Computer Communications 73 (2016) 3–11

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Predicting the attributes of social network users using a graph-based machine learning method

Yuxin Ding[a,*], Shengli Yan[a], YiBin Zhang[a], Wei Dai[a], Li Dong[b]

[a] Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, 518055, China
[b] State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences, China

ARTICLE INFO

Article history:
Received 5 May 2014
Revised 8 April 2015
Accepted 5 July 2015
Available online 18 July 2015

Keywords:
Social network analysis
Data mining
Social network privacy
Semi-supervised learning
Information inference

ABSTRACT

Attribute information from social network users can be used as a basis for grouping users, sharing content, and recommending friends. However, in practice, not all users provide their attributes. In this paper, we try to use information from both the graph structure of the social network and the known attributes of users to predict the unknown attributes of users. Considering the topological structure of a social network and the characteristics of users' data, we select a graph-based semi-supervised learning algorithm to predict users' attributes. We design different strategies for computing the relational weights between users. The experimental results on real-world data from Renren demonstrate that the semi-supervised learning method is more suitable for predicting users' attributes compared with the supervised learning models, and our strategies for computing the relational weights between users are effective. We also analyze the effect of different social relations on predicting users' attributes.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Now online social networking has become one of the most popular ways for people to interact with their friends, to grow a business and to get information. By some estimates, up to 67% of individuals aged 18–29 use social networking at some level [1]. About 22% of all people use social network sites for some reason [1]. Today social networks such as Facebook and Twitter are driving new forms of social interaction, dialogue, exchange and collaboration.

Typically, users in a social network post their own attributes on their pages, such as their universities, majors, and hobbies. This profile information can be used as a basis for grouping users, sharing content, and recommending friends. However, in practice, not all users in a social network provide their profiles. Moreover, to protect the privacy of users, some social networks allow users to hide their personal information. One interesting solution is that we can infer the hidden (unknown) attributes of users by using the known information supplied by the social network. In this paper, we try to use information from both the topological structure of the social network and the known attributes of users to predict the hidden attributes of users.

Inferring the hidden attributes of users in an online social network [2] has been a hot issue in machine learning. Although the users can hide some profile information, their friendship information and group information can be obtained directly or indirectly. For example, we can obtain a user's friendship information by searching his/her friend list, which is available in a social network, and his/her friends' attributes can also be easily obtained by accessing the links to his/her friends. We can extract a large amount of data from a social network; however, only a small part of the data is assigned attribute values. Fig. 1 shows the percentage of users who hide their attribute information in the Renren social network. Nearly 55% of users hide part of their profiles. Nearly 70% of users hide their hobbies. This means that the labeled data are considerably fewer than the unlabeled data. In general, the traditional supervised prediction models are required to be trained using a large amount of labeled data (training data) to obtain a high accuracy. Therefore, the challenge is how we can take advantage of the unlabeled data to improve the prediction accuracy.

Compared with the supervised learning algorithms, the semi-supervised learning algorithms [3–5] can use both labeled and unlabeled data to train a prediction model. The semi-supervised learning algorithms can also achieve better performance, especially when the labeled data are very limited. Therefore, the semi-supervised algorithms are more suitable for inferring the users' profiles in a social network. In this paper, our goal is to infer the users' hidden attributes by using the users' personal information, friendship information, and group information. We choose the graph-based semi-supervised learning algorithm to predict a user's hidden attributes. To improve the prediction accuracy, we propose different strategies for computing the relational weights between users according to the users' data type (labeled or unlabeled).

* Corresponding author. Tel.: +86 755 2603 2193; fax: +86 755 2603 2461.
E-mail address: yxding@hitsz.edu.cn (Y. Ding).

http://dx.doi.org/10.1016/j.comcom.2015.07.007
0140-3664/© 2015 Elsevier B.V. All rights reserved.

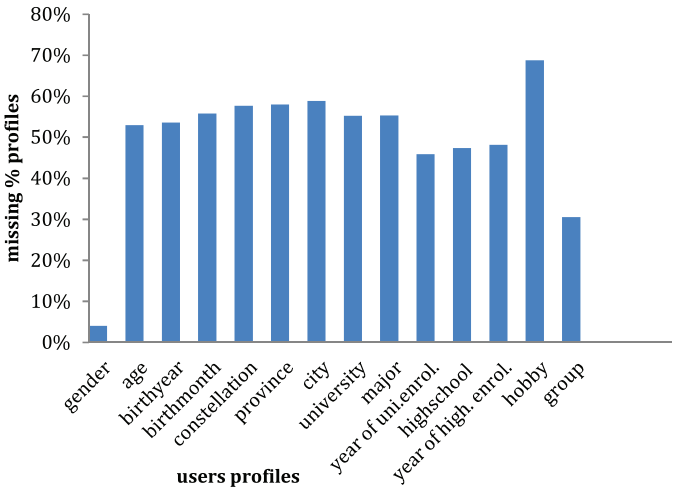

Fig. 1. The percentage of users who hide their attribute information in the Renren social network.

