Inferring Your Expertise from Twitter: Combining Multiple Types of User Activity

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

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. A range of interactions between users on Twitter can provide important information that implicitly reflects a user's expertise. This paper proposes a learning model that tries to infer a user's topical expertise from Twitter using information such as tweets posted by the user and the characteristics of their followers. The model takes various types of user-related data from Twitter as input and considers their inference consistency in the process of learning. It 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. The experiments reported in the paper were conducted on a large-scale Twitter dataset. Experimental results show that our model outperforms several baseline approaches and outperforms approaches which use only a single type of user data for inference.

KEYWORDS

User expertise, inference model, Twitter

1. INTRODUCTION

Over the past decade, Social Networking Sites (SNSs) such as Twitter and Facebook have risen to prominence in society. People share experiences, catch up on the activities of their friends, or directly communicate with them through these platforms on a daily basis. Such platforms not only facilitate communication and exchange between people but also allow people to access a wide range of information resources. The expertise of fellow users on SNSs is one such resource. For example, many companies match their employees to tasks using social media tools [1]; people often ask specific questions through their social networks [2]. To support such actions and allow people to benefit accordingly, the key challenge is obtaining expertise information of SNS users [7, 8].

However, expertise information is usually not explicitly provided by SNS users, so existing methods primarily rely on implicit inference to obtain this information [3, 4]. In other words, they try to infer a user's expertise based on the user's past behavior on a SNS. This work focuses on the most popular micro-blogging site, Twitter, and aims to infer a user's topical expertise based on the various data associated with him/her.

On Twitter, a user can interact with other users through various activities. For example, a user can send a general tweet to express their thoughts on something; they can mention another user in a tweet; a user can "follow" other users to be updated on their recent activities; a user can "heart" or "retweet" another user's tweets. These activities expose a user to multiple types of data on Twitter. In this paper, different types of user data refer to, for example, tweets from a user, lists a user is part of and the list of accounts a user follows and is followed by. Previous studies [5, 6, 7, 8] 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 the user was used to identify topic experts on the "who to follower" service of Twitter [5]; Research [6] verified the existence of *homophily* among Twitter user following relationships, i.e. Twitter users tend to follow other users with common topical beliefs or interests. Therefore, the "following" relationships between users were exploited to discover influential users on different topics. While often "noisy", user's tweets provide direct evidence about his/her expertise background. In [7] the authors proposed a learning model that uses an individual's tweets to infer his expertise on various topics.

However, previous studies tend to focus on the exploitation of a certain 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, and most tweets are generated by a small proportion of the user population [9]; Statistical analysis from about 10% of the entire Twitter population shows that on average, each user is included in less than one Twitter list [10], which is used as prime evidence to infer the user's topics of expertise in [8]. Moreover, statistics from the Twitter dataset collected as part of this research show that 24.42% of Twitter users follow less than 100 people and 24.18% have less than 100 followers. 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, such as the user's posted tweets and the followers of the user. It

WI '17, August 23-26, 2017, Leipzig, Germany

ACM ISBN 978-1-4503-4951-2/17/08…$15.00
http://dx.doi.org/10.1145/3106426.3106468

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. Meanwhile, the model considers the consistency of different types of user data in the process of inference, which means that the expertise information reflected by different types of user data should be similar. Through regularization, the model tries to penalize the differences among the inference results from different types of user data. Experiments are conducted on a Twitter dataset with over 10,000 users and 149 expertise topics. Four types of data associated with the user are considered in the experiments, namely: tweets, friends, followers and lists. Experiments first demonstrate that using each type of user data alone can effectively infer a user's expertise but with varying effectiveness. Experimental results then show that our proposed model which combines all the different types of user data outperforms the alternative inference methods which use only one type of data or the combination of fewer types of user data.

In summary, the contributions of this research are twofold:

(1) This paper proposes a learning model that infers a user's topical expertise based on multiple types of data associated with him/her on Twitter. It can deliver effective inference once there are some types of user data available, and the model can also further improve the inference accuracy by making use of multiple types of user data if available.

(2) Experiments conducted on a real-world Twitter dataset demonstrate that each of the four types of user data (tweets, friends, followers and lists) is effective for user expertise inference, and show that our model which combines multiple types of user data outperforms a number of baseline approaches.

The rest of the paper is organized as follows. Section 2 discusses related work; Section 3 defines the problem of user expertise inference on Twitter and then details our proposed model to address this problem; Section 4 describes the construction of the experimentation dataset and model input features; Section 5 gives the experimental results and analysis; Section 6 concludes.

2. RELATED WORK

Identifying the areas of expertise of people within an enterprise has been a key research challenge for a long time, as this would enhance the enterprise's ability to effectively utilize their human resources. This is also the reason why for many years an expert finding task has been included in the TREC Enterprise Track, which provides a common platform for researchers to empirically assess expert identification methods [11]. Based on TREC or other enterprise-oriented test-beds [12, 13], numerous methods have been proposed to identify domain experts based on various theories [14, 15, 16 17]. For example, several works [14, 15] applied generative language models to model the relevance between a user and an expertise topic. They estimate the expertise level of a user with respect to a topic by looking at the probability of the user documents generating the topic terms. There are also a number of studies that utilized the discriminative probabilistic models to predict a user's expertise information in the enterprise setting [16, 17]. They directly model the mapping from inputs (mainly textual information related to the user) to outputs and estimate the model parameters by optimizing the objective loss functions. Beyond these probabilistic models, there is another group of works that attempt to model a user's expertise information by exploiting social connections among users [18, 19]. These approaches determine a user's expertise on a topic by analyzing the importance of his/her connections in a topic area, and the connection strength between them.

