Top US City Clustering Based on Twitter Hashtags

Mengling Hettinger
Department of Physics and Astronomy
Michigan State University
zhangme6@msu.edu

Yanjie Zhao
Chemical Engineering and Materials Science Department
Michigan State University
zhaoyanj@msu.edu

ABSTRACT

This paper addressed the problem of similarity analysis and clustering of big cities in US based on the newly collected hashtag data from Twitter website. The problem was proposed since the ever increasing efficiency of virtual communication through internet provides a good platform for people to analyze the behavior of certain communities and use this data either for business or research purpose.

Hashtags of Tweets sent by Twitter users were collected in top 20 most densely populated cities in US for 4 weeks through the website provided API. Our goal was to cluster the cities based on how frequently the hashtags were used during the data collection period of time and analyze the similarities of the cities that fall into the same cluster. Jaccard similarity was chosen to calculate the distance, and the reason was rather than Euclidean distance calculation, Jaccard similarity gives more importance to the appearance of words than the absence while comparing the word list of multiple cities.

Both K-means and Hierarchical clustering algorithms were used to conduct the analysis. From the result, K-means did not suggest an optimal cut for the number of clusters. When Hierarchical approach was applied, the dataset with top 200 hashtags gives a more spread out results and more well defined clear-cut due to the relatively small number of attributes. Data visualization was implemented by Google Fusion Table API combined with Google Maps API. The result displayed in Google Maps implied that the location of cities was consistent with the cluster distribution, which means, cities that are close to each other tend to fall into the same cluster. This phenomenon was more obvious in the Hierarchical clustering result which indirectly supported the validity of the clustering algorithm.

Keywords

Twitter, Hashtags, Clustering, K-means, Hierarchical Clustering, Google Fusion Table

1. INTRODUCTION

Twitter is a massive social networking site tuned towards fast communication. More than 140 million active users publish over 400 million 140 character "Tweets" every day. Twitter's speed and ease of publication have made it an important communication medium for people from all walks of life. Twitter has played a prominent role in socio-political events, such as the Arab Spring and the Occupy Wall Street movement. Twitter has also been used to post damage reports and disaster preparedness information during large natural disasters, such as the Hurricane Sandy [1]. Because of the influence and power of this popular networking, we should take full advantage of the data generated by millions of Twitter users every day and try to discover the hidden information from the Tweets.

Clustering is a widely used method in social networking analysis, which is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Due to the fuzzy-group feature of social networking and the non-absolutely-correct-answer property of clustering analysis, clustering method provides a proper way to study the structure and the performance of the networking.

In this paper, we collected twitter hashtag data from 20 top cities in US with the largest populations and applied two most popular clustering methods – K-means and Hierarchical algorithm – to analyze the data. There are some issues we encountered during this project:

a. The data was collected during 4 weeks without stopping and to guarantee the continuity and the quantity of the data we had to check the task running status in the server every one or two days manually. That's because sometimes the tasks could be killed without any notification due to the maintenance or some other technical reasons;

b. The data was collected in separate text files for different cities to keep it clear. Because of this, we had to run Hadoop analysis with 20 files, and after that we had to preprocess the data to trim the hashtags and make an easily processible matrix. The whole process was done as much as possible in Hadoop so it would not be too time consuming, and this took us lots of effort on writing the Hadoop code;

c. The data visualization was implemented with Google Fusion Table/Google Maps API combined with some JavaScript code. The most difficult part was to display the menu of "Cluster Method" and enable the corresponding changes of the clusters in Google Maps.

This paper addressed the problem of city clustering based on the hashtag data collected during March 12th and April 9th from Twitter website. The contributions made in this paper include:

a. Real world big data mining was implemented using Hadoop system;

b. Multiple data processing procedure was applied with several languages and tools such as Hive, Python, R and Matlab to neaten the data and make the computation more efficiently.

c. Clustering was conducted with two methods and a comprehensive result was displayed in Google Maps in the project website.

