Volunteerism Tendency Prediction via Harvesting Multiple Social Networks

XUEMENG SONG, National University of Singapore
ZHAO-YAN MING, Digipen Institute of Technology
LIQIANG NIE, National University of Singapore
YI-LIANG ZHAO, Digipen Institute of Technology
TAT-SENG CHUA, National University of Singapore

Volunteers have always been extremely crucial and in urgent need for nonprofit organizations (NPOs) to sustain their continuing operations. However, it is expensive and time-consuming to recruit volunteers using traditional approaches. In the Web 2.0 era, abundant and ubiquitous social media data opens a door to the possibility of automatic volunteer identification. In this article, we aim to fully explore this possibility by proposing a scheme that is able to predict users' volunteerism tendency from user-generated contents collected from multiple social networks based on a conceptual volunteering decision model. We conducted comprehensive experiments to investigate the effectiveness of our proposed scheme and further discussed its generalizibility and extendability. This novel interdisciplinary research will potentially inspire more promising and important human-centered applications.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Systems**]: Information Search and Retrieval

General Terms: Algorithms, Experimentation, Performance

Additional Key Words and Phrases: Volunteerism tendency prediction, user classification, multiple sources, user-centric, network-centric

ACM Reference Format:
Xuemeng Song, Zhao-Yan Ming, Liqiang Nie, Yi-Liang Zhao, and Tat-Seng Chua. 2016. Volunteerism tendency prediction via harvesting multiple social networks. ACM Trans. Inf. Syst. 34, 2, Article 10 (February 2016), 27 pages.
DOI: http://dx.doi.org/10.1145/2832907

1. INTRODUCTION

Volunteerism was defined in Penner [2002] as long-term, planned, prosocial behaviors that can benefit strangers and occur within organizational settings. Persons exhibiting volunteerism are the so-called volunteers, serving socially and economically as an important work force in modern society. According to Renes [2005], society would face a major crisis without volunteers, especially nonprofit organizations (NPOs) because they are always in urgent need of volunteers to sustain their daily operations. Traditionally,

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.
Authors' addresses: X. Song, L. Nie, and T.-S. Chua, AS6, #5-25 Lab for Media Search, School of Computing, National University of Singapore, 13 Computing Drive, Singapore 117417; emails: sxmustc@gmail.com, nieliqiang@gmail.com, chuats@comp.nus.edu.sg; Z.-Y. Ming (corresponding author) and Y.-L. Zhao, 510 Dover Road, #03-01, Digipen Institute of Technology, Singapore 139660; emails: mingzhaoyan@gmail.com, zhaoyiliang@gmail.com.

DOI: http://dx.doi.org/10.1145/2832907

it is expensive and time-consuming for NPOs to aimlessly recruit volunteers from a huge crowd. It is thus highly desirable to develop an automatic volunteer identification system to alleviate the dilemma that a number of NPOs are facing [Song et al. 2015a].

In fact, several social researchers explored volunteerism analysis before the Web 2.0 era. These efforts are mainly based on survey data or related records of individual's volunteer activities [Wymer Jr and Samu 2002; Crosier et al. 2001]. Although great success has been achieved, these approaches suffer from two limitations: First, such approaches are hindered by limited and isolated samples, as well as by constrained individual characteristics. In particular, because the experimental data are collected via questionnaires or face-to-face interviews, only small-scale datasets and certain basic demographic information, such as gender, marital status, and income, are available. Second, they mainly focus on the correlation analysis between volunteerism and certain characteristics without quantitative volunteerism tendency prediction. For instance, Penner [2004] found that users' volunteerism tendency can be affected by four factors: demographic characteristics, personal attributes, volunteer activators, and social pressure.

On the other hand, with the popularity of social media services, there exists a large volume of User-Generated Content (UGC) that may reflect users' thoughts as well as opinions [Oh and Sheng 2011] and serve as indicators of users' attributes. Several efforts have been dedicated to research on the inference of users' attributes using these data. For example, some methods have been proposed to learn users' attributes such as gender, age, and personality from UGC [Popescu et al. 2010; Quercia et al. 2012]. Because users' demographic information and personality play a vital role in users' volunteerism tendency [Penner 2004], we believe that UGC has the potential to offer clues to the degree of a person's willingness to volunteer. Moreover, it is reported that 52% of online adults concurrently use multiple social media services.[1] This fact propels us to novelly explore users' distributed UGC from multiple social networks to approach the volunteer identification problem.

However, predicting users' volunteerism tendency by taking advantage of UGC from multiple social networks is nontrivial. First, it is not easy to generate a comprehensive overview of users from multiple heterogeneous social networks. The information about users from a single social network is often limited and incomplete [Zhu et al. 2013; Abel et al. 2013]. Thanks to the differing focus of different services, people participate in multiple social networks for different purposes. For example, people update their latest events in Facebook, search and share breaking news or interesting posts in Twitter, and construct abbreviated resumes as well as list their professional services in LinkedIn. Consequently, effectively aggregating all these different facets about users as revealed by different social media services is a challenge. Second, it is not clear how to effectively profile users from two angles, namely, user-centric and network-centric. It is a well-established fact that users on social networks are not isolated but interact with others. Therefore, users' behaviors are always affected by both intrinsic factors (those within themselves) and the extrinsic social environments they exist in. The heterogeneous natures of these two angles hinder the comprehensive understanding of users and pose a crucial challenge for us.

To tackle the task of volunteerism tendency prediction, we identify the following research problems:

(1) How can we aggregate diverse and heterogeneous user information across multiple social networks and construct comprehensive user profiles?
(2) How can we infer users' volunteerism tendency based on their profiles from both user- and network-centric angles?

[1]According to Pew Research Internet Project's Social Media Update 2014: http://www.pewinternet.org/.

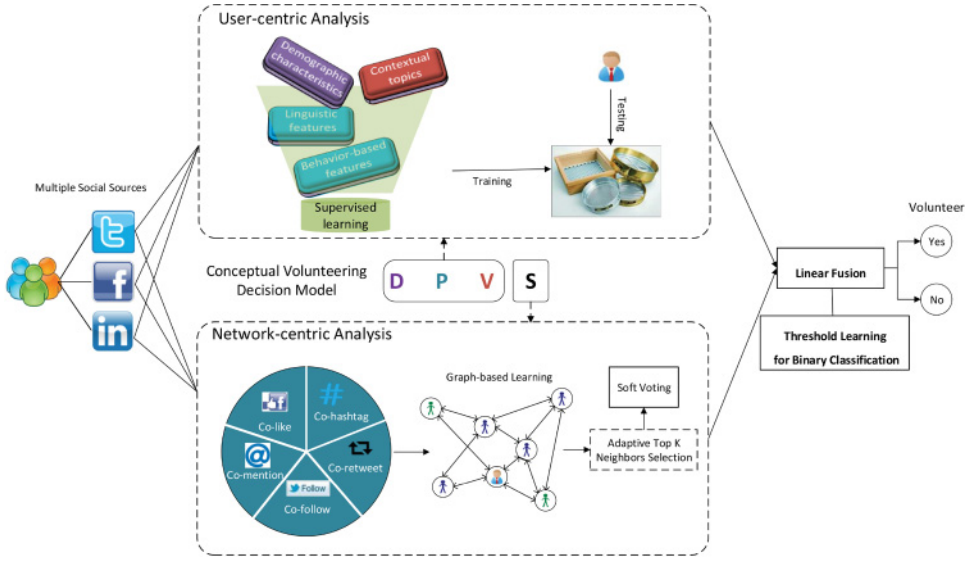


Fig. 1. The architecture of our proposed scheme. *D*: Demographic characteristics, *P*: Personal attributes, *V*: Volunteer activators, *S*: Social pressure.

(3) How can we obtain volunteers' data and construct ground truth for the study?
(4) What is the impact of utilizing multiple social networks on the performance of volunteerism tendency prediction?

We design a volunteer-discovering framework based on four key factors: demographic characteristics, personal attributes, volunteer activators, and social pressure, factors that have been well examined in social science [Penner 2004] but have not been explored in automatic information systems. In particular, our framework captures the following multifaceted user-centric information cues: demographic information, practical behaviors, historical posts, and bio descriptions of social connections, which capture the first three factors. We also propose a network-centric model to capture the social pressure factor. Figure 1 demonstrates our proposed scheme, which consists of two components from the user- and the network-centric perspectives, respectively. In the first component, the volunteer tendency prediction is framed as a binary classification task in which we categorize two classes of people: volunteers and non-volunteers. In the second component, a symmetric social graph is constructed in which vertices represent users and edges represent their heterogeneous relations. For each given user, we obtain a ranking list of his or her social neighbors via graph-based learning. Based on the ranking list, we then propose an adaptive soft voting approach for tendency prediction. Finally, we linearly fuse these two components to model the four factors completely.

Our main contributions are threefold:

—We introduce a novel task of predicting the volunteerism tendency of online users by harnessing and aggregating user information from multiple social networks.
—We propose a user profiling model from the intrinsic and extrinsic perspectives with user-centric and network-centric features. Our model covers novel volunteer prediction factors such as volunteer activators and social pressure, which are inspired by relevant research in social science.

—We conduct extensive experiments on the volunteerism tendency prediction task. To build the web-based volunteer dataset, we propose a novel strategy to piece together users' social content from multiple independent platforms.

The remainder of the article is structured as follows. Section 2 briefly reviews related work. Section 3 introduces the framework. The user-centric analysis and network-centric analysis are respectively introduced in Sections 4 and 5. Section 6 describes the experimental data. Section 7 details the experimental results and analysis. Section 8 discusses the generalizability and extendability of our scheme, followed by our concluding remarks in Section 9.

2. RELATED WORK

Our cross-discipline work is related to a broad spectrum of previous literature, including volunteerism analysis in social science study and inference of user attributes from social media contents.

2.1. Volunteerism

Volunteerism analysis has gained tremendous attention from scholars in social science in the past few years. These efforts mainly focus on exploring volunteering motivations and factors that affect the volunteering decision [Wymer Jr and Samu 2002; Davis et al. 1999; Carlo et al. 2005; Penner 2004]. Carlo et al. [2005] demonstrated that personality traits such as extraversion and agreeableness are positively associated with volunteerism. Extraversion characterizes people who are talkative, active, and keen on social activity, whereas agreeableness characterizes people who are cooperative, helpful, and sympathetic to others [Barrick and Mount 1991]. Another work in Penner [2004] presented an advanced conceptual model of factors that contribute to the decision to volunteer. The proposed factors are *Demographic Characteristics*, *Personal Attributes*, *Volunteer Activators,* and *Social Pressure*. Recently, an ongoing project for implementing a volunteer-matching service was introduced in Hitchen [2013]. This project aims to match students' specialties as well as interests with the needs of local nongovernmental organizations. It also enhances the "Town and Gown Relation" that exists between universities and the towns they reside in.