In recent years, social networking sites such as Facebook, Twitter and LinkedIn have been integrated into people's daily lives. People turn to those platforms not just for socializing but also for satisfying their various information needs such as asking questions and recruiting talent [2, 20]. As discussed in Section 1, in these application scenarios it is crucial to obtain the expertise information of SNS users in order to meet the user's information needs. This results in the demand for user expertise inference in different SNSs. For example, Wang et al. [21] proposed to predict a LinkedIn user's professional skills by analyzing the textual information in his/her public profile. A prediction model was built based on the assumption that people with similar professional and educational backgrounds tend to have common professional skills. To facilitate a user recommendation service in Instagram, Pal et al. [22] exploited the user's self-described interests and the "following" relationships among users to identify authoritative users on a given topic. They proposed an authority learning framework based on the hypothesis that an authority on a specific topic has a significantly higher proportion of followers interested in that topic. Popescu et al. [23] attempted to mine the potential domain expertise of Pinterest users. They defined a set of features that can reflect a user's expertise from their historical pinning activities and those features were then used to identify potential experts for four popular Pinterest topics.

Owing to the public availability of Twitter data, most studies in the literature use Twitter as the target research platform for the problem of user expertise inference in SNSs. These studies can be classified to two categories based on whether the user's geo-spatial information is considered in the process of inference. The first category of studies aims to find Twitter users who not only have expertise in a given topic but also be geographically close to a particular location. Li et al. [24] used the *point of interest* tags on tweets to form a possible categorization of geo-spatial expertise and investigated what potential factors would influence a person to judge a Twitter user's knowledge on a local expertise topic. It was found that the more frequently a user interacts with a topic (i.e. geo-location check-in activities), the more likely people would think this user has a good level of knowledge of that topic. Cheng et al. [25] proposed an expertise framework called LocalRank which determines a Twitter user's expertise in a topic using two key components: the user's topical expertise and his/her local authority. The work leveraged Twitter list data as the inference source and estimated the user's local authority by exploiting the geo-spatial information embedded in the Twitter lists. Most recently, Niu et al. [26] introduced a learning-based method to find local experts on Twitter. They defined multiple classes of features that could impact a user's local expertise, such as tweet content features (e.g. the TF-IDF score of a topic keyword in the candidate's tweets) and local authority features (e.g. the distance between the candidate and the query location). Given an expertise topic, the level of expertise of users is estimated by applying a learning-to-rank strategy.

The second category of studies targets the entire population on Twitter given an expertise need regardless of their location. As the user posted tweets provide direct evidence about the user's expertise information, it has often been used as the source of inference in previous studies. Bar-Haim et al. [27] attempted to locate stock experts on Twitter and aimed to assist investors in making trading decisions. The proposed inference method was based on the consistency between the user's prediction of stock prices in her/his tweets and the actual change of the stock prices. Pal et al. [4] studied the problem of authoritative user inference on three specific topics (i.e. iphone, oil spill, world cup). They took those Twitter users who explicitly mentioned a topic name in their tweets as candidates, and used a probabilistic clustering method to generate a list of the top authoritative users for a topic based on a number of predefined user features, e.g. the similarity of any two successive tweets, the number of keyword hashtags used. Most recently, Xu et al. [7] proposed a Sentiment-weighted and Topic Relation-regularized Learning (SeTRL) model to infer the topical expertise of Twitter users based on their posted tweets. This model was built based on two core assumptions which are: 1) if a person can forcefully and subjectively express their opinion on a topic, it is more likely that the person has strong knowledge of that topic; 2) if a person has knowledge of a topic, it is very likely that s/he also knows about other topics which are related to the topic. Thus, the SeTRL model uses the sentiment intensity contained in the tweets to weight the textual features defined from the user's tweets and exploits topic relatedness to regularize the learning process. Apart from the user posted tweets, there are several studies that attempt to make use of other types of Twitter data for expertise information inference. Weng et al. [6] employed the Latent Dirichlet Allocation (LDA) model [28] to generate the latent topics a Twitter user is interested in based on user posted tweets. Then a PageRank-like approach was proposed to estimate the influence of Twitter users on these latent topics, which considers both topical similarity between users and "following" connections. Ghosh et al. [8] proposed to utilize the list data of Twitter users to infer their expertise information. They observed that the list metadata, i.e. list name and list description, provides semantic cues to the expertise information of the users who are included in the list. Therefore, textual information from the list metadata associated with a user was refined to represent the topical expertise of Twitter users. Wagner et al. [3] compared the efficacy of different types of user-related Twitter data, such as tweets and list data, for inferring a user's topical expertise. They performed experiments on both specific expertise topics and latent topics using a logistical regression method and a user study. Experimental results showed that the list data achieved the best performance.

Our work belongs to the second category of studies that infers a Twitter user's topical expertise by using the user-related data without considering the user's location information. In contrast to previous studies that primarily rely on a certain type of user data for user expertise inference, our work tries to make full use of multiple types of user-related data and aims to ensure good prediction accuracy even when some types of user data are missing.

