ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A REVIEW PAPER ON PARALLELED BIG DATA ALGORITHM WITH MAPREDUCE FRAMEWORK FOR MINING TWITTER DATA

Ms.Hemlata A. Patil *, Prof.B.A.Tidke
* Computer Network, Flora institute of technology,Khopi,India

ABSTRACT

Communication is a key factor in today's human life, due to time constraints physical interaction between people is not possible. This gap is filled by the technology through 'social networking' sites it's very easy to get access to interact other based on their interests. Many applications are getting releasing with new features day-by-day from vendors, to provide efficient usability and user friendliness. This paper proposes a new system that delivers large database of Social Networking Site (SNS) called 'Twitter'. Many Third party application are building based on SNS like Twitter, they need to have processed data from their operational purpose. The main stream of the applications is visualization applications. This paper gives more beneficial solution by providing in-depth detailed information of data. In this context this implementation serves processed information of tweets accessed from Twitter Server. Here processing the tweet involves extraction of metadata of tweet, geocoding the physical address in a tweet, analyzing the sentiment of content in the tweet text and extracting the significant and key phrases from a text. This application is an integrated system used to get connect and access tweets from Twitter to get processed text analysis components. After all the Information Extracted and NER (Named Entity Recognition) text analysis from tweet, are stored into a persistence database.

KEYWORDS: Web 2.0 technology;Mapreduce framework;Big data algorithm;Social networking;Twitter.

INTRODUCTION

Twitter is classified as a micro blogging service. Micro blogging may be a variety of blogging that allows users to send transient text updates or macromedia like images or audio clips. Micro blogging services aside from Twitter embody Tumblr, Plurk, Jaiku, identi.ca, and others.All content should be written in English and should be in 1 column. An important characteristic that's common among small blogging services is their period nature. Though web log users usually update their blogs once each many days, Twitter users write tweets many times during a single day. Users will savvy alternative users do and infrequently what they're brooding about currently, users repeatedly come back to the location and check to check what people do. many necessary instances exemplify their period nature: within the case of a very sturdy earthquake in Haiti, several footages were transmitted through Twitter. Folks were thereby able to apprehend the circumstances of injury in Haiti straight off. In another instance, once associate aero plane crash- landed on the Hudson River in the big apple, the primary reports were revealed through Twitter and tumbler. In such a way, varied update leads to varied reports associated with events. They embrace social events like parties, baseball games, and presidential campaigns. They additionally embrace black events like storms, fires, traffic jams, riots, significant precipitation, and earthquakes. Actually, Twitter is employed for varied period of time notifications like that necessary for facilitate throughout a large-scale hearth emergency or live traffic updates. Adam Outgrow, the Editor in Chief at Mash able, a social media news web log, wrote in his web log regarding the attention-grabbing development of period of time media. This post well represents the motivation of our study. The analysis question of our study is, "can we tend to discover such event prevalence in period of time by watching tweets?" This paper presents associate investigation of the period of time nature of Twitter that's designed to determine whether or not we are able to extract valid data from it. we tend to propose an occurrence notification system that monitors tweets and delivers notification promptly victimization information from the investigation. during this analysis, we tend to take 3 steps: 1st, we tend to crawl varied tweets associated with target events First, to get tweets on the target event exactly, we have a tendency to apply linguistics analysis of a tweet. for instance,

