VacationFinder: A Tool for Collecting, Analyzing, and Visualizing Geotagged Twitter Data to Find Top Vacation Spots

Jalal S. Alowibdi[1], Sohaib Ghani[2], Mohamed F. Mokbel[3]

[1,2,3]KACST GIS Technology Innovation Center, Umm Al-Qura University, Makkah, KSA
[1]Dept. of Computer Sciences, , University of Illinois, Chicago, USA
[3]Dept. of Computer Science and Engineering, University of Minnesota, Minneapolis, MN
[1]jalowibd@cs.uic.edu, [2]sghani@gistic.org, [3]mokbel@cs.umn.edu

ABSTRACT

Choosing a location for vacations and weekends usually confuses many people. This concern has attracted considerable attention in recent years as currently there is no application based on actual visitors that helps people in finding out the top places for vacations. Online social networks such as Twitter are becoming very popular in last few years and can help in this regard. People nowadays generally do check-ins at new places. Also, analysis of tweets tagged with geolocation and time can provide trends of top vacation spots. In this paper, we present ***VacationFinder***; a novel location-based application that uses geotagged tweets to help people in where they should spend their holidays and weekends. We use real Twitter data crawled since October 2013. We apply indexing, spatio-temporal querying, and machine learning techniques to check, analyze, and filter the user activities in a particular country before and after a specific holiday. We then visualize the results and give our recommendations of top vacation spots for a particular holiday. The paper includes use cases on top vacation spots for Saudis in spring break of 2014 both inside as well as outside Saudi Arabia. Our application can not only help people but can also give direction to governmental agencies about promoting tourism in the country. It can also help law enforcement agencies, advertisement industry, and various businesses such as restaurants and shopping stores about where to focus during a particular holiday.

1. INTRODUCTION

With the evolution of technology, Online Social Networks (OSNs) services become very popular in the last few years. These OSNs daily generate a huge volume of highly informative data.The data is provided by end users using these services all over the world. A very famous example of OSNs is microblogging services e.g. Twitter and Facebook. Everyday, 500+ million tweets are posted by 255+ million active users [17], while 1.23+ billion Facebook users post 3.2+ billion comments [5]. With such extraordinary user activity and explosive growth in data sizes, several new applications and analysis tasks are motivated. As user-generated data, microblogs form a stream of rich data that carries different types of information including text, geolocation information, and users details. In addition, microblogs textual content is rich with user updates on real-time events, interesting keywords, news items, hyperlinks, images, and videos. This richness in data allows new queries and applications on microblogs. The examples for newly emerging applications on microblogs includes news extraction [3, 12, 13], event detection [1, 9, 11], spatial search [10], real-time keyword search [4], and analysis [6, 14, 15]. Such applications are becoming so popular that big companies are investing heavily to provide them to their customers [2, 16].

The deluge of Twitter active users enables different analysis tasks that can help in drawing fruitful conclusions for real world problems. For example, the evolution of Twitter has given a valuable spatio-temporal textual information that can be used as services for tourism [8]. Choosing a top location for vacations and weekends is usually not an easy decision for many travelers. Each family member has its own preference and generally decision is made based on a recommendation by any friend, the advertisement and reviews of vacations websites or any current famous location. Currently to our knowledge, there is no recommendation system based on actual visitors experience that is promoting services for tourism and keeping up to date information. In this paper, we are proposing a novel tool called *VacationFinder* that gives trends of most visited vacation spots. Our tool is a first attempt in this area and will promote a new era of offering vacation and tourism services by utilizing real OSNs data.

VacationFinder is based on the fact that geotagged tweets of users before and after the vacations can be used to find trends of the visited vacation spots during a particular holiday. Therefore, *VacationFinder* analyzes users profile before and after a vacation to find out trends of people visiting the vacation spots. Unlike existing approaches in finding the suitable and enjoyable sites to visit that depends heavily on the government resources, our approach in finding the top sites is based on the number of visits to that location. We leverage a varied collection of spatio-temporal features from OSN resources by using geotagged tweets. Thus, *Va-*

ACM SIGSPATIAL LBSN'14, November 4, 2014, Dallas, TX, USA
© 2014 ACM. ISBN 978-1-4503-3140-1/14/11…$15.00
http://dx.doi.org/10.1145/2755492.2755495

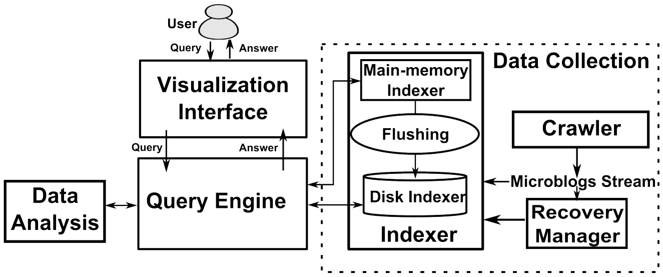

Figure 1: VacationFinder Architecture.