In spite of the compelling success achieved by these social science researchers, far too little attention has been paid to identifying volunteers from social media. Moreover, most of the existing efforts [Penner 2004; Carlo et al. 2005] employ survey or face-to-face interviews for data collection, which limits the scalability of their approaches. To bridge the gap, we propose our novel cross-discipline research aiming to enhance social welfare by exploring the large-scale information presented in social media.

2.2. Inference of User Attributes

Previous works have explored the potential of studying users' attributes from their social content and behaviors. Gender and age are the most popular personal attributes being investigated [Popescu et al. 2010; Otterbacher 2010; Rosenthal and McKeown 2011]. Otterbacher et al. [2010] showed that the gender of movie reviewers can be predicted based on stylistic, content, and metadata features. Bi et al. [2013] demonstrated that utilizing users' historical search queries can promote the inference of users' demographic characteristics such as age, gender, and political view. Furthermore, Pennacchiotti et al. [2011] described a general machine learning framework for user classification in three scenarios of political affiliation detection, ethnicity identification, and favor prediction for a particular business. Recently, Choudhury et al. [2013] studied the potential signals for the prediction of user depression from social media, ranging from a decrease in social activities and increased negative affect to

greater expressions of religious involvement. Interestingly, Song et al. [2015b] explored users' social content from multiple social networks to predict their interests. In addition to predicting an individual's attributes, Zhao et al. [2013] mined location-based social networks such as Foursquare to understand users' profiles at the community level.

Additionally, because personality has been verified to be of high relevance to volunteer behaviors [Cemalcilar 2009; Adali and Golbeck 2012], we particularly explored the literature on personality prediction. The widely approved "Big Five" personality model was first systematically introduced in McCrae and John [1998], and it represents an individual's personality on five broad dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience. Pennebaker et al. [1999] analyzed the linguistic features for each personality trait and developed a transparent text analysis tool in psychology—Linguistic Inquiry and Word Count (LIWC). Moreover, many studies have been conducted to examine personality traits over various social media, including blogs [Iacobelli et al. 2011; Yarkoni 2010], social networks [Schwartz et al. 2013; Quercia et al. 2012; Markovikj et al. 2013; Bai et al. 2012], and even the community question-and-answer forums [Bazelli et al. 2013].

The existing efforts focus on either some specific purpose, such as the prediction of depression, or on more general purposes for learning user profiles. By contrast, our work targets inferring users' volunteerism tendency based on their UGC to facilitate volunteer recruitment for social enterprises. Moreover, as far as we know, limited efforts have been dedicated to exploring users' attributes from multiple social networks, which is a major concern of our work.

3. VOLUNTEERISM TENDENCY PREDICTION FRAMEWORK

Inspired by the social science study on volunteer characteristics [Penner 2004], we propose to model four key factors: demographic characteristics (D), personal attributes (P), volunteer activators (V), and social pressure (S), as proposed in Penner [2004]. In our implementation, we model the first three factors (D, P, V) using user-centric features and the fourth (S) using network-centric features, which are all extracted from users' social media content. Overall, volunteerism tendency prediction will result in a volunteer or nonvolunteer decision.

In our user-centric analysis, we include an extensive set of personal attribute features: demographic characteristics, linguistic features, behavior-based features, and contextual topics. Several prevailing supervised machine learning models can then be applied to classify users, such as Support Vector Machine (SVM) [Järvelin and Kekäläinen 2002], Random Forests (RF) [Breiman 2001], and Gradient Boosted Regression Trees (GBRT) [Zheng et al. 2008], which will be detailed in Section 4.

In network-centric analysis, we study the factor of social pressure [Penner 2004] for volunteer characterization. We investigate users' social environments and study the effects of social influence. It is worth highlighting that, in contrast to user-centric analysis, this component performs analysis from the perspective of relations among users rather than from the contents posted by them or their social connections.

We linearly fuse these two components to enhance our final prediction. In particular, the probability of a given user u to be a volunteer is estimated as follows,

$$P_{hy}(u) = (1-\alpha)P_{user}(vol = 1|u) + \alpha P_{net}(vol = 1|u), \tag{1}$$

where $P_{user}(vol = 1|u)$ and $P_{net}(vol = 1|u)$ is the probability of user u to be a volunteer inferred from user-centric analysis and network-centric analysis, respectively. The tradeoff parameter $\alpha \in [0, 1]$ plays an important role in modulating the effects of these two models. Specifically, when α approaches zero, our scheme will be reduced to user-centric analysis only. On the contrary, when $\alpha = 1$, the optimal results will be

inferred solely from relation cues. Finally, based on the fusion score, we classify the target user as volunteer or nonvolunteer,

$$d(u) = \begin{cases} \text{volunteer} & \text{if } P_{hy}(u) \geq \gamma; \\ \text{nonvolunteer} & \text{otherwise.} \end{cases} \tag{2}$$

where γ is the volunteerism threshold parameter.

In the following two sections, we detail the two components of the prediction framework.

4. USER-CENTRIC ANALYSIS

In this section, we present a detailed and comprehensive analysis of features that concern a user's profile. In this analysis, we focus on those intrinsic features related to volunteerism. In particular, we model the basic demographic characteristics, personal attributes that involve users' post content and posting behaviors in social networks, and volunteer activators that come from users' social connections.

4.1. Demographic Characteristics

The study in Penner [2004] reports that some demographic characteristics, such as education and income level, are strong indicators for volunteerism in the United States. These studies drive us to extract demographic characteristics from users' form-based profiles. Form-based profiles correspond to the traditional methods used to organize user profiles, where a form is filled in by users. This source of information captures users' basic demographic information, such as gender, hometown, and education. These demographic characteristics serve as a strong prior indicator of whether a person will participate in volunteering activities. In our work, we explore users' demographic characteristics, including *Gender*, *Relationship status*, *Education level*, and *Number of social connections*.

4.2. Personal Attributes

Penner et al. [2004] pointed out that personal attributes, especially in the form of personality, are relatively strong predictors of volunteering behaviors. Previous studies in Yarkoni [2010]; Bai et al. [2012] have demonstrated the significant performance achieved by leveraging content-based linguistic features and behavior-based features to predict an individual's personality. Therefore, in this work, we characterize user personality by a set of linguistic features and behavior-based features.

4.2.1. Linguistic Features. Linguistic features include LIWC features as well as user topics, both extracted directly from users' own historical social posts.

LIWC Features. LIWC is a psycholinguistic transparent lexicon analysis tool that has been extensively validated as effective in users' personality inference [Markovikj et al. 2013; Bazelli et al. 2013]. The main component of LIWC is a directory containing the mapping of words to 72 categories.[2] Given a document, LIWC generates a vector to represent the percentage of words falling into each category. To capture dominance, we select the top 5 dimensions as the LIWC features according to the information gain ratio. Considering that an individual's emotion may also affect his or her volunteerism tendency, we incorporate two further categories: positive emotion and negative emotion. We use the positive-negative emotion ratio to further reflect users' emotional states. Let $L(\bullet)$ represent the percentage of users' words in a certain category. The

[2]http://www.liwc.net/.

Table I. Summary of Writing Content Patterns

Id	Feature	Definition
1	frac.emoticon	$\|\mathcal{EP}(u,n)\|/\|\mathcal{TP}(u,n)\|$
2	frac.slang	$\|\mathcal{SP}(u,n)\|/\|\mathcal{TP}(u,n)\|$
3	frac.hashtag	$\|\mathcal{HP}(u,n)\|/\|\mathcal{TP}(u,n)\|$
4	frac.url	$\|\mathcal{UP}(u,n)\|/\|\mathcal{TP}(u,n)\|$
5	frac.mention	$\|\mathcal{MP}(u,n)\|/\|\mathcal{TP}(u,n)\|$

positive-negative emotion ratio is defined as follows,

$$PN_{emo} = L(pos) log \frac{L(pos) + \xi_p}{L(neg) + \xi_n}, \tag{3}$$

where ξ_p and ξ_n are introduced to avoid the situation in which individuals have no positive or negative emotional words and both are set at 0.0001.

User Topics. Based on our observation, volunteers have, on average, a higher probability of talking about topics such as giving back or social caring, whereas non-volunteers mention other topics more often. This observation propels us to explore the topic distributions of users' historical social content. We generate topic distributions with Latent Dirichlet Allocation (LDA) [Blei et al. 2003; Li et al. 2010b], which has been widely demonstrated to be useful in latent topic modeling [Wang et al. 2008] and is able to alleviate the issue of word sparseness and vocabulary gap. It is also shown in the literature [Blei et al. 2003; Li et al. 2010b] that LDA outperforms other topic extraction methods such as LSA and PLSA.

4.2.2. Behavior-Based Features. These features are characterized by users' posting behavior patterns and egocentric network patterns. Posting behavior patterns focus on a user's writing style, whereas egocentric network patterns capture the features of his or her social connections.

Posting Behavior Patterns. Posting behavior patterns have been investigated in several scenarios spanning gender inference to age prediction [Yan and Yan 2006; Rosenthal and McKeown 2011; De Choudhury et al. 2013]. These patterns intuitively depict users' participation in information diffusion, which is closely correlated with volunteerism tendency.

On one hand, we employ a fraction of users' posts containing certain features (e.g., emoticons,[3] slang words, hashtags,[4] URLs, and user mentions[5]) to directly reflect users' engagement in topic discussions and social interactions. In particular, let $\mathcal{EP}(u,n)$, $\mathcal{SP}(u,n)$, $\mathcal{HP}(u,n)$, $\mathcal{UP}(u,n)$, $\mathcal{MP}(u,n)$ represent the set of posts that containing emoticons, slang words, hashtags, URLs, and user mentions of user u in social network n, respectively. Additionally, $\mathcal{TP}(u,n)$ stands for the set of posts of user u in social network n. Table I summarizes the writing content patterns.

On the other hand, we observe that the posting behaviors of users in social networks can be classified into several categories. For example, posts in Twitter can be easily classified into two main categories, $\mathcal{C}(tw) = \{tweets, retweets\}$, whereas those in Facebook can be roughly split into eight groups: $\mathcal{C}(fb) = \{share\_link, share\_video, share\_status,$

[3]An emoticon refers to a metacommunicative pictorial representation of users' facial expressions, such as ':)' and ':-('.

