Definition 3 (*Group*): For a given user $u \in \mathcal{U}$, a group g is the subset of u's visitors (i.e. $g \subseteq \mathcal{V}_u$). The size of group g is donated as $|g|$. A user often creates several groups according to his/her privacy requirements. The groups of u are represented as a set $\mathcal{G}_u = \{g_1, g_2, \cdots, g_{|\mathcal{G}_u|}\}$, where $|\mathcal{G}_u| \in \mathbb{N}^+$ is the size of the group set.

Definition 4 (*Visitor-Group Association*): For a given visitor $v \in \mathcal{V}_u$ and a group $g \in \mathcal{G}_u$, a visitor-group association (v, g) specifies that v belongs to group g.

Definition 5 (*Attribute*): Attributes include user identify and characteristics. Generally, a user has many attributes, such as name, sex, age, graduated university and etc. Each attribute is associated with a value.

All possible values of an attribute are called the attribute domain, which can be classified into a finite set of value ranges. For example, the value ranges of attribute *Age* could be represented as {0-20, 21-40, 41-60, 61-80, 81-100}. Attributes sometime can be extracted from user actions. For example, the frequency of keywords in user comments indicates the user's opinion.

Definition 6 (*Item*): An item refers to the data that is posted by a user in social network, such as blog, photo, and so on. In this paper, any object in a privacy policy is considered as an item. For a given user $u \in \mathcal{U}$, all items of u are represented as a set $\mathcal{I}_u = \{\iota_1, \iota_2, \cdots, \iota_{|\mathcal{I}_u|}\}$. Let $|\mathcal{I}_u| \in \mathbb{N}^+$ be the size of the item set.

Nowadays, the tag-based resource management has been widely used in social networks [13]. Many social network service providers offer sample tags for users to organize their items, such as time stamp and location mark. Some platforms allow users to annotate items with any tag they would like to describe an item, such as "family", "picnic" or "park" etc.

Definition 7 (*Tag*): A tag is a short text used to describe an item. Let $\mathcal{T}$ be the set of tags in social networks. All tags associated with u's items are represented as a set $\mathcal{T}_u = \{\tau_1, \tau_2, \cdots, \tau_{|\mathcal{T}_u|}\}$, where $|\mathcal{T}_u| \in \mathbb{N}^+$ is the size of this set.

In this paper, users express their privacy preferences on sharing items with visitors via their privacy policies.

Definition 8 (*Policy*): For a given user $u \in \mathcal{U}$, a privacy policy (Sub, Obj, Op) donates the subject Sub has the accessing Op right on the item Obj, where $Sub \in \mathcal{V}_u$ or $\mathcal{G}_u$, $Obj \in \mathcal{I}_u$ and $Op \in \{view, comment, download, share\}$.

In the group-based authorization mode, the subject refers to a friend group. For a given group $g \in \mathcal{G}_u$ and an item $\iota \in \mathcal{I}_u$, the group-item authorization (g, ι, Op) means group g is allowed to perform the operation Op on item ι. In the appointed person-based authorization mode, the subject refers to a specified person. For a given visitor $v \in \mathcal{V}_u$ and an item $\iota \in \mathcal{I}_u$, the person-item authorization (v, ι, Op)

means visitor v is allowed to perform the operation Op on item ι.

Example 1: Alice allows her family members *view* and *comment* the images in the album "*vacation*". Bob allows David to view and download a document named "*Job Data*". So Alice's privacy policies can be expressed as the group-based mode, while Bob's privacy policies can be expressed as the person-based mode, shown as follows:

Alice : [(*family; vacation; view*); (*family; vacation; comment*)]

Bob : [(*David; Job Data; view*); (*David; Job Data; download*)].

4 Personalized Privacy Policy Recommendation

In this section, we introduce the details of our personalized privacy policy recommendation algorithms. There are two core points in the algorithms: the quantitative evaluation of user privacy preferences and the probability-based personalized policy recommendation, which is illustrated in the following subsections.

4.1 User Preference Mining

The purpose of user preference mining is to find the relation between the authorizations and the characteristics of subjects or objects in the existing privacy policies. We consider the two representative authorization modes: the group authorization and the appointed person authorization, respectively. For the group based policy, we first analyze the visitor-group associations and study how properties of a subject impact user grouping. Then we study the item-group associations so as to find how properties of an object impact the authorizations of an item to groups. For the person-based policy, we analyze the person-item authorization and find out the relationship between subjects and authorized items.

