2022 International Conference for Advancement in Technology (ICONAT)
Goa, India. Jan 21-22, 2022

Event Detection using Geo-tagged Twitter Information

Srividhya Ganesan
Senior Assistant Professor, Computer Science and Engineering
New Horizon College of Engineering
Bangalore, India
v.vidhya8@gmail.com

Dr. Rachana P
Associate Professor, Computer Science and Engineering
New Horizon College of Engineering
Bangalore, India
dr.rachana@newhorizonindia.edu

Dijan Maharjan
Computer Science and Engineering
New Horizon College of Engineering
Bangalore, India
dizmaharzan@gmail.com

Sagar Panth
Computer science and Engineering
New Horizon College of Engineering
Bangalore, India
sagarpanth.np@gmail.com

Alok Prasad Kurmi
Computer science and engineering
New horizon college of engineering
Bangalore, India
patelaalok75@gmail.com

2022 International Conference for Advancement in Technology (ICONAT) | 978-1-6654-2577-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICONAT53423.2022.9726130

***Abstract*—A main challenges in finding the desired event in the locality is that there is a lot of unstructured data. It required a lot of work to make sure all those huge data to align with the output that we desire. We can say that discovering what is happening around the user is also so important for other application also. It can also help in monitoring crimes, predicting the natural disasters, protests. Many people post status and what's happening in the real time data in their social accounts, they act as a human sensor. With the increasing in population and decrease in recreational landmarks, many people undergo the path with adjustment in order to attend event they are interested in. While posting the status people not only post status but also their location along with it. In this project using that geo tag, we are extracting information about that event happening in that locality and one can get detailed information regarding the event. We are doing novel framework to find the events in current time from geo tagged twitter information and keep record of such event over time. It collects these data in huge amount from twitter and find meaningful candidates for the description of events. After collecting all those data by using clustering keyword it narrows down the data with the help of the geo tagged location similarity.**

Keywords— ***Online Event Detection, Social Network, localized events, Geo-tagging, Quad - tree***

I. INTRODUCTION

The rapid increase in numbers and adapting the GPS enabled gadgets is leading to the expansion of georeferenced social sites. In todays date almost every application requests permission of the location for their application. The reason behind this is that while posting the status these data can be use in multiple number of places and for many purposes also like monitoring crime, predicting the natural calamities before and minimize casualties. This enormous amount of data can be used as source of event related information. Everything we do nowadays, we share it in social media in real time example every concert we attend, every While sharing, we only not share text messages but also time and location of the event occurred. In this project we took twitter as a source of data. Twitter is a social media platform that millions of users use to share updates about their lives. User uploads photos and tweets in the twitter, these tweets are about localized events happening around the user. The collection of these each single tweet is clustered using keywords and narrow it down to managed and ordered data that make a lost sense than the initial data that are collected roughly. Though there are a lot of news channels that makes a report on local events but it takes time for a single reporter to investigate, collect all the data in real time and report it back to channel into broadcasting ready news especially when compared to the length of the event.

A. Event detection with specific location

There are many aspects to detect the event like statuses, time of post or the event. We use time and geo tagged location of enormous amount of data to detect and predict the event happening around. After collecting the twitter data, a large amount of data like this needs to be narrow down to increase the accuracy of detecting the event. Events are discovered by clustering the geographically tagged tweets statuses using the summarized keywords in each cluster. User can select specific locations to find the events. Though social media like twitter allows user to upload pictures, texts, and statuses using geo tag (longitude and latitude), only 1% - 3% posts or tweets are tagged with their current location per day[12]. According to twitter, around 7 million posts or tweets are posted along with geo tagged per day. This activity where small number of posts are published online with geo tag makes difficulty in accuracy of event detection in real time.

Another main task in this area of research is about the way the terrestrial area is divided for following event detection.

1. A fine method for selecting the grid size is essential, which has not been effectively covered in literature.
2. Fixed grid cells may not benefit in finding local and global events.

