Inference Models for Twitter User's Home Location Prediction

Hicham G. Elmongui*†, Hader Morsy‡, Riham Mansour§

*GIS Technology Innovation Center, Umm Al-Qura University, Makkah 21955, KSA
elmongui@gistic.org
†Computer and Systems Engineering, Alexandria University, Alexandria 21544, Egypt
elmongui@alexu.edu.eg
‡SmartCI Research Center, Alexandria University, Alexandria 21544, Egypt
hader@mena.vt.edu
§Advanced Technology Lab, Microsoft Research, Cairo 11728, Egypt
rihamma@microsoft.com

***Abstract*—Twitter has emerged as one of the most powerful micro-blogging services for real-time sharing of information on the web. A large base of Twitter users tend to post short messages of 140 characters (Tweets) reflecting a variety of topics. Location-based-services (LBSs) may be built on top of microblogs to provide for targeted advertisement, news recommendation, or even microblogs personalization. Knowing the user's home location would empower such LBSs. In this paper, we propose prediction models to infer the users' home location based on their social graph and tweets content. The problem is non trivial as the tweets are short and not many people like to share their location for privacy concerns. Our extensive performance evaluation on a publicly available dataset demonstrates the effectiveness of the proposed models. The proposed models outperform the competitive state-of-the-art home location inference techniques that are based on the social graph, tweet content, and both by a relative gain in the F-measure of up to 37.71%, 29%, and 9.06%, respectively.**

I. Introduction

Twitter has emerged as one of the most powerful micro-blogging services for real-time sharing of information on the web. As of March 2015, Twitter has more than 645 million users, with about 289 million of them described as active users [35]. The volume of tweets is rapidly increasing and has reached 500 million tweet per day. Each user receives a persistently increasing volume of tweets especially that 80% of Twitter users are on the ubiquitous mobile devices [36]. Such rich content of users and data allowed for sharing information that might be used in many situations. Numerous research efforts have spurred mining Twitter feeds to provide for different applications and monitoring services that provide useful, and sometimes critical, functionalities. For instance, personalized recommender systems were built to recommend information, possibly news articles, that are relevent to the users (e.g., [1], [16], [28]). Events are another example that are possible to be detected and to seamlessly spread as a result of this environment (e.g., [41]).

There is plenty of research efforts spent on mining Twitter feeds for different applications like finding trending topics and influential users. Some of these applications are more relevant to some cities and communities like spreading awareness around some epidemic disease or examine the shopping patterns in some cities. Many Locaiton-based services (LBSs) would benefit from the existance of the location information associated with the tweet or with its author. Unfortunately, location information is currently not sufficiently available for different reasons. only small percentage of the twitter posts has associated location information [26] as many users do not reveal their locations when they tweet for privacy concerns. Even when users opt to expose their location, the revealed locations vary from a fine granular location, such as a point of interest or a geo-location, to a coarse grain location as a country. Further, Twitter posts are short and do not provide sufficient context to infer the user's location from the post content.

The above reasons have motivated research work to infer the location of the Twitter user based on the available information like the user's posts and her social graph. Some work has investigated techniques for geo-locating Twitter users using models built with tweets originating from known locations like [4], [6]. Others have employed language modeling for the user's historical posts like [19] to infer the user's location. However, building a language model on tweets is challenging. Tweets are small in length and consequently they lack the context in which their content fall in. Other research work has been done on using the user profile location and her social graph of followers and followees to infer the location [13]. This in turn assumes that enough users would expose their location information, which is not always true. Location information entered by users are usually not complete or accurate.

In this paper, we tackle the problem of the sparseness of the user's location information by developing several algorithms to predict the *home* locations of Twitter users. Ultimately, we would like to be able to provide a source location for each incoming tweet, reply, or user action. We believe that predicting the user's home location is a major milestone towards this goal. We propose a Friends classifier, a novel graph-based location inference model. The Friends classifier uses the spatial label propagation by computing the weighted geometric median of the locations of the friends of a certain user. In addition, we propose two hybrid approaches to infer the user's location

978-1-5090-0478-2/15/$31.00 ©2015 IEEE

based on her social graph and on her tweets' content and behavior. Our first hybrid approach empowers the state-of-the-art content-based location inference technique [23] with our proposed Friends classifier. The second proposed hybrid approach empowers graph-based location inference approaches with the state-of-the-art content-based location inference technique [23]. Graph-based location inference approaches utilizes an initial set of users in the social graph with know locations to propagate their locations to other users in their graph. We use the content-base location inference in [23] to increase the number of users with known location to boost the spatial label propagation.