4.1.1 Visitor-Group Association Analysis

The analysis of user grouping preference is based on the fact that users often classify visitors into several groups according to their attributes, just as the idiom "birds of a feather flock together". We adopt the statistical methods to compute the correlations between visitor attributes and groups. The larger their correlation, the more influence of this attribute to user grouping. That is to say these attributes influence more user privacy settings.

First of all, we analyze the existing user groups. For a given user $u \in \mathcal{U}$, all the visitor-group associations $(v,\ g)$, where $v \in \mathcal{V}_u$ and $g \in \mathcal{G}_u$, are described as a binary matrix $\mathcal{GV}_u \in \{0,1\}^{|\mathcal{G}_u|\times|\mathcal{V}_u|}$, where $|\mathcal{G}_u|$ is the size of u's group set and $|\mathcal{V}_u|$ is the size of u's visitor set. Each entry of the matrix $\mathcal{GV}(k,i) \in \{0,1\}$ refers to whether visitor v_i belongs to group g_k. $\mathcal{GV}(k,i) = 1$ indicates that the visitor v_i is allocated into the group g_k. Otherwise, $\mathcal{GV}(k,i) = 0$ indicates v_i is not in g_k.

Given a visitor $v \in \mathcal{V}_u$, the probability of v belonging to group g_k is computed as:

$$P(g_k^v) = \frac{1}{|\mathcal{V}_u|} * \sum_{i=1}^{|\mathcal{V}_u|} \mathcal{GV}(k,i) \quad (1)$$

, where $|\mathcal{V}_u|$ is the size of visitor set $\mathcal{V}_u$. Similarly, the probability of $v \notin g_k$ is donated as $P(\neg g_k^v) = 1 - P(g_k^v)$.

To measure the uncertainty whether visitor v belongs to group g_k, we adopt the information entropy of $v \in g_k$ as follows:

$$\mathrm{H}(g_k^v) = -\sum_{x\in\{g_k^v, \neg g_k^v\}} P(x) * log_2 P(x) \quad (2)$$

, which is the basic evaluation on the distribution of visitors in different groups.

In the following discussion, we analyze how each attribute impacts user grouping and further analyze how each attribute value determines the grouping results. We adopt the notion feature e to denote each value associated with its attribute. All attribute values could be represented as a set $\mathcal{E}_u = \{e_1,\ e_2,\ \cdots,\ e_{|\mathcal{E}_u|}\}$, where $|\mathcal{E}_u|$ is the size of features.

Example 2: Consider the attributes of education, hobby and gender. The illustrative values of three users v_1, v_2, v_3 are given below.

v_1 :[(education, "Purdue"), (hobby, "swimming"), (gender, "male")]

v_2:[(education, "SDU"), (hobby, "basketball"), (gender, "male")]

v_3:[(education, "Harvard"), (hobby, "movie"), (gender, "female")]

In this case, there are 8 entries in the feature set, i.e. education-"SDU", education-"Purdue", education- 'Harvard", hobby-"swimming", hobby-"basketball", hobby-"movie", gender-"male" and gender-"female".

For a given visitor $v \in \mathcal{V}_u$ and a feature $e \in \mathcal{E}_u$, we adopt the visitor-feature association (v, e) to donate whether visitor v has the feature e. All the visitor-feature associations are represented as a binary matrix $\mathcal{VE} \in \{0,1\}^{|\mathcal{V}_u|\times|\mathcal{E}_u|}$. Each entry $\mathcal{VE}(i,j) \in \{0,1\}$ indicates whether v_i has feature e_j.

Given a visitor v, the probability of owning feature e_j is defined as follows:

$$P(e_j) = \frac{1}{|\mathcal{V}_u|} * \sum_{i=1}^{|\mathcal{V}_u|} \mathcal{VE}(i,j) \quad (3)$$

, where $|\mathcal{V}_u|$ is the size of visitor set.

The probability that a visitor with feature e_j belongs to group g_k is computed as follows:

$$P(g_k^v|e_j) = \frac{1}{|\mathcal{V}_u|} * \frac{\sum_{i=1}^{|\mathcal{V}_u|} \mathcal{GV}(k,i) * \mathcal{VE}(i,j)}{P(e_j)} \quad (4)$$

This indicates the probabilistic association between group g_k and feature e_j.

