Vietnam Journal of Computer Science (2018) 5:165–175
https://doi.org/10.1007/s40595-018-0112-1

REGULAR PAPER

Estimating the similarity of social network users based on behaviors

Thi Hoi Nguyen[1] · Dinh Que Tran[2] · Gia Manh Dam[1] · Manh Hung Nguyen[2,3]

Received: 25 November 2017 / Accepted: 11 May 2018 / Published online: 19 May 2018

Abstract
Recently, with the express growth of social network, users have joined more and more of these networks and live their life virtually. Consequently, they create a huge data on these social networks: their profile, interest, and behavior such as post, comment, like, joining groups or communities, etc. This brings some new challenges to researchers: do users having the same profile/interest show the same behavior? And vice versa, do users having the same behavior have interest in the same things? One of the basic issues in these challenges is the problem of estimating the similarity among users on these social networks based on their profile, interest, and behavior. This paper presents a model for estimating the similarity between users based on their behavior on social networks. The considered behaviors are activities including posting entries, liking these entries, commenting and liking the comment in these entries. The model is then evaluated with a dataset-collected users from Twitter. The results show that the model estimates correctly the similarity among users in the majority of the cases.

Keywords User similarity · Behavior similarity · Entry similarity · Social network

1 Introduction

Nowadays, with the exploration of social networks, there are more and more people joining these networks. In these digital worlds, users freely present themselves, share information about their favorites and passions, or share their personal opinion on some issues of economic, social, cultural, etc. through several activities on social network such as posting entries, sharing video clips, images, or news they read, and then leaving their comments or liking these entries or the comments of others, etc. Consequently, huge data are created on the social network. This huge data attract many researchers, businessmen, etc. to mine and exploit it. This tendency also brings some new challenges to researchers: do users having the same profile/interest show the same behavior? And vice versa, do users having the same behavior have interest in the same things? One of the basic issues in these challenges is the problem of estimating the similarity among users on these social networks based on their profile, interest, and behavior?

The problem of detecting the similarity or the difference between users is not only based on the user profile on the social network, but also based on the data about user behavior such as posting entries, commenting, and liking. This problem has been attracting many researchers. For instance, Raad et al. [16] and Peled et al. [15] proposed a model to measure the similarity between user profiles. Anderson et al. [1] calculated the similarity between user characteristics. Liu et al. [8] estimated the similarity among preferences of user behavior. Liu et al. [9] and Chen et al. [4] measured the similarity among user mobility behavior. Xu et al. [23] analyzed the user posting behavior on a popular social media website. Erlandsson et al. [5] proposed association learning to detect relationships between users. Benevenuto et al. [2] presented a kind of analysis of user workloads in online social networks. Singh et al. [18] formulated a metric based on the common words used in social networks to measure the user similarity in textual posting. Zou et al. [26] mined individual behavior patterns and study user similarity. Sun et al. [19] proposed a

✉ Manh Hung Nguyen
nmh.nguyenmanhhung@gmail.com; mhnguyen@ptit.edu.vn

Thi Hoi Nguyen
hoint2002@gmail.com

Dinh Que Tran
tdque@yahoo.com; quetd@ptit.edu.vn

Gia Manh Dam
damgiamanh@gmail.com

[1] Thuongmai University Hanoi, Hanoi, Vietnam

[2] Posts and Telecommunication Institute of Technology (PTIT), Hanoi, Vietnam

[3] UMI UMMISCO 209 (IRD/UPMC), Hanoi, Vietnam

mapping method, which integrates text and structure information for similarity computation. Guo et al. [6] developed a model to estimate continuous tie strength between users for friend recommendation with the heterogeneous data from social media community. Nguyen et al. [11] aimed to understand the strategies users employ to make retweet decision. Liu and Terzi [10] approached the privacy issues raised in online social networks from the individual users viewpoint: they proposed a framework to compute the privacy score of a user. Tang et al. [20] adopted a "microeconomics" approach to a model and predicted the individual retweet behavior. Xu et al. [22] introduced several methods to identify online communities with similar sentiments in online social networks. Zhao et al. [25] proposed to separately model users' topical interests that come from these various behavioral signals to construct better user profiles. Vedula et al. [21] detected pairwise and global trust relations between users in the context of emergent real-world crisis scenarios. Jamali and Ester [7] explored social rating networks, which record not only social relations but also user ratings for items. Bhattacharyya et al. [3] studied the relationship between semantic similarity of user profile entries and the social network topology. In the model of Zhao et al. [24], two social factors, interpersonal rating behavior similarity and interpersonal interest similarity, are fused into a consolidated personalized recommendation model based on probabilistic matrix factorization.

Most of these works try to estimate the similarity among user based on: user profile, user interests or favorites, or user relationship on social network. However, there are not many works which estimate the similarity among social network users based on their activities on social network.

In line with our previous works ([13,14]), this paper introduces a model for measuring the similarity between users based on their behavior on social network. In this model, the similarity between users is estimated from the similarity of their behaviors such as posting an entry or sharing an existing entry, liking an entry or liking a comment, commenting on a post, and joining a group or a community. The similarity of user behavior on these activities is also estimated based on the content of the entries that they post, like, or the content of their comment on these entries from social networks. The similarity among entries is estimated based on the *content*, *tags*, *category*, *sentiment*, and *emotion* included in these entries [14]. The model is then evaluated with a dataset-collected users from Twitter. The results show that the model estimates correctly the similarity among users in the majority of the cases.

The paper is organized as follows: Sect. 2 presents the similarity model. Section 3 takes some experiments to evaluate the proposed model with empirical data. Section 4 is the conclusion and perspectives.