The contributions of this paper may be summarized as follow:

- We propose a graph-based location inference model, the Friends classifier, to predict the user's home location from her social graph.
- We propose two hybrid location inference models that collect signals from both the social graph and the content of the tweets. The proposed hybrid approaches represent two different ways of empowering content-based and graph-based approaches together.
- We thoroughly evaluate the performance of the proposed models and compared them with the state of the art. Our models outperform the competitive state-of-the-art graph-based, content-based, and hybrid location inference techniques.

The rest of the paper is organized as follows. Section II highlights related work. The proposed user home location inference models are presented in Section III. In Section IV, we evaluate the proposed models using an extensive experimental study. We conclude the paper by a summary and final remarks in Section V

II. Related Work

With the scarcity in the Twitter user profiles containing a valid location entry, lots of research efforts have been spent to infer the user's location. Such efforts can be divided into three categories.

The first category uses the content of the user tweets to infer her location. In 2014, Mahmud et al. used an ensemble of statistical and heuristic classifiers to predict Twitter users' home locations. They made use of a geographic gazetteer dictionary to identify place-name entities. Their hierarchical classifiers used the tweet content and the user tweeting behavior to infer the user's home location [23]. Previously in the same year, other location inference techniques were proposed by extracting points of interests from the tweets and predicting whether the user has visited, is currently at, or will soon visit this point of interest [20]. Local words were also used to build a language model for the different locations [32]. Meanwhile, [32] worked with multiple microblogs and points of interests to predict the top-k locations for the users, whereas [29] utilized a variant of gaussian mixture models for the user location inference. Han et al. provide an extensive feature selection comparison used in location inference in [12].

In 2013, Twitter user's location inference was done using the tweet content and the user declared metadata [10], [34] or utilized insights from the social web [15]. In 2012, the location inference used location indicative words [11], local words to infer the home location [38], or to infer the tweet source location [14]. Roller et al. mapped documents on a grid using a language model [30]. In 2011, location inference was proposed using the tweet content and the social interaction [3]. Statistical methods are used to infer the user current location as well as her home location [19], [13], [6], whereas other work used to infer the location of documents and not tweets [39]. The earliest work to infer the location for a Twitter user based on the content of the tweets discovered the latent topics in 2010 [4].

The second category uses the user social graph to infer her location. The social graph connects each user with its followers and followees. In 2014, Compton et al. infer the user location from her friends by framing the geotagging problem as an optimization over a social network with a total variation-based objective using a scalable and distributed algorithm [5]. In 2013, Rout et al. formulated the problem as a classification task, where the most likely city for a user without an explicit location is chosen amongst the known locations of their social ties [31]. Meanwhile, McGee et al. proposed a location estimator, FriendlyLocation, which leverages the distance between a pair of users and the relationship between the strength of the tie between the pair [25]. The concept of landmarks, which are defined as users with a lot of friends who live in a small region, was introduced to infer the home location assuming a landmark reports her true locations [40]. The home location is also inferred by spatially propagating location assignments through the social network [18]. On the other hand, in 2012, the location of a tweet is predicted by combining two weak predictors, link prediction and location prediction [33]. In this work, friendship, *and not the location*, is predicted from the tweet content. In 2011, majority voting location inference is done from the friends' location [17]. The earliest work for location inference from the social graph, and hence is commonly compared to, assumes symmetric links and is based on Facebook [2].

The third category contains hybrid approaches that utilized both the social graph as well as the tweet content. In 2012, three hybrid techniques were introduced. First, Gu et al. proposed *GeoFind*, which uses k-means to cluster people. A logistic regression classifier is used to profide effective fussion, i.e., reranking of two ranked lists: one from the social graph using maximum likelihood estimation (MLE) of geotagged friends and one from the tweet contents using geo-sensitive textual features [8]. Li et al. focused on the problem of home location inference by proposing a unified discriminative influence model. To overcome the scarcity of the signals, they form a heterogeneous graph from the social network and from the tweets by collecting signals from both sources in a unified probabilistic model. To overcome the problem of the noisy signals, they capture how likely a user connects to a signal with respect to 1) the distance between the user and the signal, and 2) the influence scope of the signal [22]. Li et al. also proposed a multiple location profiling model [21]. In that work, their model captures that a user has multiple locations (home location and visited locations) and his following relationships and tweeted venues can be related to any of his locations.

III. Twitter User's Home Location Inference

In this section, we present our inference models to predict the user's home location. The first proposed model is the Friends classifier. It is a graph-based location inference model that may be used by itself to predict the user's home location, or may be used as a building block in the two proposed hybrid models, namely Injected Inferences and Cascaded Inferences.

To illustrate the proposed models, the next subsection briefly describes some background needed. The subsequent subsections present the novel models themselves.

