VisCAT: Spatio-Temporal Visualization and Aggregation of Categorical Attributes in Twitter Data*

Thanaa M. Ghanem*§, Amr Magdy#§, Mashaal Musleh§,
Sohaib Ghani§, Mohamed F. Mokbel#§

§KACST GIS Technology Innovation Center, Umm Al-Qura University, Makkah, KSA
#Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN
*Department of Information and Computer Sciences, Metropolitan State University, Saint Paul, MN
thanaa.ghanem@metrostate.edu,{amr,mokbel}@cs.umn.edu,
{mmusleh,sghani}@gistic.org

ABSTRACT

In the last few years, Twitter data has become so popular that it is used in a rich set of new applications, e.g., real-time event detection, demographic analysis, and news extraction. As user-generated data, the plethora of Twitter data motivates several analysis tasks that make use of activeness of 271+ Million Twitter users. This demonstration presents *VisCAT*; a tool for aggregating and visualizing categorical attributes in Twitter data. *VisCAT* outputs visual reports that provide spatial analysis through interactive map-based visualization for categorical attributes—such as tweet language or source operating system—at different zoom levels. The visual reports are built based on user-selected data in arbitrary spatial and temporal ranges. For this data, *VisCAT* employs a hierarchical spatial data structure to materialize the count of each category at multiple spatial levels. We demonstrate *VisCAT*, using real Twitter dataset. The demonstration includes use cases on tweet language and tweet source attributes in the region of Gulf Arab states, which can be used for deducing thoughtful conclusions on demographics and living levels in local societies.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: [Spatial databases and GIS]

Keywords

Microblogs, Categorical Attributes, Visualization

1. INTRODUCTION

Twitter microblogging service becomes very popular in the last few years. Everyday, 500+ Million tweets are posted by 271+ Million active users [18, 20]. With such unprecedented user activeness and huge user-generated data sizes, several new applications and analysis tasks are motivated. This includes real-time keyword search [4], spatio-temporal modeling [1], event detection [2, 9, 11, 12, 13, 15], event analysis [7, 17], news extraction [14], photo extraction [6], and general analysis [8, 16]. The plethora of Twitter active users enables meaningful analysis tasks that can deduce fruitful conclusions for actual population. For example, a recent study on geotagged tweets [10] has examined the relation between cultural diversity and Twitter language usage in different countries using ground truth real data from international organizations. The results show a localized correlation in various societies with particular demographic characteristics and standards of living. The study has also shown a strong correlation between the tweets posted in country and its first spoken language. Thus, the availability of large amount of Twitter data from a wide active user base around the world motivates a variety of more accurate potential analysis tasks.

One of the underutilized Twitter data attributes are the *categorical attributes*: the attributes that can take one of multiple discrete values. Prime example of important categorical attributes in Twitter data is the language attribute that is appended to tweets by Twitter on February 2013 [19]. The language attribute determines in which natural language the tweet is written. This single attribute, along with geolocation information, allows the whole study in [10] and it enables even more analysis, e.g., about users language usage. Another example for Twitter categorical attribute is the tweet source, which determines from which OS, device, or application the tweet is posted. This is another attribute that could enable more analysis tasks like the spread of different devices in the geo-located space, analysis of standards of living in different regions,...etc. Thus, categorical attributes are important sources of Twitter data analysis that could be exploited to draw fruitful conclusions from the continuously flowing Twitter data.

In this demonstration, we present *VisCAT*—**Vis**ualization of **C**ategorical **A**ttributes in **T**witter—as a web-based service that enables users to analyze categorical attributes of Twitter data. *VisCAT* web interface facilitates choosing Twitter data from arbitrary spatial and temporal ranges and a particular categorical attribute to analyze. Then, a web-based visual report interface is generated for aggregate

*This work is supported by KACST GIS Technology Innovation Center at Umm Al-Qura University, under project GISTIC-13-06, and was done while the first, second, and last authors were visiting the center.

SIGSPATIAL '14, Nov 04-07 2014, Dallas/Fort Worth, TX, USA

http://dx.doi.org/10.1145/2666310.2666363 ...$15.00.

counts of different categories based on pie charts, per spatial region and at different zoom levels. To this end, *VisCAT* employs a hierarchical spatial pyramid structure [3] that materializes the count of each category. All the counts are pre-computed and stored in the different pyramid levels. The pyramid structure along with it aggregate counts are stored on disk and loaded on the launch of the report interface.