users may create tweets like "Earthquake!" or "Now it's shaking," that earthquake or shaking may be keywords, however users may also create tweets like "I am attending AN Earthquake Conference," or "Someone is shaking hands with my boss." we have a tendency to prepare the coaching knowledge and devise a classifier employing a Support Vector Machine (SVM) supported options like keywords in a very tweet, the quantity of words, and also the context of target-event words. When doing thus, we have a tendency to acquire a probabilistic spatiotemporal model of an incident. We have a tendency to then create a vital assumption: every Twitter user is considered a detector and every tweet as sensory data. These virtual sensors, that we have a tendency to designate as social sensors, square measure of a large selection and have numerous characteristics: some sensors square measure terribly active; others don't seem to be. A detector may well be inoperable or defective typically, as once a user is sleeping, or busy doing one thing else. Consequently, social sensors square measure terribly creaking compared to standard physical sensors. relating to every Twitter user as a detector, the event-detection downside will be reduced to at least one of object detection and placement estimation ubiquitous/ pervasive computing surroundings during which we've various location sensors: a user features a mobile device or a full of life badge in surroundings wherever sensors square measure placed. Through infrared communication or a LAN signal, the user location is calculable as providing location-based services like navigation and deposit guides [9], [10]. We have a tendency to apply particle filters, that square measure wide used for location estimation in ubiquitous/pervasive computing [11]. As AN application, we have a tendency to develop earthquake coverage system victimization Japanese tweets. Japan has various earthquakes. Twitter users square measure equally various and Earthquake Coverage System Development by Tweet Analysis International Journal of rising Engineering analysis and Technology ninety-eight geographically spread throughout the country. Many studies have been undertaken to monitor the social situation by treating participants in social media, such as those using Twitter, as social sensors. However, most such studies are aimed at observation of long-term changes of social situations. Our research is an early approach to use Twitter as a social sensor for detection of real-time events. Additionally, it is meaningful that we apply methods for event detection using ordinal physical sensors for event detection by social sensors. The field of event detection using physical sensors has already been developed. Methods of many kinds exist in the field. Therefore, it is possible that events of many kinds can be observed from Twitter through application of those methods. Our research has produced one of the first approaches to use such methods.

Problem Statement

The approach to solve the event detection in text streams social media that has a great deal of misinformation in it and due to its ability to spread data rapidly is not without its perks. The reference systems used in the evaluation of the Topic Detection and Tracking task used reduced corpus datasets, afterward studies proved that they do not scale to larger amounts of data. The performance, effectiveness and robustness of those systems was acceptable under the specified evaluation conditions but today's online social network service volume of data (creates massive unstructured text data streams) make them obsolete systems. Apart from not being designed to handle big amounts of data, the data in online social network services is also dynamic; messages are arriving at very high data rate, so the adaption of the computing models is required to process data as it arrives. Computation of such vast amount of data needs necessarily technology that has a *highly scalable* storage platform and performs distributed concurrent parallel execution of database. Time and Cost Effectiveness is an issue, in most cases using this kind of systems it is preferable to have immediate and round about solution rather than waiting too much for an exact solution. Online social network text streams seem to be the ideal source to perform real-time event detection as they are very much Cost Effective. Performing real-time event detection using Twitter requires dealing and mining massive un-structured text data stream that has messages continuously approaching at sky-high data rates. Given this, the approach to deal with this specific problem involves providing solutions that are able to mine continuously, high-volume of open-ended data streams as they arrive. Considering that those sources of data are coming from social network users it is expected that information collected using metrics of networks analysis (nodes, connections and relations, distributions, clusters and communities) could improve the quality of the solution of the algorithm.

The reference systems used reduced corpus datasets that do not scale to larger amounts of data. The performance, effectiveness and durability of those systems was not being designed to handle big amounts of data but today's online social network service volume of data (creates massive unstructured text data streams) make them obsolete systems. Performing real-time event detection using Twitter requires dealing and mining massive un-structured text data stream that has messages continuously approaching at sky-high data rates. Given this, the approach to deal with this specific problem involves providing solutions that are able to mine continuously, high-volume of open-ended data streams as they arrive. Considering that those sources of data are coming from social network users it is expected that information collected using metrics of networks analysis (nodes, connections and relations,

distributions, clusters and communities) could improve the quality of the solution of the algorithm. Apart from, the data in online social network services is also dynamic; messages and arriving at very high data rate. Computation of such vast amount of data needs necessarily technology that has a *highly scalable* storage platform and performs distributed concurrent parallel execution of database. Time and Cost Effectiveness is an issue. Online social network text streams seem to be the ideal source to perform real-time event detection as they are very much Cost Effective.

Organization

Twitter is presently working in Japan for Earthquake detection using Twitter has been operated since August 8; 2010.Users can see the detection of past earthquakes. Also they can register to receive notices of future earthquake detections. It alerts users for imminent earthquake. It is hoped that a user receives alert before the earthquake actually affects the area. We assess various conditions under which alarms might be sent to choose better framework for our suggested system.We set alarm conditions as Ntweet (positive tweets) come in 10 minute. We evaluate those methods by Precision= Nearthquake/Nalarms And Recall = Nearthquake / Allearthquake

All earthquake (Nearthquake: Number of earthquakes detected correctly, Nalarms: number of distress signal, All earthquake: number of tremors that occurred). We must change the use of this condition vigorously to increase the accuracy of the system, particularly in terms of the repetition and intensity of earthquake. United States Centers for Disease Control and Prevention uses Twitter as a tool to gather timely health and safety information and encourages the strategic use of Twitter for effectively and inexpensively reach partners for emergency threat and to update individuals about emergency preparedness for health and safety concerns since 2009 swine flu pandemic.

