# Track Trajectory of Topic on Temporal Geo-Textual Data

Zeng Shijia
School of Physics and Mathematical Sciences

Prof. Cong Gao
Dr. Chen Zhida
Mr. Liu Shang
School of Computer Science and Engineering

***Abstract - Big data which are geo-tagged and contain textual, temporal information are being produced on a rapid scale. In this study, we study one type of query, where given a keyword, the goal is to find the spreading path of this topic on large temporal geo-textual data. This query measures spatial proximity, semantic similarity, and time recency of geo-textual objects. To sufficiently support this query, we group data based on a clustering algorithm. Based on the clustering result, we retrieve the top-k centroids and link them on temporal order to build the trajectory of the query topic. To evaluate the efficiency of our algorithm, we conduct experiments on a relatively large collection of geo-tagged tweets with timestamp. The experimental result reveals that our proposed method could achieve a high accuracy with high efficiency.***

**Keywords –** temporal-spatial data, data mining, clustering

## 1 INTRODUCTION

Amount of data are generated in a rapid speed which contain location, text and time information. This kind of data usually could provide first-hand information for a variety of local breaking news. For instance, each Tweet may include 140 characters at most with timestamp and usually are tagged with locations. Tweets are also powerful for spreading news of important events [10]. We define such data with text, temporal and geographical information as *temporal geo-textual objects*.

Considering the booming of temporal geo-textual data and their vast topic coverage, it is of great significance to mine useful information from such data. For example, Taal volcano was erupted in January 2020 in Philippines [2], which has triggered a wide discussion among the public. We could track the path of the volcanic ash by mining information from a large amount of geo-tagged Tweets such that we could inform the public to avoid these areas for safety concerns. The topic can also be extended to typhoon, flooding etc. In these applications, the public may receive faster and first-hand information regarding to some specific topics.

The contributions of this study are summarized as follows:

- Abstract and introduction. This part briefly explains the background information and objective of the project.
- Literature review. This part introduces some existing work related to this project.
- Problem definition. We define and study the problem at this part, which explains the import definition, theory and concepts of the project.
- Method. We develop an efficient algorithm to solve this problem, which consists of clustering, ranking and linking.
- Results and discussion. Our empirical study shows that our method is efficient with high accuracy.
- Conclusion and future work.

## 2 LITERATURE REVIEW

In this section, we first introduce K-Medoids Clustering, which is an unsupervised machine learning method. K-medoids clustering is a variant of K-means but it is more sensitive to noises and outliers. K-medoids uses a real point in the cluster to represent it as a centroid while using K-means, the centroid may not be an actual point in the dataset [3]. K-medoids is more applicable in our study than K-means as our objective is to find locations.

In natural language processing, algorithms such as word2vec are useful methods for computing vector representations of words from large datasets. They have been used to embed words into a vector space in order to catch semantic relationships [6]. In our study, doc2vec [4] is adopted as it could represent sentences, paragraphs and documents as a paragraph vector. This method is more align with our experiment objects – Tweets, which usually contain sentences and even paragraphs.

## 3 PROBLEM STATEMENT

In this section, we formally define the Temporal geo-textual object, Top-k retrieval query, Cluster problem and Score problem.

## 3.1 TEMPORAL GEO-TEXTUAL OBJECT:

We represent a temporal geo-textual object as o = <t, λ, p>, where t is timestamp, λ is text, p is a location point with latitude (denoted as o.p.lat) and longitude (denoted as o.p.lon).

We define temporal geo-textual objects as geo-tagged tweets in Twitter in this study.

## 3.2 TOP-K RETRIEVAL QUERY:

Define a query q = <pt, φ, k>, where pt is specified past time, φ is a set of query keywords, k is the number of results to be maintained.

## 3.3 CLUSTER PROBLEM

### 3.3.1 Temporal Geo-textual Cluster:

A temporal geo-textual cluster c is defined by a set of temporal geo-textual objects.

### 3.3.2 Cluster Problem:

With a set of temporal geo-textual objects, the Cluster problem groups these objects based on distance proximity, text relevance and time recency. The cost function is defined as Equation (1):

$$C = \sum_{m_i}\sum_{o \in m_i} \begin{matrix} (\alpha \cdot \mathrm{D}(o.p,\, m_i.p) + \beta \cdot \mathrm{T}(o.t,\, m_i.t) \\ +(1-\ \alpha - \beta) \cdot \mathrm{S}(o.\lambda,\, m_i.\lambda)) \end{matrix}$$

, where $m_i$ is an object in the cluster, i $\in 1, 2, \dots, \mathrm{k}$ and $\alpha$, $\beta$ are parameters which balance the importance between distance proximity, text relevance and time recency.

