Developing an Alert System Through Probabilistic Latent Semantic Indexing By Investigating Tweets

T. Kaviya, Pg Scholar
Dept of Computer Science & Engg,
SKP Engineering College,
Thiruvannamalai,India

A. Kumaresan, Professor
Dept of Computer Science & Engg,
SKP Engineering College,
Thiruvannamalai,India

V.Raji, Research Scholar
Assistant Professor,
Dept of Computer Science & Engg
S.K.P Engineering College,
Tiruvannamalai,India

K.Vijaya Kumar, Research Scholar
Associate Professor,
Dept of computer Science & Engg,
SKP Engineering college,
Thiruvannamala,India,

Abstract : **Nowadays, Social media such as Facebook, Twitter, you-tube, etc plays an important role to update instantly. Twitter is an Information sharing site, in which it is mainly used to share text message within 140 characters, audio links, video links, URL's,.. It got more popular because of its real time nature. Users of this site frequently checks about what others doing and updates about "what's happening?". Because of its real time character , many applications are being developed. This paper is developed to alert the users by collecting the tweets and it must be compared with the training data set. Then, the obtained data is classified by a technique called SVM and Particle filtering algorithm is used to determine the "hot spot"(ie. the location in which event occurred). This paper presents an alert system to all the registered users for real time sensitive events like Earthquake, floods, storm ,etc by investigating small text messages(tweets).The main objective is to provide the alerts by few tweets through Probabilistic latent semantic indexing(PLSI) technique and reducing the false positive rate by re tweeting analysis.**

***Index Terms*—probabilistic latent semantic analysis, support vector machine(svc), social sensors.**

I.INTRODUCTION

Although several micro blogging sites are there, Twitter has become more popular because of its real-time nature. It has wide range of users all over the world. Twitter asks a question "What's happening?"[1]for that the answer must be within 140 characters by a user. It was introduced in the year 2006,the number of users must exceeds the range 100 million in Apri2010. According to recent statistics, the number of active users must be 554,750,000. Currently, the number of users signing up per day is 135,000which increases the tweets by 190 million per day. Nowadays people will get the opportunity of getting job through Twitter. Around 2500 members got employed by this site. The range of active users per month must be 115million [2].

Researchers and developers have published their studies in twitter based on three different fields: First, some researchers have analyzing the network structure of twitter. Second, some researcher's analysis the characteristics of Twitter and the last field research is based on developing much more applications on twitter. Twitter is a micro-blogging service because it shares text within 140characters or video links or photos or audio clips. Within a short duration it got more attention because of its Follow-Follower [3] Theme. In E-commerce, customers must buy the product and rate them according to the quality. Users rapidly return to the site to know what other are thinking about and what they are doing. In such a way, numerous updates form sports like cricket, football, hockey, online games can be notified at the moment .In another event, when an airplane crashed and landed on the River of Hudson in New York, the first reports were published through Twitter and tumblr. Through the reports and updates made by tweet, we can able to detect and alert several natural disasters like flood, earthquake, storm, fires, traffics, etc.[4].

From the reference [4], we can implement the idea of developing and providing an earthquake alert system with the help of tweets. Initially, gathering and an indexing is provided for the tweets which contains keywords. Considering twitter user as a sensor and tweets as sensory information, which can have numerous characteristics and acronyms: some are active and some users are inactive, but social sensors are very noisy when compared to physical sensor. Then, the process of semantic analysis takes place of a tweet. For example , users must make tweets like, "Shaking in Pakistan", or "Now I felt earthquake here" or "Attending a conference on earthquake" or "Earthquake". These tweets can be negative as well as positive, so it can be classified using a "Support vector machine"(SVM) which was more

effective[4].only positive tweets can be allowed to predict the event. If the keyword reaches peak value then temporal detection can be done. Later, Spatial detection can be carried out using a technique named “Particle filtering”[5]. In Nature, Earthquake can travel 2-3 km/s. so it can be easy to alert the people who are at 50-100 km before the event.
With the help of Particle filtering we can alert before a minute. It helps the people to get alert about the fires, cylinders and from hazardous places.