4.1.2 Group-Item Authorization Analysis

Group item authorization analysis focuses on user preference of authorizing item access rights to which groups. A group-item authorization (g, ι, Op) indicates that group g has the accessing right Op to item ι. For simplicity of illustration, we take the *view* action as an example of actions in the following discussion. A group-item authorization is simplified to (g, ι), which represents that group g is allowed to *view* the item ι.

Given a user $u \in \mathcal{U}$, all group-item authorizations are represented as a binary matrix $\mathcal{GI}_u \in \{0,1\}^{|\mathcal{G}_u| \times |\mathcal{I}_u|}$, where $|\mathcal{G}_u|$ is the size of u's group set and $|\mathcal{I}_u|$ is the size of u's item set. Each entry $\mathcal{GI}(k,t) \in \{0,1\}$ represents whether a group-item authorization (g_k, ι_t) exists.

Given an item $\iota \in \mathcal{I}_u$, let $P(g_k^\iota)$ be the probability that group g_k can view it, defined as follows:

$$P(g_k^\iota) = \frac{1}{|\mathcal{I}_u|} * \sum_{t=1}^{|\mathcal{I}_u|} \mathcal{GI}(k,t) \quad (5)$$

Then we analyze how item properties influence such authorization. Since users are allowed to upload and share data in social network, such as photos and blogs, and adopt tags to express the content of these data, this analysis is performed on the tags associated with these items.

Given an item $\iota \in \mathcal{I}_u$ and a tag $\tau \in \mathcal{T}_u$, the item-tag association (ι, τ) indicates whether item ι has tag τ. All the item-tag associations are represented as a binary matrix $\mathcal{IT}_u \in \{0,1\}^{|\mathcal{I}_u| \times |\mathcal{T}_u|}$, where $|\mathcal{I}_u|$ is the size of u's item set and $|\mathcal{T}_u|$ is the size of u's tag set. Each entry $\mathcal{IT}(t,q)$ indicates whether item ι_t has the tag τ_q.

The probability that an item ι_t is associated with tag τ_q is defined as follows:

$$P(\tau_q) = \frac{1}{|\mathcal{I}_u|} * \sum_{t=1}^{|\mathcal{I}_u|} \mathcal{IT}(t,q) \quad (6)$$

Let $P(g_k^\iota|\tau_q)$ be the probability that group g_k could view the item ι associated with tag τ_q, defined as follows:

$$P(g_k^\iota|\tau_q) = \frac{1}{|\mathcal{I}_u|} * \frac{\sum_{t=1}^{|\mathcal{I}_u|} \mathcal{GI}(k,t) * \mathcal{IT}(t,q)}{P(\tau_q)} \quad (7)$$

This reflects the influence of tag τ_q on the authorization of group g_k.

4.1.3 Person-Item Authorization Analysis

For the appointed person authorization mode, we analyze the person-item associations. Firstly, we analyze the existing person-item authorization to explore the association between a granted visitor and an item. For a given user $u \in \mathcal{U}$, a person-item authorization (v, ι), where $v \in \mathcal{V}_u$ and $\iota \in \mathcal{I}_u$, denotes whether visitor v is allowed to view item ι. All person-item authorizations can be represented as a binary matrix $\mathcal{VI}_u \in \{0,1\}^{|\mathcal{V}_u| \times |\mathcal{I}_u|}$, where $|\mathcal{V}_u|$ is the size of u's visitor set and $|\mathcal{I}_u|$ is the size of u's item set. Each entry $\mathcal{VI}(i,t)$ indicates whether visitor v_i has the accessing right to item ι_t.

Since the preference on this type of authorizations is quite similar with the personalized recommendation systems, we adopt the well-studied collaborative filtering method to recommend the privacy setting on item authorizations.

4.2 Personalized Privacy Policy Recommendation Process

This process is to recommend appropriate privacy policy so as to help a user configure his/her privacy settings in social networks. The policy recommendation includes two aspects. When a visitor requests a friendship for a user, our method recommends some appropriate groups according to user preference. When a new item is added, some appropriate groups as well as some persons are recommended. Since the recommendation on groups in two cases are similar, we discuss these two sides together as the unified group-based policy recommendation.

