2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Location-Specific Tweet Detection and Topic Summarization in Twitter

Vineeth Rakesh, Chandan K. Reddy, Dilpreet Singh
Department of Computer Science, Wayne State University, Detroit, MI, USA

Ramachandran MS
Yahoo, Bangalore India

***Abstract*—Automatic detection of tweets that provide Location-specific information will be extremely useful in conveying geo-location based knowledge to the users. However, there is a significant challenge in retrieving such tweets due to the sparsity of geo-tag information, the short textual nature of tweets, and the lack of pre-defined set of topics. In this paper, we develop a novel framework to identify and summarize tweets that are specific to a location. First, we propose a weighting scheme called *Location Centric Word Co-occurrence* (LCWC) that uses the content of the tweets and the *network information of the twitterers* to identify tweets that are location-specific. We evaluate the proposed model using a set of annotated tweets and compare the performance with other weighting schemes studied in the literature. This paper reports three key findings: (a) top trending tweets from a location are poor descriptors of location-specific tweets, (b) ranking tweets purely based on users' geo-location cannot ascertain the location specificity of tweets, and (c) users' network information plays an important role in determining the location-specific characteristics of the tweets. Finally, we train a topic model based on Latent Dirichlet Allocation (LDA) using a large collection of local news database and tweet-based Urls to predict the topics from the location-specific tweets and present them using an interactive web-based interface.**

I. Introduction

Microblogging has quickly grown as the avatar of social interaction. Though many websites like FriendFeed, Dailybooth, and Tumblr support microblogging, Twitter is the most favored and widely used website. Boasting more than 500 million registered users, about 1 million new accounts are added and over 400 million tweets are posted every day. Twitter's ability to propagate real-time information to a wide set of users makes it a potential system for disseminating vital information, and an invaluable source of news repository.

There are many works that discuss the estimation of user's location based on their tweets' textual content [1], [2], [3]. However, these research works do not effectively utilize the network based characteristics along with the textual content in their analysis. Furthermore, the goal of their studies is to predict users' geo-location, and it does not involve the prediction of tweets that talk about a particular geographical location. Predicting a user's location from a set of tweets, and predicting whether a tweet talks about a specific location are two related ideas, but they are substantially different from the computational standpoint. Classifying tweets purely based on the users' geo-location cannot determine whether a tweet describes something about that location. Consider the following tweets tweeted by a user from New York:

1. RT @VrancoRak: Drove over Brooklyn bridge - lower Manhattan illuminated with two towers of light skyward from ground zero... Beautiful!
2. Beautiful Rhinos in #endangered list. Its time to act join the #SaveRhinos campaign http://t.co/Vrco78rtgh

From the above set of tweets we can see that the first tweet clearly talks about a topic relevant to New York. However, the second tweet talks about a different topic that is not related to the location, even though the user is from New York. Therefore, determining the geo-location of a user cannot ascertain the location specificity of tweets tweeted by that user. Hence, we classify a tweet to be *location-specific* not only based on it's geographical information, but also based on the relevancy of it's content with respect to that location. In this paper, *we aim to discover such location-specific tweets by combining the tweets' content and the network information of the user*. We then use a topic model to effectively summarize such location-specific tweets and present them using a web-based interface.

II. Related Work

There are a number of research studies that focus on predicting the users' location by either mining the content of their tweets, or by using the *twitterers'* network information. For example, Cheng et al. [1] aim to solve this problem using the textual content of tweets to estimate the location of users at city level, while [2] predicts the user's point of interests such as club or hotels by considering tweet's content and temporal information. Geo-tag information from Twitter data is utilized by [3] to build language models of locations at various levels of granularity. The work in [4] studies the Twitter network to analyze the impact of geography on user interactions; The authors in [5] infer states, cities, and time zones of the twitter users by using an ensemble of content-based statistical and heuristic classifiers. In [6], the authors propose an unsupervised measure for evaluating the usefulness of tweet words for location prediction. [4] analyze Twitter network to study the impact of geography on user interactions.

Despite such a wide range of research works proposed in the literature, most of the existing works view the location-specificness as a task of predicting users' location. However, as explained earlier, predicting users' location and predicting whether a tweet contains information about a location are two completely different and orthogonal concepts. Additionally, none of these works effectively use a combination of content based information and network based information from twitter.

