Enabling Semantic User Context to Enhance Twitter Location Prediction

Ahmed Galal and Abeer El-Korany
Department of Computing Science, Faculty of Computers and Information, Cairo University, Cairo, Egypt

Keywords: Dynamic User Modeling, Location Prediction, Semantic Context, Topical Interest, User Behavior.

Abstract: Prediction of user interest and behavior is currently an important research area in social network analysis. Most of the current prediction frameworks rely on analyzing user's published contents and user's relationships. Recently the dynamic nature of user's modelling has been introduced in the prediction frameworks. This dynamic nature would be represented by time tagged attributes such as posts or location check-ins. In this paper, we study the relationships between geo-location information published by users at different times. This geo-location information was used to model user's interest and behavior in order to enhance prediction of user locations. Furthermore, semantic features such as topics of interest and location category were extracted from this information in order to overcome sparsity of data. Several experiments on real twitter dataset showed that the proposed context-based prediction model which applies machine learning techniques outperformed traditional probabilistic location prediction model that only rely on words extracted from tweets associated with specific locations.

1 INTRODUCTION

Online social networks are very popular platforms that allow users to publish different types of contents that express their interest and ideas. Earlier in social networks, most of these published contents were merely textual posts, photos or videos. While today, online social networks have introduced location-based services that allow user to publish geo-location information from different locations and at different times of the day. Typical location-based social networking sites allow users to "check in" at a physical place, share the location with their online friends, rate, and provide tips on the visited locations. Recently, most well-known online social networks like Facebook and Twitter have incorporated the location services to allow their users to tag posts to locations. Others location-based social networking services available are: Foursquare, Gowalla, Google Places and Yelp.

The heterogeneous data in location-based social networks contain spatial-temporal social context and present new challenges and opportunities for further analysis. Those information are associated with users and was used to dynamically model users in social network (Galal and ElKorany, 2015) since both content-based and location-based information are changing over time. Generally, users profile attributes could be classified into dynamic attributes and static attributes. Dynamic time-tagged attributes such as topics used in posts and comments as well as location check-ins which are used to represent user's interest and behavior respectively. While, static attributes represent information that rarely change with time such as demographic attributes. Recently, most of research in social network analysis utilize those dynamic attributes to further be used in multiple social network analysis tasks such as recommender systems (Bobadilla et al., 2013; Abel et al., 2011), expert identification (Kleanthous and Dimitrova, 2008), location prediction (Chandra et al., 2011; Cheng et al., 2010; Ye et al., 2013), link prediction (Quercia et al., 2012) and similarity measurement between users (Galal and ElKorany, 2015; Lee and Chung, 2011).

Moreover, semantic information extracted from those dynamic attributes like topics or the category of location are used in automatic mapping of users to their "key" visited locations of interest (e.g., home, work, leisure). This mapping is done based on their online social presence and has been of great interest for the research community (Mahmud et al., 2014;

Galal, A. and El-Korany, A.
Enabling Semantic User Context to Enhance Twitter Location Prediction.
DOI: 10.5220/0005749502230230
In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence (ICAART 2016) - Volume 1*, pages 223-230
ISBN: 978-989-758-172-4

Ryoo and Moon, 2014). Furthermore, relationships between multiple dynamic attributes have emerged e.g. textual posts tagged with location check-in, images tagged with location check-in or images that contain textual description or comments. These relations between different dynamic attributes are not always realized and ignored. Although they could be mined and analyzed to add a big advantage by either to increase the accuracy (Galal and ElKorany, 2015) or better understand or represent the dynamic behavior and interest of users (Li and Chen, 2009).

Location prediction task attracted a significant amount of researchers (Wang and Prabhala, 2012). Being able to predict people's future location and hence, marketers could decide to do additional advertising at certain events or during particular TV shows. In this paper, a novel framework for predicting the category of user's current location using her/his geo-tagged Twitter activity and behavior such as topic of interest and category of previously visited places is proposed. Historical knowledge representing user activities while visiting key locations, as well as time of posting contents is used to enhance location prediction. This historical knowledge is represented by topics extracted from tweets that are posted while visiting similar location associated with the posting time of those tweets. By linking twitter user account with foursquare accounts, relationship between time, topics and location of users have been exploited.