We demonstrate *VisCAT* using real Twitter dataset, showing use cases for language and source attributes of Twitter data, in the region of Gulf Arab states, during the period from December 2013 to February 2014. The rest of this paper describes *VisCAT* service in more detail along with the demonstration scenarios.

2. *VISCAT* OVERVIEW

In this section, we present an overview of *VisCAT* service features and components. First, we describe a detailed overview about the process of generating visual reports along with the supported features. Second, we describe details of the internals of the employed data structure in *VisCAT*.

VisCAT service is a web-based asynchronous service that generates visual reports for aggregate analysis on categorical attributes of Twitter data. It allows users to submit a request for specific visual report, takes its time processing the request in the back end, and then sends the visual report to the user by email. The requested visual report is characterized by two main things: (1) the data to be included in the report, and (2) the categorical attribute to be aggregated and visualized in the report. *VisCAT* users can select data in arbitrary spatial and temporal range. The available geotagged tweets in *VisCAT* are crawled since October 2013, using Twitter streaming APIs. For the user-selected data, the user can select any categorical attribute, e.g., language or source, to generate an interactive visual report.

Once the user submits a request, the back end of *VisCAT* extracts the user-selected data through extensive scanning for geotagged tweets in the specified temporal range. Then, *VisCAT* creates an adaptive pyramid structure [3] (similar to a partial quad tree [5]) that stores counts of different attribute categories in the whole spatial range at different levels of granularity. Building the pyramid structure goes through two phases: (1) *Structuring phase*, and (2) *Computation phase*. The structuring phase determines the pyramid shape by inserting the actual individual tweets. Then, the computation phase precomputes the aggregate counts in each pyramid cell before discarding the individual tweets. The pyramid is initialized by one root cell that covers the whole spatial range and contains all the tweets of the report. The root cell is then divided into four disjoint children cells, each covering a quarter of the space. The root cell tweets are replicated in its children cells according to their spatial locations. Any cell that has number of tweets larger than a parameter *capacity* is further divided into four children cells. The process is repeated recursively for each cell until the leaf cell has tweets less than or equal to *capacity*. When the structuring process is completed, the partial pyramid structure is then fed to the computation phase.

In the computation phase, the aggregate counts of all attribute categories in each pyramid cell, either leaf or non-leaf, are precomputed and stored. Each cell stores its counts in a hashtable with attribute categories as keys and the corresponding counts inside the cell as values. For example, if a certain cell has 80 tweets from iOS, 60 tweets from Android, and 40 tweets from Windows, then the cell hashtable would contain three pairs of $<iOS,80>$, $<Android,60>$, and $<Windows,40>$. Each cell stores two hashtables: one based on distinct tweets and one based on distinct users. The distinct tweets hashtable considers every individual tweet in the cell even if multiple tweets are posted by the same user. On the contrary, the distinct users hashtable counts all tweets from the same user only once. After the computation is completed, the pyramid structure is stored on disk with its aggregate counts. Afterwards, *VisCAT* back end generates an interactive web interface that visualize the contents of the pyramid structure on a map-based interface at different zoom levels, where each map level corresponds to a pyramid level. This interface represents the output visual report. Whenever the report is launched, the stored pyramid is loaded from disk to visualize the precomputed aggregations. After the report interface is successfully generated, an email is sent to the user with a hyperlink to the report. It is important to note that we take tweet language and source OS attributes classification from Twitter and it may contain some errors. The accurate classification of such attributes is beyond the scope of this work.

3. DEMONSTRATION SCENARIOS

VisCAT user interface and functionality are demonstrated using a real dataset that is continuously being crawled from Twitter streaming APIs since October 2013. Our demo attendees would be able to interact with *VisCAT* as explained in the following scenarios.

3.1 Scenario 1: Submitting Report Request

The main user interface of *VisCAT* is shown in Figure 1, from which users would be able to submit spatio-temporal requests to generate visual reports on categorical attributes in Twitter data. The user would select the report spatial area using a map-based interface (the black rectangle in Figure 1) and the temporal range through the datepicker. Then, the user chooses an attribute to analyze. Finally, the user enters an email address to receive the output report and submit the request. *VisCAT* would process the report request in its back end to generate a report similar to those presented in the next sections. Once the report is generated, the user receives an email message with a link to the generated report. It is worth noting that the interface in Figure 1 does not let the user input the pyramid cell *capacity*. *VisCAT* sets this by default to 50 which enables the spatial granularity to be street-level.