Therefore, our work is unique from other works in two important ways: (a) We see location-specificness based on the relevancy of tweet's content to a geo-location. (b) We use both the user's network and the tweet's content to determine the location-specificness of the tweets. The main contributions of this paper are outlined as follows:

ASONAM 2013, August 25-28, 2013, Niagara Falls, Canada
978-1-4799-1498-2/13/$31.00 ©2013 IEEE

1) We propose a novel weighting scheme called *Location Centric Word Co-occurrence* (LCWC) that uses mutual information (MI) score of tweet bi-grams; the tweet's inverse document frequency (IDF); the term frequency (TF) of tweets, and the user's network score to determine the location-specific tweets.

2) We show that our method achieves better precision in predicting the location-specific tweets compared to the detection of tweets purely based on user's geo-location information or top trending tweets from a location. We also show that the location of friends in a user's network play a significant role in determining the location-specificness of tweets tweeted by that user.

3) We train the latent Dirichlet allocation (LDA) model to detect topics from our ranked set of location-specific tweets and display them using a web-based interface.

III. Location-Based Tweet Retrieval Algorithm

To identify location-specific tweets, it is important to capture those features that can determine the uniqueness of a tweet with respect to a geo-location. To achieve this, we propose a weighting scheme called *Location Centric Word Co-occurrence* (LCWC), which uses the TF-IDF score along with the point wise mutual information (PMI) and network score of users. The method for LCWC weighting is shown in Algorithm 1. The algorithm primarily consists of the following steps:

Tweet Pre-processing: The data from Twitter stream is extremely impure with wide varieties of Unicode data, symbols and numbers. The function *PreProcessTweets* in Algorithm 1 removes stop words, and performs stemming to make the data reasonably pure.

Geo-tag based querying: Lines 9-13 in the Algorithm shows the procedure for creating a secondary tweet document list TD_s that is necessary for our weighting scheme. The primary document D_{prim} is basically the dataset that needs to be weighted. In this study, we consider every state in US as a location, and we initialize the set of top tweeting states from US based on their frequencies of tweeting to Loc. Our goal is to rank the bi-gram tweets in $PrimLoc$; therefore, we remove the primary location $PrimLoc$ from the set of top tweeting states. We now create a secondary document TD_s by getting the user id u_{id} and it's corresponding location u_{loc} for every tweet t in the pre-processed database d_l. We add the tweet and it's location (t, u_{loc}) to TD_s only if u_{loc} is contained in the set Loc.

Identifying Bi-Gram Sequences: *twitterers* try to pack maximum information within 140 characters by using a combination of keywords, hash-tags and links to external news sources. Therefore, information retrieval from tweets can be made more effective by discovering such co-occurring patterns. Consider the following tweets that talk about a local event called Holiday Nights:

1) Holiday Nights was a lot of fun!! at #greenfield village
2) It was fun watching the fireworks at #Holiday Nights @Greenfield Village
3) Yipeee.. finally its a Holiday! time to go out

Algorithm 1 LCWC Weighting

```
 1: INPUT: The primary location of interest
 2: OUTPUT: Ranked location-specific tweets
 3: procedure LcwcWeighting($PrimLoc$)
 4:     $Loc \leftarrow Set(top\ tweeting\ locations) - PrimLoc$
 5:     $S \leftarrow GetStreamingData(date)$
 6:     $d_l \leftarrow PreProcessTweets(S)$
 7:     $D_{prim} \leftarrow GetAnnotatedDataset()$
 8:     $[TD_s, Tot_{B_{prim}}] \leftarrow EmptyArray()$
 9:     for $each\ t\ \epsilon\ d_l$ do
10:         if $u_{loc}\ exists\ and\ (u_{loc} \subset Loc)$ then
11:             $Add\ (t,\ u_{loc})\ to\ TD_s$
12:         end if
13:     end for
14:     $[B_{prim}, B_{sec}] \leftarrow GetBigramSeq(D_{prim}, TD_s)$
15:     for $each\ bi_t\ \epsilon B_{prim}$ do
16:         Get $Pmi(bi_t), Tf(bi_t), Idf(bi_t), N_{score}(bi_t)$
17:         $Score = Pmi * Tf * Idf * N_{score}$
18:         $Add\ (bi_t, Score)\ to\ Tot_{Bprim}$
19:     end for
20:     $return\ Rank(Tot_{Bprim})$
21: end procedure
22: PredictTopics($Tot_{Bprim}$)
```

We notice that the words like *Holiday Nights* and *#greenfiled village* are words that are commonly used by *twitterers* who describe about this event. It is worth noting that users tend to use a combination of hash-tags and words to describe the event; therefore, relying simply on a uni-gram model cannot provide the much needed information about the event. For example, though the tweet *Yipeee... finally I have my #Holidays!* contains the word *#Holidays* it does not talk anything about the event of Holiday Nights. We obtain the bi-grams of tweets from the primary and the secondary documents in line 14 of the algorithm.