Note that $\mathrm{D}(o.p, m_i.p)$ is spatial distance difference, which is defined as Equation (2):

$$\mathrm{D}(o.p,\, m_i.p) = \ |o.p.lon - m_i.p.lon\ | \\ +\ |o.p.lat - m_i.p.lat|$$

$\mathrm{T}(o.t,\, m_i.t)$ denoted as temporal difference, where $\mathrm{T}(o.t,\, m_i.t) = \ |o.t - m_i.t|$ (3).

$\mathrm{S}(o.\lambda,\, m_i.\lambda)$ is textual similarity, where we can first embed text into vector spaces using doc2vec [4] method. Then we can use textual similarity measurements such as cosine similarity to compute a similarity score ranging from 0 to 1.

## 3.4 SCORE PROBLEM:

Score function is defined as Equation (3), which calculates ranking score of each cluster:

$$R = \ \theta \cdot |C_i| + \gamma \cdot \sum_{o\ \epsilon\ C_i} e^{-\mu|curT - o.t|} \\ + (1 - \theta - \gamma) \cdot \sum_{o\ \epsilon\ C_i} S(o.\lambda, q.\varphi)$$

, where $\theta$, $\gamma$, $\mu$ are parameters. $|C_i|$ represents the size of the cluster, *curT* denotes the current time, and $e^{-\gamma|curT-o.t|}$ is an exponential decaying function which favors the more recent objects. Time recency plays an essential role in geo-textual data. For example, tweets are usually correlated with events, however, they would be less relevant to the query topic over time [9]. This exponential decaying function is widely used [1], [5], [9] to measure the recency of geo-textual data. $S(o.\lambda, q.\varphi)$ denotes textual similarity between query topic and the text of each object in each cluster.

# 4 METHOD

Given a set of data which has attributes of time, location, text and query keywords we want to track. We aim to develop an algorithm to find the spreading path of this topic from large temporal geo-textual data.

Step 1: Cluster all tweets on semantic, geographic distance and time with defined cost function using K-Medoids unsupervised machine learning method.

Step 2: Calculate average score of each cluster with query topic based on cluster size, textual similarity and time recency.

Step 3: Retrieve top-k clusters with highest scores and their centroids. Then link them on temporal order to build the trajectory of the query topic.

# 5 RESULT AND DISCUSSION

## 5.1 DATASET DESCRIPTION

A sample dataset of 16,205 geo-tagged tweets on the topic of Taal Volcano with timestamp was provided. The data spans from Jan 12, 2020 to Jan 30, 2020. Fig. 1 demonstrates the common words in this dataset by applying some visualization techniques (e.g., “Word Clouds”).

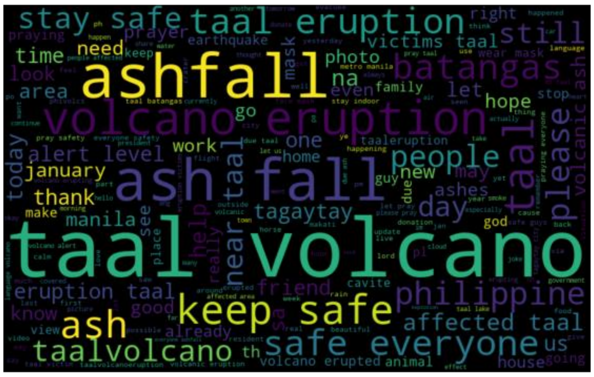

Fig. 1, Common words over temporal geo-textual data

## 5.2 RESULT ANALYSIS

Considering a query topic “volcanic ash”, our algorithm spends 1152.51s grouping 20 clusters and 0.36s retrieving 7 top locations. With a visualization of the spreading path on this topic (Fig. 2), the “volcanic ash” started from “Taal Volcano Natural Park” and spread all the way north. This is compatible with the real ash spreading path.

DBSCAN, a density-based clustering [7] was also tested in Method Step 1, however, due to the varying densities and sensitivity to parameters, it could not separate clusters well.

