Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 472917, 14 pages
http://dx.doi.org/10.1155/2015/472917

Research Article

A Novel Adaptive Conditional Probability-Based Predicting Model for User's Personality Traits

Mengmeng Wang,[1,2] Wanli Zuo,[1,2] and Ying Wang[1,2,3]

[1]*College of Computer Science and Technology, Jilin University, Changchun 130012, China*
[2]*Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, Jilin University, Changchun 130012, China*
[3]*College of Mathematics, Jilin University, Changchun 130012, China*

Correspondence should be addressed to Ying Wang; 726854768@qq.com

Received 13 March 2015; Accepted 21 June 2015

Academic Editor: Antonino Laudani

With the pervasive increase in social media use, the explosion of users' generated data provides a potentially very rich source of information, which plays an important role in helping online researchers understand user's behaviors deeply. Since user's personality traits are the driving force of user's behaviors, hence, in this paper, along with social network features, we first extract linguistic features, emotional statistical features, and topic features from user's Facebook status updates, followed by quantifying importance of features via Kendall correlation coefficient. And then, on the basis of weighted features and dynamic updated thresholds of personality traits, we deploy a novel adaptive conditional probability-based predicting model which considers prior knowledge of correlations between user's personality traits to predict user's Big Five personality traits. In the experimental work, we explore the existence of correlations between user's personality traits which provides a better theoretical support for our proposed method. Moreover, on the same Facebook dataset, compared to other methods, our method can achieve an $F1$-measure of 80.6% when taking into account correlations between user's personality traits, and there is an impressive improvement of 5.8% over other approaches.

1. Introduction

As a new medium for information dissemination, social network has become a novel means of social interactions. Hence, user's individual behaviors have gradually turned into the key factors in social networks analysis. Besides, although some users post their desirable images and lives onto social media to achieve self-presentation which reflect some sort of "untrue" self, users' contributions and activities, which can be instantly made available to entire social network [1], still provide a valuable insight into individual behaviors.

Psychologists believed that user's personality traits were the driving force of user's behaviors, and individual differences in personality traits may have an impact on user's online activities [2, 3]. Due to the accessibility of user's personality traits features, in order to avoid basic rights violation or discrimination, some researchers carried out experiments with users' psychological features and profiles which were recorded for research purposes with users' consent to make a deep understanding of online social networks via user's personality traits. User's personality traits can be used to predict early adoption about Facebook [4]: a conscientious person has sparing use of Facebook, an extroverted person has long sessions and abundant friendships, and a neurotic person has high frequency of sessions. Moreover, user's personality traits may help optimize search results [5], manifest social influence [6], and distinguish individuals who have some common properties in the crowd [7]. It also plays an important role in relationship outcomes (customer trust, satisfaction, and loyalty) [8]. To sum up, user's personality traits recognition has an important theoretical significance to mine user's behavior patterns and get user's potential needs under different contexts. Hence, analyzing and forecasting user's personality traits through mining data from online social networking sites have become a research focus. Our work on predicting user's personality traits is motivated by its broad application prospect.

Nevertheless, confronted with the same problem as stated in [9], accessibility of user's personality traits features may vary. Consequently, user's personality traits are by nature difficult to predict. Generally, psychologists regarded personality traits, which were reflected in user's attitudes towards things and actions taken by user [10], as a person's unique pattern of long-term thoughts, emotions, and behaviors [11, 12]. Therefore, leveraging long-term mode of expression and emotion behind numerous contents that a user posts to predict his/her personality traits can be a feasible measure. Furthermore, users who have diverse personality traits tend to pay attention to different types of things; in such a scenario, we expect that topics a user is concerned about would enhance the performance of user's personality traits prediction. Based on the research train of thought, in this paper, we propose a new adaptive conditional probability-based framework for user's personality traits prediction. Our main contributions are summarized next.

(1) Demonstrate the existence of interdependencies between user's personality traits.

(2) On account of social network features, linguistic features, emotional statistical features, and topic features, we put forward a novel unsupervised adaptive conditional probability-based framework for the problem of predicting user's personality traits through taking prior knowledge of correlations between user's personality traits into consideration.

(3) Exploit correlations between features and user's personality traits via Kendall correlation coefficient, so as to quantify importance of each feature.

(4) Update threshold of each personality trait dynamically rather than adopt a unified threshold.

The rest of the paper is organized as follows: Section 2 describes the related work; Section 3 defines the method we propose; details of the experimental results and dataset which is used in this study are given in Section 4; finally, conclusion appears in Section 5.

2. Related Work

As research on user's behaviors in social networks has become a hot spot, user's personality recognition has received a significant amount of attention in both theory and practice. Argamon et al. [13] and Mairesse et al. [14] were first dedicated to this research field. Generally, there were two main approaches adopted for studying user's personality traits in social networks. Merely based on social network activities, one was using machine learning algorithms to capture user's personality traits. Moore and McElroy [15] explored user's personality traits through questionnaires and log data of 219 college students, and scale reliabilities for the five personality dimensions of the IPIP were acceptable with Cronbach's alpha values of 0.90 for extraversion, 0.81 for agreeableness, 0.82 for conscientiousness, 0.83 for emotional stability, and 0.79 for openness to experience which revealed that mining user's personality traits in Facebook was feasible. Kosinski et al. [5] first presented that there were psychologically meaningful links between users' personalities, their website preferences, and Facebook profile features and then predicted individual's personality traits via multivariate linear regression. The experimental results indicated that extroversion was most highly expressed by Facebook features, followed by neuroticism, conscientiousness, and openness. Agreeableness was the hardest trait to predict (0.05 in terms of accuracy) using Facebook profile features and the simple model.

The other one extended personality-related features with linguistic cues. On the basis of the corpus which was derived from essays written by students at the University of Texas at Austin [16], Argamon et al. [13] utilized SMO [17] to determine whether each author had high or low neuroticism or extraversion, respectively. For neuroticism, their experimental results have shown clearly the usefulness of functional lexical features, in particular the appraisal lexical taxonomy, while in the case of extraversion, experimental results were less clear, but examination of indicative function words pointed the way to developing more effective features, by focusing on expressions related to norms, (in)completeness, and (un)certainty. Golbeck et al. [18] explored whether the publicly available information on users' Facebook profile can predict personality traits. In order to predict personality traits of 279 Facebook users, based on linguistic, structural, and semantic features, they used the profile data as a feature set and trained two machine learning algorithms, m5sup' Rules [19] and Gaussian Processes [20], to predict each of the five personality traits within 11% of their actual value. In addition, the experimental results have shown that user's Big Five personality traits can be predicted from the public information they shared on Facebook.

Mairesse et al. [14] leveraged classification, regression, and ranking models to recognize personality traits via LIWC and MRC features automatically. And experiments were carried out with the essays corpus and the EAR corpus, respectively. The results revealed that the LIWC features outperformed the MRC features for every trait, and the LIWC features on their own always performed slightly better than the full feature set. Concerning the algorithms, it can be found that AdaboostM1 [17] performed the best for extraversion (56.3% correct classifications), while SMO produced the best models for all other traits. They also pointed out that features were likely to vary depending on the source of language and the method of assessment of personality through analyzing the correlations between different characters. In order to forecast RenRen's 335 users' personality traits, according to the number of user's friends and a state recently released, Bai et al. [21] used many classification algorithms such as Naive Bayesian (NB) [17], Support Vector Machine (SVM), and Decision Tree [17]. And they found out that C4.5 Decision Tree can get the best results (0.697 for agreeableness, 0.749 for neuroticism, 0.824 for conscientiousness, 0.838 for extraversion, and 0.811 for openness in terms of $F1$-measure). Oberlander and Nowson [22] achieved better results (ranking on raw accuracy: agreeableness > conscientiousness > neuroticism > extraversion, the best agreeableness accuracy was 30.4% absolute over the baseline (77.2% relative)) on classification of personality

traits through leveraging Native Bayes model on account of differing sets of n-gram features. They also demonstrated that, with respect to feature selection policies, automatic selection generally outperformed "manual" selection. On the basis of automatically derived psycholinguistic and mood-based features of a user's textual messages, Nguyen et al. [23] utilized SVM classifier for examining two two-class classification problems: influential versus noninfluential and extraversion versus introversion. They experimented with three subcorpora of 10000 users each and presented the most effective predictors for each category. The best classification result, at 80%, was achieved using psycholinguistic features. However, they did not predict personality traits in finer grain. Bai et al. [24] proposed a multitask regression algorithm and an incremental regression algorithm to predict users' Big Five personality traits from their usages of Sina microblog objectively. The results indicated that personality traits can be predicted in a high accuracy through online microblog usages. Besides, the average mean absolute error of multitask regression model was 13.84% which got about 5 percentage points reduction compared to incremental regression. Sun and Wilson [25] demonstrated that, without any significant addition or modification, a cognitive architecture can serve as a generic model of personality traits. Besides, integrating personality modeling with generic computational cognitive modeling was shown to be feasible and useful.