Weighting Scheme: We now propose a new weighting scheme called Location Centric Word Co-occurrence (LCWC) that is effective in capturing the location-specific features of tweets. Since the bi-grams are representatives of the entire tweet, our aim is to assign weights to these bi-grams tweets and rank them according to their final scores. The steps 15-19 in our Algorithm shows the procedure of weighting these bi-gram tweets. For each bi-gram tweet bi_t in the bi-gram primary document B_{prim} that was obtained in the previous step, we calculate (i) the point-wise mutual information (PMI); (ii) the term frequency (TF); (iii) the inverse document frequency (IDF) (iv) the network score of each tweet and use their product to determine the final score of the bi-gram. First, we determine the mutual information of the bi-gram candidates by using the PMI score between the terms in bi_t. The PMI is a popular measure which determines the mutual information between the events x' and y' belonging to discrete random variables X and Y [11]. It is defined as follows:

$$PMI(x', y') = log\frac{P(x', y')}{P(x')P(y')} \tag{1}$$

where $P(x')$, $P(y')$ denote the probabilities of x' and y' respectively and $P(x', y')$ denotes the joint probability of x' and y'.

To capture the uniqueness of the tweets within a location, we calculate the *idf* scores of bi-grams. The *idfs* are calculated for all the bi-gram tweets $bi_t \, \epsilon \, B_{prim}$ in the primary document using the equation (2), where D_{loc} represents the

$$idf(bi_t, B_{prim}) = log \frac{|D_{loc}|}{|d \epsilon D_{loc} : bi_t \epsilon d|} \tag{2}$$

set of all documents $\{d_1, d_2 ... d_{u_{loc}}\}$ and every single document $d_{u_{loc}}$ in D_{loc} is composed of a set of bi-gram tweets from one specific user location u_{loc}. $|d \epsilon D_{loc} : bi_t \epsilon d|$ is the number documents in which the bi-gram tweet appears. Finally, we capture the frequency of the bi-gram sequence using the normalized term frequency

$$tf(bi_t) = \frac{f(bi_t, B_{prim})}{|B_{prim}|} \tag{3}$$

Where $f(bi_t, B_{prim})$ denotes the frequency of bigram bi_t in the tweet document B_{prim}, and $|B_{prim}|$ is the total number of bi-gram tweets.

Network score of tweets: We would like to see whether the posting of location-specific tweets is influenced by the geographical proximity between the user and his friends. To answer this question, we choose six different locations from US and manually select 30 tweets from each location. We choose these tweets based on two criteria: (a) the tweets should contain some information about it's location (i.e location-specific); (b) the author's geo-location must match the tweet's location. We then retrieve friends list of these users and randomly select 60 friends from each user and query their geo-location information. Interestingly, we found that more than 37% of friends are from the same location as that of the user. This clearly shows that the location-specificness in tweets are definitely impacted by the similarity of locations between the *twitterer* and his friends.

To calculate the network score, we retrieve all the friends of users in our test dataset of 10,000 tweets. We then select 60 friends for each user and retrieve their geo-location. Instead of selecting a random set of 60 friends, we track the past 200 tweets of these 10,000 users and retrieve the list of friends that the user had interacted with. We found that on an average, a user having 50 friends interacts only with 12% of his friends. Therefore, we selected the remaining friends from the pool of the user's friends list and retrieve their geo-locations to calculate the network score. The network score of a user u in location L is given by

$$N_s(u, L) = f(u, L)/f(u) \tag{4}$$

where $f(u, L)$ represents the friends of u who are from the same location L. However, not all the users have a friend count of 60. In our dataset, more than 26% of the users have a friend count less than 30 and about 500 users have less than 10 friends. Therefore, we divide $f(u, L)$ by a normalizing factor $f(u)$, which denotes the total number of friends of u.