LITERATURE SURVEY

Twitter is a noteworthy example of the foremost recent kind of social media. Varied researchers have examined Twitter. Relating to similar analysis to it conferred during this paper, some researchers have tried topic detection victimization Twitter. Cataldieal projected a unique technique to discover rising topics employing a keyword-based topic graph. They succeeded in detective work news keywords that area unit fashionable in Twitter. as an example, (a volcano in Iceland) and Samaranch (the previous President of IOC, World Health Organization died in Apr 2010). Marc et al. divided more and more fashionable keywords on Twitter into patterns of assorted type's victimization Kyrgyzstani monetary unit, thereby demonstrating that Twitter users contribute to the discussion of those trends. Other than the studies introduced in Section one and these studies, many others are done. We have a tendency to classify studies coping with Twitter or knowledge on Twitter into 3 teams. First, some researchers specifically examine the network structure of Twitter and investigate Twitter network options of assorted types. Java et al. analyzed Twitter as early as 2007. They delineated the social network of Twitter users and investigated the motivations of Twitter users [3]. Haewoon et al. crawled a massive quantity of Twitter knowledge, analyzed the Twitter follower-following topology and hierarchic users by PageRank [5]. Huberman et al. analyzed quite three hundred thousand users. They found that the relation between friends (defined as someone to whom a user has directed posts victimization associate "@" symbol) is that the key to understanding interaction in Twitter [4]. Second, some researchers have examined characteristics of Twitter as social media. Recently, Boyd et al. have continued their investigation of retweet activity, that is that the Twitter-equivalent of e-mail forwarding, by that users post messages that were originally denote by others [6]. Tumasjan et al. crawled several tweets touching on the election in Germany and tried to predict the results of the election: those political parties would win the election. Oconnor extracts vox populi from Twitter victimization sentiment analysis and reports the chance of employing a projected technique rather than polls. Third, some studies elucidate the advantages of novel applications of Twitter: Ebner and Schiefner establish a micro blogging community and studies the way to use Twitter as a tool for mobile e-learning. The combination of the linguistics net and micro blogging was delineated in a very previous report within which a distributed design is projected and also the contents area unit aggregative. We choose earthquakes in Japan as target events, supported the preliminary investigations. We tend to make a case for them during this section. First, we decide earthquakes as target events for the subsequent reasons:

1. Unstable observations are conducted worldwide, that facilitates acquisition of earthquake data, that additionally makes it straightforward to validate the accuracy of our event detection methodology; and
2. It's quite substantive and valuable to discover earthquakes in earthquake-prone regions.

Second, we decide Japan because the place supported the subsequent investigation. it's apparent that the sole intersection of the 2 maps, those regions with several earthquakes and enormous Twitter users, is Japan. Alternative regions like country, Turkey, Iran, Italy, and Pacific coastal U.S.A. cities like la and city additionally roughly meet, however their various densities are a lot of below that in Japan. Several earthquake events occur in Japan and plenty

of Twitter users observe earthquakes in Japan, which implies that social sensors are distributed throughout the country.

3. We gift a quick summary of Twitter in Japan: The Japanese version of Twitter was launched on Gregorian calendar month 2008. In February 2008, Japan was the No. a pair of country with relation to Twitter traffic.5 At the time of this writing, Japan has the second largest range of tweets (18 % of all tweets are announce from Japan) within the world. Therefore, we decide earthquakes in Japan as a target event as a result of the high density of Twitter users and earthquakes in Japan.

Validating the methods, processes and algorithms developed periodically over a span has to review with comparative study helps in concluding, and formulating assumptions for intended application.