Example, by means of a low-resolution grid for real time data might take only the global events happening around the country, while smaller events are detected by high resolution, i.e. within the locality in the ranges from 1 kilometers – 50 kilometers[1] . We can also select POIs, where each grid contains each point of interests. Next in this method, we can set the number of POIs based on tweets distribution density.

In this scenario, city center areas can have many Point of interests with small area, while places away from the city can have less Point of interests with large grid size. But having fixed Point of interests bounds the position of spotted events to the chosen Point of interests only. The physically chosen Point of interests is mandatory for every geo graphical area of investigation.

II. PROPOSED WORK

Our proposed work is different from the projects that are done before in following ways:

i) The existing projects on normal event detection goals to perceive normal events that are debated on social sites without recognising the community of these events. Furthermore, our planned method focuses to spot particularly location - specific events, that are local events happening in real time in a particular area.
ii) In addition, the supervised methods to normal event detection needs indirectly characterised events that might not generalized to new events, while our work does not need clear event tags

III. RESEARCH METHODOLOGY

A. Web scraping

Web scraping basically extracts huge amounts of data from websites for a variety of uses such as price monitoring, enriching machine learning models, financial data aggregation, monitoring consumer sentiment, news tracking, etc. Browsers show data from a website. The subsequent step is to perceive how Web Scraping can be beneficial for your business. There are quite a few conditions in which Web Scraping can be very effective, beginning with the era of income possibilities that we talked about earlier. By defining the key phrases to use, how the Internet is crawled, and adjusting your search tools, you can correctly generate leads. Data mining is no longer solely accurate for producing leads[2]. The emails you acquire from a Web Scraping mission are beneficial for advertising purposes. Now that electronic mail advertising can be entirely automated, you can switch the extracted information without delay to electronic mail advertising equipment like Active Campaign or Mail Chimp to create an extra high-quality campaign[9].

Web scraping is the method of extracting a large amount of data from various websites automatically. It also collects unstructured data and forms structured data. Web scraping with Python is just one of the ways to scrape websites. Otherwise, you can use APIs or online services. Web scraping is usually done for Research purposes. Then for social media, email lists, jobs, and competitor analysis. Some websites allow web scraping while others don't. Web scraping itself is not illegal, but these data cannot get into wrong people because it can used against them also so people should be careful on how they use the technique without offending or harming anyone[8].

B. Localized event detection

It is not necessary that all the events are celebrated or followed globally. There might also be some events that only certain community follows, or only specific country or geographic regions follows and those event details won't be available everywhere. Those events can be anything from a traffic jam, local festival, or a domestic cricket match. We can the find the events that have a small location localized event. Some events can happen considerable importance part of physical space or within a small region like Mother's Day.

C. Specific word Identification

A huge amount of data is collected and, in this segment, using certain keywords like the location of the event tagged in the tweet localized event id detected[11]. All tweets that are collected there might be a lot of tweets that are tagged with different geo location so, using location signature all the other posts and statuses are separated from the tweet that matches the location signature.

D. Clustering

Hierarchical Clustering:

Nested set of clusters is created. Each level in the hierarchy has a separate set of clusters. At the lowest level, each item is in its own unique cluster. At the highest level, all items belong to same cluster[13]. With hierarchical clustering, the desired number of clusters is not input. There are Two Types of hierarchical clustering

1. Agglomerative Hierarchical Clustering
2. Divisive Clustering

a) Agglomerative hierarchical clustering

Data objects are grouped in a bottom-up fashion. Initially each data object is in its own cluster. Then merge these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied[10]. The user can specify termination condition, as the desired number of clusters. Output is Dendrogram, which can be represent as a set of order triples <d, k, K> where d is the threshold distance, k is the number of clusters, and K is the set of clusters.

b) Divisive Hierarchical Clustering

In this data objects are grouped in a top-down manner. Initially all objects are in one cluster. Then the cluster is subdivided into smaller and smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination conditions as the desired number of clusters is obtained.