3.2 Scenario 2: Interactive Visual Reports

As described throughout the paper, *VisCAT* generates web-based visual interactive reports for tweets categorical attributes. Figure 2 shows two examples of the generated reports from *VisCAT*: Figure 2(a) shows the tweets languages in the Gulf Arab States and Figure 2(b) shows the tweet sources in the same region. In Figure 2(a) the counts and percentage of each language in each sub-region is displayed using pie charts. The size of the pie chart indicates the relative size of tweets in its corresponding region. Different languages are marked with different colors. Languages can be included/excluded selectively so that the user can compare the aggregates of any combination of the languages. In addition, changing the zoom level would give finer granularity aggregates in smaller regions. Also, the generated reports offers to show the aggregate counts based on either distinct tweets or distinct users. By default it uses the dis-

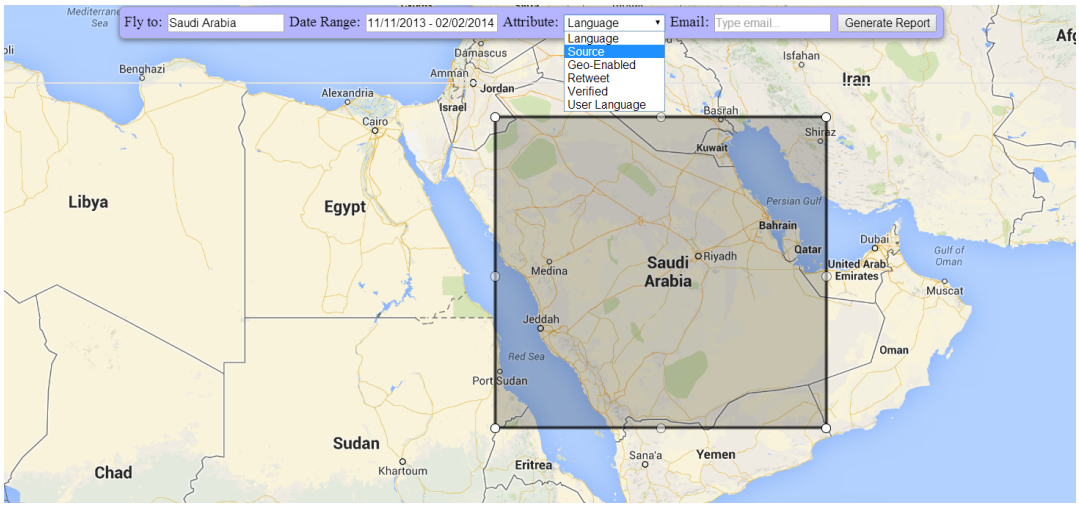
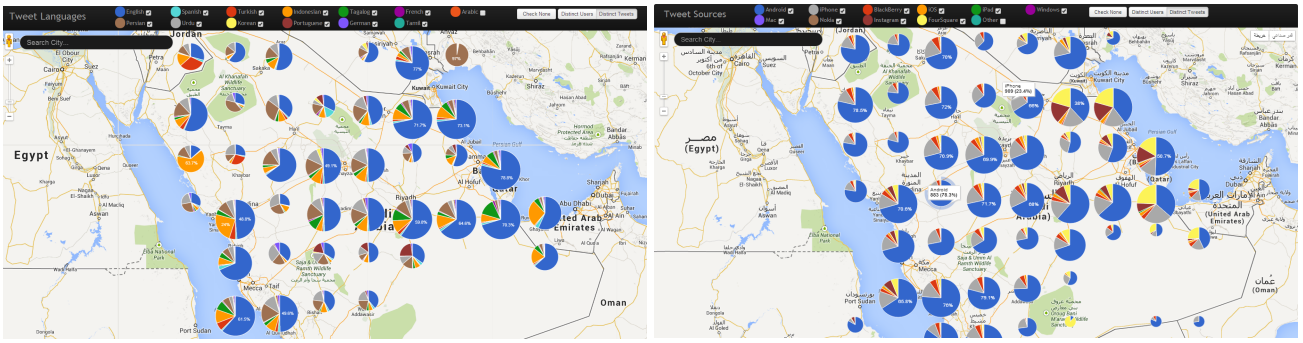


Figure 1: VisCAT Main User Interface.