Other technologies (minimum three) available to cater the same concept:-

Accessing tweets from the Twitter

Accessing Tweets from a Twitter is primary for building a database to get processed and extract information. Twitter has 3 types of API's REST API, Search API and Streaming API. Each has different usability REST API allows user to access twitter core data Search API grants methods to communicate the Twitter search. Streaming API assures long-span connection to get access huge volume of tweets. API's in Twitter is http based requests that too GET method is required in data retrieval.

Twitter API (dev twitter 2011) provides Search API and Streaming API for accessing Tweets, Search API provides recent Tweets with relevance to the search key and Tweet index of recent 6 to 9 days. Were as Streaming API gives the real time continuous stream of all Tweets, but it doesn't filter Tweets that are relevance. Limitation are laid on the users request frequency rate for both Search and Streaming API's, which not disclosed due to abuse and needless usage . The request limit can check in the response header, so that it varies over time and overall requests to get access.

Twitter API (dev twitter 2011) facilitates two ways to get access Tweets, through authenticated and unauthenticated requests. Search API supports unauthenticated and Streaming API need to have authentication. As far as authentication is concerned about types of Tweets, here we have public status and protected status Tweets. Search API present public status tweets on the other hand Streaming API present s both public and protected status Tweets. Request rate limit authenticated user-requests laid on user and for unauthenticated user-requests limit is laid on IP (ip address of the system). Client can request statuses at maximum of 3200 by REST API and 1500 statuses (response tweets) through Search API. Haewoon, lee & Housing (2010) have clearly explained about the functionality, operation and usability of twitter and also briefed about background processing to user. There is evidence (Haewoon, lee & Housung, 2010) that (1) Maximum number of requests from the user to twitter is 10,000 per hour from each IP address. (2) It is advised that tweet collector from the twitter to limit their request rate to the prescribed 10,000 requests/hour and to maintain time delay in between request for better results without any duplication.

Twitter API (dev. twitter 2011) gives scope of implementation of custom applications though broad spectrum on programming language Libraries and packages, java in particular the best in implementing object-oriented programming. Twitter4j API(twitter4j, 2011) is one of the java library for implementing custom application on Twitter , Twitter4j is feasible and flexible library for getting connected to Twitter, and communicate from custom application via Twitter to Twitter.

Twitter facilitates bifurcation of tweets into public and protected, public statuses tweets are from user accounts which are not protected, and protected is from protected user accounts. Protected statuses need user authentication credentials to get access Search API supports for public statuses. Twitter API (dev twitter 2011) gives response to requests in "JSON", "XML" and "ATOM" formats, parsing the output are in need of specific to the method you are using to extract. In twitter response, output some field are not guaranteed to return the value may it contain

Null, if value of the corresponding field value is not available to return? The http response codes may be witnessed in the output, by specifying the status of the user request. Twitter4j (twitteer4j, 2011) provides an implementation of java libraries to parse the GET responses like JSON, XML etc. Metadata of the tweet also implanted in response of a search query, it's vital in understanding the information stated in the tweet. (dev twitter 2011) have evidenced and analyzed that every tweet is not geo-tagged (geographic coordinate's latitude and longitude), but some tweets are exclusively geo-tagged in responses through Search API. It's purely optional to the user in stating the geo-location, because of user perspective and privacy to unable the disable this geo-tagging feature while tweeting through twitter.

Annotations Extraction

Our objective is to extract annotations from the Tweet text and the contemporary implement them for finding the annotations. Alias-i (2008) and Cunningham et al (2011) have proposed the corpus (document) and datasets and