cationFinder is trying to uncover tourism of the countries by discovering and recommending the top sites for the holidays and weekends based on OSNs. This is done by using indexing, spatio-temporal querying, and machine learning techniques to check, analyze, study, and filter users activities in a particular area before and after a specific holiday. We then visualize the results to provide recommendations of top and famous vacation spots for a particular holiday. This can serve government agencies for promoting tourism services, commercial organizations for promoting advertisements, law enforcement for providing security, and others for social reasons.

VacationFinder uses a novel approach of using OSNs to provide trends of top vacation spots (i.e., tourism sites) for a holiday, based on the number of visits to that place. We validate our work by examining different classifiers over a large dataset of Twitter profiles. We use real Twitter data crawled since October 2013. The paper includes use cases on top vacation spots for Saudis in spring break of 2014, both inside as well as outside of the Saudi Arabia. We also match the results found by our application with the government of Saudi Arabia statistical data and found high correlation. The rest of the paper describes *VacationFinder* architecture along with the usage scenarios in detail.

2. DESIGN OVERVIEW

VacationFinder is an application that collects Twitter data and analyzes user profiles inside the data before and after a vacation to find trends of vacation spots visited during that vacation. Our approach to find a suitable location for holidays and weekends is a location-based approach. By location-based approach we mean the user has to select a source location—either country, city, or any area—in the beginning and our application then analyzes the user profiles belonging to that location and give the trends for the vacation spots. We choose this approach as it allows us to focus on a specific country or particular area. It also allows us to filter a huge dataset. However, our technique also works without any input location information.

We study the user profiles by checking the profile's spatio-temporal activities before and after the vacation. We analyze these profiles and save the information for users who were out on vacations. Finally, we visualize the results. We also compare the results with the tourism information provided by the government of that country.

Overall, there are three steps involved in our application namely, data collection, data analysis, and data visualization. Figure 1 shows the main architecture of the *VacationFinder*. Below we describe each module in detail.

2.1 Data Collection

In general, Twitter allows us to collect about 2% of the daily available tweets. This result in about 10 million tweets collected daily. Twitter also allows to collect about 50% of the geotagged tweets. For the data collection, we write a crawler using Twitter streaming APIs. Our crawler has been collecting geotagged Twitter data since October 2013. For this work, we opt for geotagged data only as we need spatial information to identify locations visited by users.

Twitter profiles consist of about 30 features containing biographical and other personal information. However, many features in the form are optional, meaning that they are often left blank. In this work, we are only interested in saving users profile detail, tweets text, and spatio-temporal features (i.e. coordinates and the posted time) of the tweets.

The data is saved in our own database system which consists of three main components, namely, indexer, query engine, and recovery manager(Figure 1). Indexer efficiently digests incoming data in light main-memory indexes. The index consists of temporally partitioned index segments, where each segment organizes its data based on spatial attribute. We choose spatial and temporal attributes for indexing as they are most important part of the geotagged tweets and can also provide effective pruning of the search space. When the memory becomes full, a subset of the main-memory data contents are flushed to disk indexes which manage billions of microblogs. The recovery manager restores the memory contents from backup copies in case of memory failure. Query engine generates an optimized query plan to efficiently retrieve data from the indexes. It supports a wide set of generic interactive queries to answer any query with spatial, temporal, or keyword attribute.

2.2 Data Analysis

Data analysis module receives two datasets of before ($D1$) and after ($D2$) the break periods from the query engine, according to the query submit by the user on the front-end interface. It then matches the unique users from these datasets and return the results to the interface for visualization. Below, we explain the data analysis for the case to find out top vacation spots for Saudis outside Saudi Arabia during spring break, 2014.

The spring break in Saudi Arabia started on March 20th, 2014 and lasted for a week. We extract users activities 10 days before the spring break—March 10th to 19th, 2014—as well as during the spring break—March 20th, 2014 to March 26th, 2014—according to the period specified in the query. In all, the dataset consists of around 200 million tweets and about 4% to 5% unique user profiles (i.e., about 10 million user profiles).