SYSTEM OVERVIEW

Figure 1, System overview is the summary of the project. Components used as a plugin in JOSM framework are:

- Tweets repository: Tweets from the social media are gathered continuously with the support of Twitter data are continuously collected from Twitter API to collect the most recent posts and stored in a databases
- Buffer: tweets are kept in book as a record of win_{fc}^{k} in mother board. Time frames indexed by bt tweets for giving permission for fast access. All the data collected and used are kept record as usage history in main memory.
- Content Preprocessor: Numerous texts is adopted Preprocessing works to handle with the big behavior of Twitter data
- Small Event Detector: Even the small local event are possible to detect as described above in segment 2

- Visualizer: It gives users many benefits to tweets. to be precise its shows user to view detected local events on the map.

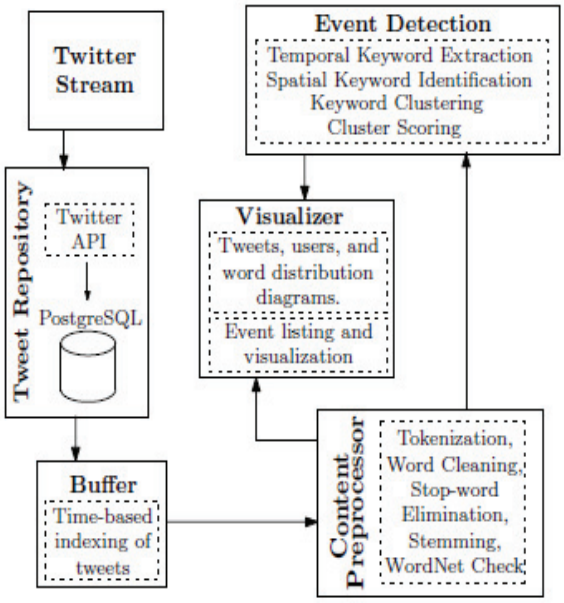

Fig 1. System overview of EvenTweet

Phase 1: build quad-tree

Here coming up to this phase, we will be using quad-tree process in order to perform spatial decomposition. Since this project is being used in variety of the applications either that is in the field of image processing or geographic evidence systems or computer graphics or a robotics and hence, we will be constructing a quad-tree at every time interval *i*. In case of two dimensional space, the quad-tree starts with a bulky rectangular range, in our work $\hat{W}_1 = \hat{W}$, that represents the root of the quad-tree[3]. $\hat{W}_1$ which is the root region is further divided into four equal sized regions as $\{\hat{W}_{11}, \hat{W}_{12}, \hat{W}_{13}, \hat{W}_{14}\}$ and every portion is recursively further divided thereby creating $\{\hat{W}_{111}, \hat{W}_{112}, \ldots\}$, and so on. If both $|W_{ix}| > \theta_{count}$ posts and the area of area *x* is at minimum θ_{area} then only the Portion of a section *x* arises. The limit of the spatial resolutions is caused by these limitations[7]. Since we created quad-tree so we can now also calculate $_{ij}$ and C_{ij} for every node, including internal nodes; where node is an identical with the area in that area j is node j.

Phase 2: Event detection

For a sliding window interval i and all regions j altogether with those at internal nodes of the quad-tree, Poisson distribution will be used to measure how probable the experimental number of posts, Cij , is used in order to perform the slide addition T that is followed by the sliding window[4]. The expected arrival rate of posts is equals to/ corresponding Pij , of observing Cij posts in time T is believed on the Eq. 3. The more unlikely the reflection, which can create result from a implicitly rise or fall in posts from the unkind, the more we may consider the posts to include an incident. Therefore, we can say that the areas with Pij <τ1, a constant threshold, can be highlighted as capable areas for actions. In order to pay off for either sparse or lacking data, we "smooth" the Poisson signal by manipulating an exponential crumbling individual event suggestion.