[4]A hashtag refers to a specially designated word in a tweet, prefixed with a '#', which usually represents the topic of this tweet (e.g., #topic).

[5]A user mention is a specially designated word in a tweet, prefixed with a '@', which usually refers to other users. E.g. @username.

share_photo, *change_photo*, *repost*, *post*, *tagged*}. Users' post distributions in these categories also reflect their participation in information diffusion, revealing whether a given user tends to share information in social media. Therefore, we compute the fraction of user posts belonging to each category c_i, defined as follows,

$$z(c_i, u, n) = |\mathcal{PP}(c_i, u, n)|/|\mathcal{TP}(u, n)|, \tag{4}$$

where $c_i \in \mathcal{C}(n)$, $\mathcal{C}(n)$ represent the categories in social network n, and $\mathcal{PP}(c_i, u, n)$ is the set of user posts in social network n falling into category c_i.

Specially, we use the profile completeness $\mathbf{k}$ to characterize users' posting behaviors in LinkedIn, which is defined by a boolean vector over six dimensions corresponding to the six most common sections in LinkedIn profiles: `summary`, `interest`, `language`, `education`, `skill`, and `honor`. We exclude the sections `experience` and `volunteer experience & causes` in order to avoid the bias introduced by the manual annotation. $\mathbf{k}$ is defined as follows,

$$k_i = \begin{cases} 1 & \text{if the corresponding section is presented,} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Egocentric Network Patterns. Apart from posting behavior patterns, we also capture users' social behaviors based on their egocentric networks. Intuitively, we believe that users from a certain class are likely to be connected with a set of class-specific accounts. Therefore, volunteers should interact with some typical accounts in social media. Let FV denote the set of typical accounts. Inspired by Pennacchiotti and Popescu [2011], we measure the degree of user correlation to volunteerism using three features: the fraction and frequency of this user's "friends" who belong to FV as well as the total number of "friends." In particular, we consider both followees and retweetings[6] as the user's "friends" in Twitter because of their direct indications of user's interests.

In order to construct FV, we take advantage of the Twitter profile repository Wefollow,[7] which allows us to find the most prominent people given a category. By crawling prominent users from Wefollow, we obtained 23,285 accounts that fall into the categories of *Nonprofit*, *Charity*, *Volunteer*, *NGO*, *Community Service*, *Social Welfare,* and *Christian*.

4.3. Volunteer Activators

Volunteer activators refer to a broad class of context stimuli that would activate the individual's desire to be a volunteer. For example, some images (e.g., a picture of a sick child) or messages evoke an individual's compassionate feelings. Moreover, these stimuli are more likely to be posted by users from volunteerism-relevant communities. Therefore, we model volunteer activators by users' contextual topics.

We define users' contextual topics as those topics extracted from their connections' social content. Particularly, we consider followee and retweeting connections in Twitter because of their intuitive reflection of topics that users are concerned about and interested in. However, considering the huge amount of Twitter followees, we choose to only consider their bio descriptions instead of their complete posts. The bio descriptions are usually provided by users upon joining Twitter to present a brief self- introduction. A bio may indicate a user's summarized interests. Therefore, we integrate the bio descriptions of a user's followees or retweeting connections. We then apply LDA on the bio documents. Similarly, the topic number is tuned based on perplexity. In this work, we only consider users' connections in Twitter since we are not able to

[6]If A broadcasted a tweet posted by B, then B is A's a retweeting user.
[7]http://wefollow.com/.

Table II. A Summary of User Centric Features

Feature	Description
D	Demographic characteristics, including gender, relationship status, education level and the number of social connections.
P	Personal attributes, characterized by linguistic features and behavior-based features.
P_ling	Linguistic features, including LIWC features and user-topics.
Liwc	LIWC features.
Topic	User-topics, extracted from users' historical social contents.
P_beha	Behavior-based features, including users' posting behavior patterns and egocentric network patterns.
Post	Posting behavior patterns.
Post_wr	Posting behavior patterns, extracted from Twitter and Facebook, including the fraction of users' posts containing certain writing styles such as emoticons.
Post_ca	Posting behavior patterns, extracted from Twitter and Facebook, including the fraction of users' post categories.
Post_in	Posting behavior patterns, extracted from LinkedIn, including profile completeness.
Net	Egocentric network patterns, extracted from users' social connections, including retweetings and followees, in Twitter.
V	Volunteer activators, characterized by users' contextual topics, extracted from users' social connections' profiles.
V_Retweeting	Users' contextual topics, extracted from users' retweetings' profiles.
V_Followee	Users' contextual topics, extracted from users' followees' profiles.

The indentation indicates the features' affiliations. The same notation will be used in our experiments.

crawl users' connections in LinkedIn, and the bio descriptions are usually missing in Facebook.

4.4. User-centric Classification

Based on the aforementioned user-centric features, we can apply several supervised machine learning models: SVM [Järvelin and Kekäläinen 2002], RF [Breiman 2001], and GBRT [Zheng et al. 2008]. SVM is a powerful machine learning method for classification tasks, especially binary classification. It aims to find the optimal separating hyperplane between two classes by maximizing the margin between the classes' closest points [Meyer and Wien 2014]. RF is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. In contrast to standard trees, where each node is split using the optimal split among all variables, RF splits each node using the optimal among a subset of predictors randomly chosen at the node [Liaw and Wiener 2002]. Consequently, RF is more robust against overfitting in the model learning stage. Similar to RF, GBRT is a machine learning model that is also based on tree averaging. However, in contrast to RF, which trains each tree separately using a random sample of data, GBRT trains one tree at a time, and each tree helps to correct the errors made by previously trained trees. Therefore, the final model becomes more expressive with each tree added.

The features used for our classification are summarized in Table II. We use some short-form notations to denote the features, which will also be used in the feature study experiment.

5. NETWORK-CENTRIC ANALYSIS

The network-centric analysis employs the graph-based soft voting approach to model the social pressure a user may experience from his or her connections. We generate a

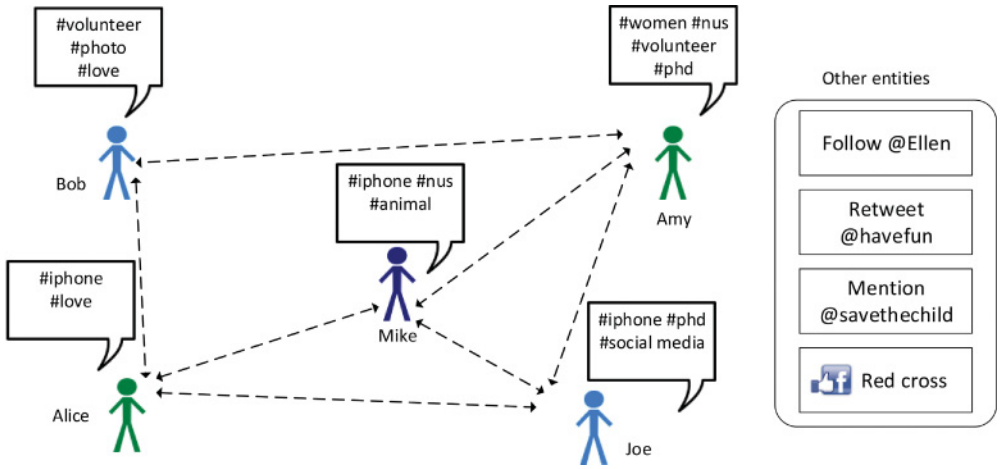


Fig. 2. Co-hashtag social graph. A black box contains all the hashtags embedded in a user's tweets. A dashed directional arrow is an implicit social connection between two users. The hashtag can be replaced by four other entities on the right: *Facebook likes*, *followees*, *mentioning users,* or *retweeting users*.

ranking list of a user's neighbors via graph-based learning. Based on this ranking list, we then conduct adaptive soft voting to predict users' volunteerism tendency.

5.1. Social Pressure

To facilitate the network-centric analysis, we first introduce the concept of social pressure. Social pressure originally refers to social influence coming from an individual's social connections. For example, some people who engage in volunteerism are subjected to direct or indirect requests from their friends. Therefore, the more volunteer friends an individual has, the higher the probability for him or her to be a volunteer. Distinguished from volunteer activators, social pressure is modeled here from the perspective of relations rather than from the contents of users' social connections. We consider two types of social relations: explicit and implicit. Explicit social relations are extremely sparse due to the limited scale of our current dataset. We observe that only 7,863 explicit follow connections exist in our dataset, whereas there are 3,065,885 *co-follow*[8] connections. This observation motivates us to explore implicit social relations because similar people may share similar behaviors, participate in similar topic discussions, and follow similar users/objects.

In this work, we categorize implicit relations into five categories: *co-follow*, *co-retweet*, *co-mention*, *co-hashtag,* and *co-like*. They can be derived from a user-object network, where the objects can be followees, retweetings (a post has been rebroadcasted once by u), mentioning users (referenced in u's posts), hashtags (embedded in u's posts), or like pages (voted for by u). As an example, Figure 2 shows the construction of a *co-hashtag* social graph. Since *Alice* and *Bob* have both discussed the topic of #*love*, there is an implicit social connection between them.

5.2. Graph-based Learning

To better capture relations among users, we use a traditional simple graph to represent their social environments. In the graph, the vertices are users and the edges reflect the strength of certain relations. The edge weight $Q(v|u)$ between users u and v on the

[8]Please refer to the later part of this section.

directional *co-hashtag* graph is computed as follows,

$$Q_{hash}(v|u) = \frac{|\mathcal{H}_v \cap \mathcal{H}_u|}{|\mathcal{H}_u|}, \tag{6}$$

where $\mathcal{H}_u$ is the set of hashtags that u has ever discussed, and $|\mathcal{H}_v \cap \mathcal{H}_u|$ is the number of hashtags discussed by both u and v. The strength of other relations, such as $Q_{rt}(v|u)$, $Q_{men}(v|u)$, $Q_{fol}(v|u)$, and $Q_{hash}(v|u)$ is estimated in the same manner. Notably, all the social relations are extracted from Twitter except the *co-like*, which is derived from Facebook. The overall strength between u and v can be aggregated as follows,

$$\begin{aligned} Q(v|u) = \beta_1 \times Q_{fol}(v|u) + \beta_2 \times Q_{rt}(v|u) + \beta_3 \times Q_{men}(v|u) + \beta_4 \times Q_{hash}(v|u) \\ + \beta_5 \times Q_{like}(v|u)), \end{aligned} \tag{7}$$

where β_i controls the contribution of corresponding relation and satisfies $\sum_i \beta_i = 1$.

