Big Data Analytics for Anomaly Detection in Online Reviews

Rashmi I Chamakeri[1], Roshnee Krishnamurthy[2], T Roja[3], Vaishnavi M[4],
Dr. Jitendranath Mungara, HOD and guide, Department of Information Science and Engineering, New Horizon College of Engineering.

***Abstract*—Big Data is an evolving term that describes any voluminous amount of heterogeneous data with respect to (3v) variety, volume and velocity of the data. In any business model, the feedback of the customers is a primary concern as their review plays a crucial role in the business development. In the existing system, the detection of anomalies in wireless mobile networks is carried out by using spatio-temporal information about a user through big data analytics. This method involves anomaly detection but does not specify the type of anomaly. We propose a system that determines if the review given by a user about a particular event is true by tracking the location of the user in a backend process. We achieve this by using truth discovery algorithm, sentimental analysis and android-web server communication using WAP. This increases the security of the user and also enhances the performance of the system, regardless of the size of the dataset along with providing accurate results.**

Keywords— Anomaly, Big Data, Big Data Analytics, Data Set, Truth Discovery, WAP.

I. Introduction

Big Data analytics is an enormous term, that incorporates methods and technologies, hardware and software for collecting, managing and analyzing large scale heterogeneous data in real-time. Big Data analytics works on the entire data as opposed to only sample data in conventional schemes. In the case of small data, analysis was performed by randomly selecting samples (partial data) that were considered as representative of the whole data. Due to analysis of only partial data, the information extracted is inaccurate and incomplete. Thus, the decisions made are suboptimal and the performance achieved is poor. Especially in the case of real network analysis and troubleshooting, precise and quick information is desired for providing exact solution, which can only be possible if whole/Big Data is analyzed. For current and the envisioned 5G mobile networks, Big Data is an evolving term that describes any voluminous amount of structured, semi structured and unstructured data.

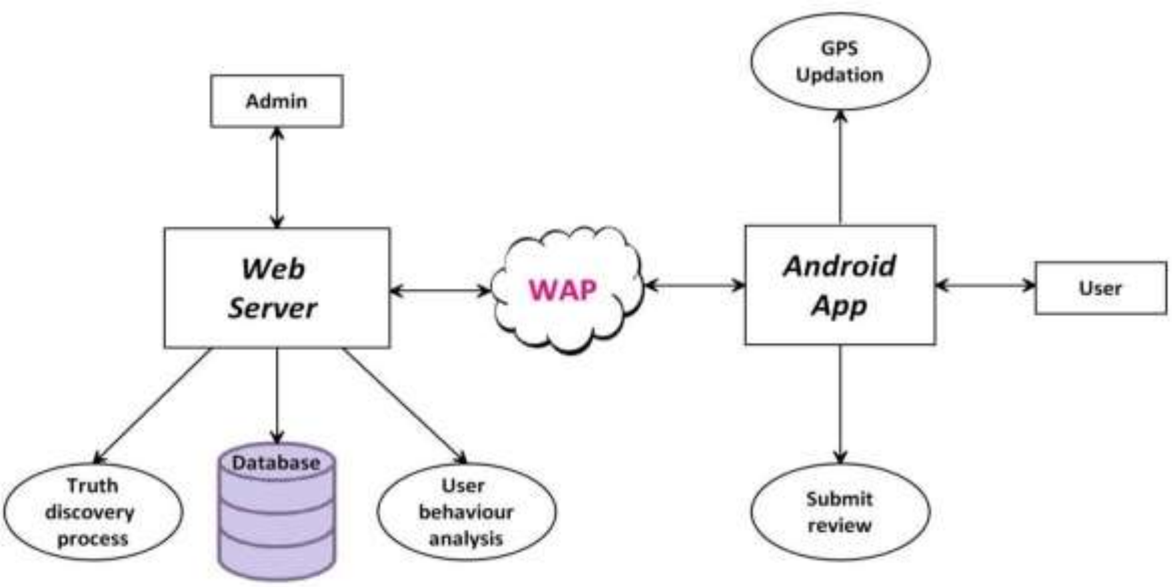
In the enormous amount of data generated in the 5G networks [5], there can be a possibility of anomalies present [12]-[15]. An anomaly is an unusual behaviour in a wireless network. These anomalies can be detected using various Big Data analytics techniques such as truth discovery, sentimental analysis, rate analysis, k-mean clustering, and behavioural analysis and so on. The anomaly detection in [3] wireless networks of the existing system doesn’t specify any particular anomaly. The growing smart phone user base has enabled mobile crowd sourcing applications on a large scale have emerged, which represent the mobile equivalent of online crowd sourcing markets. Crowd sourced detection of spatial events is one such application where participants detect events while moving around in their daily lives. These events are arbitrary phenomena that the task requester is interested in, e.g., potholes on streets, graffiti on walls and bike racks in public places .This truth discovery problem is uniquely distinct from its online counterpart in that it involves uncertainties in both participants’ mobility and reliability. Nowadays, the reviews given about any event, product or service play an important role in the business development.In our proposed system, when the user gives the feedback using wireless devices for a particular event, the system identifies if the review is true or not by Big Data analytics techniques and also verifies his presence by matching the location of the user to that of the event at the backend which provides accurate results and thus, enhancing the performance of the system. This system helps many industrial businesses to improve in their business point of view as the user reviews plays a crucial role.