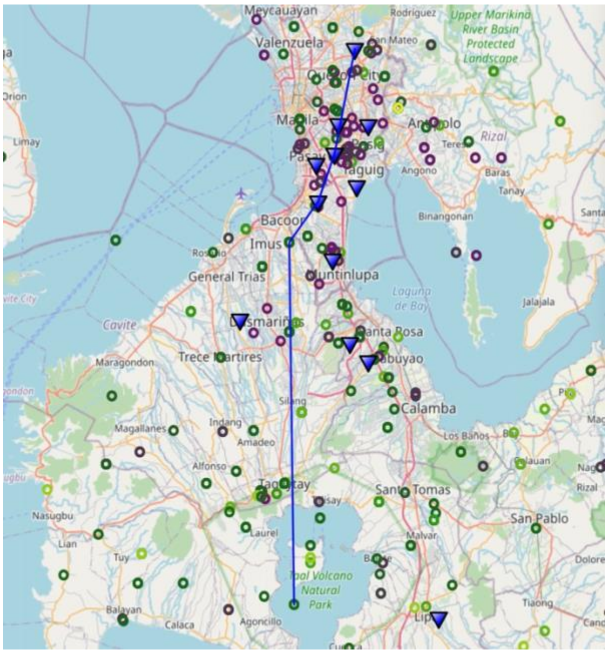

Fig. 2, Experiment result: spreading path on topic “volcanic ash”

## 6 CONCLUSION AND FUTURE WORK

In this project, we introduced a kind of algorithm which comprises of an unsupervised clustering machine learning method to find the spreading path of a given topic. The experiment result reveals that our method is able to achieve a high accuracy with high efficiency.

However, this project has some limitations: 1. This project only tested on one dataset, it can be tested on different scenarios for comparison purpose. 2. K-medoids clustering algorithm was adopted for this project, while a faster K-medoids clustering [8] could be considered for future work to further improve clustering efficiency. 3. As using K-medoids clustering method, the number of clusters need to be initialized and this parameter is sensitive to the clustering results, other clustering methods could be tested in future work.

## ACKNOWLEDGMENT

I would like to express my gratitude to my URECA supervisor, Prof. Cong Gao, who provides me with the opportunity to do this project and great guidance. I would also like to express my gratitude to Dr. Chen Zhida and Mr. Liu Shang, who offered me a lot of support and were always available for clearing my doubts.

I would like to acknowledge the funding support for this research project from Nanyang Technological University – URECA Undergraduate Research Programme.

## REFERENCES

[1] Amati, G., Amodeo, G., & Gaibisso, C. (2012). Survival analysis for freshness in microblogging search. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management - CIKM 12*. doi:10.1145/2396761.2398672

[2] International Organization for Migration. (2020, January). *Taal Volcano Eruption 2020: Philippines - Calabarzon Region Situation Report No. 2 as of 21 January 2020*. https://reliefweb.int/report/philippines/taal-volcano-eruption-2020-philippines-calabarzon-region-situation-report-no-2-21

[3] Jin X., Han J. (2011) *K*-Medoids Clustering. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_426

[4] Le, Q. & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *Proceedings of the 31st International Conference on Machine Learning, in PMLR* 32(2):1188-1196

[5] Liang, H., Xu, Y., Tjondronegoro, D., & Christen, P. (2012). Time-aware topic recommendation based on micro-blogs. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management - CIKM 12*. doi:10.1145/2396761.2398492

[6] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ICLR*.

[7] Ram, A., Jalal, S., Jalal, A. S., & Kumar, M. (2010). A Density Based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases. *International Journal of Computer Applications, 3*(6), 1-4. doi:10.5120/739-1038

[8] Schubert, E., & Rousseeuw, P. J. (2021). Fast and eager k-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms. *Information Systems*, *101*, 101804. https://doi.org/10.1016/j.is.2021.101804

[9] Shraer, A., Gurevich, M., Fontoura, M., & Josifovski, V. (2013). Top-k publish-subscribe for social annotation of news. *Proceedings of the VLDB Endowment*, *6*(6), 385–396. https://doi.org/10.14778/2536336.2536340

[10] Zhao W.X. et al. (2011) Comparing Twitter and Traditional Media Using Topic Models. In: Clough P. et al. (eds) Advances in Information Retrieval. ECIR 2011. Lecture Notes in Computer Science, vol 6611. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-20161-5_34