Inspired by Nie et al. [2014], we take advantage of the social graph-based learning framework, which can be written as follows,

$$\begin{aligned} arg\min_{\mathbf{f}} \Phi(\mathbf{f}) &= arg\min_{\mathbf{f}} \{\Omega(\mathbf{f}) + \lambda\Psi(\mathbf{f})\} \\ &= arg\min_{\mathbf{f}} \left\{ \frac{1}{2} \sum_{u,v} W(u,v) \left(\frac{f(u)}{\sqrt{D(u)}} - \frac{f(v)}{\sqrt{D(v)}} \right)^2 + \lambda \sum_u [f(u) - y(u)]^2 \right\}, \end{aligned} \tag{8}$$

where λ is the regularization parameter; $\mathbf{y}$ denotes the initial relevance score between the given user and all other users; $\mathbf{f}$ denotes the final relevance score, measuring the semantic similarity between users; and $\mathbf{W} \in \mathbb{R}^{n \times n}$ is the adjacency matrix, defined as follows,

$$\mathbf{W} = \mathbf{Q}\mathbf{Q}^T, \tag{9}$$

where $\mathbf{Q}$ is the social connection matrix, with the (u, v)-element equaling $Q(v|u)$. In order to avoid self-loops, which may override the other connections, we set the diagonal elements of $\mathbf{W}$ to 0.0001. Then $\mathbf{D}$ is the diagonal "degree" matrix with its (u, u)-element equal to the sum of the u-th row of $\mathbf{W}$. Because $\mathbf{W}$ is symmetric, we have that

$$\begin{aligned} \Phi(\mathbf{f}) &= \sum_{u,v} W(u,v) \left(\frac{f(u)^2}{D(u)} - \frac{f(u)f(v)}{\sqrt{D(u)D(v)}} \right) + \lambda \|\mathbf{f} - \mathbf{y}\|^2 \\ &= \sum_u f(u)^2 \sum_v \frac{W(u,v)}{D(u)} - \sum_{u,v} \frac{f(u)W(u,v)f(v)}{\sqrt{D(u)D(v)}} + \lambda \|\mathbf{f} - \mathbf{y}\|^2 \\ &= \mathbf{f}^T(\mathbf{I} - \tilde{\mathbf{W}})\mathbf{f} + \lambda \|\mathbf{f} - \mathbf{y}\|^2, \end{aligned} \tag{10}$$

where $\tilde{\mathbf{W}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$. Taking derivations over Equation (10), we obtain an enclosed objective function as follows,

$$\frac{\partial \Phi(\mathbf{f})}{\partial \mathbf{f}} = (\mathbf{I} - \tilde{\mathbf{W}})\mathbf{f} + \lambda(\mathbf{f} - \mathbf{y}). \tag{11}$$

Setting Equation (11) to zero, it can be derived that,

$$\mathbf{f} = (1 - \eta)(\mathbf{I} - \eta\tilde{\mathbf{W}})^{-1}\mathbf{y}, \tag{12}$$

where $\eta = \frac{1}{1+\lambda}$. In order to find the most similar neighbors, we reorder $\mathbf{f}$ the descending order and get the ordered vector of relevance score $\mathbf{f}'$. Accordingly, we can obtain the ranking list of similar neighbors $\mathcal{V} = \{v_1, v_2, \ldots, v_I\}$ for a given user, where v_1 is the

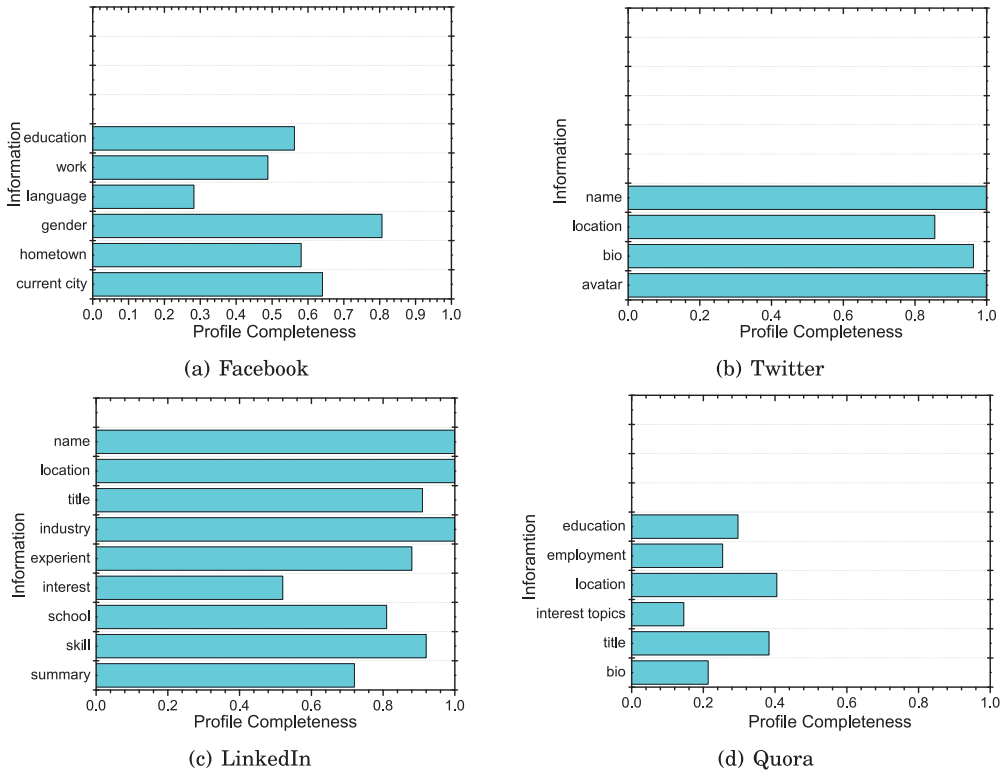


Fig. 3. Statistics of profile completeness of users over various social networks.

most similar one. According to the sorted vector $\mathbf{f}'$, we select the cutoff as the point that achieves the largest drop in $\mathbf{f}'$.

The graph-based soft voting is conducted as follows,

$$P_{net}(vol = 1|u) = \frac{1}{k}\sum_{i=1}^{k} s(v_i) f'_i, \tag{13}$$

where $s(v_i)$ is defined as follows,

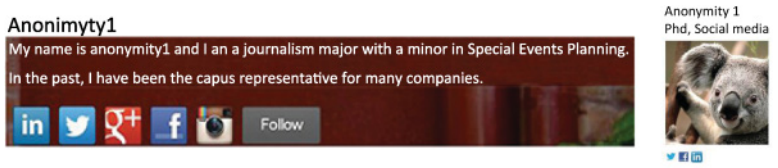
$$s(v_i) = \begin{cases} 1 & \text{if user } v_i \text{ is a volunteer;} \\ -1 & \text{otherwise.} \end{cases} \tag{14}$$

6. DATA COLLECTION

Since we aim to explore distributed UGCs to predict users' volunteerism tendency, the collection of users' social content from multiple social networks and the ground truth construction is a tough challenge for our work. In this section, we detail procedures for data collection and ground truth building.

6.1. Necessity of Multiple Social Networks

First, we provide the quantitative evidence to validate the necessity of collecting data from multiple social networks. We show the statistics of profile completeness of users over various social networks in Figure 3, based on our pilot study of 172, 235 users. We observe the following: (i) 56.2% users provide their education information in their



(a) About.me (b) Quora

Fig. 4. Screenshot of a user's About.me profile and Quora profile.

Facebook profiles, whereas 81% of LinkedIn users provide this information. This incompleteness hinders an effective similarity estimation based on users' profile data. (ii) The data distributed in different social networks is complementary. For example, Facebook profiles provide a user's gender information but fail to present a user's bio descriptions, which is alternatively given by Twitter profiles. Hence, integrating users' information distributed over various social networks is essential to derive complete user profiles. As a by-product, leveraging multiple sources increases the robustness, helps to handle the cold start problem [Schein et al. 2002], and may be beneficial to other applications, such as recommendations.

6.2. Multiple Social Accounts Alignment

Because complete user profiles that comprehensively describe users are the basis of our scheme, we start by finding users who are present on multiple social networks. In this work, we focus on three of the most popular social networks: Twitter, Facebook, and LinkedIn.

To find these users, we need to first tackle the problem of "Social Account Alignment," which aims to identify the same user across different social networks by linking one's multiple social accounts [Abel et al. 2013]. To accurately establish this alignment, we employ emerging social services such as About.me[9] and Quora[10] that encourage users to explicitly list their multiple social accounts on one profile. Figure 4 shows the screenshots of a user's profiles in About.me and Quora, respectively. From these screenshots, we can see that the bottom of each profile displays a list of external links to this user's other social network profiles. With these links, we can harvest a user's distributed social content from multiple social networks.

In particular, we proposed two strategies to collect data from About.me:

—**Keyword Search**: We first launched a search in About.me using the keyword "volunteer" and obtained 4,151 volunteer candidates.
—**Random Select**: We employed Random API,[11] provided by About.me, to collect nonvolunteers. This API returns a specified number of random user profiles. Finally, we harvested 1,867 nonvolunteer candidates. Note that volunteers may be present in these random users.

To enlarge our dataset, we also collected candidates from Quora using the breadth-first-search method. We only retained those who displayed their accounts in Twitter, Facebook, and LinkedIn.

[9]https://about.me/.
[10]http://quora.com/.
[11]http://about.me/developer/api/docs/.

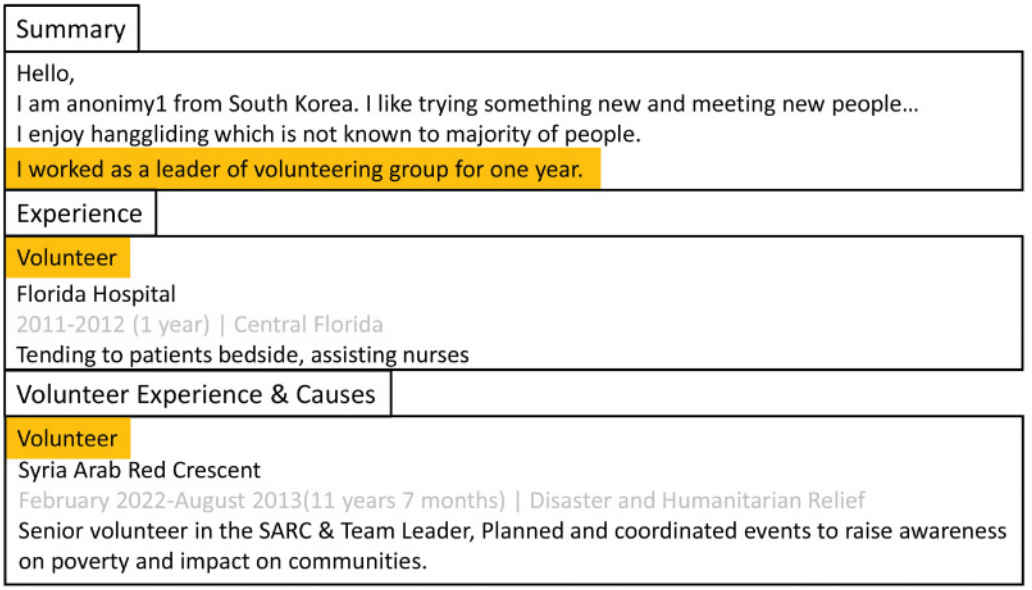


Fig. 5. Evidences of users' volunteerism services.