Finally, line 17 in the algorithm presents the LCWC scoring function that takes the product of the PMI score, the term frequency, the *idf* and the network score of the user to assign weights to every bi-gram tweet bi_t in B_{prim}. We append the bi-grams and their scores to the array Tot_{Bprim} and rank these bi-grams.

Detection of Tweet Topic: To predict topics from the location-specific tweets, we use the popular Latent Dirichlet Allocation (LDA) model [12]. Due to the sparsity of textual information, training the LDA directly on tweets results in very poor topic prediction. To overcome this problem, we train our LDA model using the RSS of local news data obtained from *Detroit Free Press* and articles crawled from tweet Urls. We use the LDA implementation from Apache Mahout, which uses field variational inference for model parameter estimation. We generate 200 topics from our input data and use it to predict the topics in the location-specific tweets (line 22).

IV. Experimental Results

A. Data Collection

To build our database, we used Twitter's streaming API to constantly fetch the streaming data. Our research is confined to locations within US; therefore, we extracted all the tweets that pertain to US, and filtered the rest. For this research, we choose two types of dataset: (a) an annotated dataset of 10,000 tweets that is used for evaluating our model and (b) a large set of unlabeled data comprising of 2,344,000 tweets that is used for presenting our results in a web-based framework.

Annotation Dataset: In this research, the primary location of interest is chosen as Michigan. Therefore, we selected only those tweets that pertain to Michigan and used a subset of 10,000 tweets for annotation. We asked the annotators to annotate the tweets based on the following instructions:

1) Based on the tweet's content, label the location-specific tweets.
2) If the tweet is location-specific, select a category for the tweet from the list of predefined topic categories.

In total the annotators classified only 289 tweets out of 10,000 tweets as location-specific. The annotators used the predefined topics such as local news, entertainment, sports, business, weather, advertisement, politics and health to label the topics of tweets. The annotation of these tweets were performed using a group of 30 graduate students. The tweets were divided into 10 sets, each containing 1000 tweets and every set was given to 3 different students. In this way, every student annotated one set (1000 tweets), and each set was annotated in 3 independent ways. The final annotation labels were decided based on the majority vote.

B. Evaluation on Twitter data

We evaluate the performance of our model by calculating the number of location-specific tweets that are successfully detected. We test our model using the annotated dataset of 10,000 tweets containing 289 location-specific tweets and compare its performance against the frequency based weighting scheme, geo-location based weighting scheme and a variant of our LCWC weighting scheme that does not involve the network score. Figure 1 shows that the LCWC weighting scheme achieves a precision of 35% for the top twenty tweets. We also notice that the LCWC outperforms the frequency based weighting scheme by over 20%. The frequency based weighting scheme simply ranks the tweets based on the frequency of trending terms. Such terms can gain popularity due to retweets or tweets that frequently talk about a popular

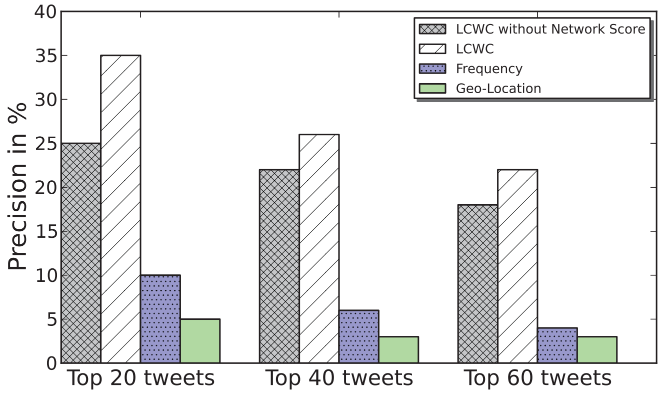

Fig. 1: Comparison of LCWC with (a) LCWC without network score, (b) frequency based weighting, and (c) geo-location based weighting of tweets.

