Limosa: A System for Geographic User Interest Analysis in Twitter

Jan Vosecky, Di Jiang, Wilfred Ng
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon
Hong Kong, China
{jvosecky,dijiang,wilfred}@cse.ust.hk

ABSTRACT

In this demonstration, we present Limosa, an interactive system mainly for visualization of geographic interests of users in Twitter. The system supports the modeling of comprehensive geographic characteristics of topics discussed in microblogs, both with respect to locations that postings originate from and also locations mentioned within the posting itself. Limosa then provides visualizations of geographic user interests, including the geographic scope of topics, terms, or the semantics associated with specific locations. Using a variety of recommendation strategies for exploration, we show that Limosa provides effective news and user recommendations.

1. INTRODUCTION

Since its foundation in 2006, Twitter has experienced an explosion in its user base, reaching over 500 million users as of 2012. At the same time, the increased popularity of mobile computing and location-based services has increased the need for effective methods for mining and organizing user interests with respect to geographic locations. Such information is increasingly used in applications such as personalization, news recommendation, the targeting of regional advertisements, and in other novel location-based services.

In this work, we develop an interactive system called Limosa, which focuses on analyzing the interplay between *geographic locations*, *users*, and *topics* in the Twitter environment and their rich implicit relationships. A facility for mining such relationships can, firstly, provide valuable global-level insights into the geographic interests of Twitter users. Secondly, it can provide a basis for topically-relevant and location-sensitive recommendation and personalization services for individual users. We now briefly discuss how the mentioned entities affect geographic user interest modeling in Limosa.

Users. Microblog users post short messages (referred to as 'tweets'), which discuss a wide range of *topics* and may relate to specific *locations*. It is thus interesting and challenging to mine geographic interests of a Twitter user.

Geographic Locations. One possible source of location information in Twitter are explicit geo-tags associated with tweets, referring to the location a tweet originates from. However, the proportion of geo-tagged tweets is still relatively low (cf. only 0.42% according to [1]). Moreover, there is unexplored potential in mining the locations discussed within the posting itself. Such location references may relate either to the user's *visited location* (e.g., "just arrived in new york") or *mention other locations* (e.g., mentioning about the Olympic Games 2012 in London).

Intuitively, semantics associated with *visited locations* may differ from semantics associated with *mentioned locations*. For example, let us consider user Bob in New York. On the one hand, Bob may be interested to know which topics are discussed by other users located in New York (e.g., business and politics). On the other hand, Bob may want to see topics that mention specifically about New York (e.g., restaurants in New York).

We focus on extracting these two location types from tweets and mining geographic user interests with respect to them. We also group the obtained locations into geographic regions, in order to obtain high-level geographic patterns.

Topics. In order to model semantic relationships between user interests and locations, we need to discover latent geographic interest topics (e.g., restaurants in New York). Each topic consists of individual terms (e.g., 'pizza', 'dinner', 'tortilla') and contains two geographic scopes: locations at which the topic is discussed (e.g., New York), and locations mentioned within the topic.

To address the above-mentioned issues, Limosa is developed to provide three facilities: (1) mining geographic Twitter user interests, (2) visualization of geographic user interests, (3) exploring content recommendation strategies based on the geographic interest model. We now highlight our approach for achieving these objectives.

How to mine geographic Twitter user interests? Firstly, we extract geographical information from Twitter users, both in terms of visited locations and mentioned locations. Secondly, we discover geographic interest regions, which correspond to clusters of dense geographic activity. Thirdly, we develop a new Geographic Twitter Topic Model (GeoTTM) that models latent geographic user interest topics. Using GeoTTM, we are able to incorporate both the semantic and the geographic scope of user interests.

How to visualize geographic user interests? Limosa presents an interactive user interface for visualization and analysis of the semantics associated with locations, geographic scope associated with users, topics and specific keywords.