2 A model for estimation of similarity among social network users based on their behaviors

2.1 Notations

Without loss of generality, we assume that:

- A social network is a 4-tuples $N =< U, G, E, B >$, in which:
  - $U = \{u_1, u_2, \ldots, u_m\}$ is a set of users,
  - $G = \{g_1, g_2, \ldots, g_n\}$ is a set of communities or groups,
  - $E = \{e_1, e_2, \ldots, e_k\}$ is a set of entries of users on the social network N,
  - $B = \{b_1, b_2, \ldots, b_l\}$ is a set of behaviors of each user $u \in U$ on the group $g \in G$ or on the entry $e \in E$ on the social network N.
- A user could post a status, an image, or a video clip that we call an entry e. An entry e could be viewed by a set of users U. Each user, within an entry, could like the entry and comment on the entry or share that entry on their homepage.
- Each user u, within an entry, could like a set of comments of the entry. A user could like a page or join a group. In this case, the user is a member of a community or group of the social network. An user could post an entry in a community or a group, like an entry, comment into an entry, or like a set of comments in an entry of a community or share an entry.

2.1.1 Entry

Generally, an entry on a social network can be a video, an image, a text, or a combination of all content. However, in this paper, we only consider entries that contain textual content. If they do not contain texts such as video or images, they are ignored. Therefore, the problem is to consider and estimate the similarity of users based on the entry that focuses on reviewing and estimating similarity between texts.

On a social network, there is a set of user $U = \{u_1, u_2, \ldots, u_m\}$. Each user u_i is characterized by a set of entries posted E and a set of behaviors B on the social network. Each user $u_i \ inU$ has a set of entries $E_i = \left\{e_1^i, e_2^i, \ldots, e_k^i\right\}$ and each entry $e_j \in E$ has a set of features: $e_j = \left\{f_1^j, f_2^j, \ldots, f_p^j\right\}$.

An entry could have several features, including explicit features such as the content, and the implicit features such as tag, category, sentiment, and emotion. As the implicit features cannot be directly extracted from an entry, the model

needs a step to extract these features before estimating the similarity on entries.

This model considers five features of an entry:

- Content of entry e^j, noted as f^j_{cont}: content is the whole text part in the entry itself. This is an explicit feature.
- Tags of entry e^j, noted as f^j_{tags}: an entry could be tagged to a set of tags. Each tag is an independent word or expression. In some cases, tags can be directly tagged by the user (explicit). In some other case, it is not explicitly tagged by the user (implicit).
- Category of entry e^j, noted as f^j_{cate}: an entry could be assigned to a category. Each category is represented by an independent word or expression.
- Sentiment of entry e^j, noted as f^j_{sent}: an entry could have a sentiment of the user. A sentiment may be agree (positive), disagree (negative), or neutral opinion.
- Emotion of entry e^j, noted as f^j_{emot}: an entry could also have some emotion of the user. Each emotion is represented by an independent word or expression

As an entry is considered as a set of features and only their textual contents are considered, the problem of estimating the similarity among entries could be considered as the computation of the similarity among texts or among sets of expressions.

2.1.2 Behavior

In this model, only five popular behaviors are considered: post an entry, like an entry, comment on an entry, share an entry, and join a group on social network. We assume that a social network has a set of user $U = \{u_1, u_2, \ldots, u_m\}$. Each user $u_i \in U$ posts a set of entries E and acts with a set of behaviors $B_i = \left\{b^i_1, b^i_2, \ldots, b^i_l\right\}$. Each behavior $b_l \in B$ may have a set of features $b_l = \left\{f^l_1, f^l_2, \ldots, f^l_q\right\}$:

- Post of entry, noted as b^l_{post}: the user writes an entry on the user homepage.
- Like an entry or like a comment, noted as b^l_{like}: the user clicks on the *like* icon of an entry or a comment.
- Comment of entry, noted as b^l_{comt}: the user writes some comments on an entry
- Share of entry, noted as b^l_{shar}: the user shares an entry on his/her wall. The shared entry could belong to different users of social media or its social network.
- Join a group, noted as b^l_{join}: the user joins a group or community. A group usually has the name of a group, description of the group and other characters of the group.

2.1.3 Group or community

As a community or a group is described by its meta-data, the similarity between two communities or groups is, thus, considered as the similarity between two multi-feature objects (Nguyen and Nguyen [12]). Each meta-data of a community or group is considered as a feature of the community or a group. In this model, we assume that a social network has a set of user $U = \{u_1, u_2, \ldots, u_m\}$. Each user $u_i \in U$ can be joined into a set of communities or groups $g_v \in G$ with features: $g_v = \left\{g^v_1, g^v_2, \ldots, g^v_w\right\}$:

- Name of the community, noted as g^v_{name}: it could be an entitle or a short brief sentence. After eliminating all stop words in the title, this feature becomes a set of words to be compared to that of other communities. So, estimating the similarity on the name of the community is to estimate the similarity between two sets of words.
- Categories of the community, noted as g^v_{camu}: on some social networks, each community is always assigned to at least one category. Each category is an independent word (or independent expression). So, estimating the similarity on the categories of the community is also to estimate the similarity between two sets of expressions.
- Description of the community, noted as g^v_{desc}: on many of the social networks, each community is also provided a short description. A description is normally a short text. After eliminating all stop words in the text, this feature becomes a set of words to be compared to that of other communities. So, estimating the similarity on the description of the community is also to estimate the similarity between two set of expressions.

As the comparison between two entries is considered as a comparison between their sets of feature and only their textual values are considered, the comparison between two behaviors and communities or groups was made by comparing only their textual values. Therefore, the problem of estimating the similarity among users based on behaviors becomes the computation of the similarity among texts or among sets of expressions.

2.2 General model

The general model is as follows:

Input: .$u_1, u_2 \in U$ with their two sets of entries $E_1, E_2 \in E$ and two sets of behaviors $B_1, B_2 \in B$
Output: Estimated similarity between the two entered users $u_1, u_2 \in U$ called $\text{sim}(u_1, u_2)$.
Inside the model, there are four main steps:

- Step 1: modeing entries E and behaviors B.

