Visualizing User-Defined, Discriminative Geo-Temporal Twitter Activity

Ingmar Weber and **Venkata Rama Kiran Garimella**
Qatar Computing Research Institute
iweber@qf.org.qa, gvrkirann@gmail.com

Abstract

We present a system that visualizes geo-temporal Twitter activity. The distinguishing features our system offers include, (i) a large degree of user freedom in specifying the subset of data to visualize and (ii) a focus on *discriminative* patterns rather than high volume patterns. Tweets with precise GPS co-ordinates are assigned to geographical cells and grouped by (i) tweet language, (ii) tweet topic, (iii) day of week, and (iv) time of day. The spatial resolutions of the cells is determined in a data-driven manner using quad-trees and recursive splitting. The user can then choose to see data for, say, English tweets on weekend evenings for the topic "party". This system has been implemented for 1.8 million geo-tagged tweets from Qatar (`http://qtr.qcri.org/`) and for 4.8 million geo-tagged tweets from New York City (`http://nyc.qcri.org/`) and can be easily extended to other cities/countries.

Introduction

Where do people tweet about doing sports? When do people check in to bars on Foursquare? Do visitors from different countries share pictures of different sites on Flickr? Creating an intuitive and interactive demo to explore such questions is challenging due to the many data dimensions involved, such as location, time, topic and language.

We present an interactive demo that gives end users a large amount of freedom in choosing the data to visualize and focuses on *relative differences* in activity patterns. Focusing on discriminative patterns avoids highlighting the generic high activity areas for all types of filters. Using Qatar and New York City as examples, we show how such a visualization can be used to discover, say, bars frequented by Western expats in Qatar or the pattern of rush hours in New York City.

What it Does

To use the demo, a user selects a set of filter criteria on the right sidebar such as a set of relevant languages (English, Arabic, ...), or a day of the week of interest (Monday, Tuesday, ...). Alternatively, they click on one of the examples suggested on the left. After clicking the "update" button on the top of the page, both the map view and the right sidebar are updated. Areas marked in red on the map show "hotspots" of higher than expected volume for the areas. Similarly, filter options with a higher/lower than expected matching volume are marked in red/cyan in the sidebar. Following feedback from test users, we chose not to show lower-than-expected "coldspots" on the map. Next, we discuss two illustrative examples.

Party life in Doha. Qatar is a Muslim country and for tourists the only points of alcohol sale are authorized night clubs and bars. These locations can be identified in our Qatar demo by selecting a filter for (i) English (as it is mostly Western tourists and expats frequenting these places), (ii) late evenings (which is the typical time for a visit), (iii) Thursdays and Fridays (which are the days preceding the non-working weekend on Friday and Saturday), and (iv) the topic "party". Manual inspection of the example tweets reveals that most of the locations found are indeed bars or night clubs. There are false positives related to sports clubs as the term "club" is in our topic dictionary for "party".

Rush hours in NYC. Selecting the topical filter "transportation" for the NYC demo, with no other filter activated, the map shows hotspots such as the area around Penn Station (See Figure 1). The sidebar on the right also shows that there is more than expected Twitter volume for this topic on weekdays and less than expected volume between 3pm and 6pm (indicated in cyan), before another relative increase during 6pm and 8pm (indicated in red). Though these rush hour patterns are to be expected, they provide evidence that our demo does indeed pick up societal level behavior.

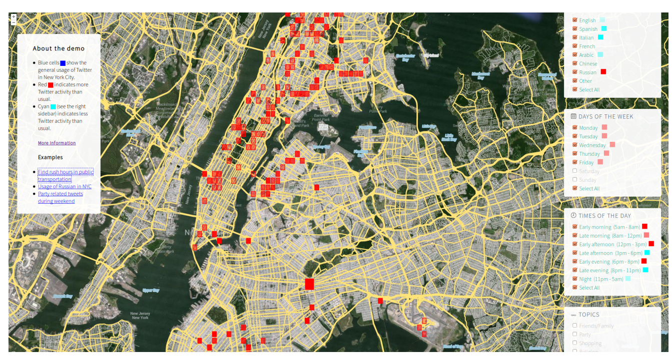

Figure 1: Screenshot of *http://nyc.qcri.org/* showing the relative distribution for tweets related to public transportation.

Data and Preprocessing

All Twitter data for the demos at `http://qtr.qcri.org/` and `http://nyc.qcri.org/` was obtained through Twitter's public APIs using both the streaming API[1]

Copyright © 2014, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]`https://dev.twitter.com/docs/streaming-apis`

for two queries defined by geographic bounding boxes and the REST API[2] to get additional historic tweets for a subset of users. In total, we obtained 1.8 million and 4.8 million tweets for Qatar and NYC respectively.