For ease of presentation, we define a visitor or an item as an object o. All objects can be presented as a set $\mathcal{O}_u$, while $\mathcal{O}_u = \mathcal{V}_u$ or $\mathcal{O}_u = \mathcal{I}_u$. A property p refers to a subject feature or an item tag. Similarly, an object-property association matrix $M(o,p)$ denotes whether the object o has the property p. If so, $M(o,p) = 1$. Otherwise, $M(o,p) = 0$. The property set is defined as $\mathcal{P}_u$. $\mathcal{P}_u = \mathcal{E}_u$ when we compute the recommendation on visitor group associations and $\mathcal{P}_u = \mathcal{T}_u$ when we compute the recommendation on item group associations.

Given a new object o' and a group g_k, let $Close(o', g_k)$ be the closeness between o' and g_k. Given a user $u \in \mathcal{U}$, our method computes $Close(o', g_k)$ between o' and each group $g_k \in \mathcal{G}_u$. The group with the highest closeness $Close(o', g_k)$ will be recommended to user u as the privacy policy. In order to calculate the closeness from the different aspects, we propose four methods to measure the closeness between new object and the group.

The first method considers the impacts of all the features or tags in an equal effect way, which is formalized as follows:

$$Close(o', g_k) = \sum_{j=1}^{|\mathcal{P}_u|} M(o', p_j) * P(g_k|p_j) \quad (8)$$

, where $|\mathcal{P}_u|$ is the size of the property set and $P(g_k|p_j)$ represents the probability that an object with property p_j belongs to group g_k.

Since the odd could facilitate to understand the relative probabilities of an event occurring, we adopt odd express the influence of a property to a group. Let $O(g_k|p_j)$ be the odd that an object associated with property p_j belongs to group g_k defined as follows:

$$O(g_k|p_j) = \frac{P(g_k|p_j)}{P(\neg g_k|p_j)} \quad (9)$$

, where $P(g_k|p_j)$ represents the probability that an object with property p_j belongs to group g_k.

The second method considers both the occurrence probability of an event and the non-occurrence probability, which enlarges the difference among probabilistic distribution. The formal definition is shown as follows:

$$Close(o', g_k) = \sum_{j=1}^{|\mathcal{P}_u|} M(o', p_j) * O(g_k|p_j) \quad (10)$$

, where $|\mathcal{P}_u|$ is the size of the property set and $O(g_k|p_j)$ represents the odd that an object associated with property p_j belongs to group g_k.

The above two methods do not consider the differences between the impacts of features. However, sometimes a few attributes may play important positions than others. For example, in the group of Mensa club, the intelligence quotient is the most important feature, while other attributes such as gender, age etc. are trivial. So the following two methods take attribute weights into consideration when computing the recommendation.

In the third method, we adopt the relative mutual information between group g_k and property p_j as the attribute weight. Each feature influence on grouping is determined by the conditional entropy, which reflects the uncertainty for grouping the objects with this property. Let $H(g_k|p_j)$ be the conditional entropy:

$$H(g_k|p_j) = -\sum_{x\in\{g_k,\neg g_k\}} P(x|p_j) * log_2 P(x|p_j) \quad (11)$$

The correlation between object features and groups can be reflected by the mutual information between g_k and p_j , which is computed as follows:

$$I(g_k; p_j) = H(g_k) - H(g_k|p_j) \quad (12)$$

This illustrates how much uncertainty are reduced by this feature. The larger this mutual information, the greater the feature impacts on grouping.

The relative mutual information between g_k and p_j is defined as follows:

$$\rho(g_k; p_j) = \frac{I(g_k;p_j)}{H(g_k)} = 1 - \frac{H(g_k|p_j)}{H(g_k)} \quad (13)$$

The third closeness evaluation method is computed as follows:

$$Close(o', g_k) = \sum_{j=1}^{|\mathcal{P}_u|} M(o', p_j) * \rho(g_k; p_j) * P(g_k|p_j) \quad (14)$$

, where $|\mathcal{P}_u|$ is the size of the property set, $\rho(g_k; p_j)$ is the relative mutual information between g_k and p_j and $P(g_k|p_j)$ represents the probability that an object associated with property p_j belongs to group g_k.

The fourth method takes an overall view of all properties and evaluates how each property influences the degree of grouping confusion. We take into account the entropy of property p_j to different groups $H(\mathcal{G}_u|p_j)$, defined as follows:

$$H(\mathcal{G}_u|p_j) = -\sum_{k\in\{k|P(g_k|p_j)\neq 0\}} P(g_k|p_j) * log_2 P(g_k|p_j) \quad (15)$$

$$Close(o', g_k) = \sum_{j=1}^{|\mathcal{P}_u|} M(o', p_j) * \frac{P(g_k|p_j)}{H(\mathcal{G}_u|p_j)} \quad (16)$$

, where $|\mathcal{P}_u|$ is the size of the property set and $P(g_k|p_j)$ represents the probability that an object associated with property p_j belongs to group g_k.