A. Background

In our hybrid home location inference models, we adopt the state-of-the-art content-based home location inference method that was proposed by Mahmud et al in [23]. In this work, a hierarchical classification is performed on the content of the tweets through a time zone classification followed by a city classification. The time zone classification consists of a dynamically weighted ensemble of statistical classifiers (nouns, entities, and hashtags), heuristic classifiers (place names and Foursquare check-ins), and behavior classifiers (frequency of user tweets). The multi-nominal city classifiers are used to classify the users to the cities in their corresponding time zone. They consist of all classifier types used in the time zone classification except for the behavior classifiers. A salient feature of this location inference technique is its extensibility property of being able to add additional classifiers to the hierarchy, which we benefit from in our Injected Inferences model.

The second piece of background needed to proceed is the "spatial label propagation". In 2002, Zhu et al. proposed a semi-supervized simple iterative algorithm to propagate labels for items connected through a network. A label is assigned to an unlabeled node as the most frequent label among its neighbors [42]. In 2013, Jurgens proposed spatial label propagation, which propagates the spatial label to infer the location of the users. Algorithm 1 illustrates the technique [18]. The key in the label propagation is how to select the label to be propagated, which is defined by the function *select*. Jurgens provided three alternative for the selection of the label to be propagated from a labeled node to an unlabeled one. Namely, 1) the geometric median of the labeled neighbor, 2) an alternative multivariate median definition, and 3) a heuristic based on social theory. The Friends classifier represents a novel variant of the spatial location propagation, which will be used in all proposed models.

Algorithm 1 Spatial Label Propagation Algorithm [18]

Procedure PropagateTheSpatialLabels
Input U: set of users in social network
N: mapping for each user to her friends
L: ground-truth mapping from users to their coordinates
Ouput E: Estimated user locations
begin
Initialize E with L;
while *Convergence criteria is not met* **do**
  Let E' be the next mapping from user to location;
  for all $u \in U - \text{domain}(L)$ **do**
    Let M be a list of locations;
    for all $n \in N(u)$ **do**
      if $E(n) \neq \phi$ **then**
        add $E(n)$ to M;
      end if
    end for
    if $M \neq \phi$ **then**
      $E'(u) = \text{select}(M)$
    end if
  end for
  $E = E'$
end while
end

B. The Friends Classifier

The proposed Friends classifier is a graph-based location inference technique that uses the social graph along with initial seed users' locations in order to infer the location of other users as depicted in Figure 1. The Friends classifier performs a variation of the spatial location propagation technique presented in [18] (see Algorithm 1). Instead of using the geometric median or its variants for the labeled neighbors as in the original spatial location propagation technique, we use the *weighted geometric median* of the labeled neighbors as the label of the current unlabeled node of the graph. The rationale behind this decision comes from the meaning of the edges of the social graph. Such edges represent relationships; i.e., friendships, among the nodes of the graph. Naturally, any user has different tie strength with his different followers or followees. The difference depends on the type of relationship between them. Therefore, the weights of the edges should not be equal over all the friends of a person. For instance, the relationship between Alice and her caring brother Bob would be stronger than that between her and an acquaintance Carol whom she met only once in an event two years ago.

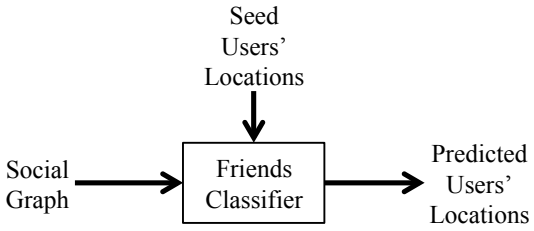

Fig. 1: The Friends Classifier

The proposed weight of an edge between a node u and its neighbor v is composed of the following five weights:

1) Number of mutual friends between u and v. Let F_u and F_v denote the set of friends of u and v respectively, then

$$w_1 = \frac{|F_u \cap F_v|}{|F_u|}$$

2) Number of mutual followers between u and v. Let F'_u and F'_v denote the set of followers of u and v

respectively, then

$$w_2 = \frac{|F'_u \cap F'_v|}{|F'_u|}$$

3) Whether the relation between u and v is symmetric; i.e., there is an edge from u to v (denoted by $u \rightarrow v$) and an edge from v to u (denoted by $v \rightarrow u$).

$$w_3 = \begin{cases} 1 & \text{if } u \rightarrow v \text{ and } v \rightarrow u, \\ 0 & \text{otherwise.} \end{cases}$$

4) Whether there is a triad relation between u and v along with w.

$$w_4 = \begin{cases} 1 & \text{if } \exists w : u \rightarrow v,\ v \rightarrow w,\ w \rightarrow u \\ 0 & \text{otherwise.} \end{cases}$$