II. Relevant Work

Cellular networks induce a large amount of data and 5G networks further increases the data and it is expected to become even more complex. In order to address upcoming challenges, SONs approach is used which leads to a number of limitations and the effectiveness in a 5G scenario would not be sufficient to achieve the global network optimization and operational cost reduction goals. Thus, in the future we can opt for Big Data Empowered SONs (BES) as the most effective solution given in [1]-[2] to autonomic network management for 5G systems. In this enormous data there could be anomalies [6]-[8] during calling activities due to large amount of users. These anomalies can be detected using advanced technologies like CADM using CDR. The solution to this problem is to apply knowledge based anomaly detection methods and set rule policies depending on network behaviour. Here, we present one variant of knowledge based technique, a rule-based technique, for detecting network anomalies for users traveling from one city to another. The method is flexible as well as robust for the detection of anomalies. We use an approach for anomaly detection by analysing call-detail records in combination with recent Big Data analytical tools (Hadoop (HDFS, Map-Reduce)).

Since this algorithm affects the performance and security [9]-[11] of the users, we can use other wide range of algorithms that can be applied for detecting anomalies with the most efficient one being clustering techniques. The most important, unsupervised learning processes such as K-means clustering algorithm which is a very simple algorithm for finding useful patterns. Its inability to escape from local optima can be overcome by combining with the PSO algorithm. PSO is a high efficient heuristic technique having the capability to escape from local optima and with low computational complexity. The anomaly detection technique combines the K-means and the PSO algorithm and it is performed by comparing real traffic and clusters centroids thus enhancing the performance and securing the privacy of theusers.

III. SYSTEM MODEL

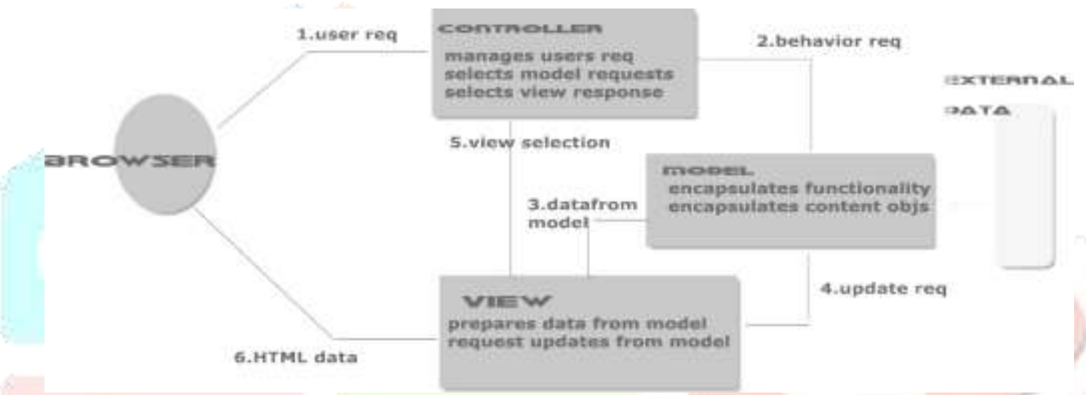


In the above figure, the architecture is divided into two modules, Web server and android application, which is connected via Wireless Application Protocol (WAP). This architecture includes a common database for both modules, MySQL. the Web Server module includes an administrator who manages the web application and also the Big Data techniques such as truth discovery, behaviour analysis and so on. The android application includes sub modules such as user log in, registration, review submission and so on.