5 Experimental Study

In this section, we present several experiments to evaluate the effectiveness and efficiency of our recommendation method over four real datasets. First we discuss the evaluation metrics. Then we introduce the experimental settings and analyze these experimental results.

5.1 Evaluation Metric

Given a test visitor set $\mathcal{V}_t$, we use the metric *Hit Rate* to evaluate the recommendation quality. For each visitor $v \in \mathcal{V}_t$, the recommended group is donated as $\bar{g}_v$, while the practical group is g_v. If the recommendation for v is correct, we donate it as $1(\bar{g}_v = g_v)$. *Hit Rate* measures the average percentage of correctly predicted members in test visitor set $\mathcal{V}_t$, which is formalized as follows:

$$Hit\ Rate = \frac{1}{|\mathcal{V}_t|} * \sum_{v\in\mathcal{V}_t} 1(\bar{g}_v = g_v) \quad (17)$$

For example, if we can correctly recommend 70 groups for 100 visitors, the *Hit Rate* is 0.7. For *Hit Rate* value of a recommendation method, the larger the better.

Another used metric is the Balanced Error Rate (BER), which is defined as the average of the errors between the predicted results and the actual results. BER is an equitable measurement of deviation from positives and negatives, and it is widely adopted [4]. BER is formalized as follows:

$$\text{BER}(\bar{g}_k, g_k) = \frac{1}{2}\left(\frac{|\bar{g}_k \setminus g_k|}{|\bar{g}_k|} + \frac{|\bar{g}_k^c \setminus g_k^c|}{|\bar{g}_k^c|}\right) \quad (18)$$

, where $|\cdot|$ is the size of a group and g^c is the complement of group g.

The accuracy of our recommendation is defined as follows:

$$Accuracy = \frac{1}{|\bar{\mathcal{G}}|} * \sum_{\bar{g}_k \in \bar{\mathcal{G}}} (1 - \mathrm{BER}(\bar{g}_k, g_k)) \quad (19)$$

, where $|\bar{\mathcal{G}}|$ is the size of the recommended group set. A larger *Accuracy* value means that the recommendation method can predict visitors' group or items' authorization more accurate.

Table 2 Dataset Statistics

Dataset	Facebook	Google+	Twitter
# of nodes	4,089	250,469	125,120
# of edges	170,174	30,230,905	2,248,406
# of groups	193	437	31400
# of node attributes	175	690	33569
avg. group size	28.76	143.51	15.54
avg. # of groups	1.36	0.25	0.39

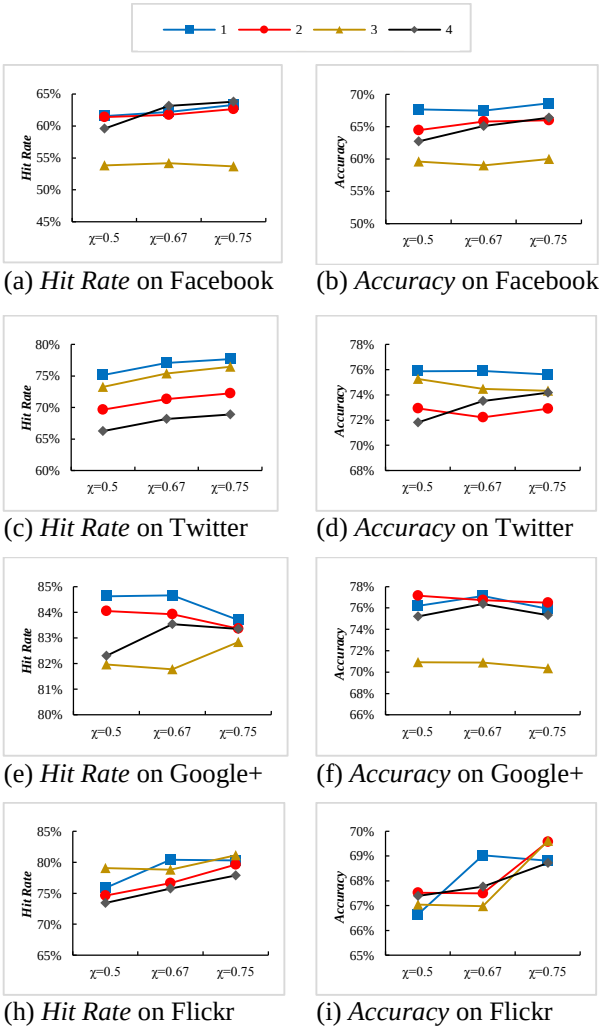
Dataset	Flickr
# of photos	10189
# of users	8698
avg. # of photos	1017
# of tags	27250
avg. # of tags	7.17
# of groups	6951