To enable an interactive user experience, we precompute certain counts for each "geographical cell". We use a data-driven approach to determine geographic boundaries as pre-existing boundaries might not match Twitter usage. Concretely, we use a recursive splitting approach based on quadtrees where, starting from a single big square encompassing all of Qatar (NYC), cells with ≥ 200 (500) tweets are recursively split while cells with < 50 (50) tweets are ignored. Recursion also stops once a minimum cell width/height of $.002$ (0.004) is reached. Our procedure ensures that, where possible, the spatial resolution is high whereas it spatial resolution is low where there is data sparsity.

For temporal analysis, tweets were mapped to a day of week and a discrete time of day (in "buckets" of 3-6 hours) using the tweets' timestamps converted to local time. The language of each tweet was identified using the Language Detection library[3]. The predominant languages in Qatar and NYC were included, but no efforts were made to detect mixed-language tweets or tweets written in transliterated form such as Arabizi or Pinyin. To detect the topic of a tweet, we manually compiled dictionaries of topic-specific word stems for each language such as "football*" (English) for sports or "desayun*" (Spanish) for food. This was done mostly for ease of deployment and other supervised, machine classification approaches could be used to improve in particular the recall of such a simple approach.

For each (geograpic cell, language, topic, day of week, time of day), combination the counts of all matching tuples are precomputed and stored in a database. We chose to use Google Fusion Tables[4] as it easily integrates with a Google Maps interface and with other web programming functionality. For each tuple we also store example tweets in a MySQL database and serve these through PHP to the web frontend. Note that the number of distinct tuples at the end does not directly depend on the number of input records but, rather, is a function of the number of distinct regions and the number of distinct values for each of the categories.

Deviation from Expectation

Our visualization compares the distribution of the data matching the user's filter criterion against an "expected baseline distribution". More-than-expected deviations are then marked in red and less-than-expected deviations in cyan. To establish the baseline distribution, we precompute for each (dimension=value) pair, such as (language=English), the total fraction of matching tweets. If, for example, globally 40% of tweets are in English or 5% pertain to a particular region, then this establishes a baseline distribution.

At query time, the user then specifies dimension=value pairs such as (language=English OR Tagalog, day of week=Friday) and a corresponding database query is issued. For the matching records and all possible (dimension=value) pairs the corresponding distributions are computed. For example, 30% of matching data records might be in English, 70% in Tagalog, 100% come from a Friday, 8% correspond to a particular geographical region and so on. These distributions are the compared against the baseline distributions. (dimension=value) pairs where the query is below the baseline are marked in cyan (30% vs. 40% English) and pairs where the query exceeds the baseline (8% vs. 5% for a particular region) are marked in red.

[2] `https://dev.twitter.com/docs/api/1.1`

[3] `https://code.google.com/p/language-detection/`

[4] `https://developers.google.com/fusiontables/`

Related Demos

The "Neighborhood Buzz"[5] interface is close to ours. Though it does not allow any temporal filtering it allows search by topic and highlights areas with large amounts of on-topic Twitter activity. However, their interface emphasizes absolute activity, rather than relative activity differences. As a consequence, a high volume area such as the "South of Market" district in San Francisco is highlighted for most topics ranging from Arts to Sports. Another closely related demo is the "Livehoods" project[6] that uses Foursquare check-in information to determine the topical geography of a city [Cranshaw et al., 2012]. Unlike our approach, they explicitly take mobility into account to determine the boundaries of the "livelihoods". Though they do have a "daily pulse" this information is for absolute activity and does not emphasize relative differences. Finally, they do not consider language distributions. Neither Neighborhood Buzz nor Livehoods allow a similar level of user input to select which data to visualize, but both are noteworthy interactive visualization alternatives. A large number of other visualizations of geo-tagged social media content exist. Though some overlap with ours in terms of functionality, none provides the same features as our demo with a focus on (i) user-defined drill-down options and (ii) highlighting deviations from a to-be-expected baseline.

Future Work

The current topic detection uses dictionaries in an attempt to achieve a high precision classification at the cost of sacrificing recall. Though the issue of topic classification of tweets is orthogonal to the actual visualization challenges, we plan to apply statistical methods such as multi-lingual topic models [Boyd-Graber and Blei, 2009], akin to the work in [Kling and Pozdnoukhov, 2012]. Our demo framework supports the integration of additional dimensions. In particular, we plan to integrate multi-lingual sentiment analysis, gender classification and nationality classification to study a multi-faceted society. Lastly, when using online data to study society one must of course not forget that younger, more affluent demographics are strongly overrepresented. In fact, the single largest nationality in Qatar – Indian – is largely missing in our data. Still, we believe that the current demo provides a useful approach to gain insights into society.

References

Boyd-Graber, J., and Blei, D. M. 2009. Multilingual topic models for unaligned text. In *UAI*, 75–82.

Cranshaw, J.; Schwartz, R.; Hong, J. I.; and Sadeh, N. 2012. The livehoods project: Utilizing social media to understand the dynamics of a city. In *ICWSM*.

Kling, F., and Pozdnoukhov, A. 2012. When a city tells a story: Urban topic analysis. In *SIGSPATIAL*, 482–485.

[5] `http://neighborhoodbuzz.knightlab.com/`

[6] `http://livehoods.org/`