However, some drawbacks can be pointed out in previous work on user's personality traits prediction: (1) some researchers have made an assumption that there had been little or no correlations between user's personality traits [26]; however, the massive approaches have been explored to examine personality psychology and have revealed the correlations between the Big Five dimensions instead of little or no correlations [27–30]. (2) Although different features played different roles in predicting user's personality traits [18, 31, 32], only a few researchers considered the correlations between features and personality traits [33]. (3) In multilabel learning task, such as PT5 method proposed by Tsoumakas and Katakis [34], thresholds were usually unified to the same value, which was not appropriate. In this light, we propose an adaptive conditional probability-based model to improve the performance of user's personality traits prediction task.

3. Adaptive Conditional Probability-Based Predicting Model for User's Personality Traits

In this section, we present predicting model adopted in our work. Initially we make a definition of preliminary features (Section 3.1), followed by measurement of distributing weights to characteristics (Section 3.2). And then, we outline framework of adaptive conditional probability-based user's personality traits prediction model (Section 3.3). Finally, in Section 3.4, we depict our algorithm and analyze time complexity of our proposed model.

3.1. Definition of Features. As a person's unique pattern of long-term thoughts, emotions, and behaviors, personality traits are reflected in user's attitudes towards things and actions taken by user. Therefore, aside from social network features which were provided in Facebook dataset [35] directly, we introduced linguistic features, emotional statistical features, and topic features to predict user's personality traits. Linguistic features, emotional statistical features, and topic features can be measured via analysis of user's Facebook status updates.

3.1.1. Social Network Features. Facebook dataset includes seven social network features, namely, date of user's register, network size, ego betweenness centrality, normalized ego betweenness centrality, density, brokerage, normalized brokerage, and transitivity, which reflect user's behavior patterns just through user's network structure. Similar to the cluster assumption [36], in this work, we took the above features into consideration to complete prediction task and assumed that the more similar were the network structures between two users, the more likely they shared the same label.

3.1.2. Linguistic Features. Since each user showed a particular mode of expression, some researchers held the view that correlations between personality traits and spoken or written linguistic cues were significant [14, 16], so language-based assessments can constitute valid personality measures [28]. Hence, in this paper, we regarded linguistic cues as factors so as to mine user's personality traits through user's means of expression.

A natural language parser is used to work out grammatical structure of sentences, such as grouping words together as "phrases" and obtaining subject or object of a verb. Probabilistic parsers try to produce the most likely analysis of new sentences via leveraging knowledge of language gained from hand-parsed sentences. Stanford Parser (http://nlp.stanford.edu/software/lex-parser.shtml#About) is a probabilistic natural language open source parser, which implements a factored product model, with separate PCFG phrase structure and lexical dependency experts, whose preferences are combined by efficient exact inference, using an A^* algorithm. Either of these yields a good performance statistical parsing system [37]. A GUI is provided for utilizing it simply as an accurate unlexicalized stochastic context-free grammar parser and viewing the phrase structure tree output of the parser. Thus, in order to learn traits of contents that user yields, we got word frequency statistics of 35 kinds of parts of speech with Stanford Parser that were conjunction, numeral, determiner, existential there, foreign word, modal auxiliary, singular or mass noun, plural noun, proper noun, plural proper noun, predeterminer, genitive marker, personal pronoun, plural personal pronoun, ordinal adverb, comparative adverb, superlative adverb, particle, symbol, interjection, verb in base form, verb in past tense, gerund, verb in past participle, verb in present tense (not 3rd person singular), verb in present tense (3rd person singular), subordinating conjunction, ordinal adjective, comparative adjective, superlative adjective, list item marker, WH-determiner, WH-pronoun, WH-plural pronoun, and WH-adverb. Besides, we defined another six linguistic features: the total number of words and frequency statistics of punctuation, comma,

period, exclamation, and question. Nevertheless, we filtered hyperlinks in users' contents as blog services tended to produce many hyperlinks for navigation and advertisement [38], which had no relationship with personality traits prediction.

3.1.3. Emotional Statistical Features. User's attitudes towards things, which show user's different personality traits, reflect user's unique pattern of long-term emotions. For instance, a neurotic person may have a tendency to experience unpleasant emotions easily, such as anger, anxiety, depression, and vulnerability. Hence, user's statistics of emotion can be characteristics in user's personality traits predicting model. In this paper, user's emotional statistical characteristics included proportion of positive words and negative words used in user's posts. On the basis of adjectives and their variants obtained in Section 3.1.2, user's emotional statistical characteristics were calculated with the corpus of HowNet Knowledge (http://www.keenage.com/download/sentiment.rar). HowNet Knowledge, which includes 8945 words and phrases, consists of six files: positive emotional words list file, negative emotional words list file, positive review words list file, negative review words list file, degree level words list file, and proposition words list file.

As user's emotional statistical features, user's positive and negative emotional statistical characteristics are defined as

$$\begin{aligned} \mathrm{PT}(i) &= \frac{pn_i}{\mathrm{sum}_i}, \\ \mathrm{NT}(i) &= \frac{nn_i}{\mathrm{sum}_i}, \end{aligned} \tag{1}$$

where PT(i) and NT(i) represent proportion of positive words and negative words used in user i's posts, respectively, pn_i and nn_i represent the number of positive emotional words and the number of negative emotional words user i used which are included in HowNet Knowledge, and sum_i represents the total number of words in user i's contents.

3.1.4. Topic Features. The things user focuses on may have an impact on actions that user has taken. Take openness which is one of the Big Five personality traits as an example: it reflects degree of intellectual curiosity, creativity, and a preference for novelty and variety a person has. Therefore, we mined a series of user's concerned themes from user's status via LDA (Latent Dirichlet Allocation) [39] for predicting user's personality traits.

Since our purpose is to extract all concerned themes of a user, rather than to extract specific themes of each post, we merged all microblogs of a user into one document and then extracted user's concerned themes; namely, each document corresponded to a user. The results of LDA model are shown as follows:

$$\mathrm{DT} = \begin{bmatrix} \mathrm{DT}_{11} & \mathrm{DT}_{12} & \cdots & \mathrm{DT}_{1n} \\ \mathrm{DT}_{21} & \mathrm{DT}_{21} & \cdots & \mathrm{DT}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{DT}_{m1} & \mathrm{DT}_{m2} & \cdots & \mathrm{DT}_{mn} \end{bmatrix}, \tag{2}$$

where DT stands for an $m \times n$ matrix, which is mainly used for storage of distribution of themes and documents, m stands for the number of documents, n stands for the number of themes, and element DT_{ij} ($i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$) in matrix DT stands for probability that the ith document belongs to the jth theme, that is, degree of attention which user i pays to the jth theme.