6.3. Ground Truth Construction

Based on these candidates, we launched a crawler to collect their social content and social relations. Note that we used Selenium[12] to simulate users' click and scroll operations on a FireFox browser and users' loading publicly available information in Facebook. Note that privacy constraints hinder us from accessing users' social relations in Facebook and LinkedIn, so we only collect users' followees in Twitter. Alternately, we collect the public pages voted on by users in Facebook.

To construct ground truth, we propose the following semi-automatic approach:

—Volunteer candidates who explicitly mention their volunteerism services in their LinkedIn summaries are tagged as volunteers. For example, a volunteer may mention "I'm working as a volunteer in rural Mozambique. . ." in the 'Summary' of his or her LinkedIn profile.
—Volunteer candidates who are approved by three annotators' majority votes based on their understanding of the candidate's LinkedIn profiles are treated as volunteers. For example, users may list their volunteer-oriented experiences in certain NPOs such as Save the Children.
—Candidates who do not satisfy the preceding two criteria are tagged as non-volunteers.

Figure 5 shows the available evidence of users' volunteerism services in LinkedIn profiles, which have been highlighted. In order to facilitate annotators in uniformly annotating volunteers, we provided them with guidelines. Given user u's LinkedIn profile, we classify him as a volunteer if and only if:

—u mentioned his or her volunteer experiences in the Summary section (e.g. ". . .I worked as a leader of a volunteering group for one year. . .'; see Figure 5).

[12]http://docs.seleniumhq.org/download/.

Table III. Statistical Summarization of the Constructed Dataset

Data	Volunteer				Nonvolunteer			
	Total	Min	Max	Std	Total	Min	Max	Std
Twitter tweets	$\sim 559k$	1	1000	337.79	$\sim 1m$	2	1000	356.28
Twitter followees' profiles	$\sim 902k$	1	5000	1138.4	$\sim 3m$	0	5000	1580.62
Facebook statuses	$\sim 83k$	0	514	106.75	$\sim 338k$	0	650	107.45
Facebook likes	$\sim 52k$	0	816	145.73	$\sim 143k$	0	815	113.87

—u listed his or her volunteer experiences in the Volunteer Experience & Causes or Experience section (see Figure 5).

We focus on LinkedIn to obtain volunteers due to the fact that volunteer experiences in LinkedIn are the most straightforward evidence by which to identify volunteers. It should be noted that those who do not mention their volunteer experiences in LinkedIn are not necessarily "nonvolunteers." However, the absence of these mentions at least suggests their limited interest in and low enthusiasm for volunteerism. Therefore, in our work, we broadly define users as "nonvolunteers" if they do not mention their relevant volunteerism experiences in LinkedIn.

Table III lists the statistics of our dataset. We obtained data for 1,425 volunteers and 4,011 nonvolunteers according to the aforementioned strategies. The crawl was conducted between August 22 and September 11, 2013. It is worth noting that we only leverage a subset of nonvolunteer data to avoid training bias. To facilitate the research community, this dataset will be released after certain privacy preservation measures are in place.

7. EXPERIMENTS

In this section, we conduct extensive experiments to verify the effectiveness of our proposed scheme and each of its components.

7.1. Data Preprocessing

We first remove obviously noisy contents through some filtering rules: Remove sentences that contain fewer than five words; remove sentences that contain more than four punctuation marks; remove sentences that contain fewer than two nouns plus verbs. For the remaining sentences that may contain a lot of noisy terms, such as URLs, user mentions, and Internet slang,[13] we did the following editing: (i) We removed the embedded URLs as well as user mentions. (ii) We replaced each slang with its corresponding formal expression. To be more specific, we first constructed a local slang dictionary containing 5,374 words obtained by crawling the Internet Slang Dictionary & Translator[14] where terms originate from various sources such as chat rooms and cell phone text. Given a UGC, we then transformed each slang to its formal expression using this dictionary. And (iii), we also performed lemmatization using the *Stanford NLP tool*[15] to link word variants.

7.2. Overall Performance

We compared the proposed scheme with each component of it. For user-centric analysis, we utilized GBRT as the learning model because it gives the best performance, as will be detailed in Section 7.3.

[13]Internet slangs refer to the variety of slang languages coined by Internet users, such as "lol," "omg," and "asap."
[14]http://www.noslang.com/.
[15]http://nlp.stanford.edu/software/tmt/tmt-0.4/.

Table IV. Performance Comparison Among Different Analysis (%)

	User-centric (D+P+V)	Network-centric (S)	Fusion (D+P+V+S)
Precision	**87.54**	64.82	87.05
Recall	87.00	82.45	**90.22**
F1-measure	87.24	72.25	**88.46**

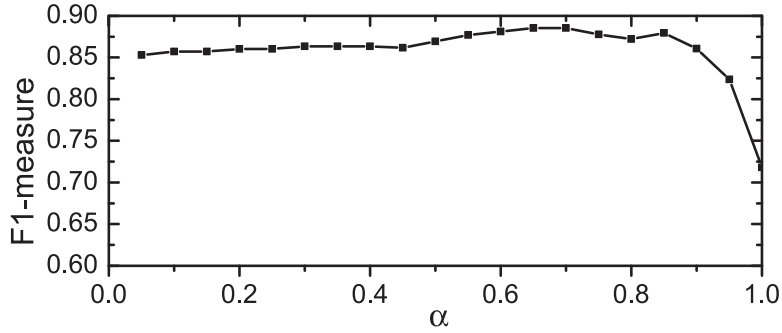


Fig. 6. Sensitivity study on the fusion parameter α in terms of F1-measure.

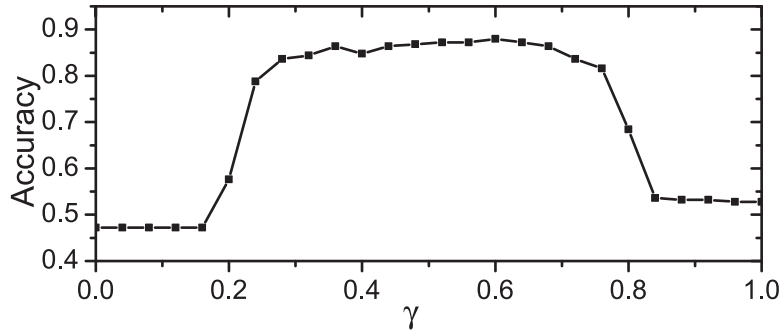


Fig. 7. Sensitivity of accuracy with respect to the threshold γ.

To demonstrate the effectiveness of the overall prediction framework, we present experimental results on the fusion of the user-centric and network-centric analyses in Table IV. From Table IV, we observe that the performance of each analysis is significantly boosted by fusion. We conducted the significance test and obtained the p-value of 0.013. The results verify that the intrinsic personal information and extrinsic social environment information of a given user are complementary to each other.

Because all the preceding analysis reported so far is based on the optimal results achieved by parameter tuning, we take a closer look at some important parameters involved in the whole scheme and explore their effects on the overall performance.

First, we studied the fusion parameter α, which controls the weight between the user-centric analysis and network-centric analysis in the whole scheme. Figure 6 shows the sensitivity curve of performance with different fusion parameters α. With the fusion parameter increasing from 0 to 1, the performance is quite steady, showing a slight increase until a sharp drop, which soon reaches the performance achieved solely by the network-centric analysis. According to Equation (1), this result shows that user-centric analysis is the main contributor of the whole prediction framework.

Second, we investigated the prediction threshold γ. Figure 7 shows the sensitivity curve of the prediction accuracy to the threshold. As can be seen, when the threshold

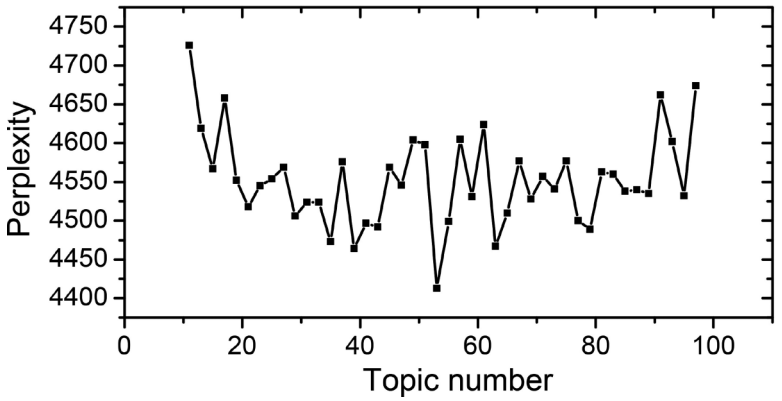


Fig. 8. Perplexity values varying over the number of topics in Twitter.

goes from 0 to 1, the accuracy increases from 0.56 first and then drops to 0.44 at the end. This corresponds to the fact that 56% of testing samples are volunteers. With the threshold's approach to the extreme minimum, all the testing samples are classified as volunteers, and the accuracy should be the percentage of volunteers, according to Equation (2).

7.3. User-Centric Analysis

7.3.1. Experimental Setting. Based on the aforementioned user-centric features, we investigated three prevailing supervised machine learning models: SVM, RF, and GBRT. First, for SVM, we chose the kernel with the Radial Basis Function (RBF) and employed the grid search method to obtain the optimal parameters, including the Gamma[16] $= 32$ and Cost[17] $= 0.0625$. Second, for RF, we set the number of trees at 100 and the number of considered features at each split at 50. Third, for GBRT, we set the learning rate $= 0.1$, number of trees $= 500$, and maximum depth of each tree $= 3$. In addition, the tradeoff parameter in the integration α, the regularization parameter λ in graph-based learning, and the threshold parameter γ are empirically set as 0.05, 0.5, and -0.14, respectively. We randomly performed 10 splits of the dataset and reported the mean over 10 trials.