Phase 3 : Merge events

In previously found event detection it has various way of location determination and same progressive scale. When many events happen in the same area then intermissions are fused. By this we can calculate approximate event during that period of time. For example, if two events event1(e1) and event2(e2) happen in same geographic area at the exact same time intermission i and i+1 then both events shared as single event with period a 2T. Both post and average are added while merging these two events.

Phase 4: Prune events

After merging these two events we trim events to get more accuracy. Event interval ≥θ. As long as the event goes the chances of getting precise and accurate result is more. In other words, flagged location for swift time it is very likely to be noise. After calculating for all quad tree sometimes lead to identified events in different tree levels . sometimes two events might happen at the same time then, it prioritizes to the event which has the strong indication. It gives the actual time and location based on quad tree.

Location-based event detection using points of interests (POIs)

Points of interests (POIs) is being mostly used in location-based agreement and various types of event detection ideas which is based upon the POIs [48]. We advance a baseline which uses comparable knowledge of classification geo-tagged social media to Point Of Interests[5]. We then surmount a bunch of known and communal POIs for each of the city from their matching Wikipedia entries. In this baseline set of information, we undergo a spatial design of tweets based on their convenience (<100m) to respective Point Of Interests. The calculating Poisson signals and essential existence break tolerate the similar as formerly defined in Segment "Proposed algorithm".

Incremental clustering for real-time event detection

Among other remaining procedures in detecting event we used web scrapping for clustering. Clustering based event is compared to our approach because it is similar to our presented problem. In this method it detects a lot of clusters that are large and used it for a long time to get desired and useful information out of it[6].

From the given data, interval occurance is equals to T, where tc = the current time & T = maximal temporal gap, the algorithm repeated extracts the data and finds a set of event circles and unions every time tick t. An event circle C = set of active with maximal radius = R. overlapped events is equals

to K- overlapped events. The values for parameters t, T , R, K, and N are given by users.

IV. PROPOSED ALGORITHM

1. Construct multiscale spatial firmness through quadtree technique.
2. Poisson model and signal smoothing is used for event detection.

3. Event integration and Event cropping.
4. Mostly, proposed algorithm upholds boundless set of spotted events E identified in limitless stream S.

Detection algorithm has been mentioned which includes quad tree construction, event detection, event merging and event pruning.

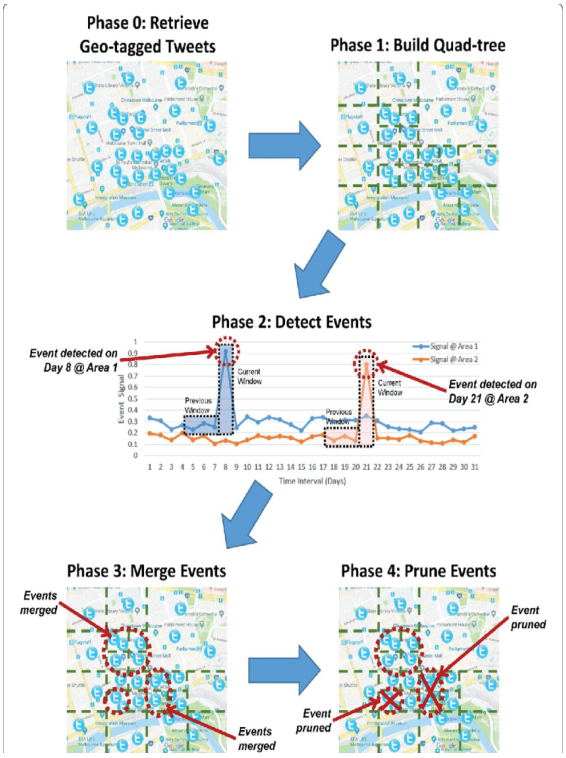

Fig 2: Overview of our proposed spatio temporal filtering

V. CONCLUSION

In this segment, we can summarize our project in four different points.First, an initial analysis of the projected technique is presented under the Section "Preliminary analysis". Second, detailed analysis is presented with baseline algorithms.