stated a mechanism for chunking text into predefined chunks based on specified regular expression or tokenizing. Cunningham et al (2011) have given a solution for NER (Name Entity Recognition) with the help of Annie gazette but input text should be a textual document. Alias- i (2008) and Cunningham et al (2011) to extract annotations we need to train the system by specifying entity trained files or
Files of gazetteer lists. The mechanism of identifying the annotation is based on the matching of the trained file content with textural words of respective annotation type of corresponding files. Alias-I (2008) has used external training files with data on annotation, whereas Cunningham et al,(2011) have used internal mechanism to mention the gazetteer index with the lists. Both Alias (2008) and Cunningham et al, (2011) stated that there is no provision of finding annotation of provided input simple text but it limits usability. Cunningham et al (2011) have said in their context that in defining the training data the usability has to analyzed first and Alias-i (2008) has mentioned that segregation entities have to be taken into different lists or files while preparing the training.
The release (Cunningham et al, 2011) specified only trained mechanism in extracting annotations from the text document, were as it not stated for untrained mechanism. In concern simple Text annotation with various discipline data, (Alias- i, 2008) (Cunningham et al, 2011) complicates the procedure of defining the training data. Nadeau & Tierney (2006) have defined the "Entity noun ambiguity" and resolved it by implementing algorithm called "Aliasing resolution algorithm", it explains entity boundary detection in the course of unsupervised system to extract annotations and stated that it is not comparative to complex system.
Stanford's (Jenny Finkel, 2006) implemented natural language processing resources for text engineering and have mainly focused on processing of natural language in to a spectrum contents like parts of speech, translators, word segmentation, classifiers etc. In comparison to (Alias- i, 2008) and (Cunningham et al, 2011) the scope is limited in (Jenny Finkel,2006). Features of (Nadeau & Tierney, 2006) and (Jenny Finkel, 2006) are relative in the context of information extraction from the corpus. Jenny Finkel (2006) has customized the implementation of the code and made reusable or user friendly in different contexts.
Extraction of annotations in a simple text is defined clearly in (Jenny Finkel, 2006) and some models have been discussed, which in general we make use of them for every textural input data. As discussed earlier there is every need to walk through the code for customization, apart from the models in the discussion. If custom implementation demands more annotation apart from models, there are alternative options to go for custom models which are mentioned in Jenny Finkel (2006). One factor that effects the performance is the training source, be cautious about the size of the training files. Main inference is the developer has to be cautious over the no entity type lists in a training file, because delay time in extracting annotation is proportional to the training data size. Query execution time crucial in designing the databases, efficient use of memory builds application efficiency so, to be selective in framing the annotation types on priority basis.

Geo-Coding of a location

Geo-coding plays an important role in representation of physical address on visual animated maps. Earth surface is divided in horizontal and vertical angles, the horizontal lines represent latitude and vertical lines represent longitude. For latitude the equator is taken a reference point as 0 Degree and towards poles end 90 Degrees, the Greenwich (prime meridian) and total 360 Degrees span of vertically into equal halves of 180 Degrees of east and 180 Degrees of west. Geo-coding coordinates are decimal values of latitude and longitude. As the objective of this work it demands for geo-coding (converting location or address in to latitude and longitude coordinates) the contemporary mechanism is to make use of the API's having functionality and huge data corresponding to the geographic coordinates.
In this context it's necessary to analyze the available resources, evaluate the relative functionality, usability and flexibility in customization of the resource. In which way the available research satisfies the user assumption in building a new system, by updating requirements of specific scenario in the available system. As per Dr. Ela Dramowicz (ela Dramowicz,2004) the address need to analyzed taken care of providing information like street name, postal code or the area name, example county, district . Which need to be conscious over providing approximate address string at least in, finding the geo-coordinates of an address? In (ela Dramowicz, 2004) there is a discussion of three methods in finding geo-coordinates they are through street address, postal codes and boundaries, which is interesting, but not briefed about the implementation.
The popular geo-coding API available in use is "Google geo-coding" (Mono marks, 2010) and "yahoo place finder" (yahoo 1.0 2010) both are providing web-services to find the geo-coordinates of the user query. Mono marks (2010) and yahoo 1.0 (2010) have provides services which require authorization and both have similarity in http request to the respective URI and response formats of JSON and XML. As the service is on commercial basis and to control load of unlimited request from users, they place restriction over the accessibility by limiting the user requests. Mono

marks (2010) is meant to have client-side purpose by limiting the 2500 requests / day for each IP address, whereas yahoo 1.0 (2010) was concerned for server-side limited 50,000 requests/day for the user application. Mono marks (2010) policy guidelines states that using geo-coding results without plotting on Google map is prohibitive. In comparison Mono marks (2010) and yahoo 1.0 (2010) both are efficient and accurate but Mono marks gives best results.

Goldberg & Wilson (2011) have explained about the Batch processing of addresses, but most worrying factor is the limitation over the request rate. In batch processing, file size and file formats are taken into consideration and the input file must follow specified guidelines, which suppress the usability here.

Using web-services in custom developmental works not only suffers from restriction imposed by the service provider but, also the dependency factor affects the functionality of the user application. Be cautious over giving unstructured input address to the system, because sub location and locations names are duplicated around the globe. Conversion of address into geographic coordinate's process requires custom database of all available addresses with their corresponding latitude and longitude coordinates. But it's expensive to buy the data from the available sources.