For latent topic modeling, perplexity [Li et al. 2010a] is frequently utilized to find the optimal number of hidden topics. Figure 8 shows perplexity over different topic numbers on users' historical content in Twitter. Owing to the noisy nature of UGCs, the perplexity distribution can only roughly monotonically decrease as it approaches the lowest point from both ends. Consequently, it is advisable to set the topic number for Twitter at 53 based on the perplexity metric. In a similar manner, we ultimately obtain 26, 3-dimensional topic-level features over users' social content in Facebook and LinkedIn,[18] respectively.

To validate the usefulness of our model being applied to the real dataset, where volunteers are a minority, we tuned the fraction of volunteer samples in our dataset. In particular, we fed $x\%$, $x \in [5, 50]$, of volunteer samples to our model. Figure 9 shows the F1-measure at different fractions of volunteer samples. As can be seen, our model can achieve satisfactory performance when the volunteer samples contribute more than 20% of the whole sample. However, in practice, the percentage of volunteers among

[16]Gamma (g) is a parameter of the RBF kernel function.
[17]Cost (c) refers to the parameter controlling the balance between the accuracy and the generalization ability of the model.
[18]The posts in LinkedIn refer to the user summary section.

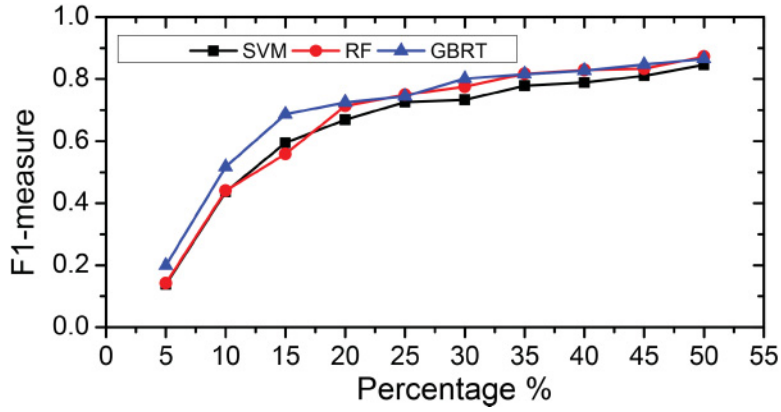


Fig. 9. F1-measure at different fraction of volunteer samples.

Table V. Overall Classification Results under Different Features and Algorithms (%)

Configuration	SVM			RF			GBRT		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
User-centric (D+P+V)	84.91	85.12	**84.95**	87.14	85.25	**86.14**	87.54	87.00	**87.24**
D	63.42	61.23	62.19	66.17	65.94	65.99	65.97	68.04	66.91
P	79.49	77.94	78.66	82.68	76.01	79.12	81.78	80.16	80.88
P_ling	74.35	70.69	72.39	80.11	62.74	70.24	74.50	68.93	71.49
Liwc	65.31	67.98	66.52	72.74	57.25	63.96	68.33	63.14	65.51
Topic	74.25	69.81	71.88	79.82	63.20	70.38	73.68	68.14	70.69
P_beha	75.60	75.62	75.56	75.87	76.84	76.28	76.59	77.20	76.83
Post	64.38	66.87	65.51	66.71	61.92	64.18	65.92	62.99	64.38
Post_wr	60.81	58.45	59.44	62.27	55.43	58.54	63.05	57.78	60.18
Post_ca	64.53	59.87	61.78	64.63	56.85	60.44	61.90	59.16	60.41
Post_in	57.99	68.15	62.59	58.07	67.76	62.45	58.16	67.68	62.48
Net	68.90	76.53	72.28	69.15	71.75	70.36	70.69	73.68	72.09
V	76.92	77.23	77.01	77.12	78.55	77.78	76.97	77.54	77.20
Retweeting	68.94	63.13	65.81	66.49	62.19	64.09	66.09	63.47	64.63
Followee	77.14	75.70	76.31	77.28	76.78	76.99	77.35	75.89	76.53

Prec: Precision; Rec: Recall; F1: F1-Measure. *D*: Demographic Characteristics; *P*: Personal Attributes; *V*: Volunteer Activators. The indentation indicates the features' affiliations. Please refer to Table II for feature descriptions.

social media users is likely much lower than 20%. Given any machine learning model, in practical cases, the distribution of the test set will be different from the training set. Ways to handle the differences are presented in the literature. For example, the outputs of a classifier can be adjusted to new a priori probability using the EM procedure if the difference lies only in the relative class frequencies in the training and test sets, which is exactly the case in our application. If the distribution of patterns within each class changes, then the problem is known as "covariate shift." Logistic regression can be used to predict whether a pattern is drawn from the training set or the test set and to weight the training data accordingly. These are general engineering issues one needs to handle when a learned model is applied to a practical case.

7.3.2. Classification Result and Feature Combinations. Table V shows the prediction performance of user-centric analysis with different feature configurations. Alternatively, we

Table VI. Feature Ablation Study: Overall Classification Results under Different Feat ures and Algorithms (%)

Configuration	SVM			RF			GBRT		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
User-centric (D+P+V)	84.91	85.12	**84.95**	87.14	85.25	**86.14**	87.54	87.00	**87.24**
X_D	83.98	84.19	84.04	87.20	84.20	85.61	86.97	85.58	86.22
X_P	77.84	78.85	78.22	78.54	81.75	80.04	80.35	81.30	80.75
X_p_ling	80.08	82.22	81.03	81.25	87.10	83.99	83.06	84.90	83.91
X_liwc	85.16	83.90	84.47	86.25	85.26	85.71	87.07	86.80	86.87
X_topic	80.01	83.08	81.43	83.06	84.31	83.62	84.32	85.16	84.66
X_p_beha	83.27	81.89	82.47	86.21	80.29	83.07	86.29	83.04	84.57
X_post	85.52	84.45	84.86	87.07	84.77	85.87	86.40	85.99	86.15
X_post_wr	85.17	84.12	85.55	86.72	85.81	86.22	87.40	86.51	86.93
X_post_ca	85.07	85.08	85.09	86.70	85.48	86.04	87.18	86.30	86.69
X_post_in	85.01	84.73	84.94	86.64	85.39	85.95	86.64	86.52	86.54
X_net	82.27	81.43	81.76	85.96	81.43	83.56	87.21	84.38	85.71
X_V	80.39	78.54	79.39	83.89	78.31	80.98	84.11	82.33	83.15
X_retweeting	85.14	84.36	84.68	86.97	85.11	85.97	86.48	85.54	85.96
X_followee	80.54	79.62	80.02	83.74	77.60	80.49	83.34	81.86	82.56

Prec: Precision; Rec: Recall; F1: F1-Measure. *D*: Demographic Characteristics; *P*: Personal Attributes; *V*: Volunteer Activators. The indentation indicates the features' affiliations. Please refer to Table II for feature descriptions. Each feature configuration prefixed by "X" means the absence of corresponding features.

compared the performance of user-centric analysis by removing each set of features in Table VI.

From the two tables, the following observations can be made:

First, removing any set of features can devastate the performance in terms of the F1-measure, more or less. This demonstrates that the set of features we developed for user-centric analysis are not redundant but complementary to each other.

Second, personal attributes (P) are the most predictive set of features, whereas the demographic characteristics (D) are the weakest. This confirms that the features related to personal attributes, including linguistic features and behavior-based features, are of significant value in user-centric analysis. Demographic characteristics contain general information that does not help much in predicting users' volunteerism tendency.

Third, the volunteer activators (V) are more prominent than the users' own topics (topic under P). We use contextual topics as the volunteer activators, and these are extracted from the bio descriptions of a user's social connections. This may be due to the fact that these bio descriptions are usually better written and more highly summarized on the user's connections compared to his or her casual posts. This observation shows the potential value of bio descriptions of users' social connections in terms of exploring users' volunteerism tendency.

Demographic characteristics are relatively straightforward. In the following, we further analyze the set of personal attribute features and the set of volunteer activator features.

7.3.3. Personal Attributes (P). We make further observations within the personal attributes. First, behavior-based features (P_beha) outperform the linguistic features (P_ling) in terms of the F1-measure. This reveals that the volunteerism tendency is better reflected by their behaviors, especially their networking behaviors, compared to the content they talk about in social media. This also implies that users with a volunteerism tendency may choose to interact frequently with some typical accounts but may talk little about the topic.

Table VII. Comparison of Profile Completeness between Volunteers and Nonvolunteers (%)

	Volunteer	Non-volunteer		Volunteer	Nonvolunteer
Interest	61.02	50.78	Honor	22.36	10.37
Education	91.58	79.84	Language	47.15	39.42
Skill	96.02	95.06	Summary	77.63	71.60

Table VIII. Comparison of the Value of LIWC Features between Volunteers and Nonvolunteers (%)

	Category	Example	Volunteer	Nonvolunteers
1	see	view, seen	1.00	0.95
2	health	clinic, flu	0.48	0.37
3	family	daughter, son	0.22	0.17
4	first person singular	i	2.52	2.26
5	body	hands, spit	0.43	0.40
6	positive	love, great	4.76	4.53
7	negative	hurt, ugly	1.36	1.37
8	PN_emo	-	7.37	6.84

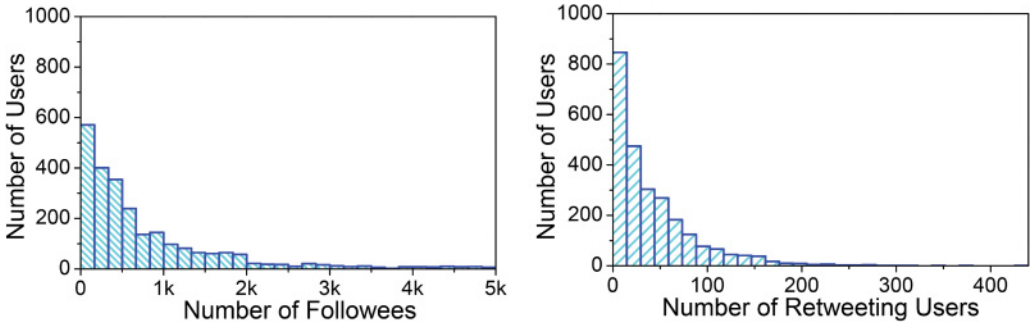
Second, profile completeness achieves the best performance among the three kinds of posting behavior patterns: written patterns (Post_wr), posting categories (Post_ca), and profile completeness (Post_in). To explore the underlying reason, we compared this feature among users belonging to different classes. Table VII shows the comparison between two classes of people in terms of their posting behavior patterns in LinkedIn. As can be seen, we found that volunteers tend to provide more information for all the sections. This not only reflects volunteers' active participation in LinkedIn but also reveals their self-confidence and openness to the public.