5.2 Experiment Setting

To evaluate our recommendation method, we select four social network datasets: Google+, Facebook, Twitter [12] and Flickr [6]. A brief description of the datasets is given in Table 2.We adopt the first three datasets to evaluate the effectiveness of predicting an appropriate group for a new visitor. Wherein, the data for Facebook were collected using a Facebook app, where people logged in and classified their friends into several lists. For Google+ and Twitter, the data were collected by their API. The Flickr data were used to measure the effectiveness of predicting an appropriate authorization for a new uploaded item. This Flickr data consist of over 10,189 images collected since 2007.

The experiments were conducted on a desktop with 2.90GHz CPU, 8GB memory and 500GB disk space installed the operating system MacOS. All experimental results are the average values of more than ten times of program running.

We verify the effectiveness and efficiency of our recommendation method. Experiments are performed on the Facebook, Twitter and Google+ data. We divide the data set into two parts, the training set and the test set. We mine the user preference based on the training set and use test set to evaluate the accuracy and performance of our recommendation results. To test the sensitivity of different scales of training set, we adopt *train/test ratio*, donated as χ, to describe the percentage of data used as the training and test sets [3]. For example, χ= 0.75 means that 75% of data are used as the training set and the other 25% of data are used as the test set. In this paper, we choose three values: 0.5, 0.67 and 0.75. On each ratio, we randomly select visitors as the training set.



(a) *Hit Rate* on Facebook (b) *Accuracy* on Facebook

(c) *Hit Rate* on Twitter (d) *Accuracy* on Twitter

(e) *Hit Rate* on Google+ (f) *Accuracy* on Google+

(h) *Hit Rate* on Flickr (i) *Accuracy* on Flickr

Fig. 2 Recommendation Quality Evaluation

5.3 Recommendation Quality Comparison

Figure 2 shows the comparison results of our four methods with different train/test ratios on *Hit Rate* and *Accuracy*, respectively. The blue lines with circle nodes denote recommendation algorithm 1, the red lines with square nodes denote algorithm 2, the yellow lines with triangular nodes denote algorithm 3, and the black lines with diamond nodes denote algorithm 4. From these figures, we find that the scale of training set has little effect on the recommendation results except method 4. This is because all these

methods are based on the association between groups and visitor features, which become already stable when χ is low. Considering method 4, the accuracy is a little sensitive to the ratio and increases as the size of training set. This illustrates the fact that it takes into account the importance of each feature for all groups.

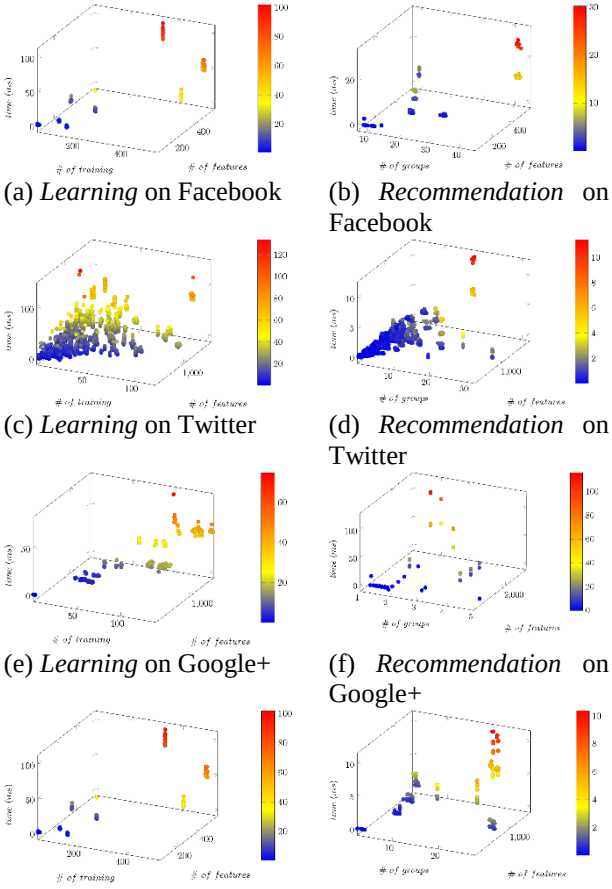
From the results, we can see that all the accuracy of four methods are more than 60 percent, which is a relative high value in recommendation. Method 1 seems better than others on all datasets. The method 3 and method 4 which take into account the importance of features do not increase the accuracy on predicting rational groups. Although the method 2 adopts a similar way, the odd adopted in this method increases the impact of feature for grouping. By further investigating the data characteristics, we find that there is not obvious bias on the user features in grouping, which also illustrates the above phenomenon.