Sentiment analysis:

Sentiment analysis become significant in today's world to analyze the corpus or bulk texts. It is evident the time constraint, high frequency of data and reports, rapid user feedbacks imposing extra burden on servicing bodies (blogging groups, market analysts, stock boards, portals). Apart from the supervision it needs an automated tool to evaluate the sentiment in a text. There is scope of study by using sentiment analysis tool in ongoing speculation in public life, customer opinion analysis, tracking the reviews of a product and to study the mass sentiment over different issues or aspects. Present it's been prioritized in research and development of certain tools to attain a better analysis over bulk data in growing economies. Rahman, Mukras & Nirmalie (2007) in their paper explained that a text or document can be analyzed and bifurcated into positive and negatives sentiment, and in order to that they have designed a procedure to evaluate the input data corpus, primary task is part-of speech tagging to each phrase of input text with predefined coding. Rahman, Mukras & Nirmalie (2007) have defined a secondary task of word/phrase frequency detection in given text, and extracting "bi-Gram" (sentiment rich phrases/words) and assign a score which is predefined for sentiment or emotion words (based on the intensity of the word). Finally by aggregation of positive and negative sets of score, the predictive score of the sentiment in the text get excavated; in this regard an algorithm was derived (Rahman, Mukras & Nirmalie, 2007).

Rudy & Mike (2009) introduced new sentiment analyzing tools for implementation and have derived a new combined approach used for single classifier for sentiment analysis. Rudy & Mike (2009) have extended the Rahman, Mukras & Nirmalie (2007) and developed a new approach using distinct classifier a two levels micro-level and macro-level, and averaging the sentiment at both levels. Let's take the scenario, we have a corpus of files each file will get analyzed by using the set of available classifiers and have taken their corresponding average score of sentiment.

Rudy & Mike (2009) have measured the accuracy of each classifier on the file and take the highest accuracy score of sentiment which is known as micro level averaging, it is important because one classifier predict a wrong score can affect the entire mechanism. Secondly by choosing the micro averages from a list and average those in macro level get overall predictive sentiment score of corpus (list of document or files) or datasets. Rudy & Mike have also stated the Rudy & Mike (2009) have made an evaluation of the contemporary available sentiment classifiers and briefed about the implementation procedure, but it was a complex implementation as the response time is high because of the complex procedure

(Rahman, Mukras & Nirmalie, 2007). Rudy & mike (2009) have defined a hybrid system by inducing a lot of rule based test which reduces adaptability and raises complexity, it influence the usability. The implementation efficiency and usability is predominant than complex theoretical procedure in choosing suitable sentiment classifiers in relative to both (Rahman, Mukras & Nirmalie, 2007) and (Rudy & mike, 2009). Now it raises the question a simple mechanism reusable sources or readily available resources to make use in sentiment mining. Always there is option to switch on alternatives like sentiment analysis API (Application Programming Interface), for the custom development programming languages like JAVA, .NET etc (libraries or API). Alias- I (2008) Provides JAVA library for semantic analysis and developed a supervised system which need to train on user specific sensitive models. Alias- i (2008) needs to train classifier with user context aggregated dataset's initially to run sentiment application. Limitation over the usability and adoptability to custom application is, (Alias- i, 2008) only operates on corpuses or datasets. Nowhere it's defined about simple text (user argument) processing apart from taking input as corpus. Cunningham et al (2011) and Alias-i, (2008) have made quite similar mechanism in mining the sentiment. A

simple and efficient classifier need to get build based on the limitations and constraints laid by, early implementation of sentiment analyzers.

Significant or key phrases extraction

Phrase is a word or a set words that form meaningful sentence, "significant Phrase" means word or set of words have significance in a statement or text. Significant phrases assist a reader or user to derive partial inference in quick review of article or text. It showcases potential idea behind the text, though highlighting the words that have potential impact on framing the sentences.

Metadata of a document or text present the key information, which elevates prominence of data provided by the document (corpus or text). Now it arises how to detect and extract significant phrases from text. Turney P.D (2000) states relative difference between human generated and machine generated key phrases, as the perspective humans vary by one another also it contradicts the machine generated ones sometimes. Turney P.D, (2000) Proposed an algorithm to extract phrases having significance, by aggregated list of common words and adverbs matches the text and extracted rest and listed separately.