The remaining of the paper is organized as follows: Section 2 describes the basic information about the structure of the data and the preprocessing work required. Section 3 provides the details about the computation process towards clustering and the analysis of the results. Section 4 discusses the results from different clustering methods and explores the potential correlation between the clusters and other information. Section 5 draws the conclusions and proposes some aspects that could be considered for further improvement of research on this topic.

2. PRELIMINARIES

To keep the data in the minimum size, the raw data only included the timestamps and the hashtags. The final objective for data processing was to generate a matrix with 22 columns: the first column would be the hashtags, the second column would be the count of the total number of cities in which each hashtag appeared, and the other 20 columns would be the exact count for the corresponding hashtags in each row from all the cities.

Data cleaning is required in this case since the data were input by users and is quite noisy. The crucial cleaning processes were listed below:

a. Remove the hashtags if they only contain stop words, punctuations or numbers. Stop words were looked up in a stopwords dictionary.

b. Remove stem endings and punctuations at the end of each hashtag. A list of ["s", "es", "ed", "er", "ly" "ing", "'s", "s'"] were used in stem endings.

c. Remove the unicodes from the hashtags.

d. Convert all the letters to lower case to avoid duplication. After these steps, there were 4019343 tweets left to be analyzed.

e. Remove the hashtags that only appear in one city or appear in all cities but one to avoid bias towards outliers.

f. Normalize the each count of hashtag.

g. Rely on the total counts of hashtags appear in the city so that cities such as New York who have much more hashtags than average will not be considered as outliers due to the large distance measurement.

Preprocessing will be conducted with Python and clustering will be computed mainly in R. K-means (k=3~5) and Hierarchical clustering algorithm (k=2~5) will be applied.

3. ANALYSIS

The cities chosen to collect data from and the geographical information of them are listed below:

1. New York (-74.00 40.71); 2. Los Angeles (-118.24 34.05); 3. Chicago (-87.62 41.87); 4. Houston (-95.37 29.76); 5. Philadelphia (-75.16 39.95); 6. Phoenix (-112.07 33.45); 7. San Antonio (-98.49 29.42); 8. San Diego (-117.15 32.71); 9. Dallas (-96.80 32.78); 10. San Jose (-121.89 37.34); 11. Austin (-97.74 30.26); 12. Jacksonville (-81.65 30.33); 13. Indianapolis (-86.15 39.76); 14. San Francisco (-122.42 37.77); 15. Columbus (-83.00 39.96); 16. Fort Worth (-97.33 32.75); 17. Charlotte (-80.84 35.22); 18. Detroit (-83.05 42.33); 19. El Paso (-106.49 31.77); 20. Memphis (-90.049 35.15).

The data collection process took 4 weeks (starting from 03/12) and overall more than 4 millions Tweets were collected. The data were collected using twitter API Python interface [2] and was stored in plain text files. The geographical region for each city was determined by using the longitude and latitude of the city center and expanding the area with a range of 0.6 degree longitude and 0.6 degree latitude.

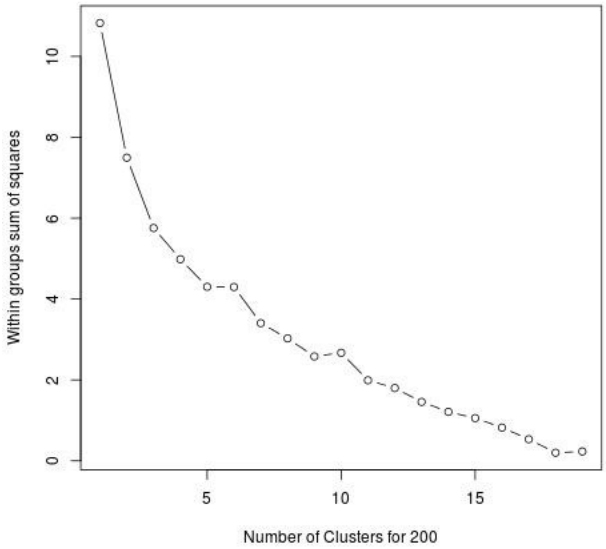

Figure 1. SSE for k means clusters using different k for top 200 hashtags from each city