3. USER EXPERTISE INFERENCE FROM TWITTER DATA

This section formally defines the problem of user expertise inference based upon multiple types of user-related Twitter data and then details the construction of our model for addressing the problem.

3.1 Problem Formulation

In [7], the authors formally defined the problem of user expertise inference on Twitter, with the focus on the utilization of the user's tweets. In contrast, our work aims to better model the problem by making use of multiple types of user data, which includes the user's tweets but is not limited to that. In particular, this work considers four different types of user data for inferring a user's expertise: (1) Tweets: the textual content posted by the user; (2) Friends: other Twitter users that the user is following; (3) Followers: other Twitter users who are following this user; (4) Lists: Twitter lists that include this user.

Formally, the input is T expertise topics and S (S equals to 4 in this work) feature matrixes of N Twitter users: X_1, X_2, …, X_S where X_s is a N-by-K_s matrix defined from the social data associated with the user through the s^{th} relationship, i.e. the s^{th} type of user data on Twitter, and K_s is the number of features defined from this data source. An entry x_{sij} of X_s denotes the feature value of the j^{th} feature of user i from the s^{th} data source. The output of the question is a T-by-N label matrix Y which represents the expertise of users on the T topics. An entry y_{ti} of Y is a binary value {+1, -1} which denotes the expertise of user i on topic t. The object of this work is to learn a predictive model for every expertise topic that can effectively infer a user's expertise on the topic based on multiple types of data associated with the user on Twitter. Table 1 gives the definition of the notations used in this work.

Table 1: Definition of Notations

Notation	Notation Description
	Data
N	Number of users
T	Number of expertise topics
S	Number of types of user data
K_s	Number of user features for the s^{th} type of user data
$\boldsymbol{x}_{si}$	Feature vector of the i^{th} user from the s^{th} type of user data
$y_{ti} \in \{+1, -1\}$	Has or doesn't have expertise
G, E	Topic relation graph
	Model
$\boldsymbol{w}_t$	Model coefficients of topic t
$\boldsymbol{w}_{st}$	Model coefficients of topic t for source S
α, β, γ	Regularization parameters

3.2 User Expertise Inference Using Multiple Types Data on Twitter

This sub-section details how we utilize multiple types of user data on Twitter to better model the problem of user expertise inference. Xu et al. [7] 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:

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

where $\boldsymbol{w}_t$ is the model parameter vector for topic t and $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_T]$ 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 related 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 is heavily dependent on abundant content posted by the user in order to perform expertise inference. This model will struggle when a user has not posted sufficient tweets. As discussed in Section 1, in practice, a large proportion of Twitter users do not actively post tweets, or have never posted any tweets. In these cases, we need to seek other information for inferring a user's expertise. In addition to posting tweets, a Twitter user could also interact with other users through various other activities. These activities may also provide effective evidence about the user's expertise information. As explained in Section 2, [6] exploited the following relationships between users to identify topically influential Twitter users and believes that a user tends to follow other users with similar topical interests; the authors in [8] observed that list data on Twitter provides valuable semantic cues to the topics of expertise of the users on the list, so they proposed exploiting this data to mine topic expertise on Twitter. Additionally, different users on Twitter exhibit different behaviours and habits. Some users like posting tweets and there are also users who mainly just follow others and read content that interests them. Therefore, it is important to jointly exploit multiple types of user data for expertise inference. This will benefit cases where only certain types of user data are available. 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:

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{2}(y_{ti}-\sum_{s=1}^{S}\frac{1}{|S|}\boldsymbol{x_{si}w_{st}})^2 \tag{2}$$

where $\boldsymbol{x_{si}}$ is the feature vector of user i defined from the data that is associated with the user through the s^{th} relationship; $\boldsymbol{w_{st}}$ is the model parameters of the s^{th} data source part of expertise topic t; $|S|$ is the total number of relationship types considered in the model.

Thus, a base model has been built that can infer a user's expertise by combining multiple types of data associated with the user. This model does not distinguish between the different types of user data in the process of expertise inference. Whereas in reality, out of the multiple types of data associated with a user, some have more related information than others to the topics we are trying to infer. So the model may only be able to identify discriminative features from certain types of user data. In this case, using different types of user data to infer a user's expertise on a topic may have inconsistent results. In fact, as discussed in Section 1, most Twitter users do not have sufficient information for all the types of data (i.e. tweets, followers, friends and lists). This could result in a low prediction accuracy in the base model. In this work, we use the following regularizer to penalize the disagreement between different sources:

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s_1=1}^{S}\sum_{s_2\neq s_1}^{S}\left|\left|\boldsymbol{x_{s_1i}w_{s_1t}}-\boldsymbol{x_{s_2i}w_{s_2t}}\right|\right|^2 \tag{3}$$

where s_1 and s_2 are any two different types of user data from S. This regularization term tries to model the inference consistency among different types of data associated with the user. It aims to ensure the model has a good prediction accuracy regardless of the user having, or not having, all four different types of data. Here, we consider a specific example to illustrate the usefulness of this regularization process. In the training data, there are Twitter users who have expertise in the area "deep learning" and posted many tweets about this topic such as:

"Theano is a good deep learning framework for researchers".