FEATURES

Scalable

Hadoop is a highly scalable storage platform, because it can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel. Unlike traditional relational database systems (RDBMS) that can't scale to process large amounts of data, Hadoop enables businesses to run applications on thousands of nodes involving thousands of terabytes of data.

Cost effective

Hadoop also offers a cost effective storage solution for businesses' exploding data sets. The problem with traditional relational database management systems is that it is extremely cost prohibitive to scale to such a degree in order to process such massive volumes of data. In an effort to reduce costs, many companies in the past would have had to down-sample data and classify it based on certain assumptions as to which data was the most valuable. The raw data would be deleted, as it would be too cost-prohibitive to keep. While this approach may have worked in the short term, this meant that when business priorities changed, the complete raw data set was not available, as it was too expensive to store. The cost savings are staggering: instead of costing thousands to tens of thousands of pounds per terabyte, Hadoop offers computing and storage capabilities for hundreds of pounds per terabyte.

Flexible

Hadoop enables businesses to easily access new data sources and tap into different types of data (both structured and unstructured) to generate value from that data. This means businesses can use Hadoop to derive valuable business insights from data sources such as social media, email conversations or clickstream data. In addition, Hadoop can be used for a wide variety of purposes, such as log processing, recommendation systems, data warehousing, market campaign analysis and fraud detection.

Fast

Hadoop's unique storage method is based on a distributed file system that basically 'maps' data wherever it is located on a cluster. The tools for data processing are often on the same servers where the data is located, resulting in much faster data processing. If you're dealing with large volumes of unstructured data, Hadoop is able to efficiently process terabytes of data in just minutes, and petabytes in hours.

PROPOSED SYSTEM

We have a tendency to investigate the period nature of Twitter, devoting specific attention to event detection. Linguistics analyses were applied to tweets to category verifies them into a positive and a negative class. We have a tendency to regard every Twitter user as a device, and set the matter as detection of a happening supported sensory observations. Location estimation strategies like particle filtering area unit are used to estimate the locations of events. As associate degree application, we have a tendency to developed associate degree earthquake coverage system, which could be a novel approach to advice folks promptly of associate degree earthquake event.we tend to take 3 steps

1) We analyze no of tweets associated with target events;

2) We got to style such a probabilistic module to research and extract events from those tweets and predict locations of events with category verifying as positive and negative class.

3) Finally developed coverage method that excerpts tremors from Tweet and shows a message to registered users.

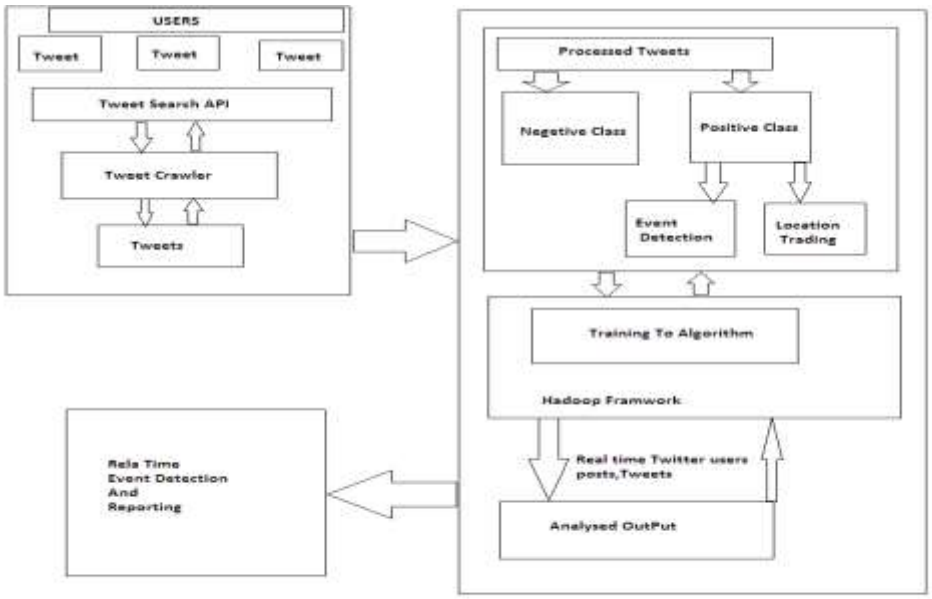

Fig 1.System architecture

System Architecture-

1) Tweet search API window collects tweets regarding events I large scale.
2) We crawl no of tweets using tweeter crawler to find out useful Tweets and scripted to processing.
3) Processed twitter distinguished between "+ class and - class" by using algorithm.
4) From positive class we find out event detection and location using Hadoop framework training algorithm.
5) Lastly we improve an actual time tweeter operator's method to report real time event detection and analysis of overall system