II. EXISTING SYSTEM

Social sensors like Facebook,Twitter,Linkedin are more powerful than physical sensors such as Seismograph so, Social sites are used now a days for providing alerts. The Keywords and its length in Indexing must be small and very few words are considered in the existing system since, It Uses Latent Semantic Indexing Technique.Detection of location can be done with the help of Particle filtering with weighting. Alerts must be provided only for the active users through E-mails.So inactive users didn’t get any alert from the Site. The main disadvantage is Physical sensor must takes 15 minutes to make a sure about event and it is inoperable or malfunctioning sometime. The alert made by Twitter may have High range of False Positive alerts .This kind of methods are only suitable for Suits only for high populated area.

III. PROBLEM STATEMENT

Existing system tends to provide Alert for Active users (Who are active on mail Messages).In India, users must be active with their Mobile phone messages like SMS,MMS,. In Order to reach the alert for all the users of Twitter in the Location which is affected by Natural disasters, it is Must to provide them by using Short Message Service. For Less populated Area it is must to consider the Keyword peak value and Increasing the sample Database words in Index With the help of Probabilistic latent semantic Indexing.

IV. WORKING OF PROPOSED SYSTEM

A. Earthquake Detection:

[6]Because of twitter’s real-time nature ,Events like earthquake, flood, storm can be detected through the information(Tweets). It involves some techniques :Semantic analysis ,Support vector machine and Particle filtering ,With the help of weighting alert can be sent more fast. It can reach the destination within a few seconds after the occurrence.

B. Semantic Analysis of Sensory Information (Existing Method For indexing):

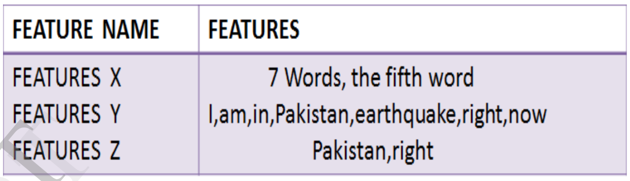
[6]To detect the occurrence of event, initially we have to find out the useful tweets. It involves the tweets that must contains keywords. For Example: ”I felt severe quake in Pakistan and got much scared about it.” Or “Earthquake !” or “ Land vibrated”. Sometimes target place can also be mentioned in the tweet, that must be “ shaking in Lahore”. Even though the tweet refers to target event it must not be an useful. Based on its characteristics it can be classified into: Positive and Negative.

Earthquake
Shaking
Vibration
flood
Storm

Table1: Keywords in Latent Semantic Index.

From the Index API, tweets can be classified with Support vector machine SVM a widely used machine learning technique. Few example for negative Tweets, “seven earthquakes this week”, “304 died in earthquake”,” I am the chief guest of international conference on Earth–Quakes and Shakes”. Classifier contains three different features to classify a information,

- *Feature* X: It contains the number of words in the tweet and place of the keyword.
- *Feature* Y: It must contains the original tweet of the person.
- *Feature* z: It have the Place and occurrence of event information.



FEATURE NAME	FEATURES
FEATURES X	7 Words, the fifth word
FEATURES Y	I,am,in,Pakistan,earthquake,right,now
FEATURES Z	Pakistan,right

Figure 1 : Tweet classifier (SVM)

These features helps the Machine learning SVM to classify and find the useful information.

C. Probabilistic Latent Semantic Indexing (Proposed Method for Indexing):

[10] It’s a machine learning technique which can be use to store many sample word and more similar words to detect the keyword Like “Earthquake”- Shuttering, violent moment , vibrating, shaking, moving, frightened , fear, panic moment, quake etc must be represented. Sometimes the words can be tweeted with spelling mistakes due to fear and time inconsistency. It must also considered and word must be extracted from the tweed to detect the quake. The Proposed system is Suggested to work on More number of related tweets at a time. Estimation of location can be made more fast by using this Probabilistic latent semantic analysis and The alert system can be made common for both active and inactive users in twitter through Short message service(SMS) .The tweet users can get the alert in their mobile device . It must works well for all users and reach maximum users of that country.