(a) Tweets Languages (Arabic is disabled) (b) Tweets Sources

Figure 2: VisCAT Reports for Tweets Languages and Sources in Gulf Arab States

tinct tweets mode. To show the results based on distinct users, the user should click the corresponding radio button in the top right corner of the report screen. In this mode, all tweets from the same user are counted only once. The same features and description apply for Figure 2(b) where the only change is the attribute and its categories. It is clear in Figure 2(b) that android (in blue) is the mostly used OS in the region. Also, Foursquare (in yellow) is popular only in some of the eastern Arab regions. The reader can check the sample visual reports on http://www.gistic.org/TwitterLanguages/ and http://www.gistic.org/TwitterOS/.

3.3 Scenario 3: Distinct Tweets vs. Distinct Users

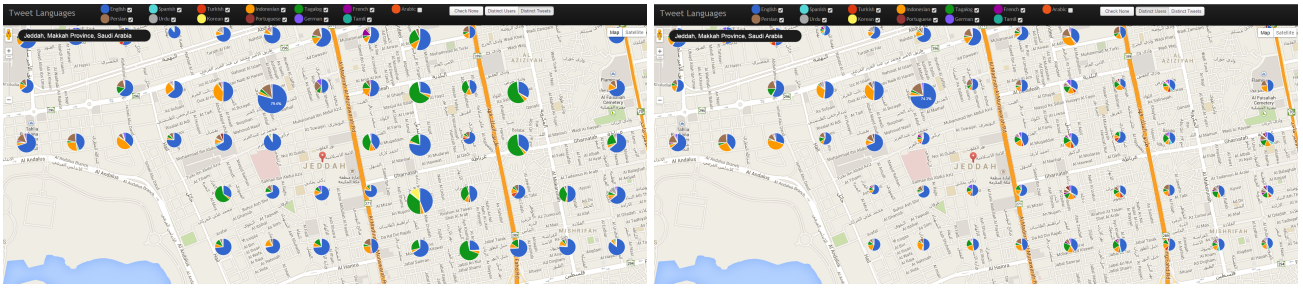
As described in Scenario 2, *VisCAT* users can display the aggregate counts by either distinct tweets or distinct users. Figure 3 shows the tweets languages distribution in city of Jeddah, Saudi Arabia using both display modes. The two modes could give insights on the activeness of certain users. For example, Figure 3 shows the activeness of Philippine users where the Tagalog language color (the green) is demoted in the distinct users mode. Also, the distinct user mode gives the portion of particular user group, e.g., Philippine users, in different regions in Jeddah. This eliminates the bias towards very active users who could mislead the report user.

3.4 Scenario 4: Localized Societal Analysis

The main purpose of enabling spatial analysis on Twitter categorical attributes is to facilitate interested users to draw fruitful conclusions by analyzing Twitter data in local areas. In this scenario, we give one example of such analysis and how it could be powerful in drawing conclusions, or even being a seed point to follow in future.

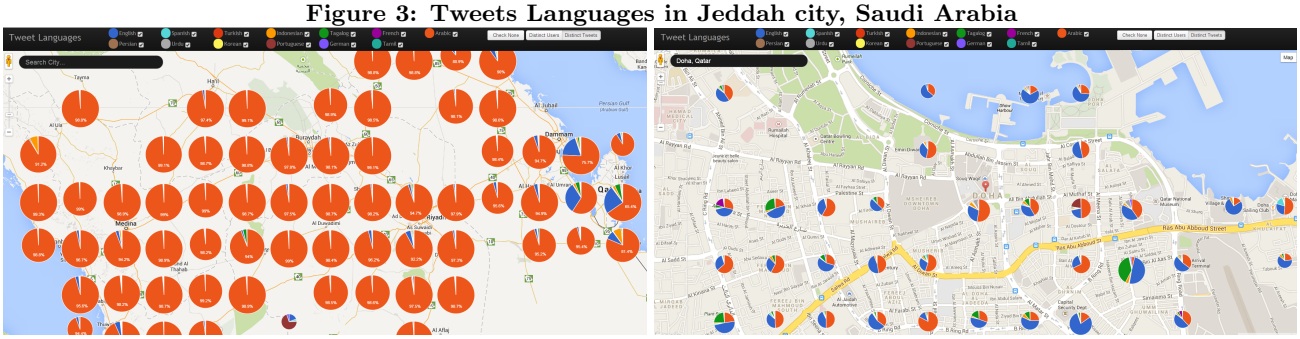
In the figures of the previous sections, whenever the tweets languages are shown in the Arab region, the Arabic tweets are excluded as it dominates all other languages which make it very hard to visualize the other languages contributions. Figure 4(a) shows the tweets language in the Arab area including the Arabic tweets. One can notice that the Arabic tweets are dominating in all regions except the eastern region near Qatar. Focusing on Doha, Qatar in Figure 4(b), one can see the lingual diversity of the city, compared to the other Arab region, where English, French, Spanish, Portuguese, Urdu, Indonesian, and Tagalog languages can be visually noticed. This gives a strong indication for high cultural diversity in the city. According to the study in [10], Qatar is the second state in terms of compliance of its Twitter data language usage to the UNESCO published language diversity index (LDI). UNESCO reported LDI of 0.608 for Qatar which means there is a 60.8% probability that any two random persons living in the country speak different languages. This can be visually noticed from Figure 4(b).