- Step 2: extracting the value for implicit features of entries.
- Step 3: estimating the similarity on each entry's features and on each user's behaviors.
- Step 4: aggregating the similarity between two sets of entries E_1, E_2 and between two sets of behaviors B_1, B_2 of users u_1, u_2.

These steps will be described in detail in the next sections.

2.3 Determination value features of entries

2.3.1 Evaluation value implicit features

Let's consider an example of a status on Twitter: "*Thank you @apple for Find My Mac - just located and wiped my stolen Air*". When we see this status, only the content is explicitly presented, which is the whole text of the status. However, we can quickly identify some other features of this status, such as category (technology), tags (Apple, Mac), sentiment (neutral—neither agree nor disagree), and emotion (gratitude, joy). The features whose value is not explicitly presented in the entry, but could be extracted from the inside of the entry, are called implicit features. Our object in this step is extracting the value of implicit features of an entry.

In this model, we apply a method to extract the value of each of four implicit features as follows (called the method to classify the texts into classes) [14]:

- Step 1: Construct a set of labeled samples (texts), called *training set*, in which each text is assigned to a set of labels. The union of all labels of all texts is called the set of labels L.
- Step 2: For each label $l_i \in L$, create two sets of text samples:
  - T_{l_i} is the set of all texts which are labeled with l_i.
  - $T_{\neg l_i}$ is the set of all texts which are not labeled with l_i.
- Step 3: For each text $t_k \in T_{l_i}$ or $(t_k \in T_{\neg l_i})$, calculate the label-oriented features as follows:
  - Split t_k into a set of n-gram or term (stop words may be removed).
  - Take the union of all terms in all texts in the set T_{l_i} and $T_{\neg l_i}$.
- Step 4: Calculate the *label-oriented term score* of each term in the corresponding set for each label l_i:

$$s_{\mathrm{LOT}}(x, l_i) = \frac{N_{l_i}^x}{N_{l_i}} \times \log\left(\frac{N_{\neg l_i}}{N_{\neg l_i}^x}\right) - \frac{N_{\neg l_i}^x}{N_{\neg l_i}} \times \log\left(\frac{N_{l_i}}{N_{l_i}^x}\right), \tag{1}$$

where N_l, $N_{\neg l}$ are the number of texts in the set T_l, $T_{\neg l}$, respectively. N_l^x, $N_{\neg l}^x$ are the number of texts in the set T_l, $T_{\neg l}$, respectively, which contains the term x.

- Step 5: For a new text t, the choice of label to assign to the text is as follows:
  - Split t into a set of n-grams or terms $X = (x_1, x_2, \ldots x_n)$.
  - Calculate the *term frequency* for each term x_i in the text t: $tf(x_i, t)$.
  - For each label $l_i \in L$, calculate the *label-oriented document score*:

$$s_{\mathrm{LOD}}(t, l_i) = \frac{1}{n_t} \times \sum_{x \in t} s_{\mathrm{LOT}}(x, l_i) * tf(x, t). \tag{2}$$

  - If $s_{\mathrm{LOD}}(t, l_i) > 0$:
    - In the multi-label problem where a text could be assigned to several labels, the text t will be labeled with l_i.
    - In the single label problem where a text could be assigned to only one label, it is needed to calculate all the final label-oriented (disoriented) scores of the text t for all the labels $l_i \in L$. The label whose *label-oriented document score* is the highest will be assigned to the text t.

2.3.2 Evaluation value sentiment features

To determine the sentiment of a short text, it is necessary to estimate the value of the point of view of the text that views the author's point of view expressed in the text. In this paper, we apply the method to classify the texts into classes in Sect. 2.3.1. Therefore, the value of the sentiment feature of a text is assigned to one of three values (classes), that is, positive, negative, or neutral.

2.3.3 Evaluation value emotion features

The emotions of the user represented in the entries are often represented by icons or images, each of which is equivalent to a term describing that emotion. Therefore, estimating the similarity between two emotions of the entry is to estimate the similarity between the two terms. In this paper, we apply the method of classifying the texts into classes in Sect. 2.3.1. Therefore, the emotion value of a text is assigned to one of the values (classes): enjoy; happy for; love; gratitude; admiration; pride; hope; sad; sorry; regret; disappointed; disgust; angry; confused; no emotion.

2.4 Estimating similarity on each feature

In this model, we distinguish two kinds of textual values of a feature:

- First, the feature value is already in the form of a set of expressions, such as the value of features *tags*, *category*, *sentiment*, and *emotion*. Their similarity is considered as the similarity among sets of expressions.
- Second, the feature value is in the form of a general text, such as the value of the feature *content*. Their similarity is considered as the similarity among texts.

2.4.1 Estimating the similarity for expression features

Since the content of the feature is in the form of a set of textual expressions, their similarity is defined as follows: suppose that $A_1 = (a_1^1, a_1^2, \ldots a_1^m)$, $A_2 = (a_2^1, a_2^2, \ldots a_2^n)$ are two sets of expressions or strings, in which, m, n are the sizes of the set A_1 and A_2, respectively. Let v be the size of the set of intersection of A_1 and A_2. The similarity between A_1 and A_2 is defined by the formula:

$$s_{\text{exp}}(A_1, A_2) = \frac{2\times \mid A_1 \cap A_2 \mid}{\mid A_1 \mid + \mid A_2 \mid} = \frac{2 \times v}{m+n}. \tag{3}$$

It is clear that all possible values of $s_{\text{exp}}(A_1, A_2)$ are in the interval [0, 1]. This formula could be applied to the features, whose value is a set of expressions.