The rest of the paper is organized as follows. In Section 2, we discuss the related work in prediction of user's behavior and interest. In Section 3, we explain the main components of the proposed location prediction framework. Results and accuracy evaluation on twitter dataset is discussed in Section 4. Finally, we draw our conclusion and discussed intended future work in Section 5.

2 RELATED WORK

Prediction of user's behavior and interests is an active area of research in social network analysis. Most of the proposed frameworks in these researches rely on user's published contents to predict his behavior or interest. A prediction framework proposed in (Jamali and Rangwala, 2009) that relies on machine learning to predict the popularity of posts in Digg based on comments statistics, users' interest which is represented by rate of commenting, users' feedback on a post, and by utilizing the users' community structure. Another framework proposed by (Weerkamp and De Rijke, 2012) that predicts future users activities based on terms extracted from tweets. This framework predicted tonight activities only without considering other timeframes within the day or whether today is a weekend or a weekday is.

Nowadays, one of the most popular features to be predicted is user's locations. These locations can be predicted using several ways either by analyzing dynamic attributes such as posts or tweets or by studying the history of users' movements.

There are two major goals in location prediction. The first one is to predict user's actual physical location such as the current city (Chandra et al., 2011; Cheng et al., 2010). While the second aims to predict the semantics of the location such as location category or type (Ye et al., 2013).

(Chandra et al., 2011) Predicted city level location of users based on the probability distribution of terms with respect to locations. These terms are extracted from tweets and reply-tweets. Another probabilistic based framework that predicts city level location of users is proposed in (Cheng et al., 2010). The main contribution in this framework is that they have managed to handle the sparsity of tweets and the nonstandard vocabulary that exist within the tweets.

The previously mentioned city level prediction frameworks are different than our proposed framework in that they did not consider the effect of time on user published content as well as the semantic relation between tweets and locations from where users tend to post specific content. On the other hand, the proposed framework predicts the category of location which is more significant when considering users actual location as users who are living in different countries or cities and may not be able to visit the same physical location.

Modeling of human mobility and a probabilistic framework for location prediction is proposed in (Cho et al., 2011). This framework relies on past user check-ins extracted from location based social networks and cell phone location tracing. (Ye et al., 2013) Proposed a framework which predicts the category of the user activities and the most likely location to be visited. By using mixed hidden Markov model to generate the activities category distribution using user's history of movement, activities and which is used further in location prediction.

Unlike all of the above mentioned works, our proposed research work makes use of some of these findings while also going further. The proposed framework predicts the category of user's current location (not the actual physical location) using topics extracted from tweets that have been posted during his stay in the location. Furthermore, taking into considering the assumption that people tends to visit some places on specific times or specific days, posting time of their tweets is considered also as a key factor in prediction of the category of user's location.

3 PROPOSED CONTEXT-BASED LOCATION PREDICTION FRAMEWORK

Location prediction is one of the most hot research topics in social network analysis. Current research in location based social network mainly focuses on two tasks: 1) predicting a user's home location; and 2) predicting a user's location at any time. The former task considers predicting the static home location of a user, while the latter considers more about predicting a user's moving trajectories (Gao & Liu, 2014). The proposed framework lends itself to second category of prediction tasks which aim to predict the category of user's current location.

Some researchers have considered the correlation between specific terms in tweets and their corresponding locations (Cheng et al., 2010; Hecht et al., 2011). Thus, for the purposes of enhancing prediction of user's current location, we propose a context-based model that integrates both content-based and location-based attributes to investigate the relationship between the published posts and the user's check-in behavior and their variation over time. The proposed framework can be divided into two major components; the first component is responsible for identifying and modeling of users' context such as locations and topics, while the second component is the prediction engine which will be explained in details in the following subsections.