More Related	Less Related	Stemming Word
Earthquake	Death Rate	#Earthquake,#eq
Vibration	Died	#vibration
Shaking	People suffered	#shaking
Quake	worries	#quake
Tsunami	occurrence	#Tsunami

Table2:Keywords in Probabilistic Latent semantic Index.

D. social Sensor:

[4]The users can play important role for detection of target event. Assumptions can be made to enable the applications of various methods :

- Tweets from the user is considered as a sensory information and user as sensor.
- An user detects event and reporting Probably.
- Each sensory information is associated with time.
- Location information can be found with the help of latitude and longitude coordinates.

The detection of location must involves Physical sensors like GPS and infrared

badge.[4]. But here we use more effective Particle filtering technique.

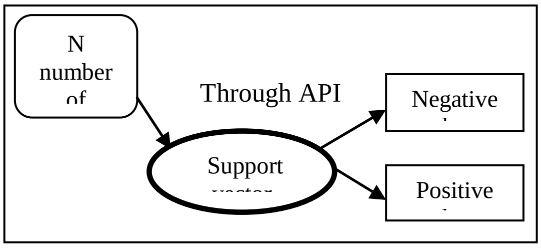


Figure 2: Function of Support Vector Machine

E. Time Evaluation:

[2]Each post is associated with time. It plays a vital role in detecting an event . Sometimes false detection can occur due to Past reported tweet like "yesterday, people suffered a lot because of earthquake". API can detect 1200 tweets/s. Time factor can flow in exponential distribution[3]and can be detected by time t.

S	Sample Data (Tweets)
T	Time to retrieve the Data
X, Y	Latitude and longitude of the User
M	Mean Value of the Location For all users
P	False- Positive ratio

Figure 3: Tweet Analysis Notations

sometimes the keyword detection can leads to false alert of earthquake and with the probability of 0.34.

F. Location Estimation:

[4,5]A particle filter is a probabilistic approximation algorithm implementing a Bayes filter, and a member of the family of sequential Monte Carlo methods. For location estimation, it maintains a probability distribution for the location estimation at time t, designated as the belief f(x) 1 . . . n. Each X(t) is a discrete hypothesis related to the object location. The which are non-negative weights, called importance factors, which sum to one. The Sequential Importance Sampling (SIS) algorithm is a Monte Carlo method that forms the basis for particle filters.

The SIS algorithm consists of recursive propagation of the weights and support points as each measurement is received sequentially.

The algorithm is presented below.

- *Generation:* Generate and weight a particle set, which means N discrete hypothesis.
- *Resampling*: Resample N particles from a particle set St using weights of respective particles and allocate them on the map. (re sampling of more than that of the same particles.).
- *Prediction*: Predict the next state of a particle set St from Newton's motion equation.
- *Weighing*. Recalculate the weight of St by measurement. Calculate the current object location
- *Iteration*. Iterate Steps 2, 3, 4, and 5 until convergence.

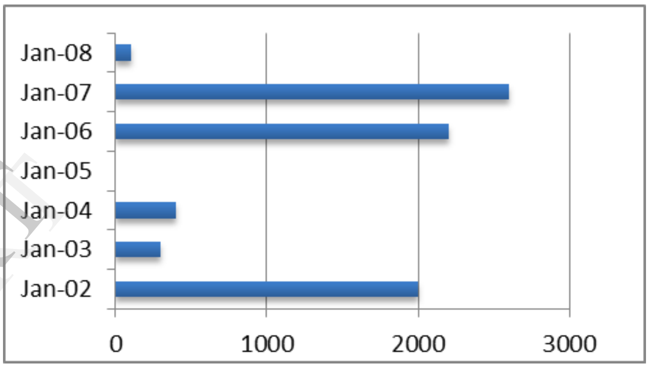


Figure 4: X-axis indicates number Occurrence of Word and more related words of "Earthquake" in the tweets and y-axis indicates the day of occurrence.