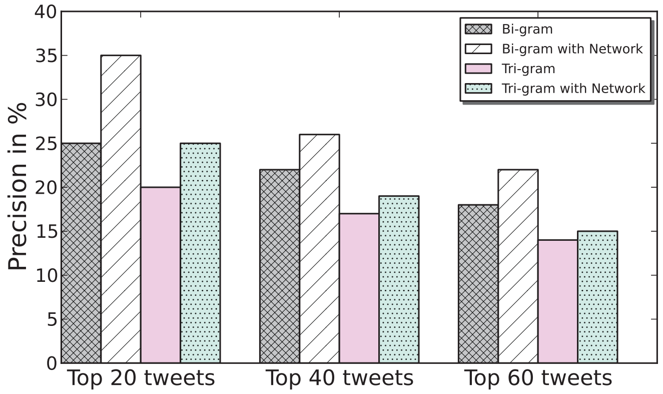

Fig. 2: Comparison of LCWC weighting using bi-grams and tri-grams.

event (bursty tweets). However, from our results we can see that the top trending tweets cannot be a good feature for predicting location-specific tweets. We also see that ranking tweets purely based on user's geo-location results in extremely poor precision since not all the tweets tweeted by a user from a location are location-specific. Finally, we see that the user's network score plays a very important role in determining the location-specific tweets. It is important to note that the precision clearly drops down when the network component from LCWC is removed. For top 40 and 60 tweets we observe a similar trend though there is a drop in precision with the increase in the number of retrieved documents.

Figure 2 shows the comparison of LCWC weighting using bi-grams and tri-grams. The weighting scheme based on tri-grams is basically an extension of the LCWC weighting scheme for tri-gram sequences. Despite the better performance of tri-gram based LCWC over the frequency based weighting scheme, and the geo-location based weighting scheme, it falls short of the bi-gram based LCWC. This clearly indicates that the *bi-gram sequences are a better descriptors of location-specific tweets*. As explained in Section III, this might be due to the fact that *the Twitter users tend to convey information using very short word sequences or hash-tags*.

Finally, we predict the topics on the annotated dataset using our LDA model that was trained using local news and tweet Urls. For our web based-interface [1], we use our unlabeled data set of 2,344,000 tweets as input to our model and categorize the tweets into different topics like sports, local news, finance, etc.

V. Conclusion

In our study, we showed the importance of detecting location-specific tweets and summarizing their topics. We developed a new model for effectively retrieving tweets that talk about a geographical location. We introduced a weighting scheme called *LCWC* that is efficient in capturing the location-specific features of tweets. Using a set of annotated tweets, we performed extensive evaluation to show the effectiveness of our model in retrieving location-specific tweets, and it's ability to outperform other models. Finally a web-based implementation was developed to provide an interactive representation of our model in a user-friendly environment.

Acknowledgement

This work was supported in part by the US National Science Foundation grant IIS -1242304.

References

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

[2] W. Li, P. Serdyukov, A. P. de Vries, C. Eickhoff, and M. Larson, "The where in the tweet," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 2473–2476.

[3] S. Kinsella, V. Murdock, and N. O'Hare, "I'm eating a sandwich in glasgow: modeling locations with tweets," in *Proceedings of the 3rd international workshop on Search and mining user-generated contents*. ACM, 2011, pp. 61–68.

[4] J. Kulshrestha, F. Kooti, A. Nikravesh, and K. P. Gummadi, "Geographic dissection of the twitter network," in *In Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.

[5] J. Mahmud, J. Nichols, and C. Drews, "Where is this tweet from? inferring home locations of twitter users," *Proc AAAI ICWSM*, vol. 12, 2012.

[6] H.-w. Chang, D. Lee, M. Eltaher, and J. Lee, "@ phillies tweeting from philly? predicting twitter user locations with spatial word usage," in *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*. IEEE, 2012, pp. 111–118.

[7] D. Ramage, S. Dumais, and D. Liebling, "Characterizing microblogs with topic models," in *International AAAI Conference on Weblogs and Social Media*, vol. 5, no. 4, 2010, pp. 130–137.

[8] O. Jin, N. Liu, K. Zhao, Y. Yu, and Q. Yang, "Transferring topical knowledge from auxiliary long texts for short text clustering," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 775–784.

[9] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, vol. 6, 2010.

[10] R. Fujino, H. Arimura, and S. Arikawa, "Discovering unordered and ordered phrase association patterns for text mining," *Knowledge Discovery and Data Mining. Current Issues and New Applications*, pp. 281–293, 2000.

[11] K. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[12] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[1] http://www.cs.wayne.edu/vineeth/location_specific_topics/index.html