3.1 Modeling of User Context

The first component in the proposed framework is responsible for creating the dynamic user model which captures user's topical interests based on his/her geographical locations. Users from different countries or cities can visit similar places that belong to similar categories (Galal and ElKorany, 2015). Thus, our proposed context-based prediction model uses novel temporal features not used by any existing work. According to (Dalvi et al., 2012), who studied the problem of matching a tweet to an object, where the object is from a list of objects in a given domain (e.g., restaurants). Their model is based on the assumption that the probability of a user tweeting about an object depends on the distance between the user's location and the object's location. Such matching can also geo-locate tweets and infer the present location of a user based on the tweets about geo-located objects. Accordingly, we assumed that a combination of geographical information and topic model could be used to discover user's current location. Furthermore, based on the hypothesis that users tend to post content related to the location they currently visit, we assumed that half an hour time frame to stay in specific place increases the likelihood that user start posting text related to the current place (different time slots values were used till the model become stable). For example, while a user is sitting in a restaurant waiting for menu, she/he usually post tweets related to type of food she/her prefer or post about trips while waiting for her/his plan in an airport. Therefore, we collect location-based users' tweets within half an hour after detecting a location check-in done by the user using foursquare social network. The topical interests will be represented as a set of vectors for each user such that one vector is used to represent set of topics posted by a user in specific location's category visited by him/her. Each topic in the vector is associated with a counter that represents the number of occurrence of this topic with respect to the location during a time window t. In the following sub-sections, modeling of user topics and locations is explained in details.

3.1.1 Identify User Locations

In order to extract users' location, we identified all tweets that contain a foursquare location check-ins. Then, we utilized the foursquare public API to identify the location categories[1] for each physical location extracted from a check-in. We utilized the Foursquare category hierarchy that consists of two kinds of nodes, location nodes and category nodes.

A location node represents a distinctive location such as Starbucks. While, category node represents a location category such as a coffee. Since members of social networks usually live in different geographic locations, they may visit different places belong to the same category. Accordingly, our proposed framework relies on distinguishing place category. Foursquare classification has given us in total 523 location category such as (hotel, stadium, etc…). Due to the sparsity of available user locations on social networks such as twitter and Facebook (Gao and Liu, 2014); we used the location category instead of the actual physical location. Location information is currently very sparse. Less than 1% of tweets are geo-

[1] https://developer.foursquare.com/categorytree

tagged and information available from the location fields in users' profiles is unreliable at best. (Cheng et al., 2010) Found that only 26% of Twitter users in a random sample of over 1 million users reported their city-level location in their profiles and only 0.42% of the tweets in their dataset were geo-tagged. In order to overcome this location sparseness problem, we further classified these location categories into higher semantic level. We utilized AlchemyAPI[2] (Gangemi, 2013) to further classify all of these location categories into 23 higher categories that represent the first tier categories provided by this API such as ('sports', 'shopping', 'food and drink', etc…).

3.1.2 Identify User Topics

For each tweet that contained a location check-in, we identified the set of adjacent tweets that have been posted by the same user within half an hour frame after the check-in. Then, we extracted the topics from each of those adjacent tweets using AlchemyAPI, which resulted in a vector of topics along with their count of occurrence as shown in Figures 1&2. Finally, these topics were linked to the parent location check-in to be used in the upcoming location prediction component.

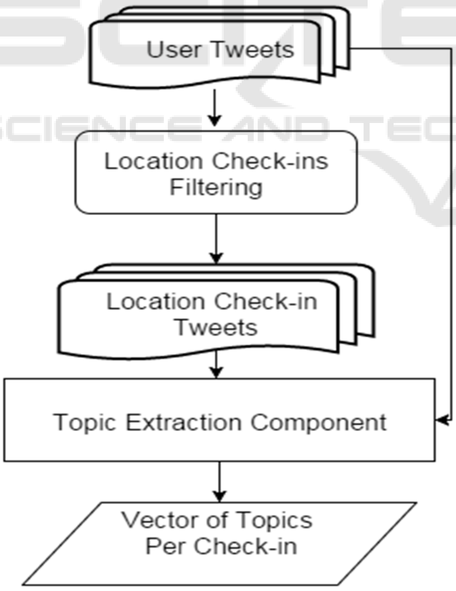

Figure 1: Topics Identification and Extraction.

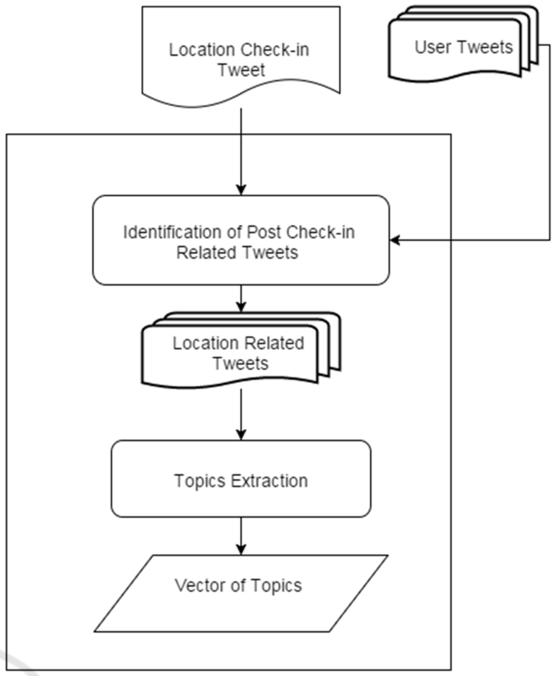

Figure 2: Topics Extraction Component.