Such visual analysis may be of interest for different types of users, e.g., administrative authorities in the country to deal with certain situation for a specific cultural group like



(a) Based on Distinct Tweets (b) Based on Distinct Users

Figure 3: Tweets Languages in Jeddah city, Saudi Arabia



(a) Arab Gulf States (b) Doha, Qatar

Figure 4: Tweets Languages in Gulf Arab States vs. Doha, Qatar

Syrian refugees, new comers to multi-cultural countries who prefer to approach a community with a similar culture, or ethnicity-specific organizations that are interested to keep track of the spatial distribution of its people of interest.

4. REFERENCES

[1] H. Abdelhaq, M. Gertz, and C. Sengstock. Spatio-temporal Characteristics of Bursty Words in Twitter Streams. In *GIS*, pages 194–203, 2013.

[2] H. Abdelhaq, C. Sengstock, and M. Gertz. EvenTweet: Online Localized Event Detection from Twitter. In *VLDB*, 2013.

[3] W. G. Aref and H. Samet. Efficient Processing of Window Queries in the Pyramid Data Structure. In *PODS*, 1990.

[4] M. Busch, K. Gade, B. Larson, P. Lok, S. Luckenbill, and J. Lin. Earlybird: Real-Time Search at Twitter. In *ICDE*, 2012.

[5] R. A. Finkel and J. L. Bentley. Quad Trees: A Data Structure for Retrieval on Composite Keys. *ACTA*, 4(1), 1974.

[6] B. C. Fruin, H. Samet, and J. Sankaranarayanan. TweetPhoto: Photos from News Tweets. In *GIS*, pages 582–585, 2012.

[7] D. Grosvenor, J. Kendall, A. Sanders, and C.-T. Lu. Kongress: A Search and Data Mining Application for U.S. Congressional Voting and Twitter Data. In *GIS*, pages 550–553, 2013.

[8] Harvard Tweet Map. http://worldmap.harvard.edu/tweetmap/, 2013.

[9] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang. TEDAS: A Twitter-based Event Detection and Analysis System. In *ICDE*, 2012.

[10] A. Magdy, T. M. Ghanem, M. Musleh, and M. F. Mokbel. Exploiting Geo-tagged Tweets to Understand Localized Language Diversity. In *Proceedings of the International ACM Workshop on Managing and Mining Enriched Geo-spatial Data, GeoRich. In conjunction with SIGMOD*, 2014.

[11] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: Aggregating and Visualizing Microblogs for Event Exploration. In *CHI*, 2011.

[12] M. Mathioudakis and N. Koudas. TwitterMonitor: Trend Detection over the Twitter Stream. In *SIGMOD*, 2010.

[13] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *WWW*, 2010.

[14] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. TwitterStand: News in Tweets. In *GIS*, 2009.

[15] V. K. Singh, M. Gao, and R. Jain. Situation Detection and Control using Spatio-temporal Analysis of Microblogs. In *WWW*, 2010.

[16] Topsy Pro Analytics: Find the insights that matter. http://topsy.com/, 2013.

[17] TweetTracker: track, analyze, and understand activity on Twitter. http://tweettracker.fulton.asu.edu/, 2013.

[18] Twitter Data Grants, 2014. https://blog.twitter.com/2014/introducing-twitter-data-grants.

[19] Twitter Metadata. https://blog.twitter.com/2013/introducing-new-metadata-for-tweets.

[20] Twitter Statistics, 2013. http://business.twitter.com/en/basics/what-is-twitter/.