Third, LIWC does not contribute much compared to the other two personal attribute features. To figure out the underlying logic, we took a close look at the comparison between users belonging to different classes. Table VIII comparatively lists the average values of these features among volunteers and nonvolunteers. According to Holtgraves [2011], Extraversion [McCrae and John 1998] was much more positively associated with the use of personal pronouns, especially the first-person singular. This offers a good explanation of volunteers' larger adoption of category "I" in that volunteers tend to be more open that nonvolunteers. Additionally, we can infer that volunteers are more concerned with health than are nonvolunteers because more of their reference words belong to the categories "health" and "body." Moreover, words from the sensory category "see" occur more in volunteers' posts. This may be due to the fact of volunteers' active participation in activities and their willingness to propagate information in social networks. After checking volunteers' posts, we found that volunteers do frequently share posts in the following patterns: *"... glad to see..."* and *"... see this proposal: URL."* Nevertheless, we observed that the difference between people of two classes is not significant.

7.3.4. Volunteer Activators (V). We make further observations within the set of volunteer activator features.

First, the profiles of users' retweetings are not strong signals to detect volunteering tendency compared to that of users' followees. To further understand this, we gather the statistics about the two types of connections. From Figure 10, users' followee profiles are much richer (about 10 times) than their retweeting profiles.

Second, utilizing a latent topic to describe the contents from users' connections outperforms the explicit use of the bag-of-word scheme. In Table IX, we give an example of



(a) Histogram of Followees (b) Histogram of Retweeting users

Fig. 10. The distribution of the number of users with respect to the number of social connections.

Table IX. Discriminative Topics Extracted from Different Profiles and Dominated by Volunteers

Data source	Topic words
Followee bios†	• public, politics, rights, development, human, government, views
	• editor, global, journalist, university, research, science, international
Retweeting bios†	• global, nonprofit, change, community, development, rights, human
	• health, education, learning, university, research, student, national
User topics♭	• woman, help, education, child, change, world, community
	• volunteer, nonprofit, support, community, service, donate

†Refers to the contextual topics corresponding to the factor (V) and ♭ refers to the user topics corresponding to the factor (P).

Table X. Overall Classification Results with Different Sources (%)

Data Source	Precision	Recall	F1-measure
LinkedIn	65.44	62.00	63.59
Twitter	80.91	83.87	82.30
Facebook	73.27	65.83	69.28
LinkedIn+Facebook	76.65	71.22	73.73
LinkedIn+Twitter	81.92	85.14	83.46
Twitter+Facebook	87.53	86.41	86.87
LinkedIn+Twitter+Facebook	**87.54**	**87.00**	**87.24**

the highly discriminative topics extracted from different profiles. We can see that the latent topics do make sense in our example.

7.3.5. Source Comparison. In addition, we incrementally integrated social sources to validate that information from multiple sources is not redundant but complementary. Table X shows the classification results with different combinations of sources. Obviously, the performance based on multiple sources is better than that based on any single source. Interestingly, we observed that LinkedIn contributes the least to the task, which may be caused by users' limited activities in LinkedIn. Usually, users update less often in LinkedIn compared to in Twitter and Facebook. Moreover, users update their professional activities instead of casual life events in LinkedIn. It is also worth noting that although the ground truth is harvested from LinkedIn Volunteer Experience & Causes and Experience data, we did not make use of these data in the proposed scheme. Therefore, the contribution of LinkedIn data is not significant.

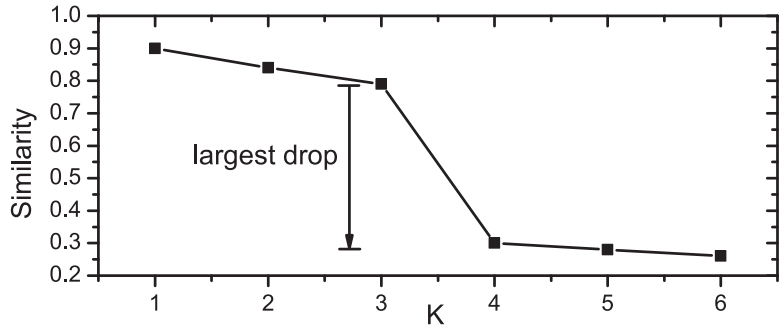


Fig. 11. Illustration procedure for parameter k in the adaptive soft voting.

Table XI. Overall Classification Results by Exploring Various Relations

	Follow	Retweet	Mention	Hashtag	Like
Precision	**63.25**	49.12	54.38	54.73	57.89
Recall	84.51	73.42	86.48	**94.39**	31.96
F1-measure	**71.75**	58.56	65.64	69.24	40.95

Table XII. Graph-based Voting Performance Comparison among Different Combinations of Sources (%)

Data Source	Precision	Recall	F1-measure
Twitter	59.70	**93.56**	**72.67**
Facebook	57.89	31.96	40.95
Twitter+Facebook	**64.82**	82.45	72.25

7.4. Network-Centric Analysis

We first tuned the parameter for the graph-based network-centric analysis. Specifically, to set the cutoff for the ranking scores in Equation (13), we select the cutoff k as the point that achieves the largest drop on the sorted scores, as illustrated in Figure 11, where the largest drop occurs when $k = 3$.

We then evaluated the effects of different relations to our proposed graph-based soft voting approach. Table XI shows the comparison results. We observed that this relation-based analysis achieves high recall but low precision. This may be attributed to the fact that, compared to nonvolunteers, volunteers tend to be more active and sociable in social networks, follow more popular accounts, and participate in the discussion of hot topics. Therefore, users are more likely connected to volunteers than to nonvolunteers, which increases the recall. Noticeably, the most prominent social connection is the *co-follow* relation, whereas the *co-like* relation fails to produce satisfactory results. A possible explanation is the limited information on users' Facebook likes because only about 50% of users' Facebook likes are available. Moreover, we check the norm of the *co-like* matrix, which is much smaller than the value of the other four matrices.

Analogous to user-centric analysis, we also validated the effectiveness of using multiple sources. As mentioned earlier, the social connections in LinkedIn are not trivial to obtain. Thus, we only considered relations in Twitter and Facebook. Table XII shows the results. Due to noise and sparseness on the *co-like* graph obtained from Facebook, only precision is boosted via the combination of multiple social sources.

8. MODEL APPLICATION AND POSSIBLE EXTENSIONS

In this section, we discuss the generalizability and extendability of our scheme.

8.1. General Observations and Model Application

In this study, we discovered some general patterns of volunteers that may shed light on the recruiting process. First, users' behaviors, especially their networking behaviors and the bio descriptions of their social connections, reveal more about volunteer potential than does the content they post on social networks. It may simply confirm the old saying that actions speak louder than words.

Second, users who are willing to reveal their personal abilities on social networks are generally good candidates. This is verified by our observation of the correlation between LinkedIn profile completeness and users' volunteerism tendency. This is understandable because volunteers tend to be more self-confident and open to the public. This tendency could be captured even earlier on social networks where users may "show off" their abilities to serve others. This is also one key motivation of this work: to facilitate the processes of both volunteerism work-seeking and volunteer-seeking.

These key observations and our volunteerism tendency prediction model can help to better bridge between NPOs and potential volunteers. According to a report from Volunteering Queensland Inc.,[19] we can broadly summarize the current processes of recruiting volunteers for NPOs as follows:

(1) Develop volunteer roles.
(2) Write volunteer job descriptions.
(3) Develop the message and advertise.
(4) Broadcast recruitment; in particular, the recruitment message can be broadcast through a variety of ways, such as the Internet, newspapers, community billboards, and even word-of-mouth.
(5) Interview.
(6) Screen and select.

Our work can be applied in Steps 4–6. For Step 4, instead of aimlessly broadcasting recruitment messages, NPOs may target their advertising toward more potential users according to activeness and openness criteria. NPOs can also contact those who publish their contact information through social networks. This would greatly increase the efficiency of volunteer recruitment. For Step 5 and 6, NPOs can also put more emphasis on the openness and activeness of candidates in order to find those who will work well and enjoy the work at the same time.

As a result, increasing our understanding of what makes people volunteer and enhancing the channels that bridge volunteer supply and demand can help volunteers and NPOs to reach each other in a more cost- and time-efficient way. More importantly, this will increase volunteer candidate quality and the satisfaction of volunteers at the same time.

In addition, the benefits of the proposed model can be measured at two stages: the promotion of volunteer spirit and the conversion of potential volunteers to actual engagement. We start off by harvesting social media users with public accounts and profiles. After classifying them using our model, we identify potential volunteers, on whom we will check actual outcomes in the two stages. For the first stage, we contact them through their publicly available social media contacts and ask if they would like to be a volunteer at any time in the future. Multiple causes are listed to encourage them to take part in any that may be interesting. The ratio of positive responses can be used as a quantitative measure, and we can compare this ratio with sending requests to unclassified public accounts.

[19] http://volunteeringqld.org.au/web/.

For the second stage, we look at the conversion rate. We will work with NPOs to get some volunteer openings and broadcast them to those who are willing to take a position. The final outcome will be measured by checking the ratio of potential volunteers who turn up and finish the tasks. As one can expect, the second stage may take months, even years to finish, because people may not be available in the near future or at the time we contact them, although they are still open to taking a volunteer job when there are suitable time slots and jobs for them.

8.2. Generalization and Extension

The user modeling scheme in the volunteerism tendency prediction task is generalizable to other application scenarios. Integrating heterogeneous information across multiple sources is beneficial to many other applications involving user modeling. For instance, our scheme can potentially tackle the problem of age group prediction and career prediction. These scenarios share a common nature—the attributes to be inferred are correlated to both intrinsic personal information and extrinsic social connection information. For these tasks, it is also reasonable to analyze users from both user-centric and network-centric angles. Regarding the task of age group prediction, there should be certain individual differences in terms of user-centric features and network-centric features among different age groups. For example, youngsters may talk more about homework as well as fashion topics, follow or retweet related accounts, and connect with more youngsters in social networks.