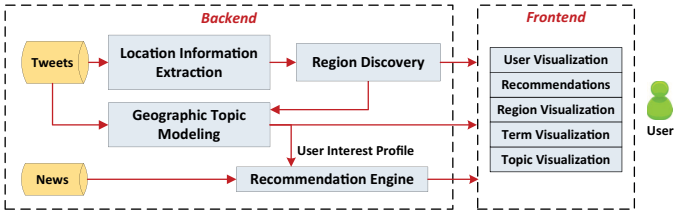

Figure 1: Architecture of the Limosa system

How to explore content recommendation strategies? Limosa adopts the user profiles obtained from GeoTTM for personalized news and user recommendations. Various recommendation strategies can then be further explored in the user interface.

Our system is novel compared with previous work on user interest mining in Twitter (e.g., [2, 3]). We are not aware of any existing system to analyze the interests of Twitter users from both topical and geographical perspectives. Limosa thus provides new insights into geographic user interest mining, visualization and location-sensitive recommendations.

2. OVERVIEW OF THE LIMOSA SYSTEM

In this section, we describe the individual components of Limosa. We broadly divide the system into *backend* components, which perform mining and modeling tasks, and the *frontend*, which is an interactive web-based application for analysis and visualizations. Figure 1 shows the architecture of Limosa.

2.1 Backend of Limosa

The backend of Limosa is implemented in Java and supports Twitter crawling and storage, location information extraction, geographic topic modeling and a recommendation engine.

Twitter crawling and storage: The system crawls Twitter user data using the Twitter REST API[1]. News articles are accessed using the Bing News API[2]. All data in the system is stored in a MS SQL Server 2008 database.

Location Information Extraction: This component extracts location information from individual tweets. Apart from utilizing GPS tags of tweets, we also extract location references from the tweet's content. Furthermore, we distinguish *visited locations* from *mentioned locations* in our extraction process. We use a feature-based supervised classification method for location extraction. The method consists of three steps: First, we perform a lookup of tweet's terms in a geo-database to identify candidate locations from the tweet. Second, we use a classifier to decide whether the candidate location refers to a real location or not. Third, we classify each location as either a visited location or a mentioned location.

The features used in both classification steps are listed in Table 1. Apart from several straight-forward features related to the tweet's characteristics and to the geographic information about the candidate location, we also propose a novel feature related to the tweet's terms. We use the intuition that certain terms are more likely to appear in tweets containing location references (e.g., 'hotel' might be likely to appear in a tweet mentioning the name of a holiday

[1]`https://dev.twitter.com/`
[2]`http://www.bing.com/toolbox/bingdeveloper/`

Table 1: The features used for Location Information Extraction

Tweet-specific features:
Tweet length (# of characters)
Tweet length (# of terms)
Position of location in tweet from the left
Position of location in tweet from the right
No. of URL links and hashtags in tweet
Specific of specific keywords (e.g. 'state', 'city')
Total no. of current mentions by user
Geographic features:
Type of location (e.g., city, state, country)
For multiple cand. locations within a tweet: Hierarchical ordering of locations? Do locations appear consecutively? Average distance between candidate locations in tweet
Term-location score (TLS) features:
Maximum TLS, Sum of TLS, Average TLS, Hashtag TLS

location). Our Term-Location Score (TLS) then captures the likeliness of a term to co-occur with location references within a tweet. To calculate the TLS, we firstly define an average term-location measure. For a set $P_t = \{p \in P : t \in p\}$ consisting of posts that contain term t, the Term-Location Score (TLS) is given by:

$$TLS(t) = \frac{1}{|P_t|} \sum_{p \in P_t} hasLoc(p) \times \log(|P_t|), \qquad (1)$$

where $hasLoc(p)$ is a function that returns 1 if post p contains a location reference, otherwise returns -1. The TLS of a term increases with more location co-occurrences and the TLS of a term decreases with more evidence of location absence.

Using the feature representation, we use a labelled dataset of tweets to train a SVM classifier for both classification steps. Our experiments have shown a 79.3% accuracy in terms of F-measure when extracting true location references and 74.4% accuracy for classifying locations as visited or mentioned.