3.2. Weight Distribution of Features. Every feature has a different impact on user's personality traits prediction. It is of great importance to allocate features' weights reasonably so as to be able to perform good prediction based only on scant knowledge of personality traits. Kendall test is a nonparametric hypothesis test which calculates correlation coefficient to test statistical dependence of two random variables. Since values of features and scores of personality traits in the dataset we used were numeric, therefore, in order to quantify importance of each feature, we analyzed relevance between user's personality traits and characteristics via Kendall correlation coefficient in which values of features and scores of personality traits were treated as two random variables. Kendall correlation coefficient is calculated as

$$\tau(X, Y) = \frac{C - D}{\sqrt{(N_3 - N_1) \times (N_3 - N_2)}}, \tag{3}$$

where $\tau(X, Y) \in [-1, 1]$, and if random variables X and Y have a positive correlation, then $\tau(X, Y) = 1$; else if random variables X and Y have a negative correlation, then $\tau(X, Y) = -1$; else if random variables X and Y are independent of each other, then $\tau(X, Y) = 0$. C represents the number of tuples whose two random variables have consistency, D represents the number of tuples whose two random variables do not have consistency, and N_1 and N_2 represent the total number of repeated elements in random variables X and Y, respectively. The calculation of N_1 is shown as

$$N_1 = \sum_{i=1}^{s} \frac{1}{2} U_i (U_i - 1), \tag{4}$$

where s denotes the number of elements that have repeated elements in random variable X and U_i denotes the number of repeated elements of the ith element. Similarly, N_2 is calculated as

$$N_2 = \sum_{j=1}^{s'} \frac{1}{2} U_j \left(U_j - 1\right), \tag{5}$$

where s' denotes the number of elements that have repeated elements in random variable Y and U_j denotes the number of repeated elements of the jth element. N_3 denotes the total number of merged sequences, which is calculated as

$$N_3 = \frac{1}{2} N (N - 1), \tag{6}$$

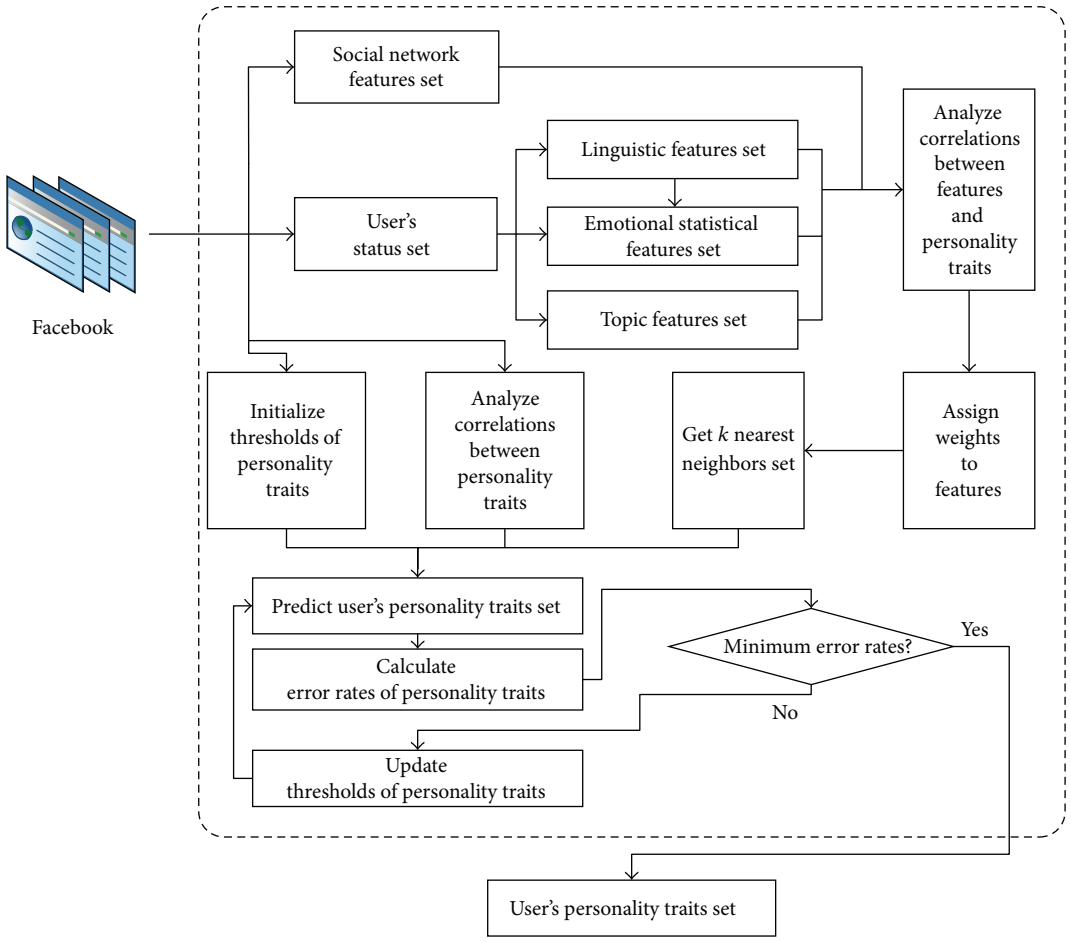

Figure 1: The architecture of adaptive conditional probability-based predicting model for user's personality traits.

where N represents dimensions of tuples. Thus, according to Kendall correlation coefficient between features and personality traits, importance of the ith feature f_i is calculated as

$$I(f_i) = \frac{\sum_{p_j \in P} \tau(f_i, p_j)}{5}, \tag{7}$$

where $\tau(f_i, p_j)$ stands for Kendall correlation coefficient between the ith feature f_i and the jth personality trait p_j and P stands for set of Big Five personality traits which will be introduced in Section 3.4. After normalizing importance of the ith feature f_i, weight of f_i is calculated as follows:

$$W(f_i) = \frac{I(f_i)}{\sum_{f_j \in F, j \neq i} I(f_j)}, \tag{8}$$

where F denotes set of personality traits predicting features.

3.3. A Framework of Predicting User's Personality Traits. Figure 1 shows the architecture of adaptive conditional probability-based predicting model for user's personality traits. An analysis of correlations between user's personality traits and features is conducted before assigning weights to linguistic, emotional statistical, and topic features which are extracted from contents that user produces, as well as social network features. Meanwhile, correlation analysis of user's personality traits is carried on. Then, according to weights of features, k nearest neighbors set is obtained. Finally, given k nearest neighbors set, dynamic updated thresholds of personality traits, and correlations between personality traits, user's personality traits set is predicted.

3.4. Algorithm of Adaptive Conditional Probability-Based User's Personality Traits Prediction. Before we proposed our predicting algorithm, we conducted experiments to analyze correlations between user's personality traits which will be shown in detail later in Section 4.2. And the results of correlation analysis revealed that there were interdependent relationships between user's personality traits. Thus, given dynamic updated thresholds of personality traits, as well as considering interdependencies between personality traits, a novel adaptive correlation-based predicting model was proposed.

In psychology, the Big Five personality traits are five broad domains or dimensions of personality traits that are used to describe human personality. The theory based on the Big Five factors is called Five Factor Model (FFM) [40]. The Big Five factors are extraversion (denoted by EXT),

neuroticism (denoted by NEU), agreeableness (denoted by AGR), conscientiousness (denoted by CON), and openness (denoted by OPN).

Our method aimed to predict user i's Big Five personality traits set PT_i. As stated in [41], here, each personality trait was classified into two degrees: namely, extraversion was mapped onto extravert and shy, neuroticism was mapped onto neurotic and secure, agreeableness was mapped onto friendly and uncooperative, conscientiousness was mapped onto precise and careless, and openness was mapped onto insightful and unimaginative. For convenience, we denoted the two degrees of each personality trait by positive level and negative level in short, respectively. First, we set initial threshold of each personality trait as 0.5 and threshold of the θth personality trait p_θ is denoted by $T(p_\theta)$, followed by calculating distance between user i and other users with (9) as follows:

$$\mathrm{Dis}(i,j) = \sum_{h=1}^{m} W(f_h) \times \left| F(i,f_h) - F(j,f_h) \right|, \quad (9)$$

where $\mathrm{Dis}(i,j)$ presents distance between user i and user j, $F(i,f_h)$ and $F(j,f_h)$ present the hth feature f_h's value of user i and user j, and m presents the number of features.