At the same time, some users who have expertise in this topic are also included in a few lists with the list names like "*BigData expert*" and "*Machine learning*". Based on this fact, the model in Eq.(2) would identify features like "deep learning", "theano" in the user's tweets as the most discriminative features for the expertise topic "deep learning" and assign high weights to them. In comparison, features like "bigdata" and "machine learning" in the user's list data would receive a lower weight but they are actually important features for this topic too. It can be expected that the learned model for the expertise topic "deep learning" would perform well if there are sufficient user posted tweets available, while it may produce unsatisfying prediction results if we only have the user's list data. The regularizer in Eq.(3) tries to balance the weights on the important features from different types of user data. In this example, it will decrease the weights for features like "deep learning" and "theano" in the user's tweets and increase the weights for features like "bigdata" and "machine learning" in the user's list data by looking at the prediction accuracy of the model using only the tweet data or list data of the training users. Thus, the learned model ensures it will deliver good prediction performance both in cases where the user has one of the two types of data, and those where they have both.

Now, we can construct our learning model based on the above formulation. By substituting Eq. (2) and Eq. (3) to the SeTRL model (Eq. (1)), we have the following optimization problem:

$$\begin{aligned}\min_{\boldsymbol{W}}&\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{2}\left(y_{ti}-\sum_{s=1}^{S}\frac{1}{|S|}\boldsymbol{x_{si}w_{st}}\right)^2\\&+\gamma\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s_1=1}^{S}\sum_{s_2\neq s_1}^{S}||\boldsymbol{x_{s_1i}w_{s_1t}}-\boldsymbol{x_{s_2i}w_{s_2t}}||^2\\&+\alpha\sum_{e=1}^{E}||\boldsymbol{w_{e(1)}}-\boldsymbol{w_{e(2)}}||_2^2\ +\ \beta\left|\left|\boldsymbol{W}\right|\right|_1\end{aligned} \tag{4}$$

where γ is the regularization parameter to control the contribution of the inference consistency of different types of user data.

Similar to [7], this work also adopts the Accelerated Proximal Gradient method [29] to learn the proposed model, i.e. to estimate a parameter configuration of $\boldsymbol{W}$ that minimizes the objective function Eq.(4). It uses a linear combination of the previous two points as the search point to achieve high convergence speed.

4. DATASET CONSTRUCTION AND EXPERIMENT PREPARATION

To evaluate the proposed model, we need a dataset that contains the expertise information of a large number of Twitter users. In [7], the authors constructed such a dataset by using the popular question and answering service Quora. This is based on the fact that many Quora users explicitly provide both their expertise topics and Twitter account information in their profiles. So they harvest the same person's Twitter account and expertise topics, and these expertise topics were used as the ground truth of the expertise of the Twitter user in the experiments. The dataset contains 10,856 Twitter users and 149 expertise topics from which each user at least has knowledge of one topic. The 149 topics are specific expertise topics that cover a wide range of knowledge areas such as "software engineering", "atheism", "religion". In our work, we reused this dataset. While in [7] 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 terms of both the follower and friend data, up to 5000 users are collected for each user. This limit is due to the API access limitation.

