User Expertise Inference on Twitter: Learning from Multiple Types of User Data

Yu Xu[1], Dong Zhou[2], Séamus Lawless[1]
[1]ADAPT Centre, Knowledge and Data Engineering Group
School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
[2]School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, China
xuyu@scss.tcd.ie, dongzhou1979@hotmail.com, seamus.lawless@scss.tcd.ie

ABSTRACT

This paper proposes a learning model that tries to infer a user's topical expertise using multiple types of user-related data from Twitter such as tweets posted by the user and the characteristics of their followers. It considers inference consistency of different types of user data in the process of learning and aims to deliver accurate and effective inference results, even in cases where some types of data are missing for a user, e.g. the user has yet to post any tweets. Experiments conducted on a large-scale Twitter dataset show that our model outperforms several baseline approaches which use only a single type of user data for inference.

KEYWORDS

User expertise, inference model, Twitter

1 INTRODUCTION

Understanding the expertise of users in social networking sites like Twitter is a key component for many applications such as user recommendation and talent seeking. Previous studies [1, 2] observed that certain user actions on Twitter could reflect that user's expertise, so they attempted to infer a user's expertise information by exploiting selected types of user-related data. For example, the short bio information provided by a user was used to identify topic experts on the "who to follow" service of Twitter [1]; In [2], the authors proposed a learning model that uses an individual's tweets to infer their expertise on various topics.

However, previous studies have tended to focus on the exploitation of a single type of user data and the potential relation between this data and the user's expertise information. Although shown to be effective in inferring a user's expertise information, these approaches ignore the fact that many Twitter users may not have a certain type of data. For example, on Twitter it is reported that approx. 44% of all registered users have never posted a tweet [3]; Statistical analysis from about 10% of the entire Twitter population shows that on average, each user is included in less than one Twitter list [4]. Therefore, approaches that rely on a single type of user data will fail when the user does not have a significant volume, or any, of this data available.

To address this issue, this work proposes a learning-based model that tries to infer a user's expertise information by jointly exploiting multiple types of data associated with the user on Twitter. The model takes multiple types of user-related data as input and considers their inference consistency in the process of learning. It aims to make the most of various data associated with the user and ensure the inference effectiveness regardless of the availability of some types of user data. Four types of user data are considered in the experiments, namely: tweets, friends, followers and lists. Experimental results demonstrate our proposed model, which combines all the four different types of user data, outperforms the alternative inference methods.

2 METHOD

This section details how we utilize multiple types of user data from Twitter to better model the problem of user expertise inference. Xu et al. [2] proposed a sentiment-weighted and topic relation-regularized learning (SeTRL) model to address this problem. The SeTRL first builds the feature vector of a user based on the user's tweets and utilizes the sentiment intensity contained in the tweets to weight the features of each user. Then by using linear regression, a base model is built to jointly learn the expertise of users on multiple topics. Meanwhile, SeTRL exploits the relatedness between expertise topics to optimize inference, which is characterized by an undirected graph G with E edges. It encodes this relatedness information in the base model through model regularization. Finally, the SeTRL is constructed by solving the following minimization problem:

$$\min_{\boldsymbol{W}} \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{1}{2} (y_{ti} - \boldsymbol{x_i} \boldsymbol{w_t})^2 + \alpha \sum_{e=1}^{E} ||\boldsymbol{w_{e(1)}} - \boldsymbol{w_{e(2)}}||_2^2 + \beta ||\boldsymbol{W}||_1 \quad (1)$$

where N is the number of Twitter users; T is the number of expertise topics; y_{ti} is a binary value {+1, -1} which denotes the expertise of user i on topic t; $\boldsymbol{x_i}$ is the feature vector of user i; $\boldsymbol{w_t}$ is the model parameter vector for topic t and $\boldsymbol{W} = [\boldsymbol{w_1}, \boldsymbol{w_2}, ..., \boldsymbol{w_T}]$ is the parameter matrix for all the T topics; the second term is the regularizer used to incorporate the relatedness information between expertise topics; e is an edge in G that connects two re-

UMAP '17, July 09-12, 2017, Bratislava, Slovakia

ACM ISBN 978-1-4503-4635-1/17/07.
http://dx.doi.org/10.1145/3079628.3079646

lated topics, and $\boldsymbol{w}_{e(1)}$ is the model parameters of a topic of e; $\|\boldsymbol{W}\|_1$ is the l_1 norm of matrix $\boldsymbol{W}$; α, β are the regularization parameters.