Secondly, we sorted user i's distance set in ascending order and selected top k users as user i's neighbors (denoted as $N(i)$). Then for each p_θ, we recorded the number of users who had positive level of p_θ in $N(i)$ as n_{p_θ}.

Thirdly, we selected a p_θ from unvisited personality traits set UP, if user i's personality traits set PT_i was empty, and then we calculated probability that user i had positive level of p_θ as follows:

$$f(i,p_\theta) = \frac{P\left(H_{p_\theta} \mid C_{p_\theta}^{n_{p_\theta}}\right)}{P\left(\sim H_{p_\theta} \mid C_{p_\theta}^{n_{p_\theta}}\right)}, \quad (10)$$

where $C_{p_\theta}^{n_{p_\theta}}$ presents event that there are exactly $C_{p_\theta}^{n_{p_\theta}}$ instances having positive level of p_θ in $N(i)$, whose k nearest neighbors also have n_{p_θ} users with positive level of p_θ. H_{p_θ} and $\sim H_{p_\theta}$ present event that user i has positive level of p_θ and event that user i has negative level of p_θ, respectively. And if user i's personality traits set PT_i was not empty, considering correlations between personality traits, we calculated probability that user i had positive level of p_θ on the basis of prior knowledge about correlations between p_θ and personality traits that have already been visited in UP as follows:

$$f(i,p_\theta) = \frac{P\left(H_{p_\theta} \mid p_{l1}, p_{l2}, \ldots\right) P\left(C_{p_\theta}^{n_{p_\theta}} \mid H_{p_\theta}\right)}{P\left(\sim H_{p_\theta} \mid p_{l1}, p_{l2}, \ldots\right) P\left(C_{p_\theta}^{n_{p_\theta}} \mid \sim H_{p_\theta}\right)}, \quad (11)$$

where $p_{l1}, p_{l2}, \ldots \in \mathrm{PT}_i$ present personality traits which have already been visited in UP. If $f(i,p_\theta)$ was greater than or equal to $T(p_\theta)$, then positive level of p_θ was added to PT_i; else negative level of p_θ was added to PT_i.

And then, we calculated error rate of each personality trait, and error rate of p_θ is denoted by $\mathrm{Err}(p_\theta)$ which is calculated as

$$\mathrm{Err}(p_\theta) = 1 - \frac{r_\theta}{\mathrm{to}}, \quad (12)$$

where r_θ denotes the number of instances which are classified correctly to denote the total number of instances.

Finally, if unvisited personality traits set UP was empty and error rate of each personality trait fell within a specified range, then algorithm was terminated; else if unvisited personality traits set UP was not empty, then we continued to predict another personality trait in UP; else if there was a personality trait p_θ whose error rate did not reach defined limits, then we updated $T(p_\theta)$ with (12) and added it to personality traits set UP to predict it again:

$$T(p_\theta) = T(p_\theta) + \varepsilon \times \mathrm{Err}(p_\theta), \quad (13)$$

where ε denotes a monotonic decreasing learning rate. Our algorithm can now be defined as Algorithm 1. Assume that size of dataset is Tr, dimension of features is d, and size of the nearest neighbors set is k. The complexity of conditional probability-based predicting model for user's personality traits is analyzed as follows. From step (10) to step (12), computation time of Kendall correlation coefficient between features and personality traits is taking $O(Tr^2)$ time, calculating weights of features takes $O(d)$ time, and then distributing weights to features can be computed in $O(dTr)$ time; the total time is $O(Tr^2)$ as dimension of features d is far less than size of dataset Tr. Calculating distance between user i and other users from step (13) to step (15) will take $O(dTr)$ time. In step (16), it will take $O(Tr\log_2 Tr)$ time for sorting set of distances between user i and other users. Step (18) to step (20) take $O(k)$ time to select k nearest neighbors of user i. If algorithm goes from step (23) to step (25), it will take $O(k^2)$ time. Else if algorithm goes from step (26) to step (28), it will take $O(Trk^2)$ time. It takes $O(Tr)$ time from step (37) to step (45). Assume that the number of iterations is t. Since steps (10), (11), and (12) can be done offline and do not need to be calculated repeatedly every time, hence, from step (13) to step (52), the overall online complexity of our proposed method is $O(Tr\log_2 Tr)$ as k and t are far less than Tr. From here we see that our proposed method is feasible in a big data environment as in Facebook monitoring.

4. Experimental Evaluation

In this section, we first describe dataset used in our experiments. And then we analyze the interdependent relationships between user's personality traits. Finally, we conduct experiments on different kinds of features and make a comparison with other methods based on the same dataset.

4.1. Dataset. MyPersonality (http://mypersonality.org) is a popular Facebook application which is a 336-question test measuring the 30 facets underlying the Big Five traits that allows users to take real psychometric tests. The respondents who come from various age groups, backgrounds, and cultures are highly motivated to answer honestly and carefully. Additionally, users' psychological and Facebook profiles are "verified" by individuals' social circles; hence, it is hard to lie to the people that know you best. Also, there is no pressure, and one does not have to share his/her profile information; thus, myPersonality avoids deliberate faking

Input: user i; users set U; size of the nearest neighbors set k; unvisited personality traits set UP
Output: user i's personality traits set PT_i
(1) $f \leftarrow false$
/* f stands for a flag, if any of the thresholds of personality traits is updated, then f equals to true, otherwise f equals to false */
(2) For each personality trait $p_\theta \in$ UP do
(3) set temporary error rate $\mathrm{Temp}(p_\theta)$ as 0
(4) End for
(5) For each personality trait $p_\theta \in$ UP do
(6) set initial threshold $T(p_\theta)$ as 0.5
(7) End for
(8) $\mathrm{PT}_i \leftarrow \emptyset$
(9) UP $\leftarrow$ all of the Big Five personality traits
(10) For each feature $f_h \in$ user i's feature set do
(11) calculate f_h's weight according to (8)
(12) End for
(13) For user $j \in U$ and $j \neq i$ do
(14) calculate $\mathrm{Dis}(i, j)$ according to (9)
(15) End for
(16) sort $\{\mathrm{Dis}(i, 1), \mathrm{Dis}(i, 2), \ldots\}$ in ascending order
(17) $N(i) \leftarrow$ select top k users in $\{\mathrm{Dis}(i, 1), \mathrm{Dis}(i, 2), \ldots\}$
(18) For each personality trait $p_\theta \in$ UP do
(19) $n_{p_\theta} \leftarrow$ the number of users who have positive level of personality trait p_θ in $N(i)$
(20) End for
(21) While UP $\neq \emptyset$ do
(22) $p_\theta \leftarrow$ UP.poll()
(23) If $\mathrm{PT}_i = \emptyset$ then
(24) calculate $f(i, p_\theta)$ according to (10)
(25) End if
(26) Else if $\mathrm{PT}_i \neq \emptyset$ then
(27) calculate $f(i, p_\theta)$ according to (11)
(28) End if
(29) If $f(i, p_\theta) \geq T(p_\theta)$ then
(30) $PT_i \leftarrow$ positive level of p_θ
(31) End if
(32) Else if $f(i, p_\theta) < T(p_\theta)$ then
(33) $\mathrm{PT}_i \leftarrow$negative level of p_θ
(34) End if
(35) End while
(36) $f \leftarrow false$
(37) For each personality trait $p_\theta \in \mathrm{PT}_i$ do
(38) calculate p_θ's error rate $\mathrm{Err}(p_\theta)$ according to (12)
(39) If $|\mathrm{Err}(p_\theta) - \mathrm{Temp}(p_\theta)| > \delta$ then //δ denotes a constant which is small enough
(40) $f \leftarrow true$
(41) $\mathrm{Temp}(p_\theta) \leftarrow \mathrm{Err}(p_\theta)$
(42) update p_θ's threshold $T(p_\theta)$ according to (13)
(43) UP $\leftarrow p_\theta$
(44) End if
(45) End for
(46) If $f == true$ then
(47) Go to step (21)
(48) End if
(49) Else if $f == false$ then
(50) Go to step (52)
(51) End if
(52) Return PT_i