Suppose that $e^i = (f_1^i, f_2^i, \ldots f_n^i), e^j = (f_1^j, f_2^j, \ldots f_n^j)$ are two entries represented by their features. Let us consider the feature k whose value is a set of expressions. The similarity between entries e^i and e^j on the feature k is defined by the formula:

$$s_k(e^i, e^j) = s_{\text{exp}}(f_k^i, f_k^j), \tag{4}$$

where f_k^i, f_k^j are the expression values of the feature k of the two entries e^i and e^j.

2.4.2 Estimating similarity for text features

The problem of estimating the similarity among textual values becomes the estimation of the similarity among texts. We can apply the technique TF–IDF (term frequency–inverse document frequency) [17] to characterize the texts as follows:

- Split the text into a set of n-gram $t^1 = (g_1^1, g_2^1, \ldots g_n^1)$ and $t^2 = (g_1^2, g_2^2, \ldots g_m^2)$.
- Calculate the TF–IDF of each n-gram in the text. Then, represent the feature value by a vector in which each element is a pair $< n\text{-gram, td-idf} >$: $v^1 = (< g_1^1, v_1^1 >, < g_2^1, v_2^1 > , \cdots < g_n^1, v_n^1 >)$ and $v^2 = (< g_1^2, v_1^2 >, < g_2^2, v_1^2 > , \cdots < g_m^2, v_m^2 >)$.
- Calculate the distance between the two vectors:

$$D(v^1, v^2) = \frac{1}{N}\sum_1^N d_k, \tag{5}$$

where N is the number of different n-grams considered in both $t^1 \cup t^2$ and d_k is the distance on each element $< g_i^1, v_i^1 >$ of v^1 (or element $< g_j^2, v_j^2 >$ of v^2, respectively):

  - If there is an element $< g_l^2, v_l^2 >$ of v^2 (or element $< g_l^1, v_l^1 >$ of v^l, respectively) such that $g_l^2 = g_i^1$, then:

$$d_k = \frac{\mid v_i^1 - v_l^2 \mid}{\max(v_i^1, v_l^2)}. \tag{6}$$

  - Otherwise, $d_k = 1$.

- It is clear that the value of $D(v^1, v^2)$ is in the interval [0, 1]. Similarity between the two features is then:

$$s_{\text{txt}}(t^1, t^2) = 1 - D(v^1, v^2). \tag{7}$$

2.4.3 Estimating similarity between two entries

We considered an entry via five features including: content, tag, category, sentiment and emotion. In this case, there are four expression features of entry including: tags, category, sentiment and emotion. So they are estimated as the similarity on expression feature as follows:

$$s_{\text{cate}}(e^i, e^j) = s_{\text{exp}}(f_{\text{cate}}^i, f_{\text{cate}}^j), \tag{8}$$

$$s_{\text{tags}}(e^i, e^j) = s_{\text{exp}}(f_{\text{tags}}^i, f_{\text{tags}}^j), \tag{9}$$

$$s_{\text{sent}}(e^i, e^j) = s_{\text{exp}}(f_{\text{sent}}^i, f_{\text{sent}}^j), \tag{10}$$

$$s_{\text{emot}}(e^i, e^j) = s_{\text{exp}}(f_{\text{emot}}^i, f_{\text{emot}}^j). \tag{11}$$

One text feature of entry is content, so it is estimated as the text feature similarity, calculated as follows:

$$s_{\text{cont}}(e^i, e^j) = s_{\text{txt}}(f_{\text{cont}}^i, f_{\text{cont}}^j). \tag{12}$$

Let e^i and e^j be two considered entries whose content, tags, categories, sentiment and emotion are features of entries: e_{cont}^i; e_{cont}^j; e_{tags}^i; e_{tags}^j; e_{cate}^i; e_{cate}^j; e_{sent}^i; e_{sent}^j; e_{emot}^i; e_{emot}^j. Based on the approach of multi-attribute similarity of two objects [12], the similarity between the two entries e^i

and e^j is estimated as follows:

$$s_{\text{entry}}(e^i, e^j) = f_{\text{ent}}(s_{\text{cont}}(e^i, e^j), s_{\text{tags}}(e^i, e^j), s_{\text{cate}}(e^i, e^j), s_{\text{sent}}(e^i, e^j), s_{\text{emot}}(e^i, e^j)), \tag{13}$$

where $f_{\text{ent}} : [0, 1]^5 \rightarrow [0, 1]$ is similarity is a similar function between two entries, which satisfies the following conditions:

$$\begin{aligned}
&(i)\ \ f_{\text{ent0}}(v_1, w, x, y, z) \leqslant f_{\text{ent}}(v_2, w, x, y, z)\ if\ v_1 \leqslant v_2;\\
&(ii)\ \ f_{\text{ent}}(v, w_1, x, y, z) \leqslant f_{\text{ent}}(v, w_2, x, y, z)\ if\ w_1 \leqslant w_2;\\
&(iii)\ \ f_{\text{ent}}(v, w, x_1, y, z) \leqslant f_{\text{ent}}(v, w, x_2, y, z)\ if\ x_1 \leqslant x_2;\\
&(iv)\ \ f_{\text{ent}}(v, w, x, y_1, z) \leqslant f_{\text{ent}}(v, w, x, y_2, z)\ if\ y_1 \leqslant y_2;\\
&(v)\ \ f_{\text{ent}}(v, w, x, y, z_1) \leqslant f_{\text{ent}}(v, w, x, y, z_2)\ if\ z_1 \leqslant z_2.
\end{aligned} \tag{14}$$

2.4.4 Estimating the similarity between two groups

Once the similarity between two groups on each feature is estimated, the similarity between two groups is then estimated by a weighted average aggregation of the similarity between them on all considered features as follows:

- Let w_1, w_2, w_3 be the weight of features *Name*, *Description* and *Category*, respectively. They have to satisfy this condition: $w_1 + w_2 + w_3 = 1$.
- The similarity between group g_i and group g_j is:

$$\begin{aligned}
s_{\text{group}}(g_i, g_j) = &w_1 \times s_{\text{exp}}(g^i_{\text{name}}, g^j_{\text{name}})\\
&+ w_2 \times s_{\text{exp}}(g^i_{\text{desc}}, g^j_{\text{desc}})\\
&+ w_3 \times s_{\text{exp}}(g^i_{\text{camu}}, g^j_{\text{camu}}),
\end{aligned} \tag{15}$$

where w_1, w_2, w_3 are, respectively, the weight of the features *Name*, *Description*, and *Category*. $s_{exp}(A, B)$ is the similarity between the two sets of expressions A and B.