At last, the data is collected for about a year and calculate proposed algorithm using the precision, recall and Strength index.

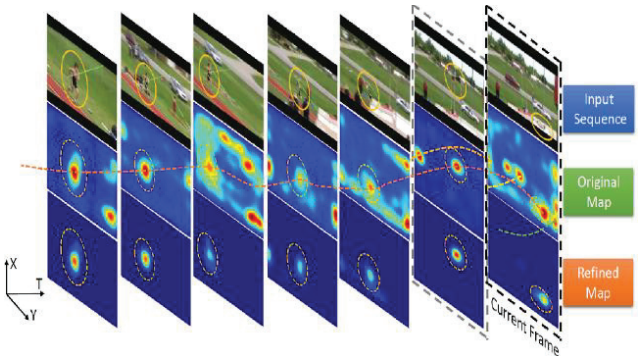

Fig 3: Overview of our proposed spatio temporal filtering

Algorithm 1 Spatio-temporal Online Event Detection Algorithm

Global constants (inputs): $\Gamma, T, \Delta T, \theta_{area}, \theta_{count}, \tau_1, \tau_2, \alpha, \theta_{duration}, \theta_{entity}$
Output: $\mathcal{E}$
Initialise: $i \leftarrow 1, \mathcal{E} \leftarrow []$
Repeat:
  $qTree \leftarrow \text{QUADTREE}(i, \Gamma_1 = \Gamma)$ ▷ Phase 1: Build quad-tree
  $\mathcal{E}_i \leftarrow \text{EVENTDETECTION}(i, qTree)$ ▷ Phase 2: Detect events
  $\mathcal{E} \leftarrow \text{EVENTMERGE}(\mathcal{E}, \mathcal{E}_i)$ ▷ Phase 3: Merge events
  $\mathcal{E} \leftarrow \text{EVENTPRUNE}(\mathcal{E})$ ▷ Phase 4: Prune events
  $i \leftarrow i + 1$

Phase 1: Build Quad-tree

function $\text{QUADTREE}(i, \Gamma_x)$
  $N \leftarrow \{(\lambda_{ix}, C_{ix})\}$ ▷ This node is synonomous with region x
  if $|\mathcal{W}_{ix}| > \theta_{count}$ and area of $\Gamma_x \geq \theta_{area}$ **then**
    Subdivide Γ_x into $\{\Gamma_{x1}, \Gamma_{x2}, \Gamma_{x3}, \Gamma_{x4}\}$
    $N \leftarrow N \cup \text{QUADTREE}(i, \Gamma_{x1})$
    $N \leftarrow N \cup \text{QUADTREE}(i, \Gamma_{x2})$
    $N \leftarrow N \cup \text{QUADTREE}(i, \Gamma_{x3})$
    $N \leftarrow N \cup \text{QUADTREE}(i, \Gamma_{x4})$
  return N
end function

Phase 2: Event Detection

function $\text{EVENTDETECTION}(i, qTree)$
  $\mathcal{E} \leftarrow \{\}$
  for node (C_{ix}, λ_{ix}) in $qTree$ **do**
    Compute P_{ix} using Eq. 3
    Compute F_{ix} using Eq. 4
    if $F_{ix} \geq \tau_2$ **then**
      $e \leftarrow (\Gamma_{ix}, T_i, T_i + \Delta T, \Delta T, \mathcal{W}_{ix}, P_{ix})$
      ▷ region, start, end, period, posts, signal
      $\mathcal{E} \leftarrow \mathcal{E} \cup \{e\}$
  return $\mathcal{E}$
end function