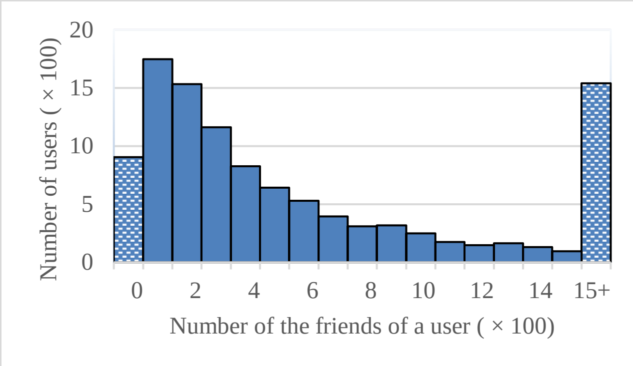

Figure 1: Distribution of the friend numbers of Twitter users in the dataset

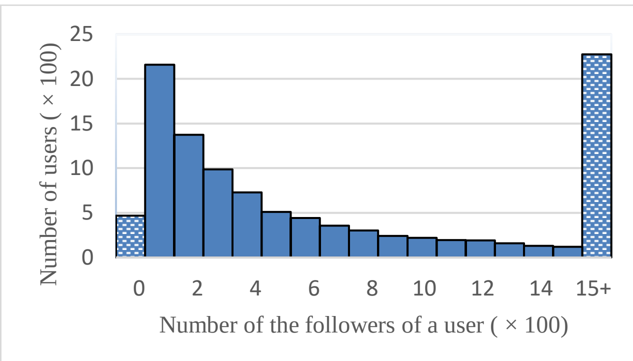

Figure 2: Distribution of the follower numbers of Twitter users in the dataset

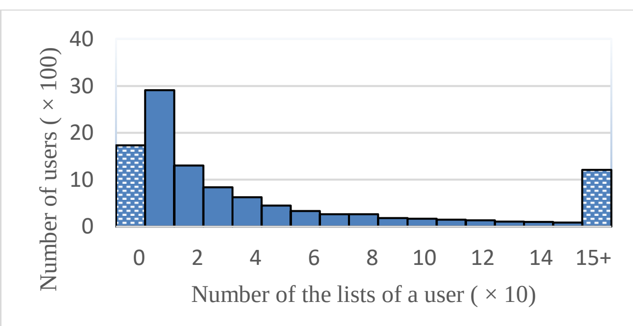

Figure 3: Distribution of the list numbers of Twitter users in the dataset

Figs.1, 2 and 3 show the distribution of the number of collected friends, followers and lists of all the users in the dataset respectively. For example, as shown in Fig. 1, we failed to collect any friend data for 904 users (8.3%) from the 10,856 Twitter users. This could either be due to these users not following any other users, or that they do not make this data publicly available. There are 1,539 users (14.2 %) who have over 1500 friends and most of the users (5,905 users, 54.4%) have between 1 and 500 friends in the dataset. In comparison, Twitter users have much less list data as shown in Fig. 3. There are 1,733 users (15%) who do not have any list data and over half of users are included in less than 20 lists in the dataset.

In the experiment, the profile information of the friends (or followers) of a Twitter user is used as an input for the inference task. A user profile on Twitter includes attributes such as user name, location and a short bio. There are two main reasons why we did not include the tweets posted by the friends (or followers) for inference. First, there are millions of friends (or follower) users in the dataset. The Twitter API limitation prevents us from harvesting the tweets of that many users within a reasonable time-frame. Second, a user's own tweets are already noisy with regard to reflecting his/her expertise [3]. Therefore, combining the tweets of their friends (followers) is likely to introduce even more noise to the process. In comparison, a user's profile provides direct personal information about him/her. Although, at an individual level, the profile may not contain any expertise information, or may even have missing elements, aggregating the profiles of a group of users could capture the main characteristics of the group [30]. Thus, 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. In the experiments, we combine multiple documents of a user, such as *tweet document* and *friend document*, as one input document for testing the effectiveness of the combination of multiple types of user data in user expertise inference.

This research considers two methods for constructing the user features with the input document. 1) **Unigram features**: This method uses the bag-of-words features of the document as the user feature space. 2) **Latent topic features**: LDA is applied to generate the latent topic distributions of the input documents of users, which are used as the user features. In the construction of experiment samples for each expertise topic, we only take the users with documents of over 100 terms as valid samples. Then balanced positive and negative samples of each topic are randomly chosen from the dataset, two thirds of which are used as the training set. The remaining one third is the test set for the topic. In the experiment, we do not consider expertise topics with less than 50 samples. As there are a large number of users in the dataset who do not have any, or have limited, list data, 64 expertise topics meet the experiment requirements when testing with the user list data alone. While experiments with tweets, friends and followers data respectively cover all the 149 topics.

5. EXPERIMENTAL RESULTS AND ANALYSIS

This section first outlines the metrics used to measure the performance of the various methods in the experiments. It then analyzes in detail the experimental results, both of methods which use a single type of user data and those which combine multiple types of user data.

5.1 Evaluation Metrics

The four metrics: accuracy, precision, recall and F1-score are used to measure the performance of methods in the work. Specifically, we use the averaged score of each of the four 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.

Table 2: Performance of SVM method with each type of user data for user expertise inference (%)

Data Type	Method	Recall	Precision	Accuracy	F1
Tweets	SVM-Sen	**72.10**	70.88	69.75	69.76
Friends	SVM-BI	67.84	67.11	65.46	64.75
	SVM-TF	68.37	**75.43**	**72.23**	**70.77**
	SVM-TFIDF	66.09	71.89	67.18	65.53
Followers	SVM-BI	68.55	61.54	62.67	62.72
	SVM-TF	61.64	74.23	69.37	66.08
	SVM-TFIDF	57.16	71.16	61.73	56.45
Lists	SVM-BI	64.61	69.76	66.89	64.46
	SVM-TF	66.04	74.47	70.18	68.29
	SVM-TFIDF	60.13	77.72	69.79	64.98

Table 3: Performance of TRL method with each type of user data and user unigram features (%)

Data Type	Method	Recall	Precision	Accuracy	F1
Tweets	SeTRL	**82.49**	78.76	79.65	80.08
Friends	TRL-TF	80.99	**80.24**	**79.82**	**80.09**
Followers	TRL-TF	76.33	78.92	77.51	77.32
Lists	TRL-TF	78.21	79.95	78.90	78.89

5.2 Performance of Different Types of User Data

The problem of user expertise inference has been studied [7] using the user's tweets. This work [7] examined the performance of the Support Vector Machine (SVM) method with unigram features of user tweet documents. They experimented on various schemes of weighting the user unigram features and discovered that tweet sentiment-based weighting scheme performed best. Using this weighting scheme, they then take into consideration the topic relatedness for user expertise inference, which is their proposed SeTRL model. In our work, we use a similar method to examine the performance of each of the other three types of user data (i.e. friends, followers and lists). Specifically, we first use SVM to examine the performance of each type of user data with different user features (i.e. unigram features and latent topic features) and different weighting schemes. The aim is to identify an appropriate user feature space and the optimal feature weighting scheme. Then we incorporate the topic relatedness into the inference model (i.e. model Eq. (1) but with the identified method of constructing user feature vectors) to further examine the performance of each type of user data. Note that the topic co-occurrence based method in work [7] is applied to construct the topic relation graph in the experiments. Below, we will present and analyze the experimental results using two methods of user feature construction: unigram features and latent topic features.