Algorithm 1: Adaptive conditional probability-based predicting model for user's personality traits.

of data. With consent, users' psychological and Facebook profiles are recorded for research purposes. Currently, the database contains more than 6,000,000 test results together with more than 4,000,000 individual Facebook profiles. In this paper, we adopted the dataset that was provided in the Workshop on Computational Personality Recognition (Shared Task) (http://mypersonality.org/wiki/lib/exe/fetch.php?media=wiki:mypersonality_final.zip) [35]. The dataset was a subset of myPersonality database including 250 users which had both information about personality traits (annotated with users' personality scores and gold standard personality labels which were self-assessments obtained using a 100-item long version of the IPIP personality questionnaire (http://ipip.ori.org/newFinding_Labeling_IPIP_Scales.htm)) and social network structure (network size, betweenness centrality, density, brokerage, and transitivity) along with their 9900 status updates in raw text. With the aid of this standard dataset, we can compare the performance of our proposed method with others' personality recognition systems on a common benchmark.

4.2. Analysis of Correlations between Personality Traits. Previous works were usually predicting user's personality traits without considering interdependencies between them. In this context, we investigated whether there had been relationships between user's personality traits. Since not all users had the same level of interactions, consequently, we first grouped users according to different degrees of user's personality traits. As an example, if a user has positive level of agreeableness, another user has positive level of agreeableness as well; then they will be divided into a group; else if a user has positive level of agreeableness, another user has negative level of agreeableness, and they will not be divided into a group. Then, we explored cooccurrences between personality traits in each group simultaneously, which is shown in Figure 2.

It can be observed intuitively that a significant proportion of users have negative level of extraversion, positive level of openness, or positive level of agreeableness. It is also noteworthy that positive level of extraversion has less overlap with negative level of openness, negative level of conscientiousness, and negative level of agreeableness. Besides, positive level of neuroticism has less overlap with positive level of conscientiousness, positive level of agreeableness, and negative level of openness. Furthermore, there are bits of users that have positive level of extraversion and positive level of neuroticism simultaneously.

However, the above cooccurrences may be due to the a priori statistics of each trait. For instance, if there are more users with negative level of neuroticism, positive level of openness, and positive level of agreeableness than others, it is more likely to have more users with cooccurrence between negative level of neuroticism and positive level of openness, negative level of neuroticism and positive level of agreeableness, and positive level of openness and positive level of agreeableness than others, as it is shown in Figure 2. Therefore, since users were grouped according to different degrees of user's personality traits, in order to investigate whether the above phenomenon meant that there were positive or negative correlations between different degrees of personality traits, in each group, we treated scores of a certain personality trait, according to which users were divided into the group, and scores of another personality trait with a certain degree as two random variables, followed by employing Kendall correlation coefficient which was calculated in (3) to analyze real correlations between them. The results are presented in Table 1.

As it can be seen from Table 1, we can observe a negative correlation between negative level of extraversion and positive level of neuroticism; in other words, people that score high on extraversion have less possibility to score high on neuroticism as well. In psychology, a person with negative level of extraversion can be explained as solitary and a person with negative level of conscientiousness means being careless. Consequently, negative level of extraversion and negative level of conscientiousness are not in line with positive level of neuroticism which tends to experience unpleasant emotions easily. Another factor is social desirability. Moreover, agreeableness, extraversion, conscientiousness, and openness are more or less "desirable," whereas neuroticism is quite clearly negative. Thus, generally, there are negative correlations between neuroticism and other personality traits.

What is more, it is inconsistent with Figure 2 that there is a positive correlation between positive level of neuroticism and positive level of extraversion. It may be explained that a person with positive level of extraversion tends to be enthusiastic and talkative, which is compatible with positive level of neuroticism. In addition, in line with Figure 2, since negative level of openness shows cautiousness, it has a positive correlation with negative level of extraversion and positive level of agreeableness which is a tendency to be compassionate. However, there is a positive correlation between negative level of openness and positive level of neuroticism which is not in keeping with Figure 2.

Since there are some contradictions between Figure 2 and Table 1, we leveraged Jensen-Shannon divergence [45] to further analyze correlations between user's personality traits. Jensen-Shannon divergence between two probability distributions is calculated as

$$D_{\mathrm{JS}}(P,Q) = \frac{1}{2}\left(D_{\mathrm{KL}}(P \parallel R) + D_{\mathrm{KL}}(Q \parallel R)\right), \quad (14)$$

where P and Q denote two probability distributions, R denotes an average distribution of P and Q, and $D_{\mathrm{KL}}(P \parallel R)$ denotes Kullback-Leibler divergence [46] between P and R, which is calculated with

$$D_{\mathrm{KL}}(P \parallel R) = \sum_i P(i) \log \frac{P(i)}{R(i)}, \quad (15)$$

where $P(i)$ and $R(i)$ denote the ith value of P and R. The results are presented in Table 2.

In keeping with Figure 2 and Table 1, Jensen-Shannon divergences between negative level of neuroticism and positive level of agreeableness, positive level of conscientiousness, and positive level of openness are larger. It is because of the fact that a person with negative level of neuroticism shows confidence on everything, and positive level of openness reflects degree of intellectual curiosity, creativity, and

Table 1: Kendall correlation coefficients between personality traits. p and n stand for positive level and negative level of a certain personality trait. The bold number is the highest Kendall correlation coefficients and the italic number is the lowest Kendall correlation coefficients. The bigger Kendall correlation coefficient is, the better the consistency is.

		EXT		NEU		AGR		CON		OPN	
		p	n	p	n	p	n	p	n	p	n
EXT	p			0.22	−0.03	0.09	−0.15	−0.06	−0.13	0.07	0.01
	n			*−0.23*	−0.18	0.04	0.04	0.04	0.18	0.07	0.22
NEU	p					−0.09	−0.17	0.10	−0.21	−0.03	0.27
	n					−0.14	−0.13	0.02	−0.01	−0.06	−0.10
AGR	p							−0.07	0.07	0.16	**0.29**
	n							−0.08	0.16	0.06	0.02
CON	p									−0.01	−0.10
	n									0.07	−0.07
OPN	p										
	n										