Region Discovery: In this component, we discover geographic regions from discrete locations obtained from our dataset. First, we translate all location references to latitude and longitude coordinates. We then take a statistical approach to region discovery, instead of a taxonomy-based approach (e.g., using existing geographical classification into boroughs, cities, states and countries). For example, Washington D.C. is a city that is closely surrounded by Arlington city from its west side, located in a neighboring state. Our approach considers this agglomeration as a single region.

Specifically, we assume that each region follows a bivariate gaussian distribution over latitude/longitude points. We then use the widely adopted EM algorithm to estimate the parameters of a gaussian mixture model from the locations in our dataset. We discover two types of regions: (1) visited regions, obtained using geo-tagged tweets, and (2) mentioned regions, discovered using all mentioned locations extracted by the Location information extraction component.

Geographic Topic Modeling: Using the tweets and the associated geographic regions, we build a novel Geographic Twitter Topic Model (GeoTTM), which jointly models latent user interest topics with two geographic dimensions, incorporating both visited and mentioned geographic regions. The model also considers the social links between users as

implicit evidence of user interests. Although we do not provide the technical details of the topic model here, the generative process of a tweet can be summarized as follows:

1. For each user-document d, draw a document specific distribution θ.
2. For each tweet in d:
   - Sample a linked document c_i with proportion to weight of $link(d, c_i)$, then draw a document specific distribution θ_{c_i};
   - Combine θ and θ_{c_i} by tuning parameter λ to generate a document distribution θ_c
   - Sample a topic z according to the combined topic distribution θ_c;
   - Generate words $w \sim$ Multinomial (ϕ_z);
   - Generate the mentioned regions $r_m \sim$ Multinomial (Ω_z) if $X_i = 1$;
   - Generate the visited regions $r_v \sim$ Multinomial (π_z) if $Y_i = 1$;

The model serves two main purposes. On a global-level, the model captures user interest patterns with a geographical perspective. On the user-level, the model obtains the *user interest profile* of the i-th user as a topic vector $\{\theta_{i1}, \theta_{i2}, \theta_{i3}, ..., \theta_{in}\}$, where θ_{ij} is a real number indicating user i's endorsement for the j-th topic.

To train the GeoTTM, we collect a dataset of 143,284 tweets from 908 Twitter users selected from locations across the United States.

To aid interpretability when displaying latent topics from GeoTTM in the user interface, we use the following approach to assign names to latent geographic topics. From each topic, we select the 50 most important terms to form a 'topic-document'. We then calculate the inverse-document frequency (IDF) of all terms. From each topic, we then select 8 terms with the highest IDF value as the topic's name. We observe that this method results in more distinctive terms being selected as topic names.

Recommendation Engine: Based on the user interest profile from GeoTTM, Limosa can recommend news articles and similar users. The following strategies are adopted in the news recommendation algorithm:

- *Geographic proximity:* ranking based on the geographical distance between the user's visited locations and the news article's location.
- *Content similarity:* ranking based on cosine similarity between news articles and the user document, represented in the vector-space model by grouping all user's tweets.
- *GeoTTM relevance:* incorporating both topical and geographical relevance to rank news based on GeoTTM.

For user recommendation, the following strategies are implemented in Limosa:

- *Based on the location profile:* users ranked based on the similarity of (1) visited-location profiles, or (2) mentioned-location profiles, measured by the inverse KL-divergence of the location profile vectors.
- *GeoTTM relevance:* ranking based on the inverse KL-divergence of user interest profiles from GeoTTM.

2.2 Frontend of Limosa

The frontend is an interactive web-based interface for users to browse the geographical interests obtained by Limosa, visualizing the semantics associated with specific users, regions, terms, or topics. Users are able to choose different personalized recommendation strategies. The interface is implemented using Java Server Pages (JSP) and hosted on a Apache Tomcat server. Geographic visualizations are implemented using JavaScript, GoogleMaps API[3] and the HeatmapJS library[4] for heatmap rendering.

User Visualization (cf. Figure 2): This window shows the profile of an individual Twitter user. The user's distribution over topics is shown using a pie-chart, the user's visited and mentioned regions are shown using heatmap visualization. This window also provides news and user recommendations, according to the recommendation strategies (cf. Section 2.1).