Unigram Features: The following schemes are used to weight unigram features of a user document in this work:

(1) BI: This scheme uses a binary value to represent the feature value. It means that the feature value is set to 1 if this user document has this feature (term), otherwise it is set to 0;

(2) TF: This scheme uses the frequency of the feature terms occurring in the user document as its weight;

(3) TFIDF: This scheme uses the Term Frequency – Inverse Document Frequency (TF-IDF) algorithm [31] to assign weights to the user features, where all the user documents are taken as a document set.

For each type of user data, we use the above three schemes to build user feature vectors, and employ SVM for user expertise inference. Table 2 presents the performance of the SVM method with the input of each type of user data and each weighting scheme. The SVM-Sen method adopts the tweet sentiment based weighting scheme proposed in [7] for user expertise inference using tweets. The experimental results first show that the TF weighting scheme outperforms the other two schemes on all the three types of user data. For example, when using friend data for user expertise inference, the SVM-TF achieves about 5% (F1-score) significant improvement, compared with SVM-TFIDF. This demonstrates that the commonly used terms in all the documents are also useful for user expertise inference on Twitter and decreasing their importance will harm the prediction accuracy. So the TF weighting scheme is applied to the experiments below that consider the topic relatedness or the use of multiple types of user data for expertise inference.

In addition, it is also observed from Table 2 that using friends data or lists data can achieve better performance than using followers data for user expertise inference. For example, SVM-TF achieves a 70.77% F1-score with friends data, while it gets a 66.08% F1-score with followers data. This difference indicates that information from a Twitter user's friends or lists can more effectively reflect the user's expertise. This confirms the intuition that a user tends to follow other users to reach the content s/he is interested in, and that a user is usually included in lists which contain some information about her/him. However, a user usually has no control over his/her followers and anyone can follow a user without his/her permission. This could introduce significant noise, such as spam users or advertisement users, to the expertise inference process. Furthermore, we use the TF weighting scheme to build user feature vectors and consider the topic relation information for user expertise inference, which is topic relation-regularized learning (TRL) model [7]. As shown in Table 3, it validates again that the friends data and lists data are more effective in inferring a Twitter user's expertise information.

Latent Topic Features (LTF): The LDA algorithm generates a probability distribution over the latent topics to represent the topic distribution of a user document. The probability score can be naturally used as the weight of a latent topic feature of a user. Although there is no need for selecting an appropriate feature weighting scheme, we still need to identify an optimal latent topic dimension. Therefore, we observe the performance of SVM with different topic dimensions for the optimal dimension selection. Figs. 4, 5, 6, 7 show the accuracy and F1 performance of SVM with the varying dimension of latent topics for each type of user data respectively. The results illustrate that its performance rises as the topic dimension increases and reaches the best performance when the dimension is set to 100. It then starts to decline for each type of user data. This means that about 100 latent topics can best represent the content richness of the input documents for our expertise inference task. Less or more topics could result in an underrepresentation or overrepresentation of the input content respectively. Thus, we select 100 as the optimal topic dimension and use it for the next experiments. From the experimental results, it can also be observed that the SVM with latent topic features achieves a similar or slightly better performance for each type of user data (considering the optimal topic dimension), compared with the SVM with unigram features. For example, when using followers data for user expertise inference, SVM-TF achieves a 66.08% of F1 score, while the SVM with latent topic features gets a slightly higher 67.3% of F1 score (topic dimension is set as 100). This improvement could benefit from the representation of a general topic distribution of the user. It may help to alleviate the impact of the noise in hundreds of thousands of unigram features but it also loses much important information about the user as you will see from the experiment analysis below.

Similar to the unigram features, we also examine the performance of TRL with the user latent topic features. As shown in Table 4, the import of the topic relation information does not remarkably improve its performance, compared with the SVM method. This result indicates that the latent topic distribution of a user can effectively reflect the knowledge information of a user on a specific expertise topic. But this feature representation may lose some detailed information about the user from the original user document, which could limit its potential for exploitation to further improve user expertise inference. Specifically, the relation information is most useful in the case where there is only limited evidence available to infer a user's expertise on a certain topic. However, this limited information about the user usually cannot be reflected from the latent topic distribution.