However, the SeTRL will struggle when a user has not posted sufficient tweets. In this research, we propose incorporating multiple types of data associated with the user into the process of user expertise inference through the loss function. Meanwhile, we use a regularization term to model the inference consistency among different types of user data. Thus, we can construct our learning model and have the following optimization problem:

$$\min_{\boldsymbol{W}} \sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{2}\left(y_{ti} - \sum_{s=1}^{S}\frac{1}{|S|}\boldsymbol{x}_{\boldsymbol{si}}\boldsymbol{w}_{\boldsymbol{st}}\right)^2 + \gamma\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s_1=1}^{S}\sum_{s_2\neq s_1}^{S}||\boldsymbol{x}_{\boldsymbol{s_1 i}}\boldsymbol{w}_{\boldsymbol{s_1 t}} - \boldsymbol{x}_{\boldsymbol{s_2 i}}\boldsymbol{w}_{\boldsymbol{s_2 t}}||^2 + \alpha\sum_{e=1}^{E}||\boldsymbol{w}_{\boldsymbol{e(1)}} - \boldsymbol{w}_{\boldsymbol{e(2)}}||_2^2 + \beta\big|\|\boldsymbol{W}\|\big|_1 \quad (2)$$

where S is multiple types of user data and |S| is the total number of data types considered in the model (|S|=4 in this work); $\boldsymbol{x}_{\boldsymbol{si}}$ is the feature vector of user i defined from the s^{th} type of user data; $\boldsymbol{w}_{\boldsymbol{st}}$ is the model parameters of the s^{th} data source part of expertise topic t; the second term is the regularizer used to model the inference consistency among different types of user data; s_1 and s_2 are any two different types of user data from S; γ is the regularization parameter to control the contribution of the inference consistency of different types of user data.

3 EVALUATION

In this work, we reused the dataset from [2] to evaluate our proposed model. The dataset contains 10,856 Twitter users and 149 expertise topics from which each user has knowledge of at least one topic. While in [2] they only used the user's tweets for expertise inference, this research needs to use other types of data of the user, i.e. friends, followers and list data. So we harvested this additional data of each user in the dataset, if it was available, using the official Twitter API. In the experiment, the text information in the profiles of all friends (or followers) of a user are combined as an input document (called *friend document* or *follower document*) for inferring the user's expertise. In terms of the list data, the name and description information of the lists of the user are combined as the input document (*list document*). Correspondingly, the combination of the user posted tweets is called the *tweet document* of the user. The unigram features are used as the user feature space.

The two metrics: accuracy, and F1-score are used to measure the performance of methods in the work. Specifically, we use the averaged score of each of the two measurements on all the tested topics to examine the performance of various inference methods. In the experiments, a standard 5-fold cross validation on the training data is performed to select the regularization parameters α, β and γ. In addition, the frequency of the feature terms occurring in the user document is used to weight the user unigram features for friend, follower and list documents. In terms of the tweet document of a user, the tweet sentiment-based weighting scheme proposed in [2] is applied.

Table 1: Performance of different methods (%)

Methods	Data Used	Accuracy	F1
SVM	Friends	72.23	70.77
	Followers	69.37	66.08
	Lists	70.18	68.29
	All	67.67	62.64
SeTRL	Tweets	79.65	80.08
Our Model	All	**85.72**	**86.80**

We compared the performance of our proposed model with the two baseline approaches: Support Vector Machine (SVM) and SeTRL. The three approaches are based on either only a single type of user data or the combination of multiple types of user data, as shown in Table 1. Note that for SVM with all the four types of user data, the features generated from the four types of user data are directly concatenated as a single feature vector. The experimental results show that each type of user data is useful for user expertise inference and using friend data or list data can achieve better performance than using follower data for user expertise inference. It also shows that our model using the four types of user data for expertise inference significantly outperforms the baseline approaches. In particular, the SVM approach using all the four types of user data achieves the worst performance, which is even lower than that of SVM using one type of user data alone. This could be due to the over-fitting problem, as too much inconsistent information is considered during the learning process. It verifies the significance of taking into consideration the source consistency when using multiple types of user data for expertise inference.

4 CONCLUSIONS

This paper studies the problem of inferring a user's expertise based on various data associated with the user on Twitter. A learning model is proposed that can infer the user's topical expertise under the influence of multiple types of user data. Experiments on a real-world Twitter dataset show that our model using multiple types of user data for expertise inference outperforms several baseline approaches that are based on a single type of user data.

ACKNOWLEDGMENTS

This research is supported by the ADAPT Centre for Digital Content Technology, which is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. The work is also supported by the Scientific Research Fund of Hunan Provincial Education Department of China under Project No. 16K030, and Hunan Provincial Natural Science Foundation of China under Project No. 2017JJ2101.

REFERENCES

[1] Twitter Improves "Who To Follow" Results & Gains Advanced Search Page. http://selnd.com/wtfdesc.

[2] Xu, Y., Zhou, D., and Lawless S. 2016. Inferring Your Expertise from Twitter: Integrating Sentiment and Topic Relatedness. In *the proceedings of the 2016 IEEE/WIC/ACM International Conference on Web Intelligence*, IEEE, pp.121-128.

[3] Many Twitter users don't tweet, finds report. http://www.cbsnews.com/news/many-twitter-users-dont-tweet-finds-report/

[4] Kim, D., Jo, Y., Moon, I. C., and Oh, A. 2010. Analysis of twitter lists as a potential source for discovering latent chara-cteristics of users. In *ACM CHI workshop on microblogging.*