Region Visualization (cf. Figure 4): This window shows the semantics associated with a selected geographic region. We display the region's scope using a heatmap visualization, top topics mentioning the region, and top topics of users located in the region. The top users associated with the region are also listed, both from the viewpoint of mentioning the region and visiting the region.

Topic Visualization (cf. Figure 3): Given a selected topic, the system displays its most important terms and associated visited and mentioned geographic regions. We also list top users associated with the topic and top related topics in the window.

Term Visualization (cf. Figure 5): Given an input term (e.g. 'apple'), we list the related geographic topics, thus providing information about the semantic and geographic patterns of the term's usage.

3. DEMONSTRATION PLAN

During the demonstration, attendees will be able to view all sections of the Limosa web interface, which are fully interconnected to provide an interactive user experience.

To walk through the functionalities of Limosa, we adopt a case-study approach: we pick user 'Jenna' as an example Twitter user and introduce the individual Limosa functions on the following scenario.

User Visualization. Firstly, we show how to learn more details about Jenna. We launch the User Visualization window in Limosa and select user Jenna from the list. Limosa then displays the basic information about Jenna, retrieved from her Twitter page: she is located in Manhattan, New York and works as a news reporter. We observe that Jenna's top topics are mostly related to politics and news-related issues. We then examine Jenna's geographic spectrum, shown using a heatmap visualization. On the map, we can toggle between visited and mentioned regions by Jenna. Visited regions include Manhattan, New York on the first place. Among the mentioned regions, the Middle East region has the highest importance.

The news recommendation function provides three lists of recommended news for Jenna. The *geographic proximity* list contains 10 news articles from New York. The *content similarity* list contains a mixture of political news from areas around the USA. The *GeoTTM similarity* list contains news, some related to New York and others related to politics.

[3] `https://developers.google.com/maps/`
[4] `http://www.patrick-wied.at/static/heatmapjs/`

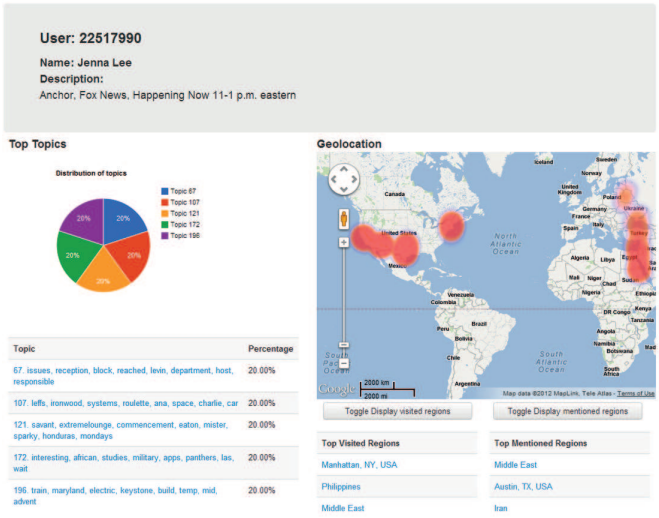

Figure 2: User visualization in Limosa, showing the user's topic distribution, visited and mentioned geographic regions.

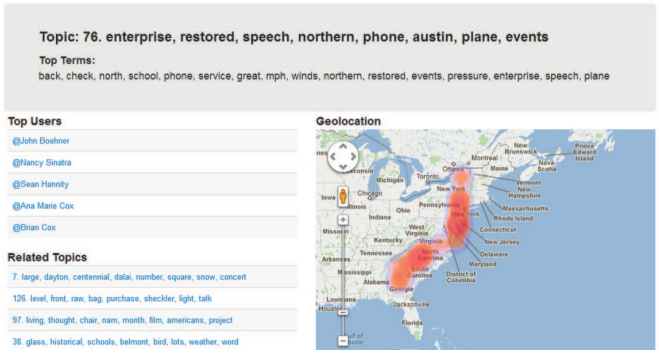

Figure 3: Topic visualization in Limosa, showing top terms, visited and mentioned geographic regions, top users and related topics.