A. *Web Application*

In our proposed system, we use the MVC architecture, model is the MySQL database, the view is the user interface which is the android and web application and the controller is the server that connects both the database to both the applications.

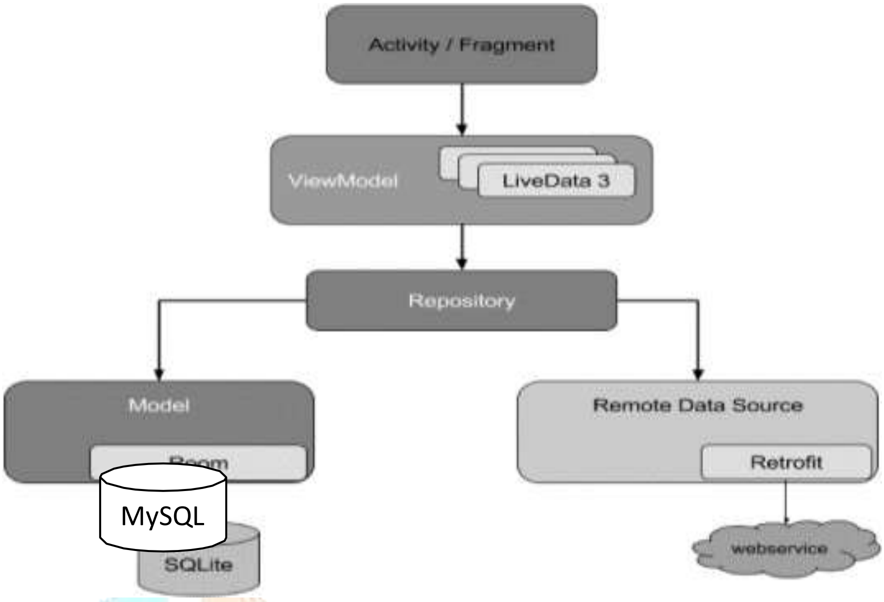
Administrator controls the Web application by performing various activities like logging into the homepage in the application where it has the option of changing the password, adding of an event, adding particulars like the location (latitude and longitude) and date of the event. Administrator has access to the user details obtained during registration. To detect if the review given by the user for the event added by the administrator is true, first performs a survey for that particular event. The reviews are the preprocessed where noisy words are removed. The admin then performs rate analysis by comparing the review given by the user with the inbuilt values in the positive and negative tables. If the positive count is more than the negative count, the chance of the feedback given by the user is more. To confirm this, we apply the truth discovery algorithm. Now, track the user location and match their latitude and longitude with that of the event to ensure the presence of the user only if it is within the Euclidean distance (200 meters). If the location matches, the rating of the event is considered true.



B. Android Application

In the android system, each application in the system is called an activity/fragment and these applications can be accessed by the user with the help of the view model (interface). The user can give the review with the help of the android application installed in the device using APK file. This information is stored in MySQL database which can be accessed by the administrator of the Web application. The web application and the android application share the same database.

Once the user rates the event by logging into the application, the review and the location of the user will be stored in the database and will be cross verified from the actual location of the event using various techniques of big data analytics. This gives the accurate results with the output being a true review or not.



IV. ALGORITHMS

A. K-means Algorithm

K-means [4] is one of the simplest, unsupervised clustering methods used to solve clustering problems, especially for large datasets. It involves partitioning methods in which a database containing n objects is partitioned into a set of k clusters. It assumes fixed number of cluster k known a priori. Determining the optimal value of k itself differs from dataset to dataset and method to method. After determining k, the main task involves finding partition of k clusters that optimizes the chosen partitioning criterion. Given the input to the algorithm is k, the k- means algorithm partitions a given dataset into k clusters so that the resulting intracluster homogeneity is high and intercluster homogeneity is low. Cluster similarity is measured in terms of mean value of the objects in a cluster, which is usually cluster's centroid or centre of gravity. The k-means algorithms clustering can be summarized in following few steps.