In the case of two sets of communities, let $G_1 = g^1_1, g^1_2, \ldots, c^1_m$ and $G_2 = g^2_1, g^2_2, \ldots, g^2_n$ be the two considered sets of communities. We create a common set of these two sets $G_{12} = G_1 + G_2 = g_1, g_2, \ldots, g_{m+n}$ and then construct their non-ordered semantic vectors $T = (t_1, t_2, \ldots, t_{m+n})$ as:

$$\begin{aligned}
&t_i = \min(\max(s_{\text{group}}(g_i; g^1_k)), \max(s_{\text{group}}(g_i; g^2_v)))\\
&k = 1 \ldots m; v = 1 \ldots n,
\end{aligned} \tag{16}$$

where $s_{\text{group}}(x, y)$ is the similarity between two groups x and y. The similarity between two non-ordered sets of communities G_1 and G_2 is defined by the formula:

$$s_{\text{css}}(G_1, G_2) = f_{\text{set}}(T) = f_{\text{set}}(t_1, t_2, \ldots, t_{p+q}), \tag{17}$$

where $f_{\text{set}} : [0, 1]^k \rightarrow [0, 1]$ is a similar function between two sets.

2.5 Estimating each behavior of the user

In this paper, we consider five behaviors of the user on social networks including: post an entry, like, comment, share an entry, and join a group or a community.

2.5.1 The similarity between post or share behavior

In the case of post or share an entry, the similarity post or share an entry is estimated by estimating the similarity of two sets of posted or shared entries as follows:

Let $E_1 = e^1_1, e^1_2, \ldots, e^1_p$ and $E_2 = e^2_1, e^2_2, \ldots, e^2_q$ be two considered sets of posted or shared entries. We create a common set of these two sets $E_{12} = E_1 + E_2 = e_1, e_2, \ldots, e_{p+q}$ and then construct their non-ordered semantic vectors $T = (t_1, t_2, \ldots, t_{p+q})$ as:

$$\begin{aligned}
&t_i = \min(\max(s_{\text{entry}}(e_i, e^1_k)), \max(s_{\text{entry}}(e_i, e^2_v)))\\
&k = 1 \ldots p; v = 1 \ldots q,
\end{aligned} \tag{18}$$

where $s_{\text{entry}}(x, y)$ is the similarity between the two entries x and y. To measure the similarity between two sets of entries E_1 and E_2, we make use of the following assumptions: The bigger the magnitude of the vector T, the higher is the similarity between E_1 and E_2. The similarity between two non-ordered sets of entries E_1 and E_2 is defined by the formula:

$$s_{\text{ess}}(E_1, E_2) = f_{\text{set}}(T) = f_{\text{set}}(t_1, t_2, \ldots, t_{p+q}), \tag{19}$$

where $f_{\text{set}} : [0, 1]^k \rightarrow [01]$ is a similar function between two sets, which satisfies the following conditions:

$$\begin{aligned}
&(i)\ \ f_{\text{set}}(0, 0, \ldots, 0) = 0;\\
&(ii)\ \ f_{\text{set}}(1, 1, \ldots, 1) = 1;\\
&(iii)\ \ f_{\text{set}}(X_1) \leqslant f_{\text{set}}(X_2) \text{ if } \|X_1\| \leqslant \|X_2\|.
\end{aligned} \tag{20}$$

For example, the following functions are similar function between two sets of entries:

$$\begin{aligned}
&(1)\ \ f(x_1, x_2, \ldots, x_n) = \frac{\sum_{i=1}^{n} x_i}{n},\\
&(2)\ \ f(x_1, x_2, \ldots, x_n) = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}.
\end{aligned} \tag{21}$$

In the case of similarity between two sets of posted or shared entries, let $E_{\text{post}}^{u_i}$ and $E_{\text{post}}^{u_j}$ be, respectively, two sets of posted or shared entries of user u_i and user u_j. The posting-based (or sharing-based) behavior similarity of user u_i and user u_j is defined by the formula:

$$s_{\text{post}}(u_i, u_j) = s_{\text{ess}}(E_{\text{post}}^{i}, E_{\text{post}}^{j}), \tag{22}$$

where $s_{\text{ess}}(A, B)$ is the similarity between two sets of entries A and B.

2.5.2 The similarity on behavior of joining a group

Let's $G_{\text{join}}^{u_i}$ and $G_{\text{join}}^{u_j}$ are respectively the two sets of communities or groups which were joined by user u_i and user u_j. The *joining a group* behavior similarity of user u_i and user u_j is defined by the formula:

$$s_{\text{join}}(u_i, u_j) = s_{\text{css}}(G_{\text{join}}^{i}, G_{\text{join}}^{j}), \tag{23}$$

where $s_{\text{css}}(A, B)$ is the similarity between two sets of communities A and B.