Region Visualization: To learn about the interests of other users in Manhattan, New York, we turn to the Region Visualization Window. When the region 'Manhattan, New York' is selected, Limosa displays the semantics associated with this region. We see the top topics of users located in this region and top topics in which Manhattan is mentioned. We then examine the top users located in New York (ranked by the number of followers), which includes Jenna. The top users who mention Manhattan include other business and political figures.

Topic Visualization: To gain a further insight of the topics discussed by Jenna, we select the first topic in the list: 'enterprise, restored, speech, northern, phone, austin, plane, events'. Firstly, we examine the top terms from this topic and its geographic spectrum. The top visited regions are located at the East Coast of the United States, including New York. The top mentioned regions include the region of Germany, Italy and Switzerland, the US East Coast and California. Among the top 5 users in this topic, we note two news reporters, a politician and a political activist.

Term Visualization: To show Term visualization, we would ask attendees to investigate semantics associated with a given

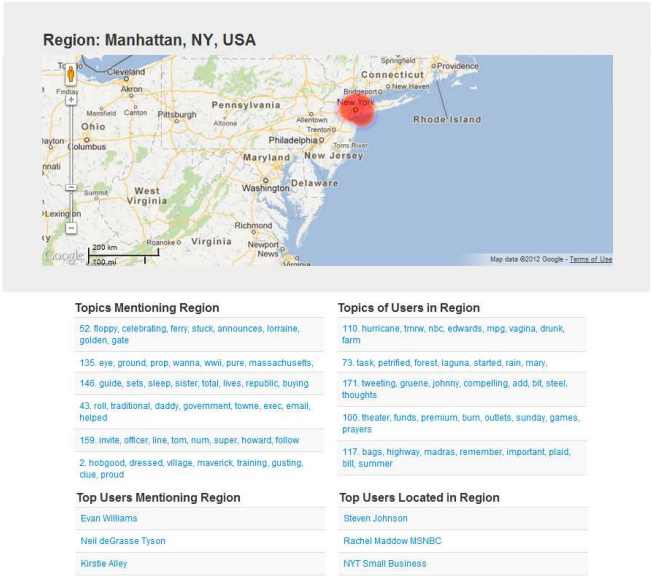

Figure 4: Region visualization in Limosa, showing top topics and related users

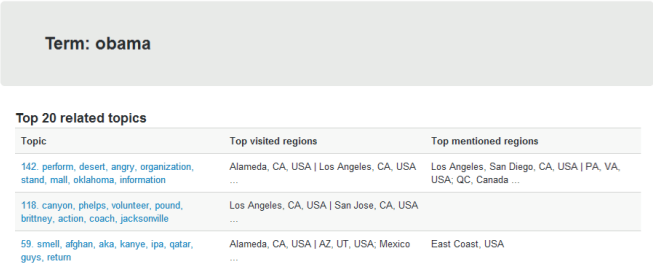

Figure 5: Term visualization in Limosa, showing top geographic topics related to a term

term. As an example, the profile of 'obama' lists the most relevant user interest topics associated with this term, which include a topic mentioning 'afghan, qatar, guys, return', discussed within California, a topic mentioning 'perform, angry, organization, oklahoma' (potentially referring to Pres. Obama's visit to Oklahoma), and others.

4. CONCLUSIONS

In this paper, we present Limosa, a system for analyzing geographic interests of Twitter users. The system mines geographic interest topics in Twitter and provides various visualizations via an interactive user interface, enabling users to explore geographic semantics from the perspective of users, topics, regions or specific terms. We aim to provide an interesting demonstration that allows users to explore various geographic recommendation strategies and their impacts.

5. REFERENCES

[1] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: A content-based approach to geo-locating Twitter users. In *Proc. of ACM CIKM*, 2010.

[2] P. Kapanipathi, F. Orlandi, A. Sheth, and A. Passant. Personalized Filtering of the Twitter Stream. In *Proc. of ISWC*, pages 1–8, 2011.

[3] K. Tao, F. Abel, Q. Gao, and G.-j. Houben. TUMS: Twitter-based User Modeling Service. In *Proc. of UWEB Conference*, pages 1–15, 2011.