5) The dispersion of the followers of the friend. Let v denote u's friend with known location. If v has followers from around the globe (e.g., a celebrity), thus he is not indicative of u's location. Let w denote any follower of v. Also, let d_{vw} denote the Haversine distance between the coordinates of v and w. Let $\tilde{d}_v$ denote the median of the distances between v and its followers. Then

$$\tilde{d}_v = \text{median}\big(d_{vw}\big)$$

Let D_v denote the dispersion of the followers of v and is computed as follows.

$$D_v = \frac{1}{|F_v|} \sum_{w \in F_v} \left(d_{vw} - \tilde{d}_v\right)^2$$

The lower the dispersion is, the higher value takes the weight w_5. Let D_{max} and D_{min} denote the maximum and minimum dispersions among all the nodes. Therefore,

$$w_5 = \frac{D_{\text{max}} - D_v}{D_{\text{max}} - D_{\text{min}}}$$

While computing the weighted geometric median of the labeled neighbors, we combine the above five weights by using their root mean square as the weight W_{uv} of the edge between a node u and its neighbor v. Thus,

$$W_{uv} = \sqrt{\frac{1}{5}\big(w_1^2 + w_2^2 + w_3^2 + w_4^2 + w_5^2\big)}$$

C. The Injected Inferences Model

Injected Inferences model is our first hybrid approach; i.e., it leverages signals from both the user's tweet text as well as her social network. This model takes any content-based location inference technique that uses machine learning classification as its underlying technique, and empowers it with an extra feature. That feature is the tweet author's home location, if known, or null otherwise.

In our work, we take the state-of-the-art content-based location inference technique of Mahmud et al [23], and empower it with our Friends classifier, which is described in Section III-B. As illustrated in Figure 2, the Friends classifier uses the already-known users' locations as seed locations. The produced predicted users' locations along with the initial seed users' locations are used as a feature in the classification task.

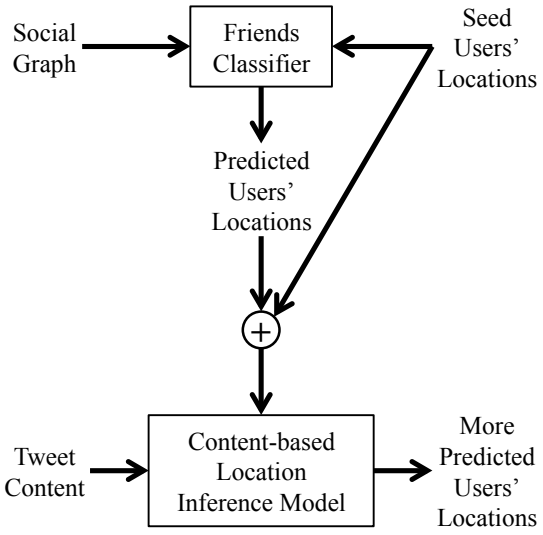

Fig. 2: The Injected Inferences Model

The Friends classifier's output is *injected* into the list of statistical, heuristic, and behavior classifiers in the user's time zone classification. In the city classification, the Friends classifier is also added to the list of classifiers. However, we take into consideration the friends that belong to the same time zone as friends from other time zones would be unneeded outliers.

The adopted state-of-the-art content-based approach uses dynamically weighted ensemble of classifiers, where each classifier is given a weight inversely proportional to the number of classes it discriminates from. We empirically give the Friends classifier a weight of 1. The votes of the classifiers are then merged using those weights.

D. The Cascaded Inferences Model

The Cascaded Inferences model is our second hybrid model to infer the home location of Twitter users. It also leverage signals from both the user's social network along with the content of her posted tweets. In contrast to Injected Inferences, Cascaded Inferences empowers any other graph-based home location predictor with the state-of-the-art content-based location inference technique.

In our work, Cascaded Inferences *cascades* the Friends classifier, described in Section III-B, after the state-of-the-art content-based location inference technique of Mahmud et al [23]. This cascading operation is depicted in Figure 3. Like any other graph-based home location inference model, the Friends classifier gets a set of seed users' locations as ground truth. The spatial label propagation algorithm uses this seed set in order to label the rest of the social graph. By increasing the number of users with known location, the spatial label propagation labeling power would increase, and additional correctly predicted users' location can be reached.