2.5.3 The similarity on behavior of liking an entry

Let's $L_{\text{like}}^{u_i}$ and $L_{\text{like}}^{u_j}$ are respectively the two sets of entries were liked by user u_i and user u_j. To measure the like-based behavior similarity of user u_i and user u_j, the following is defined: the more the two sets $L_{\text{like}}^{u_i}$ and $L_{\text{like}}^{u_j}$ are similar, the higher is the like/dislike-based behavior similarity of user u_i and user u_j is. The like-based behavior similarity of user i and user j is defined by the formula:

$$s_{\text{like}}(u_i, u_j) = s_{\text{ess}}(L_{\text{like}}^{u_i}, L_{\text{like}}^{u_j}), \tag{24}$$

where $s_{\text{ess}}(A, B)$ is the similarity between the two sets of entries A and B

2.5.4 The similarity between two comment likes in comment-based behaviors

Although this behavior is obviously a confirmation of what the user already liked or disliked, sometimes some user could comment or like some comment without liking or disliking the entry. In these cases, we take it into account to measure the similarity among users on the following principles:

- The value of each comment is detected as: positive, negative, or neutral. This determination could be done by applying the method of classifying the texts into classes in Sect. 2.3.1.
- In each entry, only the positive or negative comments are counted. The neutral comment will be removed.
- In each entry, if the number of positive comments of an user is greater than that of the negative comments, then the entry is considered as positive for the user. Vice versa, if the number of positive comments of a user is smaller than that of negative comments, then the entry is considered as negative for the user.
- In the case where the numbers of positive comments and the negative comments are equal, we will consider the comments as liked by the user:
  - If the number of positive comments liked by an user is greater than that of negative comments, then the entry is considered as positive for the user.
  - If the number of positive comments liked by a user is smaller than that of negative comments, then the entry is considered as negative for the user.
  - If the numbers of positive comments and negative comments liked by a user are equal, then the entry is considered as neutral for the user and it will be removed from the considering set of entries for the user.

Let $C_p^{u_i}$ and $C_n^{u_i}$ be, respectively, the set of positive and negative entries for user u_i. $C_p^{u_j}$ and $C_n^{u_j}$ are, respectively, the set of positive and negative entries for user u_j. To measure the comment/like comment-based behavior similarity of user u_i and user u_j, the following is defined:

- The more the two sets $C_p^{u_i}$ and $C_p^{u_j}$ are similar, the higher is the comment/like comment-based behavior similarity of user u_i and user u_j.
- The more the two sets $C_n^{u_i}$ and $C_n^{u_j}$ are similar, the higher is the comment/like comment-based behavior similarity of user u_i and user u_j is.
- The less the two sets $C_p^{u_i}$ and $C_n^{u_j}$ are similar, the higher is the comment/like comment-based behavior similarity of user u_i and user u_j.
- The less the two sets $C_n^{u_i}$ and $C_p^{u_j}$ are similar, the higher is the comment/like comment-based behavior similarity of user u_i and user u_j.

The comment/like comment-based behavior similarity of user u_i and user u_j is defined by the formula:

$$\begin{aligned} s_{\text{comt}}(u_i, u_j) = \min(1, \max(0, & s_{\text{ess}}(C_p^{u_i}, C_p^{u_j}) \\ & + s_{\text{ess}}(C_n^{u_i}, C_n^{u_j}) - s_{\text{ess}}(C_p^{u_i}, C_n^{u_j}) \\ & - s_{\text{ess}}(C_n^{u_i}, C_p^{u_j}))), \end{aligned} \tag{25}$$

where $s_{\text{ess}}(A, B)$ is the similarity between two sets of entries A and B.

2.5.5 Estimating the similarity between two users

Once the similarities between two users on each kind of behavior are estimated, the similarity between the two users is then estimated by a weighted average aggregation, and the similarity between them on all considered kinds of behaviors are as follows:

- Let w_1, w_2, w_3, w_4 be the weight of the similarity based on posting/sharing, joining a group, liking entries, and comment/like comment respectively. They have to satisfy this condition: $w_1 + w_2 + w_3 + w_4 = 1$.
- The similarity between user u_i and user u_j is:

$$s(u_i, u_j) = w_1 \times s_{\text{post}}(u_i, u_j) + w_2 \times s_{\text{join}}(u_i, u_j) + w_3 \times s_{\text{like}}(u_i, u_j) + w_4 \times s_{\text{comt}}(u_i, u_j), \quad (26)$$

where w_1, w_2, w_3, w_4 are, respectively, the weight of the similarity based on posting/sharing, joining a group, liking entries, and comment/like comment. $s_{\text{post}}(u_i, u_j)$; $s_{\text{join}}(u_i, u_j)$; $s_{\text{like}}(u_i, u_j)$; $s_{\text{comt}}(u_i, u_j)$ are the similarity between the two users u_i and u_j based on posting entries, joining a group, liking entries, commenting/liking comment behaviors, respectively.

3 Experiments and evaluation

3.1 Method

3.1.1 Collection of data

To evaluate the proposed model, we collected data from Twitter.com sources (Table 1): we could directly apply the model to estimate the similarity among Twitter users. Each tweet is considered in five features: *content, tags, category, sentiment and emotion* as in the model. The considered activities of Twitter user are: *post/share, like, comment, and list of groups of user*. Note that in Twitter, there are no explicit activities as *like* and *join a group* as in the model. Therefore, we have to map some similar activities in Twitter to these two activities as follows:

- *Like*: the like activity of a user in Twitter is considered as the favorite tweets list of the user.
- *Join a group*: in the case of Twitter, a group could be considered as a list that some users subscribed to.

3.1.2 Construction of sample set

Each sample is constructed as follows:

Table 1 Collected data from Twitter.com

Criteria	Twitter
Collected data	User: 1000
	Posts: 150000
	Activities: 150000
Criteria of entry	Content
	Tags
	Category
	Sentiment
	Emotions
Behavior of user	Post
	Like
	Comment
	Share
	Join a group

Table 2 Sample constructed from Twitter

Source	Number of samples
Twitter	500

- Each sample contains three users collected from Twitter.com. These users are called as user A, B, and C, respectively.
- We ask a number of selected volunteers to answer the question: Which user, B or C, is more similar to user A than the other?
- Then, we compare the number of people who chooses B, and that of people who chooses C. If the number of answer B is greater than that of C, then the value of this sample is 1. It means that user B is more similar to user A than C. On the contrary, if the number of answer C is greater than that of B, then the value of this sample is 2. It means that user C is more similar to user A than B. If the number of the answers B and C are not significantly different, this sample will be removed from the sample set.