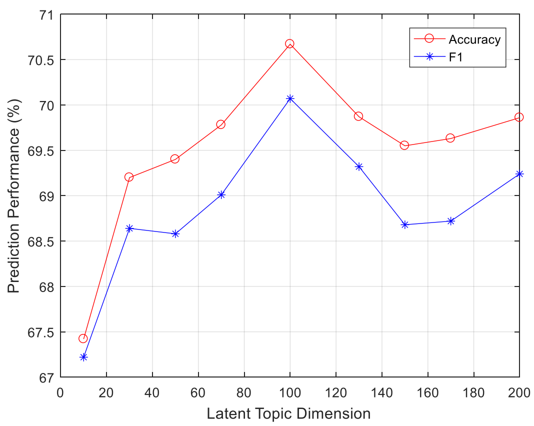

Figure 4: Performance of SVM with tweets data and latent topic features

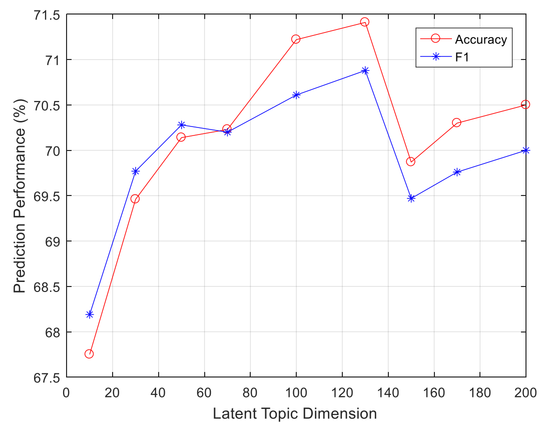

Figure 5: Performance of SVM with friends data and latent topic features

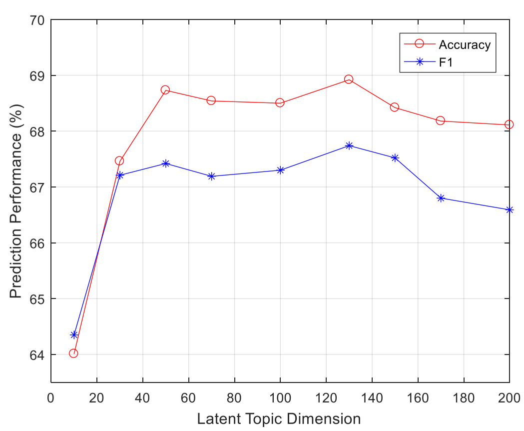

Figure 6: Performance of SVM with followers data and latent topic features

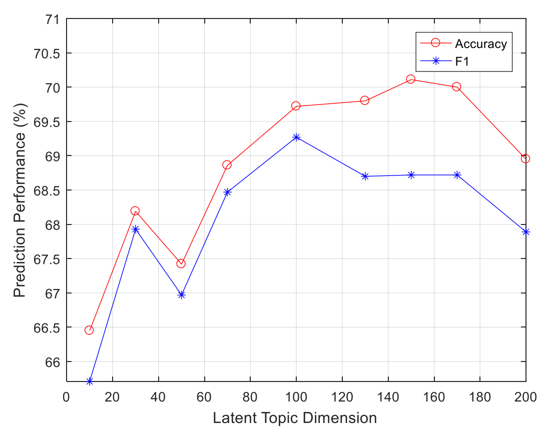

Figure 7: Performance of SVM with lists data and latent topic features

Table 4: Performance of TRL method with each type of user data and user latent topic features (%)

Data Type	Method	Recall	Precision	Accuracy	F1
Tweets	TRL-LTF	74.56	70.30	71.51	72.06
Friends	TRL-LTF	73.95	71.16	71.89	72.22
Followers	TRL-LTF	72.31	69.98	70.50	70.87
Lists	TRL-LTF	73.90	70.35	71.27	71.53

Table 5: Performance of different methods with the combination of the four types of user data (%)

Methods	Recall	Precision	Accuracy	F1
SVM	57.26	73.87	67.67	62.64
SeTRL	83.93	80.56	81.01	82.13
Our Model	**91.75**	**83.35**	**85.72**	**86.80**

Table 6: Performance of our model with various combinations of different types of user data (%)

Data Type Combinations	Recall	Precision	Accuracy	F1
Tweets+Friends	87.23	82.65	84.24	85.16
Tweets+Followers	84.62	80.71	81.76	82.67
Tweets+Lists	85.93	81.23	83.10	83.93
Tweets+Friends+Followers	88.09	81.87	84.76	84.90
Tweets+Friends+Lists	90.73	83.28	85.11	86.32
Tweets+Followers+Lists	86.58	82.47	84.03	84.15
Tweets+Friends+Followers+Lists	**91.75**	**83.35**	**85.72**	**86.80**

Overall, the above experimental results and analyses demonstrate that each type of user data is useful for user expertise inference but with varying performance. In terms of the construction of user features, although both schemes can deliver good inference performance, the unigram features keep more information about the user from the original data and leave space for advanced methods to further improve the performance. So we will use the unigram features for the experiments in the next subsection that describes the combination of multiple types of user data for user expertise inference.

5.3 Performance of Combinations of Different Types of User Data

This sub-section examines the performance of different methods using the combination of different types of user data for expertise inference. Table 5 compares the performance of our model with that of SVM and SeTRL using all four types of user data. Note that for SVM and SeTRL, the features generated from different data sources are directly concatenated as a single feature vector. Experimental results show that SVM 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 much inconsistent information is considered in the learning process. It is also shown that our model significantly outperforms SeTRL. It verifies the significance of taking into consideration the source consistency when using multiple types of user data for expertise inference.

Furthermore, we also conduct experiments to observe the performance of our model with various combinations of the four types of user data, as shown in Table 6. The results show that aggregating multiple types of user data by our model for expertise inference always can achieve a better performance than only using a single type of user data, and the more data about the user we incorporate, the better performance it can achieve. For instance, our model achieved the best F1 score of 86.8% in the experiment that exploits all four types of user data. It is also noted that the combination of the two types of user data, tweets and friends, can obtain satisfying performance, while the combination of the three types of user data: tweets, followers and lists achieves lower performance. This could be due to the facts we discussed previously that tweets and friends data are more effective in reflecting a user's expertise, while followers and list data are either more noisy or sparse for many users.