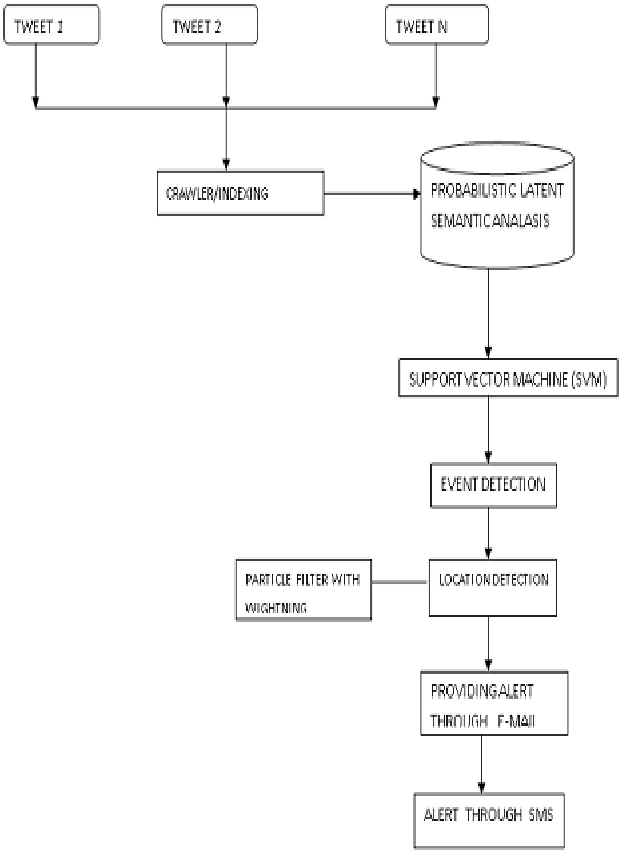

TWEET 1
TWEET 2
TWEET N
CRAWLER/INDEXING
PROBABILISTIC LATENT SEMANTIC ANALASIS
SUPPORT VECTOR MACHINE (SVM)
EVENT DETECTION
PARTICLE FILTER WITH WIGHTNING
LOCATION DETECTION
PROVIDING ALERT THROUGH E-MAIL
ALERT THROUGH SMS

Figure 5: Alerting through SMS using Probabilistic latent semantic analysis

Definition 1:

An active user of tweet are those who will " updates the information during a week if He or She updates at least one tweet or re-tweets something" during that week.

Alerting Active users of twitter :

The users who can use to check the mail frequently can be alerted about the event and other users cannot get this alert message. Most of the users who can tweet frequently can check the mail messages, but others can't. Through this System, Only users who can have twitter account and frequent E-mail users in that region will get the alert about the event. This cannot reach much people in that region to overcome this we can determine and conclude a technique that must be more powerful to get an alert about the earthquake.

Definition 2:

If an user can updating and checking monthly once in Twitter then the user is said to be inactive to that site.

V. EVALUATION

Through the help of semantic analysis and particle filtering the events like earthquake and other hazarders events can be easily detected in dense populated country .Because more user can use to update their current affairs. It helps well to predict the events before its occurrence. Analysis of tweet can be made through the Application programming interface. An event alert can be much more valuable, if it can be provided in Real-time. An event can cause Economic loss to the country and Emotions of the people of the country due to death rate. To overcome this an alert system can be made in Japan by Japan metrological [3] agent since 2007 with the help of P-waves(Primary waves).But they are not much effective for early alert. Predicting events and alerting must helps to

- Turn off a stove or heater in our house.
- Hide ourselves under a desk or table
- Coming to roadsides.
- Slowing trains.
- Controlling Elevators.

It can be only possible if we can get the alert a few seconds or minutes before an event actually hits. This can be done through Tweets and alerted through e-mail for users. The users who are using the mail and checking it frequently will get the alert about and other users can't. so providing an SMS alert must reach all the users and thy got alerted easily(reaching both active and inactive users of twitter).