1) Choose randomly the initial value of k from the space represented by the objects being clustered. The values of k represent the initial values of the centroids of k clusters.

2) Compare the distance of each of the objects to each of the centroid, and assign the object to the cluster with closest centroid.

3) Recalculate the centroid of clusters by finding the mean value of the new cluster formed in step 2).

4) Repeat steps 2) and 3) until the centroids become stationary.

After clustering through k-means, the clusters containing the fewest number of objects are considered to be anomaly. There are many other ways to detect an anomaly after grouping them into k clusters. Since the activity of the user at a grid is always changing based on events and occasion, date, and time, anomaly will be detected only when there is unexpectedly high rise in their activity and such activities are observed to be fewer and will be grouped into a different cluster by k-means.

B. Preprocessing of the data

In this paper, the data is preprocessed before continuing any further activities. The data that is preprocessed is the user reviews where all special characters are removed and only meaningful words are retained.

Step 1: Let n be the number of reviews fetched from the excel sheet.

Step2: Each review will be considered as the new input for the process and can be given by i Step 3: the

reviews are converted to lower case initially and can b given as j

Step 4: the j reviews are again cleared by the removal of Special characters and can be called as k

Step 5: the k reviews are again sent for removing the stop words as well as removing the nouns and can be kept as keywords

C. Sentiment analysis

Sentiment analysis or opinion analysis refers to processing natural language, text analysis and computational linguistics, to systematically identify, extract, express, and study subjective information. Sentiment analysis aims to determine the perspective of a speaker, writer, or other subject with respect to some topic or the overall content or emotional reaction to a document, interaction, or event. The attitude may be a judgment or evaluation. The algorithm used in this paper is as follows

Step 1: Collect the preprocessed keywords from the data base which are seperated by “,” Step2:

Each keyword will be considered as i.

Step 3: I keyword will be checked in the noun as well as the standardistion table to check the sentiment of the keyword.

Step 4: I will be increasing by i+1 as well as compring and aagin storing in database .

Step 5: Collect all the keywords of ith row and check the instance of the positveness (P) as well as negativeness(N)

If(P(i)>N(i)) then, the sentiment of review will be Positive Else

Negative Else neutral

D. Truth Discovery

In the era of information explosion, data is continuously being generated through a variety of channels, such as social networks, blogs, discussion forums, crowdsourcing platforms, etc. Motivated by the strong need to resolve conflicts among multi-source data, truth discovery has gained more and more attention. When choosing a truth discovery approach for a particular task, users and developers can refer to this comparison as guidelines.

Input: Information from sources {v s o} o∈O,s∈S

Output: Identified truths {v ∗o }o∈O and the estimated source weights {ws}s∈S . 1: Initialize

source weights {ws}s∈S

2: repeat

3: for o ← 1 to |O| do

4: Truth computation: infer the truth for object o based on the current estimation of source weights;

5: end for

6: Source weight estimation: update source weights based on the current identified truths; 7: until Stop

criterion is satisfied;

8: return Identified truths {v ∗o }o∈O and the estimated source weights {ws}s∈S .

V. CONCLUSION

Big Data analytics can be utilized to understand the users' contextual information such as mobility pattern.The information extracted from this may contain anomalies and the anomaly detection techniques will completely enhance the performance of the system. Reviews play an important role for the development of an organization. So, when the user attempts to add the review on particular event, our proposed system may help in detecting whether the review is real or fake by matching user location with the event location at the backend through the various Big Data analytic techniques such as sentimental analysis, truth discovery and rate analysis. Hence, it enhances the security of the user regardless of size of the dataset along with providing accurate results helping in industrial growth.

VI. FUTURE WORK

In this system, if the user location is within a certain radius of the event location, the reviews given by that user is considered to be true. Future work would include trying to expand this radius to be able to apply this system to much larger events. Future enhancements can be done with various factors such as users can provide reviews later where history of the location can be verified for his presence. It can also avoid location tracking of the user to provide complete security. With the complex data being generated, system can be implemented on the big data platform with wide range of data storage option.