The analysis is started with dataset $D1$ and tweets that belong to the Saudi Arabia region are filtered. The new dataset DA contains 1.2 million tweets. There are around 300,000 unique users extracted from DA. The collected 300,000 unique users are then applied to the dataset $D2$ in order to match and extract all the tweets that come from the same users but in different time (i.e., the spring break) and without restrictions (i.e., all the tweets with no particular restriction to specific country's coordinates). The new matched dataset is named as $D2_B$. Statistically, there are around 100,000 unique users in $D2$ that matched DA. Furthermore, the dataset of $D2_B$ contains around 1.5 million tweets. All the tweets in the dataset $D2_B$ that contains Saudi Arabia's coordinates are then removed—as for the cur-

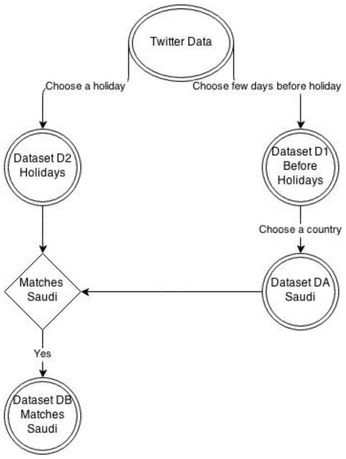

Figure 2: Data Analysis Information Flow.

rent case we need vacation spots outside of Saudi Arabia—and the resulting dataset is called DB. The dataset DB contains around 35,000 unique users. Figure 2 shows the information flow of the data analysis. From the datasets DA and DB, users location before and during the break are saved. We use GraphML format to save source and destination locations. For users with multiple destinations, we chose only the final destination.

2.3 Data Visualization

VacationFinder has an interactive front-end interface that takes queries from the user. Users in the query choose the source and optional destination country, or they are also allowed to interactively choose the source or destination areas they are interested in over the map. The query also contains two temporal periods of before and after any vacation. The query is dispatched to the query engine through Java based APIs that allow efficient interaction. The query engine then extracts the require data from the indexes and give it to data analysis module. After the analysis, the results are provided to the interface in the form of GraphML file.

The next step is to visualize this information. We use simple node-link diagram with curved edges to visualize this data. Our visualization displays source and destination locations and an edge between them to show where the user went during the holiday. It is important to note here that for large datasets, our visualization may get cluttered. For this problem, we propose to use edge bundling technique [7] to remove the clutter and make visualization more appealing. We are also planning to allow users to group the data based on spatial areas such as regions, continents, etc.

3. USAGE SCENARIOS

In this section, we present two usage scenarios to explain how the user can find top vacation spots for any country and for any particular holiday. The section includes use cases on top vacation spots for Saudis in spring break of 2014 both inside as well as outside of the Saudi Arabia.

3.1 Scenario 1: Top Vacation Spots for Saudis Outside Saudi Arabia

This scenario is about finding top vacation spots outside of Saudi Arabia visited by Saudis during spring break of 2014. The user first issues a query with Saudi Arabia as a source location and temporal periods of before and after the spring break. The resulting data is then visualized in the interface. Figure 3 shows the visualization of countries where the Saudis spent their spring break of 2014. It shows that Gulf countries along with United Kingdom, Indonesia, and Turkey are the most visited countries during the spring break. For people visiting multiple countries, we only pick the final destination for simplicity.

According to study by Saudi Tourist Information and Research Centre[1], published by SABQ Online Newspaper[2], the top 10 destinations for about 6 million Saudis during 2013 are: United States of America, United Kingdom, Malaysia, Gulf Cooperation Council Countries excluded Saudi Arabia, Indonesia, Philippines, Turkey, Morocco, Australia, and Switzerland. Similarly, our application almost matches the study by the Saudi Tourist Information and Research Centre and shows that the top 10 destinations are: Gulf Cooperation Council Countries excluded Saudi Arabia, United Kingdom, Indonesia, Turkey, United States of America, Egypt, Australia, Malaysia, France, and Spain. This comparison leads us to better understanding where the Saudis are spending their vacation. Also, this high correlation shows that our technique could be efficiently used to find top vacation spots for any particular holiday.

Moreover, we listed all the countries that were visited by Saudis in ascending order. We found 215 unique countries. Also, there are around 34 countries that have been visited by at least 10 unique Saudis. There are 1482 Saudis that visited unknown location because we identified the location to be located in the ocean. Furthermore, countries issue travel advisories for their citizens. For example, currently according to the government of Canada[3], there are 12 potentially dangerous destinations. Likewise, our experiment found that there are 62 unique Saudis that visited these places during spring break of 2014. Such scenarios could be very useful for various government agencies.