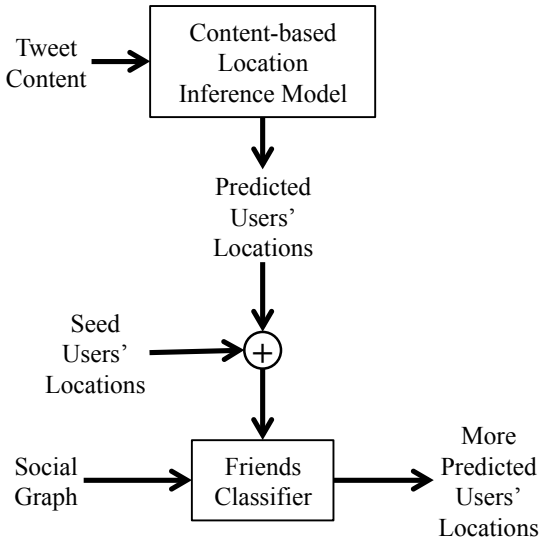

Fig. 3: The Cascaded Inferences Model

Not all locations infered by the content-based location inference technique are used as seeds since many of them might have low confidence. We only pass those locations that have a a confidence score larger than a picked threshold. This threshold is the confidence score that gives the maximum F-measure for the content-based location classifier.

IV. Experimental Evaluation

We perform extensive experiments to evaluate the quality performance of the proposed Twitter user's home location inference techniques. We compare the proposed techniques against the state-of-the-art content-based, graph-based, and hybrid approaches. All used machine-learning algorithms were executed from the WEKA suite [9].

To calculate the weighted geometric median in the location inference component, we got the coordinates using the Open Geocoding API of Mapquest [27]. We retrieved the time zone of each location using the Geonames Timezone API [7].

A. Dataset

We adopted the dataset used in the state-of-the-art hybrid location inference approach [22]. The dataset is publicly available at [37]. This dataset contains 284 million following relationships, 3 million user profiles and 50 million tweets. We extracted the ground truth locations of users by searching the "Location" field in their profiles for the patterns "CityName, StateName" or "CityName, StateAbbreviation" from USA as was done in [22]. We sampled this dataset to retrieve 20,000 users, 9.1 million following relationships and 10 million tweets. The sampling consisted of the first 20K users from a breadth first search that started from a random user (id = 25582718). Working with a publicly available dataset is used as a benchmark to guarantee results repeatability.

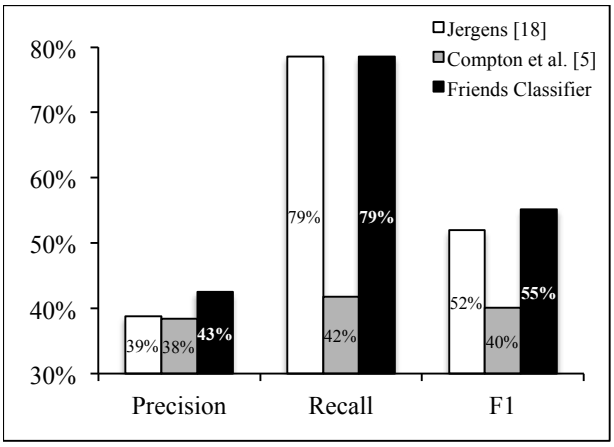

Fig. 4: Performance of Proposed Graph-Based Approach

B. Evaluation Metric

In addition to the precision and recall of the system, we use the micro-averaged F-measure, which considers predictions from all instances [24]. It calculates the F-measure across all labels as

$$F1 = \frac{2PR}{P+R}$$

where P is the precision and R is the recall of the system.

C. Performance of Proposed Friends Classifier

We compared the proposed graph-based location inference component with Jurgens [18], who proposed the spatial location propagation, and Compton et al. [5] representing the state-of-the-art graph-based location inference approaches.

Figure 4 gives the output of the 10-fold cross validation of the graph-based techniques. The experiments shows that the using the weighted geometric median is superior to using the geometric median, which comforms with our rationale that not edges in the social graph should be equally treated. The edges represent relationships among the people and should be different as the tie strength between the user and his followers and followees. Not only the proposed graph-based component outperforms the original spatial label propagation [18], but also the picked Friends' features let it outperform the competitive state-of-the-art techniques. The Friends classifier produces an F-measure of 55.14%, which is superior to the state-of-the-art techniques: Jurgens [18] produces an F-measure of 51.94% representing the geometric median choice for the spatial label propagation. Last, Compton et al. [5] produces an F-measure of 40.04% on the same benchmark dataset; i.e., Friends produces a relative gain of 37.71%.

D. Performance of Proposed Cascaded Inferences

In this section, we compare the proposed Cascaded Inferences approach with its baseline, the proposed graph-based location inference technique. This comparison is to show the benefit the Friends classifier gets from increasing the seed locations.