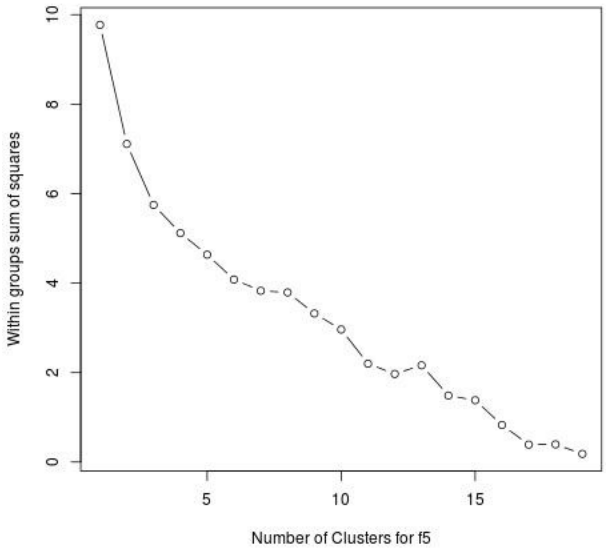

Figure 2. SSE for k means clusters using different k for top 200 hashtags from each city

We first tried to create a subset of the hashtags where at least one of the city's frequency of a specific hashtag has to be large than 5 for each word. This gave us an overall result of 26804 hashtags. We then only used the top 200 most frequently appeared hashtags of each city to represent each city. This approach gave us 1210 hashtags. After removing the hashtags that only appear in one city and appears in all cities but one, the former set had 19095 words left while the later set had only 430 left.

Due to the large dataset, the prototype-based approach was first used. In order to obtain the number of clusters, we looped over

the number of clusters and measured the SSE to determine the optimal solution. Neither of the datasets showed an obvious plateau to suggest a best cut of number of clusters (Fig 1&2).

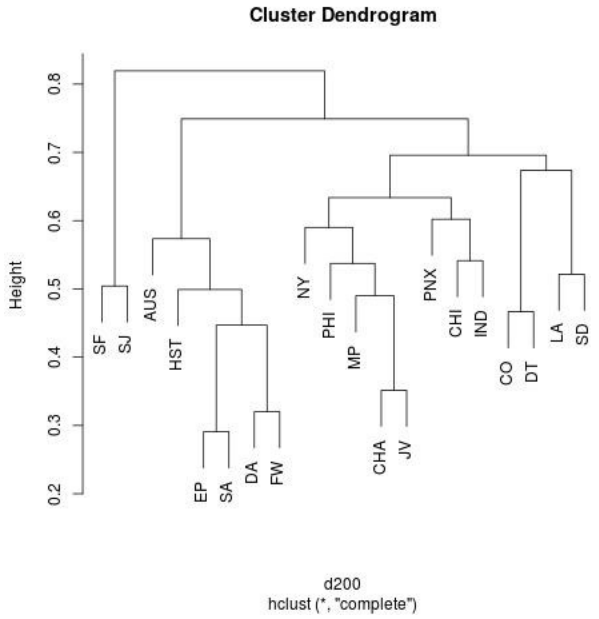

Figure 3. Clusters using top 200 hashtags from each city (complete measurement)