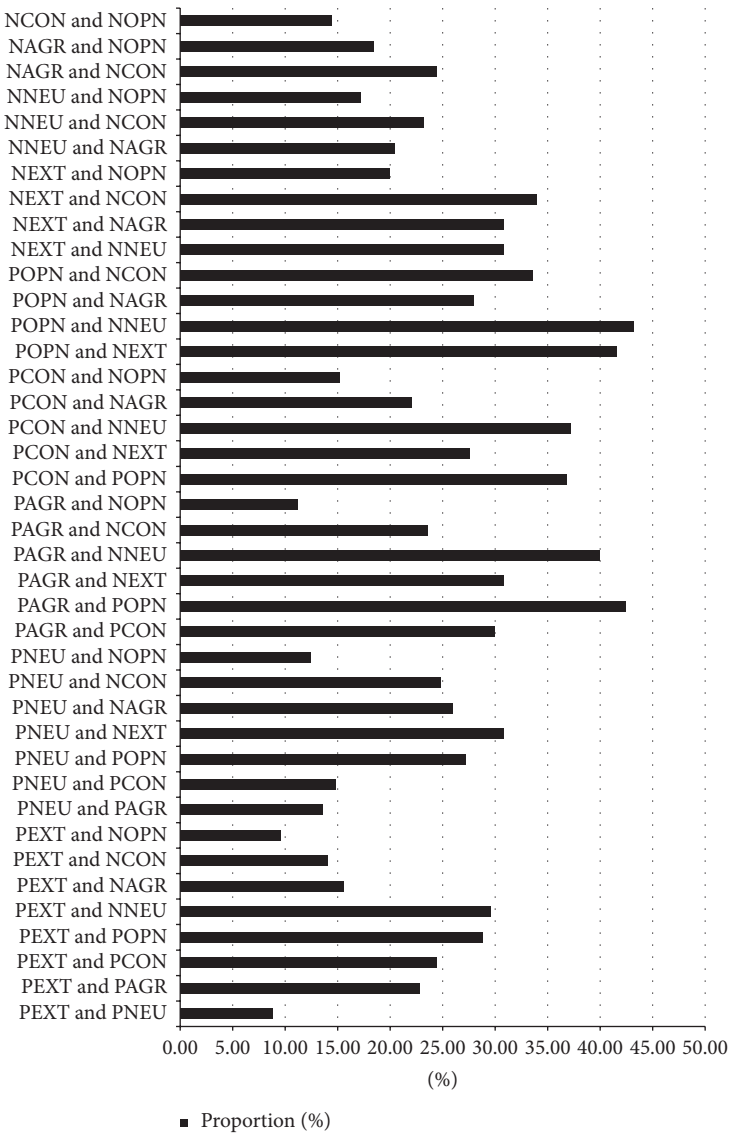

Figure 2: The proportion of users who have a certain pair of personality traits simultaneously. PEXT, PNEU, PAGR, PCON, and POPN stand for positive level of extraversion, neuroticism, agreeableness, conscientiousness, and openness. NEXT, NNEU, NAGR, NCON, and NOPN stand for negative level of extraversion, neuroticism, agreeableness, conscientiousness, and openness.

Table 2: Jensen-Shannon divergence between personality traits. p and n stand for positive level and negative level of a certain personality trait. The bold number is the highest Jensen-Shannon divergence and the italic number is the lowest Jensen-Shannon divergence. Smaller Jensen-Shannon divergence means better consistency.

		EXT		NEU		AGR		CON		OPN	
		p	n	p	n	p	n	p	n	p	n
EXT	p			0.58	14.79	0.44	2.19	0.70	2.36	0.73	0.84
	n			4.79	4.49	5.21	2.36	5.61	1.91	10.43	2.11
NEU	p					0.80	2.12	1.10	2.46	2.98	*0.40*
	n					19.78	2.94	16.97	3.05	**24.73**	3.61
AGR	p							0.77	3.29	0.82	0.91
	n							3.40	0.86	4.63	1.25
CON	p									1.15	1.17
	n									6.16	1.40
OPN	p										
	n										

a preference for novelty and variety a person has, along with positive level of agreeableness and positive level of conscientiousness, which are compatible. Furthermore, it is in conformity with Table 1 that positive level of extraversion has smaller Jensen-Shannon divergences with positive level of neuroticism and positive level of agreeableness; besides, positive level of neuroticism has a smaller Jensen-Shannon divergence with negative level of openness.

In addition, Soto et al. [29] demonstrated convergent and discriminate correlations for the Big Five Inventory. Soto and John [30] also illustrated that there was strong convergence between each BFI facet scale and corresponding NEO PI-R facet. Moreover, Park et al. [28] demonstrated convergence with self-reports of personality at the domain- and facet-level. In summary, although utilizing different Big Five measures, it can be found that correlations in myPersonality dataset are also in keeping with the findings from other samples. Hence, correlations between personality traits are expected and actually prove that myPersonality results are valid. So leveraging prior knowledge about relationships between personality traits may contribute to a better prediction on personality traits. The results provide a better theoretical support for our proposed method.

4.3. Evaluation Metric. Since most researchers exploring on personality recognition leveraged different measures to evaluate their experiments on various datasets, it was hard to completely appraise their performance and quality. However, a common dataset which was annotated with gold standard personality labels was available in the Workshop on Computational Personality Recognition (Shared Task), users were free to split the training and test sets as they wish, and precision, recall, and $F1$-measure were suggested to be used to evaluate results of predictions. So in this paper, for more reliable results, the proposed method was also evaluated with stratified 5-fold cross-validation. Since there was no training set in unsupervised learning, the common dataset was split into folds and in order to avoid overfitting, the same author did not appear twofold at the same time. The mean precision, recall, and $F1$-measure are calculated as follows:

$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}},$$
$$\text{Recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}, \tag{16}$$
$$F1\text{-measure} = \frac{2 \times P \times R}{P + R},$$

where tp stands for the number of positive samples that are classified correctly, fp stands for the number of positive samples that are classified incorrectly, and fn stands for the number of negative samples that are classified incorrectly.

Table 3: Precision, recall, and $F1$-measure of adaptive conditional probability-based predicting model (in bold, best performance).

	Precision	Recall	$F1$-measure
EXT	0.93	0.76	0.83
NEU	0.86	0.67	0.75
AGR	**0.96**	0.74	0.83
CON	0.89	0.70	0.78
OPN	0.91	**0.79**	**0.84**
Mean	0.91	0.732	0.806

4.4. Analysis of Impacts of Different Factors. First, we carried on experiments with our proposed method. Scores of precision, recall, and $F1$-measure are shown in Table 3 including mean scores of all personality traits.

4.4.1. Impacts of Different Categories of Feature Sets. In this section, we conducted experiments on social network features, linguistic features, emotional statistical features, topic features, and all features, respectively. Due to space restrictions, Figure 3 only illustrates $F1$-measure of all personality traits along with mean $F1$-measure of them.

It can be observed that experiments which are conducted on linguistic features, emotional statistical features, and topic features result in unsatisfactory performance with respect to social network features, and social network features have the best classification performance for extraversion which is consistent with the conclusion in [33]. Since social network features reflect a user's pattern of social status and habits, which are relatively important decisive factors in user's personality traits, as a result, they provide more information than other types of features. Nevertheless, whether more or less, each kind of characteristic provides unique information in user's personality traits prediction. Therefore, carrying on experiment with combination of social network features, linguistic features, emotional statistical features, and topic features can achieve better performance.

4.4.2. Impact of Weighted Features. Additionally, we conducted experiments on unweighted features. Here, we only presented the most successful results in Table 4, including mean scores of all personality traits.

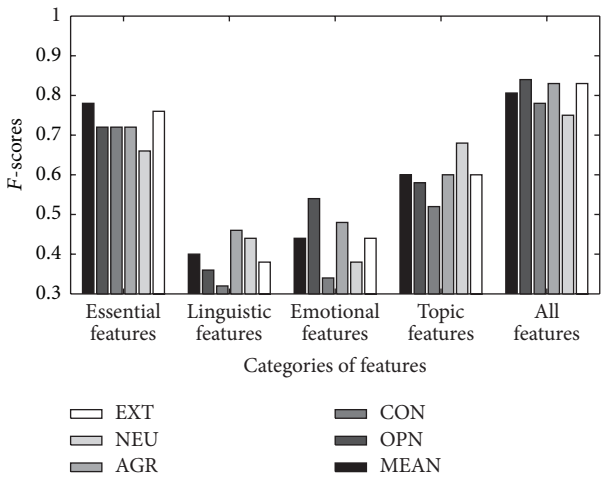

Figure 3: $F1$-measures of five personality traits along with mean $F1$-measures of them with different categories of feature sets.

Table 4: Precision, recall, and $F1$-measure of predicting model with unweighted features (in bold, best performance).

	Precision	Recall	$F1$-measure
EXT	0.85	0.76	0.82
NEU	0.92	0.68	0.76
AGR	**0.94**	0.64	0.76
CON	0.91	0.57	0.68
OPN	0.87	**0.78**	**0.83**
Mean	0.90	0.686	0.77

Table 5: Precision, recall, and $F1$-measure of predicting model with unified thresholds (in bold, best performance).

	Precision	Recall	$F1$-measure
EXT	0.62	0.54	0.57
NEU	**0.80**	0.16	0.25
AGR	0.57	0.73	0.63
CON	0.57	0.77	0.60
OPN	0.70	**1.00**	**0.82**
Mean	0.652	0.64	0.574

From Tables 3 and 4, we can observe that the best results for myPersonality dataset are achieved after distributing weights to features. It illustrates that considering correlations between features and user's personality traits can bring performance gain to the prediction task since there are limited features that may remain useful for user's personality traits prediction.