Comparing the accuracy for different datasets, we could see that the *Hit Rate* and *Accuracy* are much high in Google+ dataset than in Facebook dataset. For example, when χ= 0.76, the *Hit Rate* under method 1 are 62.18% in Figure 2(a), 77.07% in Figure 2(c) and 84.66% in Figure 2(e); the *Accuracy* under method 1 are 67.50% in Figure 2(b), 75.89% in Figure 2(d) and 77.12% in Figure 2(f). There are three reasons. Firstly, there is a larger average of groups for each user in Facebook dataset than other two. This increases the probability of distributing a user into multiple groups. Secondly, the groups in Facebook data are manually organized by actual users, while the other two datasets are the public-visible groups about tweet follower relationships. Thirdly, there is less overlapping in Facebook data, while the groups overlap a lot in other two datasets.

5.4 Efficiency Analysis

Figure 3 shows the comparison of performance in generating the recommended groups. Since the training/test ratio has little effect for the recommendation results, we adopt χ = 0.5 in these experiments. We evaluate the elapsed time for two parts according to our method. The first part is mining user preference according to the training set (i.e. learning process). The second part of recommending a rational group for a visitor need run multiple times for different visitors in the test set. To investigate how different factors affect the performance, we adopt 3D scatter plot for analysis.

The figures in the first column of Fig.3 denote the learning algorithm, in which the x-axis is the size of training set, the y-axis is the size of feature set, and the z-axis is the elapsed time. The figures in the second column show the experiments on privacy setting recommendation. Differently the x-axis in these figures is the group size. The color of nodes in all these figures shows the difference of time, the red color denotes longer time while the blue color indicates shorter time. The figures (a)-(f) in Fig.3 show the experiments on visitor group recommendation. Overall, the performance is efficient. Although the learning time is much longer than the recommendation time, the learning process only requires run once in practice. The size of feature set has more impact on performance than other factors. For example, in Figure (c) it takes 35ms when x=102 and y=1153, while it takes 98m when x=102 and y=1979. Specially, the learning process on Google+ spends more time than others. This is because there are more attribute features than in other dataset. The figures (h) and (i) are the experiments on tag item grouping, the results in which are similar with the above discussion.



(a) *Learning* on Facebook (b) *Recommendation* on Facebook

(c) *Learning* on Twitter (d) *Recommendation* on Twitter

(e) *Learning* on Google+ (f) *Recommendation* on Google+

(h) *Learning* on Flickr (i) *Recommendation* on Flickr

Fig. 3 Performance Evaluation

6 Conclusion

In this paper, we investigate the problem of privacy setting for users in social networks. Considering the popular authorization modes in social platforms,

we propose a user preference based privacy policy recommendation approach so as to help a user better manage private data. We investigate user privacy preference from his/her own current privacy policies and recommend similar settings when a new friend is added or a new item is uploaded. To evaluate our methods, we propose several criteria and perform a lot of experiments on some practical datasets. The experimental results show that our algorithms are applicable for both person assignments and item management.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (61173140), Special Program on Independent Innovation & Achievements Transformation of Shandong Province (2014ZZCX03301) and Science & Technology Development Program of Shandong Province (2014GGX101046).

References

[1] Mohd Anwar, Philip WL Fong, et al. Visualizing privacy implications of access control policies in social network systems. *Data Privacy Management and Autonomous Spontaneous Security*, PP. 106–120. Springer, 2010.

[2] Yan Tang, Lili Lin, et al. Effective Social Circle Prediction Based on Bayesian Network. *Web Information System and Application Conference*, PP.131,135, 2014.