Despite the comprehensiveness of social media information, the current framework can be extended by the user's offline data. In order to gain a more holistic view of users, we can potentially use some offline data sources, such as a user's sensor data. User sensor data is gaining increasing researcher attention [Tjondronegoro and Chua 2012; Singh et al. 2010, 2013] and is recorded by several novel wearable sensors, such as Fit-bit, Google glass, and Apple iWatch. Unfortunately, due to the lack of sensor data, we fail to conduct relevant experiments, although we believe that incorporating these personal sensor data can facilitate the process of learning users' offline activities and boost the performance of our scheme. For example, individuals who are enthusiastic about outdoor activities or keen on sports may possess a healthy body and tend to be energetic and extraverted. Consequently, these individuals are more likely to be volunteers. These personal physical attributes, which are not accurately captured by social media, are definitely of significant importance for the task of volunteer identification. It is worth mentioning that this kind of information can be naturally embedded into the user-centric analysis component (especially the behavior-based features) of our proposed scheme.

Our model can be further extended to classify volunteers by causes. We believe that the task of classifying volunteers by causes can be divided into two subtasks: (i) classify volunteers generally to determine whether this user is keen on volunteerism and (ii) identify their skills and route them to different causes or NPOs. We currently focus on the first subtask. It is also worth highlighting that different volunteer causes may require volunteers to possess different skills, and users' volunteerism tendency may vary depending on the causes. This consideration sheds light on our future work direction, where we need to identify users' specific volunteerism interests instead of generally predicting their volunteerism tendency.

9. CONCLUSION

This article presented a novel scheme to infer users' volunteerism tendency based on UGCs from multiple social networks, which casts the task of volunteerism tendency prediction as a binary classification problem. According to a conceptual volunteer decision model, we measured users' volunteerism tendency by user-centric analysis and

network-centric analysis in subtle ways. Within the user-centric analysis, we designed and extracted a set of application-oriented features from users' social contents. In contrast, the network-centric analysis utilized a graph-based model to integrate various social connections among users.

In addition, to comprehensively learn users' profiles, we introduced strategies for collecting users' data across multiple social networks. We finally constructed our own dataset—AQV—which consists of about 5,000 online users. Based on AQV, we thus developed comprehensive experiments to evaluate the performance of our proposed scheme. Experimental results demonstrate the effectiveness of our proposed approach and verify the advantages of utilizing multiple sources over a single source.

In addition, we discussed the possible generalization and extension of our scheme. The proposed scheme for the identification of users' volunteerism tendency can be used in application scenarios such as age group prediction. Furthermore, we can extend our scheme to include users' offline behaviors that are recorded by sensors, such as those on mobile devices.

It is worth highlighting that different volunteer causes may require volunteers to possess different skills and that the volunteerism tendency of users may vary depending on the causes. This consideration propels us to further identify users' specific volunteerism interests rather than generally predicting their volunteerism tendency in the future.

REFERENCES

Fabian Abel, Eelco Herder, Geert-Jan Houben, Nicola Henze, and Daniel Krause. 2013. Cross-system user modeling and personalization on the social web. *User Modeling and User-Adapted Interaction* 23, 2–3 (2013), 169–209.

Sibel Adali and Jennifer Golbeck. 2012. Predicting personality with social behavior. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*. IEEE Computer Society, 302–309.

Shuotian Bai, Tingshao Zhu, and Li Cheng. 2012. Big-five personality prediction based on user behaviors at social network sites. *arXiv:1204.4809* (2012).

Murray R. Barrick and Michael K. Mount. 1991. The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology* 44, 1 (1991), 1–26.

Blerina Bazelli, Adrian Hindle, and Eleni Stroulia. 2013. On the personality traits of stackoverflow users. In *Proceedings of the IEEE International Conference on Software Maintenance*. IEEE, 460–463.

Bin Bi, Milad Shokouhi, Michal Kosinski, and Thore Graepel. 2013. Inferring the demographics of search users: Social data meets search queries. In *Proceedings of the International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 131–140.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32.

Gustavo Carlo, Morris A. Okun, George P. Knight, and Maria Rosario T. de Guzman. 2005. The interplay of traits and motives on volunteering: Agreeableness, extraversion and prosocial value motivation. *Personality and Individual Differences* 38, 6 (2005), 1293–1305.

Zeynep Cemalcilar. 2009. Understanding individual characteristics of adolescents who volunteer. *Personality and Individual Differences* 46, 4 (2009), 432–436.

Tim Crosier, Jeni Warburton, and others. 2001. Are we too busy to volunteer?: the relationship between time and volunteering using the 1997 ABS time use data. (2001).

Mark H. Davis, Kyle V. Mitchell, Jennifer A. Hall, Jennifer Lothert, Tyra Snapp, and Marnee Meyer. 1999. Empathy, expectations, and situational preferences: Personality influences on the decision to participate in volunteer helping behaviors. *Journal of Personality* 67, 3 (1999), 469–503.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the International AAAI Conference on Web and Social Media*.

James Hitchen. 2013. *Implementing a Volunteer–Match Service*. Ph.D. Dissertation. Al Akhawayn University.

Thomas Holtgraves. 2011. Text messaging, personality, and the social context. *Journal of Research in Personality* 45, 1 (2011), 92–99.

Francisco Iacobelli, Alastair J. Gill, Scott Nowson, and Jon Oberlander. 2011. Large scale personality classification of bloggers. In *Proceedings of the Affective Computing and Intelligent Interaction*. Springer, 568–577.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20, 4 (2002), 422–446.

Daifeng Li, Bing He, Ying Ding, Jie Tang, Cassidy Sugimoto, Zheng Qin, Erjia Yan, Juanzi Li, and Tianxi Dong. 2010a. Community-based topic modeling for social tagging. In *Proceedings of the ACM International Conference on Information and Knowledge Management.* ACM, 1565–1568.

Fangtao Li, Minlie Huang, and Xiaoyan Zhu. 2010b. Sentiment analysis with global topics and local dependency. In *AAAI Conference on Artificial Intelligence*, Vol. 10. 1371–1376.

Andy Liaw and Matthew Wiener. 2002. Classification and regression by randomForest. *R News* 2, 3 (2002), 18–22.

Dejan Markovikj, Sonja Gievska, Michal Kosinski, and David Stillwell. 2013. Mining facebook data for predictive personality modeling. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

Robert R. McCrae and Oliver P. John. 1998. An introduction to the five-factor model and its applications. *Personality: Critical Concepts in Psychology* 60 (1998), 295.

David Meyer and F. H. Technikum Wien. 2014. Support vector machines: The Interface to libsvm in package e1071. *Technische Universitat Wien, Austria* (2014).

Liqiang Nie, Yi-Liang Zhao, Xiangyu Wang, Jialie Shen, and Tat-Seng Chua. 2014. Learning to recommend descriptive tags for questions in social forums. *ACM Transactions on Information Systems* 32, 1 (2014), 5.

Chong Oh and Olivia Sheng. 2011. Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. In *ICIS*.

Jahna Otterbacher. 2010. Inferring gender of movie reviewers: Exploiting writing style, content and metadata. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. ACM, 369–378.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. Democrats, republicans and starbucks afficionados: User classification in twitter. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 430–438.

James W. Pennebaker and Laura A. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology* 77, 6 (1999), 1296.

Louis A. Penner. 2002. Dispositional and organizational influences on sustained volunteerism: An interactionist perspective. *Journal of Social Issues* 58, 3 (2002), 447–467.

Louis A. Penner. 2004. Volunteerism and social problems: Making things better or worse? *Journal of Social Issues* 60, 3 (2004), 645–666.

Adrian Popescu, Gregory Grefenstette, and others. 2010. Mining user home location and gender from flickr tags. In *The International AAAI Conference on Web and Social Media*.

Daniele Quercia, Renaud Lambiotte, David Stillwell, Michal Kosinski, and Jon Crowcroft. 2012. The personality of popular facebook users. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work.* ACM, 955–964.

Reint Jan Renes. 2005. Sustained volunteerism: Justification, motivation and management. (2005).

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.* Association for Computational Linguistics, 763–772.

Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. 2002. Methods and metrics for cold-start recommendations. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 253–260.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One* 8, 9 (2013), e73791.

Vivek K. Singh, Tat-Seng Chua, Ramesh Jain, and Alex Sandy Pentland. 2013. Summary abstract for the 1st ACM international workshop on personal data meets distributed multimedia. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1105–1106.

Vivek K. Singh, Mingyan Gao, and Ramesh Jain. 2010. Social pixels: Genesis and evaluation. In *ACM International Conference on Multimedia*. ACM, 481–490.

Xuemeng Song, Liqiang Nie, Luming Zhang, Mohammad Akbari, and Tat-Seng Chua. 2015a. Multiple social network learning and its application in volunteerism tendency prediction. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 213–222.

Xuemeng Song, Liqiang Nie, Luming Zhang, Maofu Liu, and Tat-Seng Chua. 2015b. Interest inference via structure-constrained multi-source multi-task learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 2371–2377.

Dian Tjondronegoro and Tat-Seng Chua. 2012. Transforming mobile personal life log into autobiographical multimedia eChronicles. In *Proceedings of the International Conference on Advances in Mobile Computing & Multimedia*. ACM, 57–63.

Hongning Wang, Minlie Huang, and Xiaoyan Zhu. 2008. A generative probabilistic model for multi-label classification. In *Proceedings of the IEEE International Conference on Data Mining*. IEEE, 628–637.

Walter W. Wymer Jr and Sridhar Samu. 2002. Volunteer service as symbolic consumption: Gender and occupational differences in volunteering. *Journal of Marketing Management* 18, 9–10 (2002), 971–989.

Xiang Yan and Ling Yan. 2006. Gender classification of weblog authors. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. 228–230.

Tal Yarkoni. 2010. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality* 44, 3 (2010), 363–373.

Yi-Liang Zhao, Qiang Chen, Shuicheng Yan, Tat-Seng Chua, and Daqing Zhang. 2013. Detecting profilable and overlapping communities with user-generated multimedia contents in LBSNs. *ACM Transactions on Multimedia Computing, Communications, and Applications* 10, 1 (2013), 3.

Zhaohui Zheng, Hongyuan Zha, Tong Zhang, Olivier Chapelle, Keke Chen, and Gordon Sun. 2008. A general boosting method and its application to learning ranking functions for web search. In *Advances in Neural Information Processing Systems*. 1697–1704.

Xingwei Zhu, Zhao-Yan Ming, Xiaoyan Zhu, and Tat-Seng Chua. 2013. Topic hierarchy construction for the organization of multi-source user generated contents. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 233–242.

Received October 2014; revised August 2015; accepted September 2015