4.4.3. Impact of Dynamic Updated Thresholds of Personality Traits. In addition, we predicted user's personality traits with unified thresholds as well. The results are summarized in Table 5.

Multilabel learning task is to predict one or more categories for each instance. In the existing algorithms, such as PT5 method proposed by Tsoumakas and Katakis [34], generally, thresholds were unified to the same value. However, all categories employing a unified minimum threshold are not appropriate. If threshold is set too high, not all categories tags will be predicted; on the other hand, if threshold is set too low, too many classes will be predicted in results. Hence, in this paper, we updated thresholds of personality traits dynamically according to error rates of them. Compared to Table 3, based on thresholds obtained through updating dynamically, our method can achieve better results.

Table 6: Precision of different methods on Facebook dataset. The best performance per personality trait appears boldfaced.

Related works	Methods	EXT	NEU	AGR	CON	OPN	Mean
Our work	CP	**0.93**	**0.86**	**0.96**	**0.89**	**0.91**	**0.910**
Verhoeven et al. [42]	SVM	0.79	0.71	0.67	0.72	0.87	0.752
Farnadi et al. [33]	SVM, kNN, NB	0.58	0.54	0.50	0.55	0.60	0.554
Alam et al. [43]	SVM, BLR, mNB	0.58	0.59	0.59	0.59	0.60	0.590
Tomlinson et al. [44]	LR	NA	NA	NA	NA	NA	NA

4.5. Performance Evaluation with Different Methods. As mentioned in Section 4.1, in the Workshop on Computational Personality Recognition (Shared Task), different systems for personality recognition from text and network features on a common benchmark have been compared. Verhoeven et al. [42] proposed a meta-learning approach which can be extended to certain component classifiers from other genres with other class systems or even from other languages. Farnadi et al. [33] leveraged SVM, k nearest neighbors (kNN) [17], and Naïve Bayes (NB) for automatic recognition of personality traits from users' Facebook statues updates, respectively; even with a small set of training dataset, it could achieve better results than most baseline algorithms. On the basis of a set of features extracted from Facebook dataset, Alam et al. [43] explored suitability and performance of several classification techniques, which were SVM, Bayesian Logistic Regression (BLR) [47], and Multinomial Naïve Bayes (mNB) [48]. Tomlinson et al. [44] used ranking algorithms for feature selection and Logistic Regression (LR) [49] as learning algorithms, which achieved a high performance. In this paper, we tested our proposed method with the same set than the participants in the Workshop on Computational Personality Recognition (Shared Task). In terms of precision, recall, and $F1$-measure, a comparison of the works described above, as well as our proposed method (denoted by CP), is summarized in Tables 6, 7, and 8.

From Tables 6, 7, and 8, the experimental results indicate that openness is the easiest personality trait to be predicted by Facebook features with the above methods because there are more users with positive level of openness than other personality traits in Facebook dataset. Since Verhoeven et al. [42] trained an ensemble SVM model based on both Facebook and Essays personality traits dataset, as a consequence,

Table 7: Recall of different methods on Facebook dataset. The best performance per personality trait appears boldfaced.

Related works	Methods	EXT	NEU	AGR	CON	OPN	Mean
Our work	CP	0.76	0.67	**0.74**	0.70	0.79	0.732
Verhoeven et al. [42]	SVM	**0.79**	**0.72**	0.68	**0.72**	**0.87**	**0.756**
Farnadi et al. [33]	SVM, kNN, NB	0.61	0.53	0.50	0.54	0.70	0.576
Alam et al. [43]	SVM, BLR, mNB	0.58	0.58	0.59	0.59	0.60	0.588
Tomlinson et al. [44]	LR	NA	NA	NA	NA	NA	NA

Table 8: $F1$-measure of different methods on Facebook dataset. The best performance per personality trait appears boldfaced.

Related works	Methods	EXT	NEU	AGR	CON	OPN	Mean
Our work	CP	**0.83**	**0.75**	**0.83**	**0.78**	0.84	**0.806**
Verhoeven et al. [42]	SVM	0.79	0.70	0.67	0.72	**0.86**	0.748
Farnadi et al. [33]	SVM, kNN, NB	0.56	0.52	0.50	0.54	0.61	0.546
Alam et al. [43]	SVM, BLR, mNB	0.58	0.58	0.58	0.59	0.60	0.586
Tomlinson et al. [44]	LR	NA	NA	NA	NA	NA	0.630

when tested on the same Facebook dataset, it can achieve better results than our proposed method. Furthermore, when trained and tested on the same Facebook dataset, our method outperforms the methods in [33, 43, 44] which do not take weighted features, dynamic updated thresholds of personality traits, and correlations between personality traits into consideration.

5. Conclusion

In this paper, we studied the problem of exploiting interdependencies between user's personality traits for predicting user's personality traits. First, after analyzing importance of features, we conducted experiments on Facebook dataset to demonstrate the existence of correlations between user's personality traits. Bearing this in mind, an unsupervised framework, adaptive conditional probability-based predicting model, was then proposed to predict user's Big Five personality traits based on importance of features, dynamic updated thresholds of personality traits, and prior knowledge about correlations between personality traits. Furthermore, we compared our results with the ones achieved by others in the Workshop on Computational Personality Recognition (Shared Task) on the same dataset. In general, the experimental results demonstrated the effectiveness of our proposed framework.

In future work, we will speculate on what directions can be undertaken to ameliorate its performance with respect to time complexity so as to better apply it to a big data environment as in Facebook monitoring. Besides, in order to make our framework applicable to dynamic networks better, we will explore combining time series analysis with personality traits, predicting algorithm to capture dynamic evolution process of information and network structure. Furthermore, since social network features (described in Section 3.1.1) are specific for our dataset, they may be difficult (or even impossible) to obtain in another dataset; hence, we will conduct experiments on other datasets, such as the Tweeter corpus of PAN shared task (http://pan.webis.de/) on author profiling where also personality is considered, to further validate the effectiveness of the proposed framework based on different features.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant no. 61300148; the Scientific and Technological Break-Through Program of Jilin Province under Grant no. 20130206051GX; the Science and Technology Development Program of Jilin Province under Grant no. 20130522112JH; the Science Foundation for China Postdoctor under Grant no. 2012M510879; and the Basic Scientific Research Foundation for the Interdisciplinary Research and Innovation Project of Jilin University under Grant no. 201103129.

References

[1] M. Oussalah, F. Bhat, K. Challis, and T. Schnier, "A software architecture for Twitter collection, search and geolocation services," *Knowledge-Based Systems*, vol. 37, pp. 105–120, 2013.

[2] R. R. McCrae and P. T. Costa, "Validation of the 5-factor model of personality across instruments and observers," *Journal of Personality and Social Psychology*, vol. 52, no. 1, pp. 81–90, 1987.

[3] G. Saucier and L. R. Goldberg, "The structure of personality attributes," in *Personality and Work: Reconsidering the Role of Personality in Organizations*, M. R. Barrick and A. M. Ryan, Eds., pp. 1–29, Jossey-Bass, 2003.

[4] B. Caci, M. Cardaci, M. E. Tabacchi, and F. Scrima, "Personality variables as predictors of facebook usage," *Psychological Reports*, vol. 114, no. 2, pp. 528–539, 2014.

[5] M. Kosinski, Y. Bachrach, P. Kohli, D. Stillwell, and T. Graepel, "Manifestations of user personality in website choice and behaviour on online social networks," *Machine Learning*, vol. 95, no. 3, pp. 357–380, 2014.

[6] M.-G. Cojocaru, H. Thille, E. Thommes, D. Nelson, and S. Greenhalgh, "Social influence and dynamic demand for new products," *Environmental Modelling & Software*, vol. 50, pp. 169–185, 2013.