[3] Yi Cai, Ho-fung Leung, et al. Typicality-based collaborative filtering recommendation. 2013.

[4] Yi-Wei Chen and Chih-Jen Lin. Combining svms with various feature selection strategies. *Feature Extraction*, PP. 315–324, 2006.

[5] Scott H. Burton, et al. 2014. Discovering Social Circles in Directed Graphs. *ACM Transactions on Knowledge Discovery from Data*, Vol. 8, No. 4, Article 21, 2014.

[6] Mark Everingham, Luc Van Gool, et al. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[7] Lujun Fang and Kristen LeFevre. Privacy wizards for social networking sites. *19th international conference on World Wide Web*, PP. 351–360. ACM, 2010.

[8] Ai Ho, Abdou Maiga, and Esma A¨ımeur. Privacy protection issues in social networking sites. *Computer Systems and Applications*, PP. 271–278. IEEE, 2009.

[9] Hemank Lamba and Ramasuri Narayanam. Circle based community detection. *5th IBM Collaborative Academia Research Exchange Workshop*, PP. 16. ACM, 2013.

[10] Yabing Liu, Krishna P. Gummadi, et al. Facebook privacy settings: User expectations vs. reality. *2011 ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '11, PP. 61–70 , 2011. ACM.

[11] Alessandra Mazzia, Kristen LeFevre, and Eytan Adar. The pviz comprehension tool for social network privacy settings. *8th Symposium on Usable Privacy and Security*, PP. 13. ACM, 2012.

[12] Julian J McAuley and Jure Leskovec. Learning to discover social circles in ego networks. NIPS, Vol. 272, PP. 548–556, 2012.

[13] Mucheol Kim, Sang Oh Park, et al. Tag Based Collaborative Knowledge Management System with Crowdsourcing. *Journal of Internet Technology*, Vol. 14 No. 5, PP. 859-866, 2013.

[14] Yun Wei Zhao, et al. A Framework for Multi-Faceted Analytics of User Behaviors in Social Networks. *Journal of Internet Technology*, Vol. 15 No. 6, PP. 985-994, 11 2014

[15] Yuhao Yang, et al. 2014. Automatic Social Circle Detection Using Multi-View Clustering. *23rd ACM International Conference on Conference on Information and Knowledge Management*, PP. 1019-1028, 2014.

[16] Xueming Qian, He Feng, Guoshuai Zhao, and Tao Mei. Personalized recommendation combining user interest and social circle. 2013.

[17] Mohamed Shehab, Gorrell Cheek, et al. User centric policy management in online social networks. *2010 IEEE International Symposium on Policies for Distributed Systems and Network*, PP. 9–13. IEEE, 2010.

[18] Mohamed Shehab and Hakim Touati. Semi-supervised policy recommendation for online social networks. *2012 International Conference on Advances in Social Networks Analysis and Mining*, PP. 360–367, 2012.

[19] Anna Squicciarini, Sushama Karumanchi, Dan Lin, and Nicole DeSisto. Identifying hidden social circles for advanced privacy configuration. *Computers & Security*, 2013.

[20] Anna Squicciarini, Dan Lin, et al.. Automatic social group organization and privacy management. I Collaborative Computing: *8th International Conference on Networking, Applications and Worksharing*, PP. 89–96, 2012.

[21] Anna Cinzia Squicciarini, et al. A3p: adaptive policy prediction for shared images over popular content sharing sites. *22nd conference on Hypertext and hypermedia*, PP. 261–270, 2011.

[22] Niran Subramaniam, Joe Nandhakumar, et al. Exploring social network interactions in enterprise systems: the role of virtual co-presence. *Information Systems Journal*, 23(6):475–499, 2013.

[23] Nitya Vyas, Anna Cinzia Squicciarini, Chih-Cheng Chang, and Danfeng Yao. Towards automatic privacy management in web 2.0 with semantic analysis on annotations. *5th International Conference on Collaborative Computing: Networking, Applications and Worksharing*, PP. 1–10. IEEE, 2009.

[24] Ting Wang, Mudhakar Srivatsa, and Ling Liu. Fine-grained acces control of personal data. *17th ACM symposiu on Access Control Models and Technologies*, PP. 145–156. ACM, 2012.

[25] Xiwang Yang, Harald Steck, and Yong Liu. Circle-based recommendation in online social networks. *18th International Conference on Knowledge Discovery and Data mining*, ACM SIGKDD, PP. 1267–1275. ACM, 2012.