VII. REFERENCES

[1] MdSalikParwez, Student Member, IEEE, Danda B. Rawat, Senior Member, IEEE, and Moses Garuba, Member, IEEEBig, "Data Analytics for User-Activity Analysis and User Anomaly Detection in Mobile Wireless Network", IEEE transactions on industrial informatics(vol. 13, no. 4, August 2017),Pages: 2058 – 2065.

[2] Nicola Baldo, LorenzaGiupponi, JosepMangues-Bafalluy,"Big data empowered self organized networks", 20th European Wireless Conference, European Wireless2014 Pages:1–8.

[3] IlyasAlperKaratepe, EnginZeydan,"Anomaly detection in cellular network data using big data analytics", 20th European Wireless Conference, European Wireless 2014,Pages: 1 – 5.

[4] Mois´es F. Lima, Bruno B. Zarpel˜ao, Lucas D. H. Sampaio, Joel J. P. C. Rodrigues, TaufikAbr˜ao and Mario LemesProenc¸aJr,"Anomaly detection using baseline and K-means clustering",18th International Conference on Software, Telecommunications and Computer Networks,SoftCOM 2010,Pages: 305 – 309.

[5] DialaNaboulsi, RazvanStanica, Marco Fiore,"Classifying Call Profiles in Large-scale Mobile Traffic Datasets",IEEE Conference on Computer Communications, IEEE INFOCOM 2014,Pages: 1806 – 1814.

[6] Kai Yang; Ruilin Liu, Yanjia Sun, Jin Yang, Xin Chen,"Deep Network Analyzer (DNA): A Big Data Analytics Platform for Cellular Networks",IEEE INTERNET OF THINGS JOURNAL, VOL. 4, NO. 6, DECEMBER 2017,Pages: 2019 – 2027.

[7] PedroCasas; Juan Vanerio,"Super Learning for Anomaly Detection in Cellular Networks",IEEE 13th International Conference on Wireless and Mobile Computing, Networking and Communications

(WiMob),2017,Pages: 1 – 8.

[8] DuyguSinancTerzi, RamazanTerzi, SerefSagiroglu,”Big Data Analytics for Network Anomaly Detection from Netflow Data”, International Conference on Computer Science and Engineering (UBMK),2017,Pages: 592 – 597.

[9] B. Balasingam, M. S. Sankavaram, K. Choi, D. F. M. Ayala, D. Sidoti, K. Pattipat, P. Willett, C. Lintz, G. Commeau, F. Dorigo, J. Fahrny”Online anomaly detection in big data”,17th International Conference on Information Fusion (FUSION),2014,Pages: 1 – 8.

[10] Pedro Casas, Francesca Soro, Juan Vanerio, Giuseppe Settanni, Alessandro D'Alconzo,”Network security and anomaly detection with Big-DAMA, a big data analytics framework”,IEEE 6th International Conference on Cloud Networking (CloudNet),2017,Pages: 1 – 7.

[11] Abdul Razaq,HuagloryTianfield, Peter Barrie,”A Big Data Analytics Based Approach to Anomaly Detection”,IEEE/ACM 3rd International Conference on Big Data Computing Applications and Technologies (BDCAT),2016,Pages: 187 – 193.

[12] Laura Rettig, MouradKhayati, Philippe Cudré-Mauroux, Michal Piórkowski,”Online Anomaly Detection over Big Data Streams”,IEEE International Conference on Big Data (Big Data),2015,Pages: 1113 – 1122.

[13] RoimahDollah, HazleenAris,”A Review of Sector-Specific Big Data Analytics Models”, IEEE Conference on Big Data and Analytics (ICBDA),2017,Pages: 72 – 80.

[14] ShengjieXu, Yi Qian, Rose Qingyang Hu,”A Data-driven Preprocessing Scheme on Anomaly Detection in Big Data Applications”, IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS): BigSecurity 16: The Fourth International Workshop on Security and Privacy in Big Data,2017,Pages: 814 – 819.

[15] Annie Ibrahim Rana, Giovani Estrada, Marc Solé, Victor Muntés,”Anomaly Detection Guidelines for Data Streams in Big Data", 3rd International Conference on Soft Computing & Machine Intelligence,2016,Pages: 94 – 98.