6. CONCLUSION

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. The model considers the inference consistency of different types of user data in the process of learning. It aims to optimize the model learning based on the assumption that a user's expertise information reflected by each type of data associated with him/her should be similar. Experiments are conducted on a large-scale real-world Twitter dataset with over 10,000 Twitter users and 149 expertise topics. In the experiments, four different types of user-related data are tested: tweets, friends, followers and lists. Detailed experimental analysis demonstrates that each type of user data is effective for user expertise inference, with variation in performance. Experimental results show that our model can better make use of the various data sources associated with the user, when compared with several baseline approaches, and that combining the data sources in a single model produces the best expertise inference performance.

ACKNOWLEDGEMENT

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 National Natural Science Foundation of China under Project No. 61300129, and a project Sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, China under grant number [2013] 1792.

REFERENCES

[1] Bughin, J., Byers, A. H., and Chui, M. 2011. How social technologies are extending the organization. *McKinsey Quarterly*, 20(11): 1-10, 2011.

[2] Nichols, J., and Kang, J. H. 2012. Asking questions of targeted strangers on social networks. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, pp. 999-1002.

[3] Wagner, C., Liao, V., Pirolli, P., Nelson, L., and Strohmaier, M. 2012. It's not in their tweets: modeling topical expertise

of Twitter users." In *Proceedings of 2012 International Confernece on Social Computing*. IEEE, pp. 91-100.

[4] Pal, A., and Counts, S. 2011. Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, pp. 45-54.

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

[6] Weng, J., Lim, E. P., Jiang, J., and He, Q. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, pp. 261-270.

[7] 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.

[8] Ghosh, S., Sharma, N., Benevenuto, F., Ganguly, N., and Gummadi, K. 2012. Cognos: crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 575-590.

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

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

[11] Craswell, N., de Vries, A. P., and Soboroff. I. 2005. Overview of the TREC 2005 enterprise track. In Text Retrieval Conference. Vol. 5. 199–205.

[12] Balog, K., Bogers, T., Azzopardi, L., de Rijke, M., and Van den Bosch, A. 2007. Broad expertise retrieval in sparse data environments. in *Proceedings of the 2007 International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp. 551–558.

[13] Deng, H., King, I., and Lyu, M. R. 2009. Enhancing expertise retrieval using community-aware strategies. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. ACM. pp. 1733–1736.

[14] Balog, K., Azzopardi, L., and De Rijke, M. 2006. Formal models for expert finding in enterprise corpora. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 43–50.

[15] Fang, H., and Zhai, C. 2007. Probabilistic models for expert finding. In *Proceedings of the European Conference on IR Research*. pp. 418–430.

[16] Fang, Y., Si, L., and Mathur, A. P. 2010. Discriminative models of integrating document evidence and document-candidate associations for expert search. In *Proceeding of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 683–690.

[17] Macdonald, C., and Ounis, I. 2011. Learning models for ranking aggregates. In *Proceedings of the European Conference on IR Research*. pp. 517–529.

[18] Dom, B., Eiron, I., Cozzi, A., and Zhang, Y. 2003. Graph-based ranking algorithms for e-mail expertise analysis. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*. pp. 42-48.

[19] Yeniterzi, R., and Callan, J. 2014. Constructing effective and efficient topic-specific authority networks for expert finding in social media. In *Proceedings of the 1st International Workshop on Social Media Retrieval and Analysis*. pp. 45-50.

[20] Sumbaly, R., Kreps, J., and Shah, S. 2013. The big data ecosystem at LinkedIn. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. ACM, pp. 1125-1134.

[21] Wang, Z., Li, S., Shi, H., and Zhou, G. 2014. Skill inference with personal and skill connections. In *Proceeding of the 25th International Conference on Computational Linguistics*. pp.520-529.

[22] Pal, A., Herdagdelen, A., Chatterji, S., et al. 2016. Discovery of Topical Authorities in Instagram. In *Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee*. pp.1203-1213.

[23] Popescu, A. M., Kamath, K. Y., and Caverlee, J. 2013. Mining Potential Domain Expertise in Pinterest. In *Proceedings of UMAP Workshops*.

[24] Li, W., Eickhoff, C., and de Vries, A. P. 2014. Geo-spatial domain expertise in microblogs. In Proceedings of *European Conference on Information Retrieval.* pp.487-492.

[25] Cheng, Z., Caverlee, J., Barthwal, H., et al. 2014. Who is the barbecue king of texas?: a geo-spatial approach to finding local experts on twitter. In *Proceedings of the 37th international ACM SIGIR conference on Research and development in information retrieval.* ACM, pp. 335-344.

[26] Niu, W., Liu, Z., Caverlee. J. 2016. On Local Expert Discovery via Geo-Located Crowds, Queries, and Candidates. *ACM Transactions on Spatial Algorithms and Systems,* 2(4): 14.

[27] Bar-Haim, R., Dinur, E., Feldman, R., et al. Identifying and following expert investors in stock microblog. 2011. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, pp.1310-1319.

[28] Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*. 3(2003): 993-1022.

[29] Ji, S., and Ye, J. 2009. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th annual international conference on machine learning*. ACM, pp. 457-464.

[30] HerdaĞdelen, A., Zuo, W., et al. 2013. An exploration of social identity: The geography and politics of news-sharing communities in twitter. *Complexity.* 19(2)10-20.

[31] Salton, G. and Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5), pp.513-523.