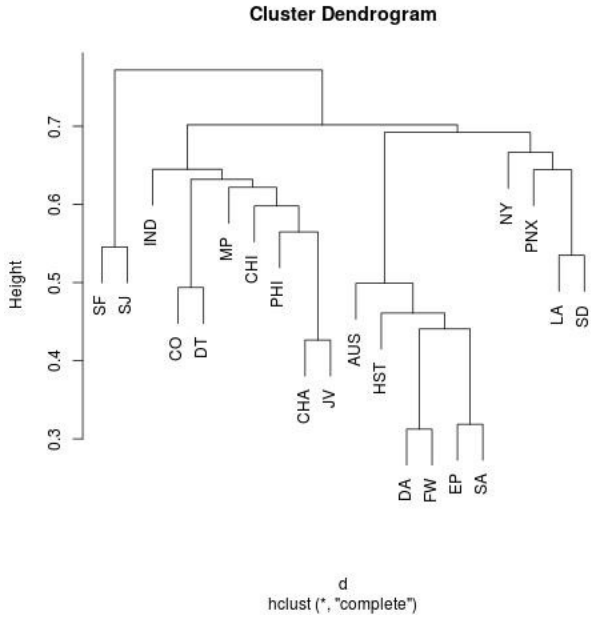

Figure 4. Clusters using hashtags that appear at least 5 times in at least one city (complete measurement)

Due to the uncertainty of the analysis by using K-means, Hierarchical clustering method was also applied. Both parts of the analysis were completed by using the statistics language R. R is a free software programming language and software environment for statistical computing and graphics. The stat package in R includes both K-means implementation and Hierarchical cluster implementation. The Jaccard distance measurement was used. Fig 3 is the result for the top 200 words in each city dataset. Fig 4&5 are the results for the dataset where words appear at least 5 times in at least one of the cities. Both complete and single measurements were used. The results show that the dataset with top 200 hashtags gives a more spread out results and more well defined clear-cut due to the relatively small number of attributes.

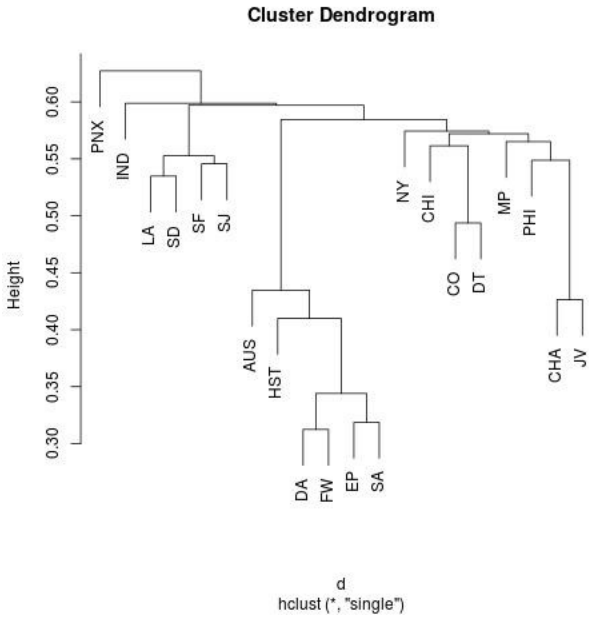

Figure 5. Clusters using hashtags that appear at least 5 times in at least one city (single measurement)

4. DISCUSSION

The results based on K-means gave a similar geographical separation as the ones shown in hclust. We looped over k from 2-19 for our datasets and found that k = 3 to k =5 illustrate the most interesting results. For the k=3 K-means, cities in California all grouped together, all the east and cities in the middle as well as Phoenix, AZ grouped together, the rest cities mainly in Texas fell into the third group. While k = 4 K-means showed a similar result but further split north Texas (Dallas and Fort Worth) from South Texas(San Antonio, Austin, El Paso and Houston). When increasing k to 5, K-means showed a more messy result in terms of geographical location. Cities with larger populations, such as New York, Chicago, Phoenix and Los Angeles tend to group together. Detroit and Columbus their own were classified as a small group. Cities in the Middle East belong to one cluster (e.g. Charlotte, Jacksonville, and Memphis).