3.2 Scenario 2: Best Vacation Spots Inside Saudi Arabia

In this scenario, we present vacation spots inside Saudi Arabia visited by Saudis during spring break of 2014. We queried and analyzed the data as described previously. The only additional thing is that we provided destination source as Saudi Arabia. For such scenarios, matching is done between two datasets of before and after the break, based on the fact that the user tweeted before and during the spring break from within Saudi Arabia and his locations are different. We also group the source locations that are present within the same city boundaries. We found out that Riyadh, Makkah, Madina, Jeddah, Dammam, and Abha are the most visited cities during the spring break. Also, we found various attraction places located outside the cities. Some of those attractions are located in the north desert, wildlife sanctuary, and northeastern desert. We think these places are of interest to travelers because during spring season people go to such places to enjoy the greenery that is difficult to find in Saudi Arabia.

4. DISCUSSION

We present *VacationFinder* as a tool to find out top vacation spots during a particular holiday. Unlike tourism websites, which gives a list of vacation spots to be visited all

[1] http://www.mas.gov.sa/

[2] http://sabq.org/yX6fde

[3] http://travel.gc.ca/travelling/advisories

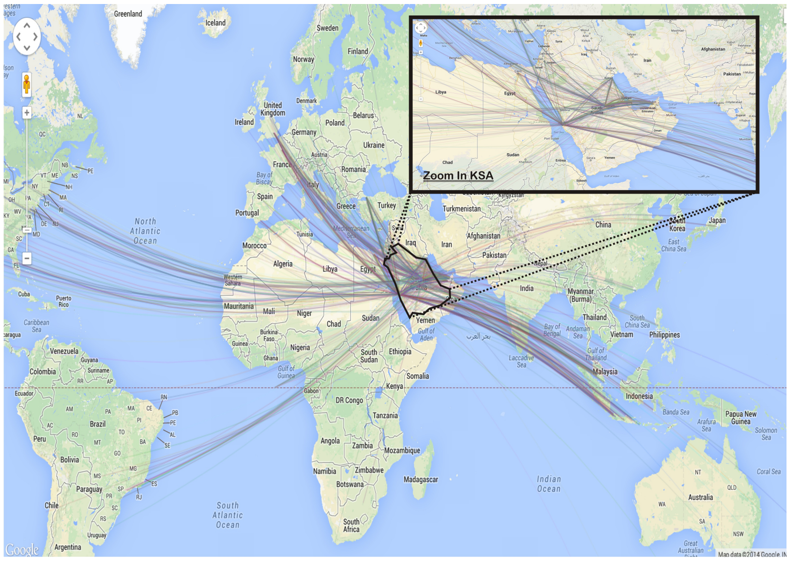

Figure 3: Where did the Saudis Spent the Spring Break of 2014.

round the year, our tool gives top vacation spots based on a particular holiday by using Twitter data. *VacationFinder* can be used by the government agencies to promote tourism in the country. It could also be used to do planning before any vacation. Similarly, media industry can use this application to find out where to focus for a particular holiday. Other businesses such as restaurants and shopping stores can use the information to open new stores and also to find out how many workers needed during a specific holiday.

In this work, we have not focused on the validation of the data. It may be possible that the user has falsely mentioned his spatial location. As some people check-in places to just show that they are visiting that place. The validation of the user spatial location is outside the scope of this work. We have also not currently used tweet text in our application. It is an important feature, and we will use it in the future.

Also, we used streaming APIs to collect 2% of the Twitter data. For geotagged tweets, we only collect 50% of the total geotagged tweets. It may be said that these numbers do not represent the whole dataset and may not give accurate results. However, we are only giving trends of visited spots, and as shown by the usage scenarios, when we compare the results with ground truth data, our results show good trends of top vacation spots.

5. REFERENCES

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

[2] Apple buys social media analytics firm Topsy Labs. http://www.bbc.co.uk/news/business-25195534, 2013.

[3] After Boston Explosions, People Rush to Twitter for Breaking News, 2013. http://www.latimes.com/business/technology/la-fi-tn-after-boston-explosions-people-rush-to-twitter-for-breaking-news-20130415,0,3729783.story.

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

[5] Facebook Statistics, 2012. http://newsroom.fb.com/Key-Facts/Statistics-8b.aspx.

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

[7] D. Holten. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):741–748, 2006.

[8] J. Krumm, R. Caruana, and S. Counts. Learning likely locations. In *User Modeling, Adaptation, and Personalization - 21th International Conference, UMAP 2013, Rome, Italy, June 10-14,*, pages 64–76, 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, M. F. Mokbel, S. Elnikety, S. Nath, and Y. He. Mercury: A Memory-Constrained Spatio-temporal Real-time Search on Microblogs. In *ICDE*, pages 172–183, 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] O. Phelan, K. McCarthy, and B. Smyth. Using twitter to recommend real-time topical news. In *RecSys*, 2009.

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

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

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

[16] New features on Twitter for Windows Phone 3.0. https://blog.twitter.com/2013/new-features-on-twitter-for-windows-phone-30.

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