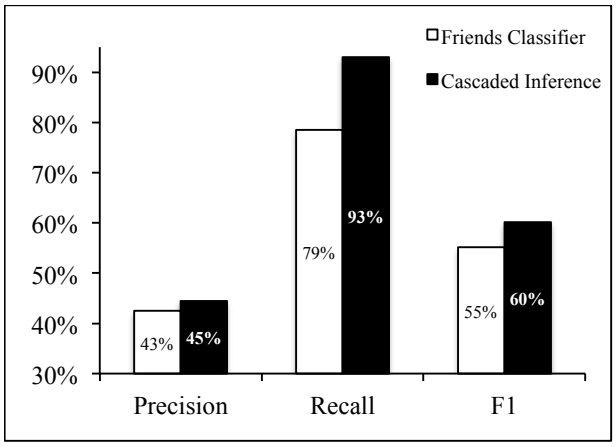

Fig. 5: Performance of Proposed Cascaded Inferences

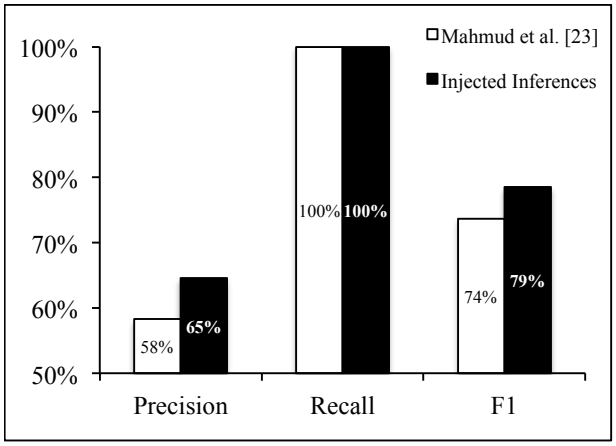

Fig. 6: Performance of Proposed Injected Inferences (Time-zone Location Granularity)

Figure 5 gives the output of the 10-fold cross validation of Cascaded Inferences. For comparison, we restate the results of its baseline. The experiments shows that Cascaded Inferences model produces an F-measure of 60.20%; i.e., there is a relative increase of 9.18% in the F-measure when increasing the input labels of the proposed graph-based approach. This is done when it was seeded with the locations inferred with high confidence using the state-of-the-art content-based technique as discussed in Section III-D.

E. Performance of Proposed Injected Inferences

Here, we compare the proposed Injected Inferences, described in Section III-C, with the competitive state-of-the-art content-based location inference approach by Mahmud et al. [23]. The comparison shows the effect of adding the *Friends* classifier to the other classifiers used in the hierarchical classifiers used in [23]. For both techniques, a set of naive Bayes classifiers is used. This is why a prediction is made for each user, and hence the recall is 100%.

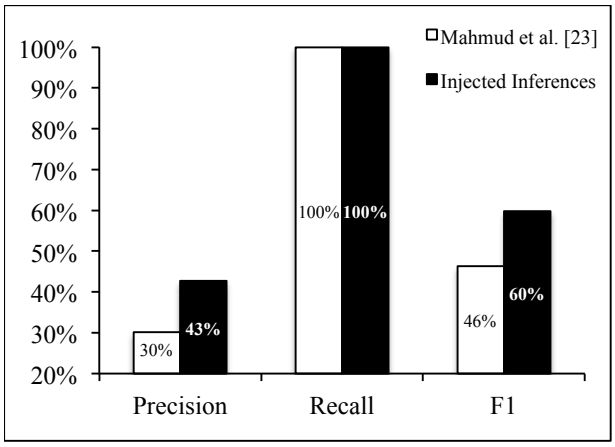

Fig. 7: Performance of Proposed Injected Inferences (City Location Granularity)

Figure 6 shows the performance of injecting the Friends classifier's output into the list of statistical, heuristic, and behavior classifiers in the user's time zone classification. The figure shows the result of the 10-fold cross validation before and after the injection. The injection made the resulting Injected Inferences model outperform the classifiers of [23]. The F-measure is 73.66% for the original predictor before the injection, whereas the Injected Inferences result into an F-measure of 78.50%, which is a relative gain of 6.57%.

The city level location predictor naturally gets lower F-measure for both techniques. However, the relative gain of Injected Inferences is higher. Before the injection, the F-measure is 46.27% for the state-of-the-art content-based location inference technique. After the injection, the F-measure is 59.81%. This is a relative gain of 29% in the F-measure. Figure 7 shows the performance of injecting the Friends classifier's output into the list of classifiers in the user's city classification.

F. Comparing with State-of-the-Art Hybrid Model

The previous experiments showed that the proposed hybrid approaches outperformed the competitive content-based state-of-the-art as well as the Friends classifier, which in turn outperformed the competitive graph-based state-of-the-art. In this section, we present the results of our experiments to compare Injected Inferences and Cascaded Inferences against the competitive state-of-the-art hybrid location inference model [22].

Figure 8 shows the results of the 10-fold cross validation of the two proposed hybrid models as well as the state-of-the-art hybrid model. Both Cascaded Inferences and Injected Inferences outperform the state-of-the-art hybrid model. The state-of-the-art model gets an F-measure of 55.20%. Cascaded Inferences get an F-measure of 60.20% - a relative gain of 9.06% in the F-measure with respect to [22], whereas Injected Inferences get an F-measure of 59.81% - a relative gain of 8.35% in the same evaluation metric.