After this step, we have a set of samples. We use the samples and save them in a set of samples. In experiments, we use the sample with the size of each sample set as described in Table. 2.

3.1.3 Scenario

The experiment is performed as follows:

- For each sample, we use the model proposed in this paper to estimate the similarity between user B and user A, and that between user C and user A.

Table 3 Correct ratio CR of the sample set

Sample set	Number of correct samples	Correct ratio CR
Twitter	438	87.60

- If B is more similar to A than C is, then the result of this sample is 1. On the contrary, If C is more similar to A than B is, then the result of this sample is 2.
- We then compare the result and the value of each sample. If they are identical, we increase the variable number of correct samples by 1.

3.1.4 Output parameters

The correct ratio (CR) of the model over the given sample set is calculated as follows:

$$\mathrm{CR} = \frac{\text{number of correct sample}}{\text{total of sample}} \times 100\%. \tag{27}$$

The more the CR value is close to 100%, the more is the model correct. We expect that the obtained value of CR would be as high as possible.

3.2 Results

The results are presented in Table 3. In total, the correct ratio of the model over all samples is about 438/500 (87.60%).

For more details, we run experiments with several combinations of weights from criteria of an entry, and weights from behavior of user with the following detailed scenario:

- At the level of entry, we run the experiment with only 1/5, 2/5, 3/5, 4/5, and 5/5 criteria to detect the similarity among entries: 1/5 and 4/5 criteria have five possible combinations; 2/5 and 3/5 criteria have ten possible combinations, 5/5 criteria have only 1 combination.
- For each combination, we run the experiment with different weights of each selected criteria. The changing step for each weight is 0.05. Therefore, each criteria weight runs from 0.05 to 1.00 as long as the sum of all criteria weights in the experiment is equal to 1.
- The same principle is applied at the level of behavior: we run the experiment with 1/4, 2/4, 3/4, and 4/4 behaviors. Each combination is also applied in the same manner as the previous level.

The results are presented in Table 4 (for entry) and Table 5 (for behavior). At the level of entry, the best weight combination is that: 0.30 of *content*, 0.25 for *tags*, 0.20 for *emotion*, 0.15 for *sentiment*, and 0.10 for *category*. These results are reasonable: in Twitter, the *content* and *tags* are explicit value of entry, so they are most important in the results. Meanwhile, the three remaining criteria *emotion*, *sentiment*, and *category* are implicit values from entry. Their value possibly depends on the classifying method and, therefore, their importance may be reduced in the final results. The criterion *category* is less important possibly because this has only small possible different values. A value of this criterion could represent a big number of different entries. Therefore, this criteria does not very well classify the entries as the other criteria.

Table 4 Best weight of the entry criteria for the sample set

	Cont.	Tags	Cate.	Sent.	Emot.	Accuracy (%)
1/5 criteria				1.00		34.00
			1.00			47.20
					1.00	55.00
		1.00				57.00
	1.00					67.60
2/5 criteria			0.80		0.20	45.40
				0.35	0.65	52.20
		0.85	0.15			57.00
		0.90		0.10		58.80
		0.80			0.20	65.00
			0.60	0.40		69.20
	0.70	0.30				70.00
	0.85			0.15		74.00
	0.85		0.15			75.00
	0.80				0.20	76.00
3/5 criteria		0.60		0.20	0.20	64.00
		0.35	0.35		0.30	65.80
	0.40	0.30	0.30			69.00
		0.70	0.15	0.15		70.00
	0.45			0.10	0.45	71.40
	0.20	0.70		0.10		73.00
		0.20	0.50	0.30		75.20
	0.65		0.15		0.20	79.20
	0.65		0.20	0.15		82.00
	0.40	0.30			0.30	83.00
4/5 criteria		0.45	0.25	0.15	0.15	77.80
	0.15	0.60		0.10	0.15	78.60
	0.20	0.55	0.15	0.10		83.20
	0.35		0.25	0.10	0.30	85.20
	0.35	0.25	0.10		0.30	86.20
5/5 criteria	**0.30**	**0.25**	**0.10**	**0.15**	**0.20**	**87.60**

Bold values indicate the best case of experiment result

At the level of behavior, the best weight combination is: 0.35 of *post*, 0.30 for *comment*, 0.25 for *like*, and 0.10 for *join a group*. As mentioned in the scenario, in Twitter, there is no real data about *like* and *join a group*; we have to map the *favorite list* in Twitter to the activity *like*, and map the *subscribed to list* in Twitter to the activity *join a group* of

Table 5 Best weight of behavior for the sample set

	Post	Like	Comm.	Join	Accuracy (%)
1/4 behavior				1.00	38.00
		1.00			39.80
			1.00		54.00
	1.00				65.00
2/4 behavior		0.60		0.40	57.00
		0.45	0.55		59.00
	0.60			0.40	64.20
			0.65	0.35	74.80
	0.70	0.30			81.20
	0.75		0.25		83.20
3/4 behavior		0.30	0.45	0.25	82.20
	0.40	0.30		0.30	84.20
	0.40		0.30	0.30	84.80
	0.40	0.25	0.35		85.20
4/4 behavior	**0.35**	**0.25**	**0.30**	**0.10**	**87.60**

Bold values indicate the best case of experiment result

the model. This may be the main reason why in this experiment these two activities are less important than the two real activities of *post* and *comment*.

4 Conclusions

In this paper, we present a model for estimating the similarity between users based on their entries and behavior on social network. The considered behaviors are based on the activity of posting an entry or sharing an existing entry, liking an entry or liking a comment on an entry, commenting on an entry, and joining a group or community. The model is applied to estimate the similarity among users of Twitter. The results show that the model could estimate correctly the similarity among users in the majority of cases.