3.2 Location Prediction

In order to predict current user location category type, we utilized both tweets' posting time and user topics as features for the prediction model. This classification problem will be explained in detail in the following subsection.

3.2.1 Features Selection

We used three main features to predict the category of user current location which are: the topics posted during half an hour time window stay in a location, the posting time of the check-in, and the type of the posting day (either weekend or weekday). In the following, each of those features is explained in details.

Topics. The first feature is the vector of topics that posted by user during her/his stay in specific location category. Each field of the vector represents the frequency of occurrence of each topic using AlchemyAPI topics categorization. In order to overcome sparsity problem, a threshold variable beta β is used such that we eliminate this topic from vector if the frequency of occurrence of a topic is less than this threshold. This threshold is used to exclude any noisy topics that may be irrelevant to the location and hence improve the accuracy of prediction. Finally,

[2] www.alchemyapi.com/

this threshold will also eliminate location check-ins that have very few topics.

Day Time. According to (Mahmud, Nichols, & Drews, 2014), user's posting pattern in twitter changes over day time. As early morning hours show less activity than hours in the morning, afternoon, and evening hours. Therefore, we split day time into 3 categories, morning, working hours and Night. We assumed the morning check-ins to be any check-in that has been posted between 12 am to 8 am. The working hour check-ins will be any check-in posted from 8 am to 4 pm, and night check-ins will be from 4 pm to 12 am.

Type of Day. A slight shift in the tweeting activity of the users during weekends is noticed compared to weekday (Mahmud et al., 2014). Thus, type of day where the user check-in was considered as third feature. Type of day is either a weekend if posted on Saturday or Sunday or it can be a working day if posted on any other day.

3.2.2 Location Prediction Model

Based on the above three mentioned features extracted for each check-in we proceed to use traditional classifiers to predict user current location. Several classical classifiers are used and applied using Weka toolkit (Witten and Frank, 2005) such as Naïve Bayes (NB), C4.5, k-Nearest Neighbor (k-NN). For KNN different value of K was applied till 10 neighbors with 1/d distance weighting which provided better accuracy value.

In order to evaluate the accuracy of the proposed framework we used a baseline probabilistic model (Cheng et al., 2010). This model relies on the probability distribution of actual words published by user in each location. Thus, all words that are related to a specific location category are extracted and aggregated from all tweets that have been posted within half an hour time window after any check-in. These extracted words are further filtered to remove any mention tags or stop words.

In this model, prediction of current user's location category can be divided into three main steps. First step is to generate the probability distribution of words for each location category by calculating the probability of each word w given the location category c as shown in equation (1).

$$\boldsymbol{P}(w|c) = \frac{\boldsymbol{count}(\boldsymbol{w})}{\boldsymbol{n}} \tag{1}$$

Where *count(w)* is the frequency of occurrences of the word w in all tweets that have been posted within half an hour timer frame after any location check-in of type category *c*. *n* represent the total number of words in these tweets. The second step is to calculate the probability that a user *u* exist in a specific location category *c* based on words *ws* extracted from his/her tweets that are posted half an hour after the check-in. this is done using equation (2).

$$\boldsymbol{P}(\boldsymbol{c}|\boldsymbol{u}) = \boldsymbol{P}(ws|c) * \boldsymbol{P}(\boldsymbol{ws}) \tag{2}$$

Where *P(ws|c)* represent the total probability of words *ws* to exist in location category *c*. *P(ws)* represent the total probability of words *ws* in the whole dataset of tweets which represent all textual tweets extracted for all location categories. Finally, the probability of set of locations are ranked in order to select the highest predicted location category based on the extracted words from half an hour time window.