The remaining part of this paper is organized as follows. Section 2 gives an introduction of the related work. Section 3 gives the definition of the problem. In Section 4 we discuss how to use semi-supervised method to predict the users' attributes. The experimental results are shown in Section 5. The conclusion and future work are discussed in the last section.

2. Related work

To protect privacy information, many social network users provide only partial information or deliberately hide some of their personal information. Can failing to complete or hiding personal information protect the privacy of users? The answer is no. Fond and Neville [6] proposed the principle of social influence, which states that users who are linked are likely to adopt similar attributes, and suggest that the network structure should inform attribute inference. Other studies [7, 8] show that users with similar attributes are likely to link to one another, motivating the use of attribute information for link prediction. In our study, we mainly focus on attribute inference.

Rao et al. [9, 10] inferred user attributes (e.g., ethnicity and political orientation) using supervised learning methods with features extracted from user names and user-generated text. Strictly speaking, their study belongs to text classification. They did not consider the topological information and social relationships of a social network.

Yin et al. [11, 12] proposed an attribute-augmented social network model, which is called the social-attribute network (SAN). In the model, attributes are represented as a network node. Link information and attribute link information are extracted to infer user attributes. In their work, an unsupervised method (random walk with restart) was used to infer user attributes. Gong et al. [13, 14] also used SAN models to infer user attributes. They extracted a set of topological features for each positive and negative attribute link. The positive attribute links are taken as positive examples, while the negative attribute links are taken as negative examples. They trained a binary classifier (support vector machine, SVM) to infer the missing attributes.

In addition to topological information, social relations can be used to infer user attributes. In [15], the authors discussed how to infer user profiles by using friendship relations. They used the Bayesian network to model the cause and effect relationships among people in a social network and analyzed the effect of the prior probability, influence strength, and society openness to the inference accuracy on the online weblog service provider "live journal social network". In their work, they did not consider the social relations when inferring user attributes. In [16], the authors noted that different social relations such as the relationship between students and staff relations, suggesting the properties of some users, can be used to infer different attributes. For example, the relations between classmates can be used to infer users' ages. In [17], the authors used the naive Bayesian classifier to infer users' attributes in online social networks. They predicted the political affiliation of each user by using the users' personal information and friendship information. In [2], the authors noted the importance of group information. They predict users' attributes by using the available public group information and friendship information. Experimental results show that a public group contains a large amount of potential information for predicting user's attributes.

In [2,15–17], the authors discussed how to use different types of information available in a social network to infer the user's private information. The information they considered included personal information, friendship relations, and group relations. In [18], the authors proposed that users that have the same attributes are more likely to become friends or to form a dense community, and they employed a community-based method to infer users' attributes in an online social network. In [19], the authors inferred users' demographics based on their daily mobile communication patterns. They extracted the individual features, friend features and circle features of users and used these features to infer users' age and gender.

Most of the above mentioned papers employed supervised learning methods to infer users' attributes. However, in practice, the online social network users often supply only a small amount of information. Therefore, we want a learning algorithm that can use not only the labeled data but also large amount of unlabeled data to train classifiers. The semi-supervised learning method can learn from both labeled data and unlabeled data. In [20], the authors used a semi-supervised learning algorithm to infer users' profiles in a social network. They proposed a general semi-supervised learning framework and built two prediction models, the graph-based model and co-training model, to infer users' attributes. In contrast to the supervised learning algorithms, the semi-supervised learning algorithm predicts the hidden attributes of social network users more accurately. In this paper, we also use a graph-based semi-supervised learning algorithm to infer users' hidden attributes. We focus on the following problems:

(1) The principle of social influence [6] is the base assumption for inferring user's attributes. Presently, researchers proposed different algorithms to predict the attributes of social network users. However, most of them did not interpret if the data they used to infer attributes supports the base assumption. Moreover, what types of attributes can be derived more accurately? These problems are the main motivation of our study.
(2) Reference [20] uses a semi-supervised learning algorithm to predict the attributes of social network users. The semi-supervised leaning algorithm can learn from both labeled data and unlabeled data. However, the potential information supplied by the two types of data is different. How can we use the different potential information to improve the performance of the semi-supervised learning algorithm?
(3) In our work, we use different social relations, such as a group relation and friendship relation to calculate the similarity of two users. Different from [20], we try different combinations of social relations to calculate user similarity and evaluate their effect on the accuracy of the semi-supervised learning algorithm.

3. Problem definition

In general, a social network can be represented as a graph. A user can be viewed as a node in the graph, and the edge can be viewed as the relationship between users. The weight of the edge can be used to measure the similarity of two users. Fig. 2 shows a simple network of

Download English Version:

https://daneshyari.com/en/article/448453

Download Persian Version:

https://daneshyari.com/article/448453

Daneshyari.com