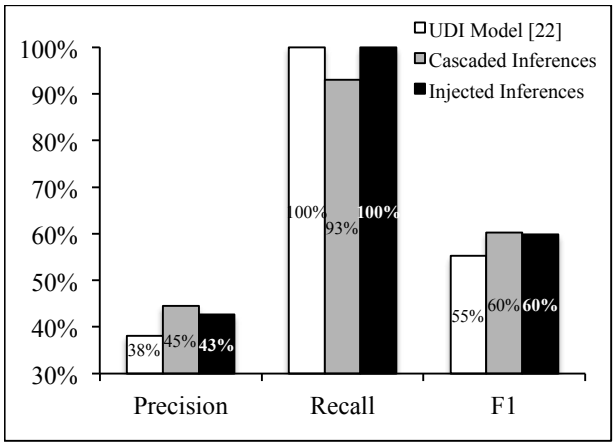

Fig. 8: Comparing with State-of-the-Art Hybrid Model

V. Conclusion

In this paper, we proposed three models to infer Twitter user's home location First, the Friends classifier was proposed to predict the location based on the user's social network. Next, Cascaded Inferences and Injected Inferences were introduced as a way to get more signals from the users' tweet content beside the social graph. Despite the scarcity and noisy signals of the short tweets, the extensive performance evaluation on a publicly available dataset demonstrates the effectiveness of the proposed models. The proposed models outperform the competitive state-of-the-art home location inference techniques that are based on the social graph, tweet content, and both. The Friends classifier produces a relative gain in the F-measure of up to 37.71%. The proposed hybrid models, which outperform Friends, also produce a relative gain of 29%, and 9.06% over the content-based and hybrid models, respectively.

Acknowledgement

This material is based on work supported in part by (1) Research Sponsorship from Microsoft Research, (2) the KACST National Science and Technology and Innovation Plan under grant 14-INF2461-10, and (3) the KACST GIS Technology Innovation Center at Umm Al-Qura University.

References

[1] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing user modeling on twitter for personalized news recommendations. In Joseph A. Konstan, Ricardo Conejo, Jos L. Marzo, and Nuria Oliver, editors, *User Modeling, Adaption and Personalization*, volume 6787 of *Lecture Notes in Computer Science*, pages 1–12. Springer Berlin Heidelberg, 2011.

[2] Lars Backstrom, Eric Sun, and Cameron Marlow. Find Me if You Can: Improving Geographical Prediction with Social and Spatial Proximity. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*, pages 61–70, 2010.

[3] Swarup Chandra, Latifur Khan, and Fahad Bin Muhaya. Estimating Twitter User Location Using Social Interactions-A Content Based Approach. In *Proceedings of the 2011 IEEE Third International Conference on Social Computing (SocialCom'11)*, pages 838–843, 2011.

[4] Zhiyuan Cheng, James Caverlee, and Kyumin 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 (CIKM'10)*, pages 759–768, 2010.

[5] Ryan Compton, David Jurgens, and David Allen. Geotagging One Hundred Million Twitter Accounts with Total Variation Minimization. In *Proceedings of the 2014 IEEE International Conference on Big Data (BigData'14)*, 2014.

[6] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP'10)*, pages 1277–1287, 2010.

[7] Geonames Timezone API. http://www.geonames.org/export/web-services.html\#timezone, 2014. (Last accessed 2014/06/01).

[8] Hansu Gu, Haojie Hang, Qin Lv, and Dirk Grunwald. Fusing Text and Frienships for Location Inference in Online Social Networks. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT'12)*, pages 158–165, 2012.

[9] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[10] Bo Han and Paul Cook. A stacking-based approach to twitter user geolocation prediction. In *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13): System Demonstrations*, pages 7–12, 2013.

[11] Bo Han, Paul Cook, and Tim Baldwin. Geo-location Prediction in Social Media Data by Finding Location Indicative Words. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING'12)*, 2012.

[12] Bo Han, Paul Cook, and Timothy Baldwin. Text-based Twitter User Geolocation Prediction. *Journal of Artificial Intelligence Research*, 49(1):451–500, 2014.

[13] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*, pages 237–246, 2011.

[14] Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. Discovering Geographical Topics in the Twitter Stream. In *Proceedings of the 21st International Conference on World Wide Web (WWW'12)*, pages 769–778, 2012.

[15] Yohei Ikawa, Maja Yohei, Jakob Rogstadius, and Akiko Murakami. Location-based Insights from the Social Web. In *Proceedings of the 22nd International Conference on World Wide Web Companion (WWW'13 Companion)*, pages 1013–1016, 2013.