4 EXPERIMENT

4.1 Experimental Set Up

Initial dataset that has been used in our previous research work that represent 1452 public twitter users with about half million (524,000) tweets (Galal & ElKorany, 2015). This dataset was prepared for the following experiments through the following steps. The first step was the identification of tweets that contain embedded location check-ins and the second step was the extraction of the set of adjacent textual tweets that are posted right after the location check-in tweets.

4.1.1 Identifying Location Check-in Tweets

In the first step we extracted all tweets that contained an embedded foursquare location check-ins. This step gave us in total 16,400 check-in tweets that have been posted by 1074 users. For each one of these tweets we extracted its posting time and the embedded check-in URL.

4.1.2 Extraction of the Set of Adjacent Textual Tweets

The second step was the extraction of the set of all adjacent textual tweets for each user that have been posted within a specific time frame after posting a location check-in tweet. It is significant to mention that in order to identify the amount of time which is considerable enough to post content relevant to the current place, different time windows have been tried (an hour and half an hour after check-in). However, prediction accuracy enhanced with half an hour time frame which is used as our hypothesis that this time could be considered is the average time for people to stay in a specific location. Thus, we aggregated tweets for half an hour after every check-in. This step gave us a total 30,000 textual tweets.

4.2 Utilizing Semantic of Topics and Locations Categories

Since location-based social networks suffer from data sparsity problem, we utilize location categories using foursquare API. These foursquare locations categories are further classified using AlchemyAPI into 23 locations categories. Also we use AlchemyAPI to identify set of topic of interest per location category by extracting it from text tweets that have been posted within half an hour after every check-in.

4.3 Classification

After the extraction of location categories and the topics from all aggregated tweets, we started to perform our classification by using the features to train our prediction model.

4.3.1 Classification Features

By utilizing the posting time of tweets and the topics extracted from the adjacent tweets, the features vector is built as follows:-

1. Posting time of tweets which is either morning, working hours or night.
2. The type of the posting day whether it was a weekday or a weekend.
3. The frequency of occurrence of each one of the 23 AlchecmyAPI topics that have occurred in aggregated tweets.

These features will be used to predict the location category of the corresponding location check-in tweet. In order to estimate the correct number of labeled classes, we collect all location visited by users and all tweets posted in each location category. Then, we calculate the coverage of tweets for each location. Accordingly, and as shown in Figure 3, the top 2 classes (location categories) covers 52 % of the total check-ins in the dataset while the top 4 and 6 classes cover 70% and 80% of the total check-ins respectively. Therefore, we consider only the top most used 2, 4 and 6 location categories as labeled classes in the prediction problem. Furthermore, we compare the prediction accuracy for those classes with the whole 23-class available in the dataset.

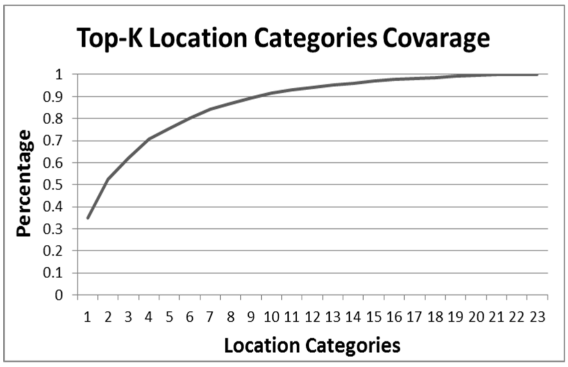

Figure 3: Top-K location Categories Coverage Percentage.

4.3.2 Classification Results

We applied the location prediction using the C4.5 (Mahmud et al., 2014) decision tree algorithm, the nearest neighbor classifier and Naïve Bayes classifier with 10 fold cross validation on the four classification problems. We also used three different values for threshold beta (which is used to eliminate topics which are 1 (no threshold), 2 and 3 number of occurrences) of topic. This threshold is used to remove any noise topics that have only occurred once or twice within the half an hour time frame in order to reduce sparsity problem. Those numbers are used as further increasing the threshold will lead to empty training instances that contain all topics with zero occurrences.

Finally, in order to evaluate our proposed prediction model we compared it with the baseline traditional probabilistic classifier which was applied on words' probability distribution of tweets.