[7] F. Durupinar, N. Pelechano, J. Allbeck, U. Gudukbay, and N. Badler, "The impact of the OCEAN personality model on the perception of crowds," *IEEE Computer Graphics and Applications*, vol. 31, no. 3, pp. 22–31, 2011.

[8] X. Feng, "Study on customer personality characteristics and relationship outcomes using SEM analysis," in *Proceedings of the International Conference on Mechatronics and Information Technology*, pp. 841–844, 2013.

[9] J. G. Lee, S. Moon, and K. Salamatian, "Modeling and predicting the popularity of online contents with Cox proportional hazard regression model," *Neurocomputing*, vol. 76, no. 1, pp. 134–145, 2012.

[10] J. W. Stoughton, L. F. Thompson, and A. W. Meade, "Big five personality traits reflected in job applicants' social media postings," *Cyberpsychology, Behavior, and Social Networking*, vol. 16, no. 11, pp. 800–805, 2013.

[11] D. A. Cobb-Clark and S. Schurer, "The stability of big-five personality traits," *Economics Letters*, vol. 115, no. 1, pp. 11–15, 2012.

[12] W. Mischel, "Toward an integrative science of the person," *Annual Review of Psychology*, vol. 55, pp. 1–22, 2004.

[13] S. Argamon, S. Dhawle, M. Koppel, and J. Pennebaker, "Lexical predictors 19 of personality type," in *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*, 2005.

[14] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of Artificial Intelligence Research*, vol. 30, pp. 457–500, 2007.

[15] K. Moore and J. C. McElroy, "The influence of personality on Facebook usage, wall postings, and regret," *Computers in Human Behavior*, vol. 28, no. 1, pp. 267–274, 2012.

[16] J. W. Pennebaker and L. A. King, "Linguistic styles: language use as an individual difference," *Journal of Personality and Social Psychology*, vol. 77, no. 6, pp. 1296–1312, 1999.

[17] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, The Morgan Kaufmann Series in Data Management Systems, edited by J. Gray, Morgan Kaufmann, Boston, Mass, USA, 3rd edition, 2011.

[18] J. Golbeck, C. Robles, and K. Turner, "Predicting personality with social media," in *Proceedings of the 29th Annual CHI Conference on Human Factors in Computing Systems (CHI '11)*, pp. 253–262, ACM, May 2011.

[19] J. R. Quinlan, "Learning with continuous classes," in *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence (Ai '92)*, pp. 343–348, Hobart, Australia, November 1992.

[20] S. Roweis and Z. Ghahramani, "A unifying review of linear gaussian models," *Neural Computation*, vol. 11, no. 2, pp. 305–345, 1999.

[21] S. Bai, T. Zhu, and L. Cheng, "Big-five personality prediction based on user behaviors at social network sites," http://arxiv.org/pdf/1204.4809v1.pdf.

[22] J. Oberlander and S. Nowson, "Whose thumb is it anyway? classifying author personality from weblog text," in *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*, pp. 627–634, 2006.

[23] T. Nguyen, D. Q. Phung, B. Adams, and S. Venkatesh, "Towards discovery of influence and personality traits through social link prediction," in *Proceedings of the International Conference on Weblogs and Social Media*, pp. 566–569, Barcelona, Spain, July 2011.

[24] S. T. Bai, B. B. Hao, A. Li, S. Yuan, R. Gao, and T. S. Zhu, "Predicting big five personality traits of microblog users," in *Proceedings of the 12th IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technolog (WI '13)*, pp. 501–508, November 2013.

[25] R. Sun and N. Wilson, "A model of personality should be a cognitive architecture itself," *Cognitive Systems Research*, vol. 29-30, no. 1, pp. 1–30, 2014.

[26] A. Ortigosa, R. M. Carro, and J. I. Quiroga, "Predicting user personality by mining social interactions in Facebook," *Journal of Computer and System Sciences*, vol. 80, no. 1, pp. 57–71, 2014.

[27] O. P. John, L. P. Naumann, and C. J. Soto, "Paradigm shift to the integrative Big Five trait taxonomy: history, measurement, and conceptualissues," in *Handbook of Personality: Theory and Research*, O. P. John, R. W. Robins, and L. A. Pervin, Eds., pp. 114–158, Guilford Press, New York, NY, USA, 3rd edition, 2008.

[28] G. Park, H. A. Schwartz, J. C. Eichstaedt et al., "Automatic personality assessment through social media language," *Journal of Personality and Social Psychology*, vol. 108, no. 6, pp. 934–952, 2015.

[29] C. J. Soto, O. P. John, S. D. Gosling, and J. Potter, "Age differences in personality traits from 10 to 65: big five domains and facets in a large cross-sectional sample," *Journal of Personality and Social Psychology*, vol. 100, no. 2, pp. 330–348, 2011.

[30] C. J. Soto and O. P. John, "Ten facet scales for the Big Five Inventory: convergence with NEO PI-R facets, self-peer agreement, and discriminant validity," *Journal of Research in Personality*, vol. 43, no. 1, pp. 84–90, 2009.

[31] M. V. Kosti, R. Feldt, and L. Angelis, "Personality, emotional intelligence and work preferences in software engineering: an empirical study," *Information and Software Technology*, vol. 56, no. 8, pp. 973–990, 2014.

[32] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern et al., "Personality, gender, and age in the language of social media: the open-vocabulary approach," *PLoS ONE*, vol. 8, no. 9, Article ID e73791, 2013.

[33] G. Farnadi, S. Zoghbi, M. F. Moens, and M. D. Cock, "Recognising personality traits using facebook status updates," in *Proceedings of the Workshop on Computational Personality Recognition*, July 2013, http://clic.cimec.unitn.it/fabio/wcpr13/farnadi_wcpr-13.pdf.

[34] G. Tsoumakas and I. Katakis, "Multi-label classification: an overview," *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, pp. 1–13, 2007.

[35] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 15, pp. 5802–5805, 2013.

[36] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[37] R. Socher, J. Bauer, C. D. Manning, and Y. N. Andrew, "Parsing with compositional vector grammars," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL '13)*, pp. 455–465, Sofia, Bulgaria, August 2013.

[38] K. Kazama, M. Imada, and K. Kashiwagi, "Characteristics of information diffusion in blogs, in relation to information source type," *Neurocomputing*, vol. 76, no. 1, pp. 84–92, 2012.

[39] M. B. David, Y. N. Andrew, and I. J. Michael, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993–1022, 2003.

[40] P. T. Costa and R. R. McCrae, *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional Manual*, Psychological Assessment Resources, 1992.

[41] R. L. Atkinson, C. A. Richard, E. S. Edward, J. B. Daryl, and S. Nolen-Hoeksema, *Hilgard's Introduction to Psychology*,

Harcourt College Publishers, Orlando, Fla, USA, 13th edition, 2000.

[42] B. Verhoeven, W. Daelemans, and T. D. Smedt, "Ensemble methods for personality recognition," in *Proceedings of the Workshop on Computational Personality Recognition*, pp. 35–38, Boston, Mass, USA, July 2013.

[43] F. Alam, A. Stepanov, and G. Riccardi, "Personality traits recognition on social network—facebook," in *Proceedings of the Workshop on Computational Personality Recognition*, pp. 6–9, Boston, Mass, USA, July 2013.

[44] M. T. Tomlinson, D. Hinote, and D. B. Bracewell, "Predicting conscientiousness through semantic analysis of facebook posts," in *Proceedings of the Workshop on Computational Personality Recognition*, pp. 31–34, July 2013.

[45] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1858–1860, 2003.

[46] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[47] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale Bayesian logistic regression for text categorization," *Technometrics*, vol. 49, no. 3, pp. 291–304, 2007.

[48] A. K. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*, pp. 137–142, Madison, Wis, USA, July 1998.

[49] I. Ruczinski, C. Kooperberg, and M. LeBlanc, "Logic regression," *Journal of Computational and Graphical Statistics*, vol. 12, no. 3, pp. 475–511, 2003.