[16] Nirmal Jonnalagedda and Susan Gauch. Personalized news recommendation using twitter. In *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT),*, volume 3, pages 21–25, Nov 2013.

[17] Clodoveu Augusto Davis Jr., Gisele Lobo Pappa, Diogo Renn Rocha de Oliveira, and Filipe de Lima Arcanjo. Inferring the Location of Twitter Messages Based on User Relationships. *Transactions in GIS*, 15(6):735–751, 2011.

[18] David Jurgens. That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM'13)*, 2013.

[19] Sheila Kinsella, Vanessa Murdock, and Neil 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 (SMUC'11)*, pages 61–68, 2011.

[20] Chenliang Li and Aixin Sun. Fine-grained Location Extraction from Tweets with Temporal Awareness. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'14)*, pages 43–52, 2014.

[21] Rui Li, Shengjie Wang, and Kevin Chen-Chuan Chang. Multiple Location Profiling for Users and Relationships from Social Network and Content. *Proceedings of the VLDB Endowment (PVLDB)*, 5(11):1603–1614, 2012.

[22] Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. Towards Social User Profiling: Unified and Discriminative Influence Model for Inferring Home Locations. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12)*, pages 1023–1031, 2012.

[23] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Home Location Identification of Twitter Users. *ACM Transactions on Intelligent Systems and Technology (ACM TIST)*, 5(3):47:1–47:21, 2014.

[24] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[25] Jeffrey McGee, James Caverlee, and Zhiyuan Cheng. Location Prediction in Social Media Based on Tie Strength. In *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management (CIKM'13)*, pages 459–468, 2013.

[26] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM'13)*, pages 400–408, 2013.

[27] Open Geocoding API of Mapquest. http://open.mapquestapi.com/geocoding/\#batch, 2014. (Last accessed 2014/06/01).

[28] Owen Phelan, Kevin McCarthy, and Barry Smyth. Using twitter to recommend real-time topical news. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 385–388, New York, NY, USA, 2009. ACM.

[29] Reid Priedhorsky, Aron Culotta, and Sara Y. Del Valle. Inferring the Origin Locations of Tweets with Quantitative Confidence. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW'14)*, pages 1523–1536, 2014.

[30] Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised Text-based Geolocation Using Language Models on an Adaptive Grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'12)*, pages 1500–1510, 2012.

[31] Dominic Rout, Kalina Bontcheva, Daniel Preoţiuc-Pietro, and Trevor Cohn. Where's @Wally?: A Classification Approach to Geolocating Users Based on Their Social Ties. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media (HT'13)*, pages 11–20, 2013.

[32] KyoungMin Ryoo and Sue Moon. Inferring Twitter User Locations with 10 Km Accuracy. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion (WWW Companion'14)*, pages 643–648, 2014.

[33] Adam Sadilek, Henry Kautz, and Jeffrey P. Bigham. Finding Your Friends and Following Them to Where You Are. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM'12)*, pages 723–732, 2012.

[34] Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Mhlhuser. A Multi-Indicator Approach for Geolocalization of Tweets. In *Proceedings of the 7th International AAAI Conference on Web and Social Media (ICWSM'13)*, 2013.

[35] Twitter Statistics from Statistic Brain. http://www.statisticbrain.com/twitter-statistics/, 2015. (Last accessed 2015/07/15).

[36] Twitter Usage. http://about.twitter.com/company, 2014. (Last accessed 2014/06/01).

[37] UDI-TwitterCrawl-Aug2012. https://wiki.cites.illinois.edu/wiki/display/forward/Dataset-UDI-TwitterCrawl-Aug2012, 2012. (Last accessed 2014/06/01).

[38] Hau wen Chang, Dongwon Lee, Mohammed Eltaher, and Jeongkyu Lee. @Phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM'12)*, pages 111–118, 2012.

[39] Benjamin P. Wing and Jason Baldridge. Simple Supervised Document Geolocation with Geodesic Grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT'11)*, pages 955–964, 2011.

[40] Yuto Yamaguchi, Toshiyuki Amagasa, and Hiroyuki Kitagawa. Landmark-based User Location Inference in Social Media. In *Proceedings of the First ACM Conference on Online Social Networks (COSN'13)*, pages 223–234, 2013.

[41] Xiangmin Zhou and Lei Chen. Event detection over twitter social media streams. *The VLDB Journal*, 23(3):381–400, June 2014.

[42] Xiaojin Zhu and Zoubin Ghahramani. Learning from Labeled and Unlabeled Data with Label Propagation. Technical Report CMU-CALD-02-106, Center for Automated Learning and Discovery (CALD), Carnegie Mellon University, 2002.