Phase 3: Merge Events

function $\text{EVENTMERGE}(\mathcal{E}, \mathcal{E}_i)$
  for event $e = (rg, st, en, pe, po, si) \in \mathcal{E}_i$ **do**
    if $e' = (rg', st', en', pe', po', si') \in \mathcal{E}$ where $rg' = rg$ and $en' = st$ **then**
      $e' \leftarrow \left(rg', st', en' + \Delta T, pe' + \Delta T, po \cup po', \frac{si + si'}{2}\right)$
      ▷ Updates the existing event in $\mathcal{E}$
    else
      $\mathcal{E} \leftarrow \mathcal{E} \cup \{e\}$
  return $\mathcal{E}$
end function

Phase 4: Prune Events

function $\text{EVENTPRUNE}(\mathcal{E})$
  for event $e = (rg, st, en, pe, po, si) \in \mathcal{E}$ **do**
    if $pe < \theta_{duration}$ **then**
      $\mathcal{E} \leftarrow \mathcal{E} / \{e\}$ ▷ The duration of the event is insufficient
      Continue
    if $\exists e' = (rg', st', en', pe', po', si') \in \mathcal{E} \mid si' < si, rg' \cap rg \neq \emptyset$ **then**
      $\mathcal{E} \leftarrow \mathcal{E} / \{e\}$ ▷ The event is subsumed by a stronger event
      Continue
    if $|uniqueEntities(e)| < \theta_{entity}$ **then**
      $\mathcal{E} \leftarrow \mathcal{E} / \{e\}$ ▷ The event is not semantically consistent
      Continue
  return $\mathcal{E}$
end function

Alg 1: Spatio temporal online event detection Algorithm

REFERENCES

[1] Yasmeen George, Shanika Karunasekera , Aaron Harwood and Kwan Hui Lim, Realtime spatio-temporal event detection on geotagged social media, 28 january 2021: vol no – 8, issued no. 91, page 4-5

[2] Patil S., Jayadharmarajan A.R. ,Clustering with modified mutation strategy in differential evolution, Pertanika Journal of Science and Technology, 2020

[3] Han, Y., Karunasekera, S., Leckie, C., Harwood, A.: Multi-spatial scale event detection from geo-tagged tweet streams via power-law verification. In: Accepted by IEEE Big Data 2019(2019)

[4] Clara Kanmani A., Vinitha P. Ontologies in semantic web: A comprehensive analysis, Journal of Advanced Research in Dynamical and Control Systems ,11, 2019

[5] Hong Wei , Hao Zhou , Jagan Sankaranarayanan , Sudipta Sengupta and Hanan, Detecting latest local-events from geo tagged tweets streams november 6,2018: vol no – 9, issued no. 18, page 2

[6] Yugian huang, Yue Li, Jie Shan, Spatial temporal event detection from geotagged tweets,15 april 2018: vol no – 7, issued no. 4, page 8

[7] Venugopal S., Nagraj G.: A proficient web recommender system using hybrid possiblistic fuzzy clustering and Bayesian model approach, International Journal of Intelligent Engineering and Systems, 11, 2018.

[8] Chao Zhang, Guangyu Zhou, Quan Yuan, Honglei Zhuang, Yu Zheng, Lance Kaplan, Shaowen Wang1, and Jiawei Han , GeoBurst:Real time local event Detection in geo tagged TweetStreams: published 21 july 2016, page 8

[9] Singhal R., Deepika N., Detecting fraudulent words: Using PFCM clustering, 2016 IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, RTEICT 2016 – Proceedings.

[10] Alqhtani, M, Luo, S., Regan, B.: Fusing text and image for event detection in Twitter. Int. J. Multimedia Appl. 7(1), 27-35 (2015).

[11] Ilango V., Subramanian R., Vasudevan V., Outlier detection and influential point observation in linear regression using clustering techniques in financial time series data., Journal of Theoretical and Applied Information Technology, 2013.

[12] Hamed Abdelhaq, Christian Sengstock, and Michael Gertz , EvenTweet locaized event Detection From Twitter,26 Aug 2013: vol no – 6, issued no. 12, page 2,3

[13] Becker, H., Naaman, M, Gravano, L.:Beyond trending topics:real world event identification on twitter. In:ICWSM 2011