LIMITATIONS

There are other parameters and strategies for performance metric of information retrieval system, such as the area under the precision-recall curve (AUC) For web document retrieval, if the user's objectives are not clear, the precision and recall can't be optimized.

CONCLUSION

In this Project we aim to serve a processed twitter tweet database to frontend third party visualization applications leads to Locate & detect real time event like Earthquake. Text analysis focused on processing the tweets to extract information from the raw data of tweet, which can benefit the frontend application in projecting more information to the user, in terms of usability and exploring text-engineered data. Here we focused on Earthquake related datasets which after preprocessing and sorting as positive & negative given to training algorithm as input while real time tweets coming from different users related such event will detect event & location. And at same time as per our proposed system module it will report the event. We have given name to this project as 'CrisisCall' and it will suits this architecture while our system will do very fast processing on huge datasets at very short time as we used Hadoop framework for it. And it will definitely very helpful to fastest detection of events like happed in Nepal.

FUTURE SCOPE

In future we can work for datasets for various datasets & connectivity between our system and Disaster management team of nation.

REFERENCES

1. LI Bing Keith, C.C. Chan," A Paralleled Big Data Algorithm with MapReduce Framework for Mining Twitter Data", IEEE Fourth International Conference on Big Data and Cloud Computing, 2014
2. M. Sarah, C. Abdur, H. Gregor, L. Ben, and M. Roger, "Twitter and the Micro-Messaging Revolution," technical report, O'Reilly Radar, 2008.
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 (WebKDD/SNA-KDD '07), pp. 56-65, 2007.
4. B. Huberman, D. Romero, and F. Wu, "Social Networks that Matter: Twitter Under the Microscope," ArXiv E-Prints, http:// arxiv.org/abs/0812.1045, Dec. 2008.
5. H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, A Social Network or A News Media?" Proc. 19th Int'l Conf. World Wide Web (WWW '10), pp. 591-600, 2010.
6. G.L. Danah Boyd and S. Golder, "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter," Proc. 43rd Hawaii Int'l Conf. System Sciences (HICSS-43), 2010.
7. A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment," Proc. Fourth Int'l AAAI Conf. Weblogs and Social Media (ICWSM), 2010.
8. P. Galagan, "Twitter as a Learning Tool Really" ASTD Learning Circuits, p. 13, 2009.
9. W. Zhu, C. Chen, and R.B. Allen, "Analyzing the Propagation of Influence and Concept Evolution in Enterprise Social Networks Through Centrality and Latent Semantic Analysis," Proc. 12th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD '08), pp. 1090-1098, 2008.
10. E. Scordilis, C. Papazachos, G. Karakaisis, and V. Karakostas, "Accelerating Seismic Crustal Deformation before Strong Mainshocks in Adriatic and Its Importance for Earthquake Prediction," J. Seismology, vol. 8, pp. 57-70, http://dx.doi.org/10.1023/ B:JOSE.0000009504.69449.48, 2004.
11. T. Bleier and F. Freund, "Earthquake [earthquake warning systems]," IEEE Spectrum, vol. 42, no. 12, pp. 22-27, Dec. 2005.
12. M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging Topic Detection on Twitter Based on Temporal and Social Terms Evaluation," Proc. 10th Int'l Workshop Multimedia Data Mining (MDMKDD '10), pp. 1-10, 2010.
13. B. O'Connor, R. Balasubramanyan, B.R. Routledge, and N.A. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," Proc. Int'l AAAI Conf. Weblogs and Social Media, 2010.
14. M. Ebner and M. Schiefner, "Microblogging - More than Fun?" Proc. IADIS Mobile Learning Conf., pp. 155-159, 2008.