It leads to

Decrease in Death rate.

More Supportive for Economic growth of a Country.

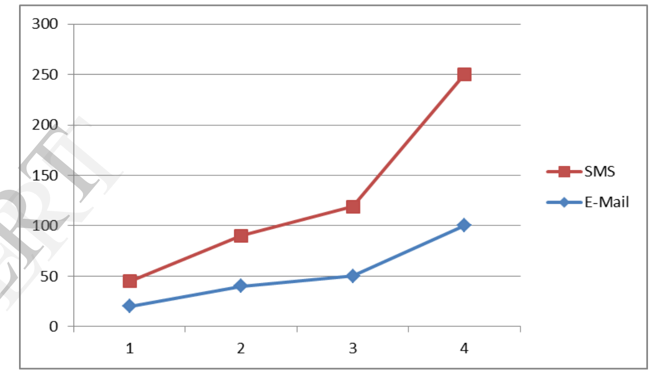


Figure 6 : SMS alert Saved More number of people compared to Mail.

VI.RELATED WORK

A. Predicting elections with Twitter : what 140 characters reveal about political sentiment:

[9] This paper reveals with the result of election in online with the help of around 1,04,003 tweets from twitter through API. The Main Goal is to Predict the leader of the Germany based on tweet. With the help of collected Sensory information the election result can be evaluated and result can be published in media's and it can be compared with the offline result. It leads to the researchers to get an idea about predicting periodic events like rainbow and other hazarders events like earthquake, flood, storm, Tsunami, etc.

B. Text categorization with support vector machine: Learning with many relevant features:

[7]It deals with the usage , characteristics and features of support vector machine. Comparison and Experiments can be made to separate the documents to its corresponding category. Support vector machine(SVM) involves four steps to detect the document type:

- High dimensional input space,

- Few irrelevant features,
- Document vector are sparse
- Most text categorization problem are linearly separable.

The aim is to calculate the mean value of Precision-recall Point .Although several categorization techniques like K-NN,bayes,ricchio,c4.5 are used, but quicker performance can be provided by SVM.

C. Bayesian Filter for location Estimation:

[8]This reference can be made to get an idea about the Location estimation. At initial stage location can be detected with the help of sensors such as GPS and infrared badges. But nowadays due to the dramatic change in technology, it must be simple and more quick to estimate the location of devices. Developers can made it more easy for mobile devices too. Methods include

- Bayes filters
- Multi hypothesis tracking
- Grid –based approaches
- Topological approaches
- Particle filter.

Among these methods Particle Filtering Technique is more effective. It can be Mathematically derived and experimentally proven in this paper. Further more Evaluation is going on in Particle filter technique to detect the location more fast

.

VII.CONCLUSION

Studies can be made to prevent the false detection and to alert early as soon as possible. Increasing the keywords and detecting the approximate location must leads to improve and enhancing the system from the death of people in natural disasters.

REFERENCES

1. en.wikipedia.org/wiki/twitter.com
2. www.statisticbrain.com/twitter-statistics/
3. A. Java, X. Song, T. Finin, and B. Tseng, “Why We Twitter: Understanding Micro blogging Usage and Communities,” Proc. Ninth WebKDD and First SNA-KDD Workshop Web Mining and Social Network Analysis
4. T.Sakaki, M.Okazaki, and Y.Matsuo, “Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development”.
5. J. Hightower and G. Borriello, “Particle Filters for Location Estimation in Ubiquitous Computing: A Case Study,” Proc. Int’l Conf. Ubiquitous Computing (UbiComp ’04), pp. 88-106, 2004.
6. T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors,”
7. T.Joachims ,“Text Categorization with Support Vector Machines: Learning with Many Relevant”.
8. V. Fox, J. Hightower, L. Liao, D. Schulz, and G. Borriello, “Bayesian Filtering for Location Estimation,”
9. A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Welpe,“Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment.
10. “Probabilistic Latent Semantic Indexing” Thomas Hofmann.