Hierarchical structure by using hclust can be further cut into different k's. For k=2, clearly, San Francisco and San Jose grouped together. It is likely that due to the high density of high technology companies located in Silicon Valley area. When increased to k = 3, Texas cities were all split from the rest of the cities (other than San Francisco and San Jose). Interestingly, Phoenix, Los Angeles and San Diego grouped with the cities in the East instead of Texas or San Francisco area. Further increasing k to 4, South California and Michigan/Ohio area teamed up and the result stayed the same as the previous case. When increasing k to 5, South California and Michigan/Ohio further split into the 5th group.

Throughout the analysis, we did not see a strong correlation between the population and topics in the 20 cities we selected.

Data visualization was implemented by Google Fusion Table API combined with Google Maps and webpage was written by JavaScript. The demonstration can be found in this website: www.cse.msu.edu/~zhaoyanj/citycluster

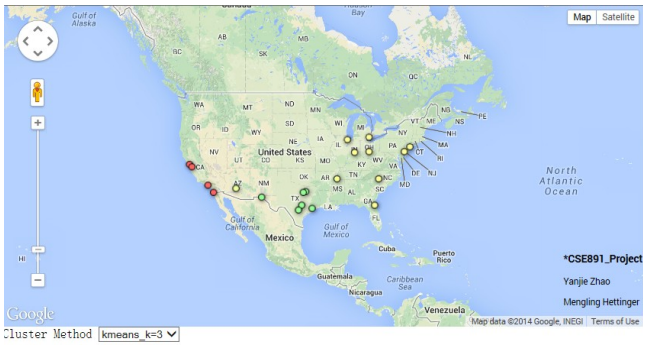

Figure 6. Clustering result visualization in Google Maps

5. CONCLUSIONS AND FUTURE WORK

In this paper, the big city clustering problem based on the hashtag data collected from Twitter is proposed. We first preprocessed the data with Python followed by word count tasks conducted in Hadoop MapReduce system, and then constructed the final matrix with Hive query language for the subsequent similarity calculation and clusters computation in R. The clustering implementation was accomplished with two algorithms – K-means and Hierarchical clustering. Finally the comprehensive results were displayed in Google Maps through the connection with Google Fusion Tables. Correlation between clusters and geographical locations was observed however no obvious impact of population of cities was found on the clustering result.

For the future work, temporal data can be used to present the clustering analysis. Timestamps for weekdays and weekends can be analyzed separately since it is very likely that people are interested in different topics during these two periods of times. We can even further split the weekdays' data into morning, afternoon and evening and split weekends' data into these three categories as well. This may give us a valuable result because people do different things during those 6 different times. For example, people may talk about TV shows during the weekday's evenings and we can further discuss if different TV shows have the different popular level cross the United States. Wester TV shows which involves cowboys, cattle ranchers, miners and farmers may be popular in the mid-west area like Texas, while shows take place in New York like How I Met Your Mother may be more popular in new England area.

More cities can be used to improve our results. There are many cities with their unique features can be used, e.g. Seattle, WA, Portland, OR, Orlando, FL and Miami, FL. Another direction of this project leads to a collection of hashtags all around the world. People in the cities of Europe may surprisingly be interested in total different subjects as the ones in US.

Last but not the least, the project can be further developed into a streaming analysis. Tweets and hashtags are constantly collected from Twitter API for different cities, fed into the algorithms and update the Google Maps every few minutes. In this way, people can even check our website and see the topics people are talking about in the different areas in US.

6. REFERENCES

[1] Shamanth Kumar, Fred Morstatter, Huan Liu. *Twitter Data Analytics.* SpringerBriefs in Computer Science. 2014

[2] Twitter API python interface used in this paper. More detail, please see: https://github.com/tweepy/tweepy

[3] For K means implementation in R, please see: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html

[4] For Hierarchical clustering implementation in R, please see: http://stat.ethz.ch/R-manual/R-patched/library/stats/html/hclust.html