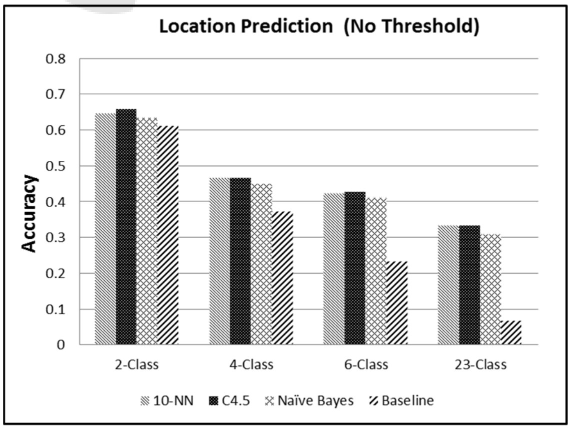

Figure 4: Comparison between the accuracy of the location category prediction for each classification problem and without using threshold for topics count of occurrence.

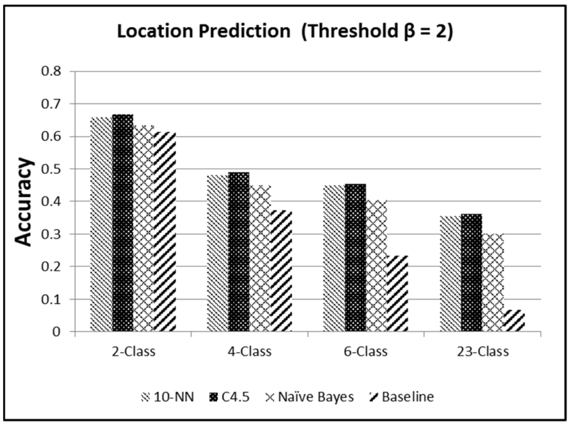

Figure 5: Comparison between the accuracy of the location category prediction for each classification problem and using threshold of minimum two occurrences for any topic.

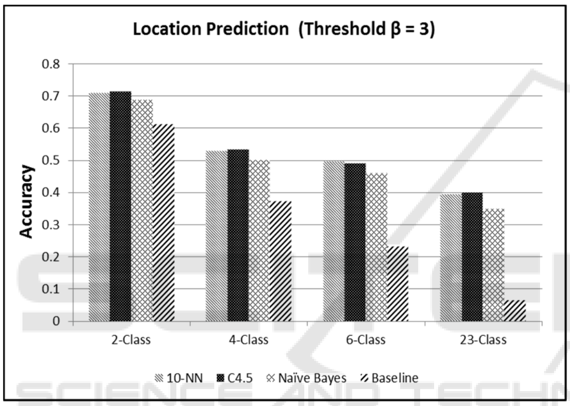

Figure 6: Comparison between the accuracy of the location category prediction for each classification problem and using threshold of minimum three occurrences for any topic.

By analyzing Table 1 and Figures from 4 to 6, the accuracy of our context-based prediction model is extremely high when applied to classification that discriminates between fewer classes (2 classes). This coincides with assumptions of researchers who tried to differentiate between two main locations in users life time (a user's home location and work). However, our proposed location prediction framework also provided an acceptable accuracy value for 4 and 6 classes representing other locations.

Furthermore, the proposed framework outperformed the baseline prediction method on all classification problems especially when increasing the number of predicted location categories. The results demonstrates the advantage of utilizing the semantic of user published content by considering topics rather than words as well as the significant of the proposed model to overcome the sparsity of data.

Table 1: Detailed comparison between the different classification problems using C4.5, NB, 10-NN and the baseline classifier.

No Threshold				
	2-Class	4-Class	6-Class	23-Class
10-NN	0.648	0.468	0.425	0.335
Naïve Bayes	0.636	0.452	0.412	0.31
C4.5	0.661	0.467	0.428	0.334
Baseline	0.613	0.373	0.234	0.066
Threshold (β) = 2				
	2-Class	4-Class	6-Class	23-Class
10-NN	0.659	0.48	0.449	0.356
Naïve Bayes	0.633	0.449	0.405	0.3
C4.5	0.668	0.49	0.454	0.362
Baseline	0.613	0.373	0.234	0.066
Threshold (β) = 3				
	2-Class	4-Class	6-Class	23-Class
10-NN	0.71	0.53	0.498	0.395
Naïve Bayes	0.688	0.501	0.46	0.35
C4.5	0.713	0.534	0.49	0.399
Baseline	0.613	0.373	0.234	0.066

The results showed that the threshold improved the accuracy of the predication as it removes any noisy topics that may be irrelevant to the location. Also a prediction accuracy of 49.8 % is achieved using threshold β=3 when considering the top-6 location categories that cover 80% of all location check-ins and an accuracy of 40% % is achieved when considering all location categories. Also even without using any threshold an accuracy of 42.8% is achieved when considering the top-6 location categories and 33.5% when considering all location categories in the classification.