This model could be applied to several applications such as to predict the behavior of a social network user in commenting or liking some kind of status; to recommend some new entries which could be appropriate to a given user; to cluster the user based on some criteria.

References

1. Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J.: Effects of user similarity in social media. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM '12, pp. 703–712. ACM, New York, NY, USA (2012)
2. Benevenuto, F., Rodrigues, T., Cha, M., Almeida, V.: Characterizing user behavior in online social networks. In: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement, IMC '09, pp. 49–62. ACM, New York, NY, USA (2009)
3. Bhattacharyya, P., Garg, A., Wu, S.F.: Analysis of user keyword similarity in online social networks. Soc. Netw. Anal. Min. **1**(3), 143–158 (2011)
4. Chen, X., Pang, J., Xue, R.: Constructing and comparing user mobility profiles for location-based services. In: Proceedings of the 28th Annual ACM Symposium on Applied Computing, SAC '13, pp. 261–266. ACM, New York, NY, USA (2013)
5. Erlandsson, F., Bródka, P., Borg, A., Johnson, H.: Finding influential users in social media using association rule learning. CoRR arXiv:1604.08075 (2016)
6. Guo, C., Tian, X., Mei, T.: User specific friend recommendation in social media community. In: 2014 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2014)
7. Jamali, M., Ester, M.: Modeling and comparing the influence of neighbors on the behavior of users in social and similarity networks. In: 2010 IEEE International Conference on Data Mining Workshops, pp. 336–343 (2010)
8. Liu, H., Hu, Z., Mian, A., Tian, H., Zhu, X.: A new user similarity model to improve the accuracy of collaborative filtering. Knowl. Based Syst. **56**, 156–166 (2014)
9. Liu, H., Schneider, M.: Similarity measurement of moving object trajectories. In: Proceedings of the Third ACM SIGSPATIAL International Workshop on GeoStreaming, IWGS '12, pp. 19–22. ACM, New York, NY, USA (2012)
10. Liu, K., Terzi, E.: A framework for computing the privacy scores of users in online social networks. ACM Trans. Knowl. Discov. Data **5**(1), 6:1–6:30 (2010)
11. Nguyen, D.A., Tan, S., Ramanathan, R., Yan, X.: Analyzing information sharing strategies of users in online social networks. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 247–254 (2016)
12. Nguyen, M.H., Nguyen, T.H.: A general model for similarity measurement between objects. Int. J. Adv. Comput. Sci. Appl. **6**(2), 235–239 (2015)
13. Nguyen, T.H., Tran, D.Q., Dam, G.M., Nguyen, M.H.: Multifeature based similarity among entries on media portals. In: Akagi, M., Nguyen, T.T., Vu, D.T., Phung, T.N., Huynh, V.N. (eds.) Advances in Information and Communication Technology. Proceedings of the International Conference on Advances in Information and Communication Technology (ICTA 2016), pp. 373–382. Springer, Thai Nguyen, Viet Nam (2016)
14. Nguyen, T.H., Tran, D.Q., Dam, G.M., Nguyen, M.H.: Integrated sentiment and emotion into estimating the similarity among entries on social network. In: Chen, Y., Duong, T.Q. (eds.) Industrial Networks and Intelligent Systems, pp. 242–253. Springer, Cham (2018)
15. Peled, O., Fire, M., Rokach, L., Elovici, Y.: Entity Matching in Online Social Networks. Social Computing/IEEE International Conference on Privacy, Security, Risk and Trust, 2010 IEEE International Conference on 0, pp. 339–344 (2013)
16. Raad, E., Chbeir, R., Dipanda, A.: User profile matching in social networks. In: Proceedings of the 2010 13th International Conference on Network-Based Information Systems, NBIS '10, pp. 297–304. IEEE Computer Society, Washington, DC, USA (2010)

17. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill Inc, New York (1986)
18. Singh, K., Shakya, H.K., Biswas, B.: Clustering of people in social network based on textual similarity. Recent Trends in engineering and material sciences. Perspect. Sci. **8**(Supplement C), 570–573 (2016)
19. Sun, S., Li, Q., Yan, P., Zeng, D.D.: Mapping users across social media platforms by integrating text and structure information. In: 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), pp. 113–118 (2017)
20. Tang, X., Miao, Q., Quan, Y., Tang, J., Deng, K.: Predicting individual retweet behavior by user similarity. Know. Based Syst. **89**(C), 681–688 (2015)
21. Vedula, N., Parthasarathy, S., Shalin, V.L.: Predicting trust relations within a social network: A case study on emergency response. In: Proceedings of the 2017 ACM on Web Science Conference, WebSci '17, pp. 53–62. ACM, New York, NY, USA (2017)
22. Xu, K., Li, J., Liao, S.S.: Sentiment community detection in social networks. In: Proceedings of the 2011 iConference, iConference '11, pp. 804–805. ACM, New York, NY, USA (2011)
23. Xu, Z., Zhang, Y., Wu, Y., Yang, Q.: Modeling user posting behavior on social media. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12, pp. 545–554. ACM, New York, NY, USA (2012)
24. Zhao, G., Qian, X., Feng, H.: Personalized Recommendation by Exploring Social Users' Behaviors, pp. 181–191. Springer, Cham (2014)
25. Zhao, Z., Cheng, Z., Hong, L., Chi, E.H.: Improving user topic interest profiles by behavior factorization. In: Proceedings of the 24th International Conference on World Wide Web, WWW '15, pp. 1406–1416. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2015)
26. Zou, Z., Xie, X., Sha, C.: Mining user behavior and similarity in location-based social networks. In: 2015 Seventh International Symposium on Parallel Architectures, Algorithms and Programming (PAAP), pp. 167–171 (2015)