5 CONCLUSION

In this paper the semantic relation between topics, location and time is explored and utilized in a framework for location category prediction. The results of the experiment proved the significance of such relation and how by simple utilization of this relation can achieve high accuracy in classification problems for location prediction. Also the experiment shows the importance of considering the semantic information rather than terms or words in location prediction.

In future work we consider enhancing this framework by utilizing more advanced features such as user's friendships and the past history of user's check-ins. Also it is considered to use this advanced version of the framework in prediction of the actual physical location because majority of users tend to stay in their city or country for long periods hence their visited location will not be changed drastically over short periods especially if we utilized their past check-ins.

REFERENCES

Abel, F., Gao, Q., Houben, G.-j., & Tao, K. (2011). Analyzing User Modeling on Twitter for Personalized News Recommendations. Em *User Modeling, Adaption and Personalization* (pp. 1-12).

Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems, 46*, 109-132.

Chandra, S., Khan, L., & Muhaya, F. (2011). Estimating twitter user location using social interactions--a content based approach. *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing* (pp. 838-843). IEEE.

Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 759-768). ACM.

Cho, E., Myers, S., & Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1082-1090). ACM.

Dalvi, N., Kumar, R., & Pang, B. (2012). Object matching in tweets with spatial models. *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 43-52). ACM.

Galal, A., & ElKorany, A. (2015). Dynamic Modeling of Twitter Users. *Proceedings of the 17th International Conference on Enterprise Information Systems*, *2*, pp. 585-593. Barcelona, Spain.

Gao, H., & Liu, H. (2014). Data analysis on location-based social networks. Em *Mobile social networking* (pp. 165-194). Springer.

Hecht, B., Hong, L., Suh, B., & Chi, E. (2011). Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 237-246). ACM.

Jamali, S., & Rangwala, H. (2009). Digging digg: Comment mining, popularity prediction, and social network analysis. *WISM 2009. International Conference on Web Information Systems and Mining* (pp. 32-38). IEEE.

Kleanthous, S., & Dimitrova, V. (2008). Modelling Semantic Relationships and Centrality to Facilitate Community Knowledge Sharing. Em *Adaptive Hypermedia and Adaptive Web-Based Systems* (pp. 123-132). Springer Berlin Heidelberg.

Lee, M.-j., & Chung, C.-w. (2011). A User Similarity Calculation Based on the Location for Social Network Services. *16th international conference on Database systems for advanced applications*, (pp. 38-52).

Li, N., & Chen, G. (2009). Analysis of a location-based social network. *CSE'09. International Conference on Computational Science and Engineering. 4*, pp. 263-270. IEEE.

Mahmud, J., Nichols, J., & Drews, C. (2014). Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST), 5*(3), 47.

Quercia, D., Askham, H., & Crowcroft, J. (2012). TweetLDA : Supervised Topic Classification and Link Prediction in Twitter. *Proceedings of the 3rd Annual ACM Web Science Conference*, (pp. 247-250).

Ryoo, K., & Moon, S. (2014). Inferring Twitter user locations with 10 km accuracy. *Proceedings of the companion publication of the 23rd international conference on World wide web companion* (pp. 643-648). International World Wide Web Conferences Steering Committee.

Wang, J., & Prabhala, B. (2012). Periodicity based next place prediction. *Nokia Mobile Data Challenge 2012 Workshop. p. Dedicated task, Vol. 2. No. 2.* .

Weerkamp, W., & De Rijke, M. (2012). Activity prediction: A twitter-based exploration. *SIGIR Workshop on Time-aware Information Access.*

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann.

Ye, J., Zhu, Z., & Cheng, H. (2013). What's your next move: User activity prediction in location-based social networks. *Proceedings of the SIAM International Conference on Data Mining. SIAM.*

Gangemi, A. (2013) 'A Comparison of Knowledge Extraction Tools for the Semantic Web', in The Semantic Web: Semantics and Big Data, Springer Berlin Heidelberg, pp.351-366.