A Survey of Call Detail Records (CDR) Analysis

Mohamed A. Zawra
Faculty of Computers and Artificial Intelligence, Helwan University, Egypt

O.E. Emam
Faculty of Computers and Artificial Intelligence, Helwan University, Egypt

M. Elemam Shehab
Faculty of Computer Science
AASTMT, Egypt

ABSTRACT

Call Detail Records (CDR) is one of the most important sources of information that could be used in many different aspects and ways to help the community. It is one of basic criminal activity detection techniques and it also creates new prospects for the telecommunication field. Never the less the analysis of CDR is not an easy task as it contains a huge amount of data with different varieties making it seen as a big data source. Consequently, big-data techniques and tools need to be employed to extract insights from such data. Not only that but also different graphical tools could be used for analyzing the huge amount of extracted data. There are many different challenges needed to be addressed when handling CDR. This survey paper presents what is CDR, how and why is it being used. Finally, it reviews the proposed frameworks in the literature that analyze the CDR to help prevent criminal acts.

Keywords

Call Detail Records (CDR), Mobile Communication, Big data, anomalies, analysis.

1. INTRODUCTION

Recently, the mobile communication is becoming one of the most important things in our lives and forms a huge part of our daily lives. This is because people use their phones doing many activities starting from chatting with friends to business calls and meetings. Analyzing the data generated from the use of these mobile communications is not an easy task as there are almost more mobiles than humans. Furthermore, the amount of data generated, including calls, messages, texts and voice, is very huge and can be used for many different purposes. For example, the police could track down someone or predict a criminal's next move through understanding the behavior of the user by analyzing that data [1]. Unfortunately, studying and analyzing such huge amount of different frequently interchanging data is a very challenging task [2]. The Call Data Records (CDR's) are records that include a lot of different information concerning the mobile or telecommunication transactions like the accessed websites, different called parties, usual call durations and the cell's ID. One of the main reasons why CDR's is very popular is the fact that it provides an accurate statistical demonstration of large heterogeneous group of people through their data in a three-dimensional and time based manner [3]. Even though CDR is the best way to analyze and study this type of gathered data. It is still considered an extremely hard task including a lot of tedious tasks where a lot of different tools could be used. The rest of the paper is organized as follows. Section 2 presents a background about the CDRs. A literature review on different proposed CDR's systems is presented in section 3. Finally, section 4 concludes the paper.

2. BACKGROUND

According to [3], CDR is the produced data records from the use of different telecommunication equipment to document and store all the details of the transaction done through telecommunication. The details or attributes of a transaction include the calls duration, start time, end time, called number, caller ID. The authors highlighted the fact that analyzing these data is a very difficult and tough task to be carried on due to its extremely huge size.

The importance of CDR's either for telecom industries or for other industries and how could they be used has been discussed by olanrewaju et al. For example, the uses of CDR inside the telecom industry includes real-time decision and analysis based on dynamic monitoring of the network, improvement of customer's experience, profit analysis of either services or customers, and marketing. While CDRs outside a telecom industry is used in targeted marketing , social studies, emergencies , and last but not least by police and governmental agencies to help them fight crime and capture criminals. The authors have discussed the fact that CDR is basically a big data source since it has the same basic characteristics including volume, velocity, and variety and stated that the different technologies of big data are extremely applicable when it comes to CDRs either as parallel DBMS solutions or MapReduced based solution. Through their survey they studied the strength and shortages present in the current research related to CDR solutions. The findings of the study stated that different solutions suffer from long average response time, scalability issues, some of the methods discussed had problems with real time processing and others had a problem with handling or investigating different aspects. It was highlighted that the CDRs based on big data are more cost effective with better performance and also the ones based on parallel DBMS had better measures when it came to predefined processes and querying while the ones based on MapReduced is more flexible when it comes to more complicated analysis algorithms and high time process consumptions [2].

Another study discussed analyzing the calling patterns using real-time recognition data mining and other evolving algorithms [4]. They highlighted that CDRs is one of the most helpful information sources for multidimensional analysis in the telecom industries that could be based on user-centric grid based technique which helps in the utilization of the location of information when its being retrieved, or multi scale based dynamic networks, patter cube algorithm, and user centric based approaches. They also discussed different new algorithms that would help in the real time analysis that could be used in data analysis including evolving fuzzy systems where they are basically self-developing, neuro-fuzzy systems, and self-learning fuzzy rule based systems.

Another research has discussed the analysis of CDRs to observe the activities and actions of a specific user [5], where and when they started or even get a call whether it's a voice, or data, or even sms. A mediator collects the instance of the specific coordinates, time, and location from different nodes and then distributes them using ftp to their specific related departments. This is considered the generation process of the CDR data that is then being analyzed through specific algorithms. The authors stated that one of the most challenging aspects in the analysis of such huge amount of data is managing the same structure through the whole life cycle of the service as the data being analyzed has multiple heterogeneous modules.

The authors in [6] have studied anomaly based CDR mentioned that the fact that gathering the information of different mobile users at any given instance is a very vital aspect as it provides instant information that could help in criminal fighting, health and other areas. But the fact that a sudden increase in the collected data could happen at any given point might have a negative impact on the Quality of Service (QoS) being provided to the users and on the network's performance. They concluded that in order to handle the reduction in the QoS, the network's admin should be able to add resources for the following time frames to mitigate the network congestions.

3. LITERATURE REVIEW

A CDR based model using PageRank algorithm called "CDR corpus" is proposed in [3]. The authors have explained how that algorithm would help in the understanding of the effect of any occurring communication stream on the entire network leading to better detection of abnormalities in the communication process. Their model uses only the calls and text message communication. The implementation of the PageRank algorithm helps in the capturing of any base station with higher flow than normal or even the highest traffic at any given time frame. Then, these captured behavioral abnormalities in the network are associated with different patterns and behaviors that are used later to detect malicious activities on the network. The CDR corpus comprises information about communication flows between pair-wise connected base stations, such as the number of active connections, the total duration of these flows, and the number of users currently connected to them, as well as their geographic location. For each base station throughout each time slot, the CDR corpus is used to produce a PageRank score. It enables for the detection of base stations that have a high volume of communication flows. If the PageRank score on a certain base station increases spontaneously, it means that activity on that base station is also higher than normal. This can be reported as anomalous behavior when differentiated to the behavior of earlier time frames. Then the impact of the amount of calls and the duration is assessed by the PageRank score calculus. The model was tested in real life on the data set based on a country-level size and the results were successfully promising but the authors believe that they still need to make more real-time testing with bigger datasets.

An anomaly based and traffic prediction CDR mobile networks model has been presented in [6] with the help of machine learning algorithms. The first step in the proposed framework is to discover network anomalies using the CDR data. They employed k-means clustering which is an unsupervised machine learning approach to authenticate and verify abnormalities. The detection of the anomaly based techniques is very vital for appropriate resource distribution design as well as defect identification and avoidance through effective anomaly detection. Second, they remove anomalous activities from the data and train a neural network model on it. Then the effect of anomalous activities in model training as well as mean square error of anomaly and anomaly-free data by running anomaly and anomaly-free data through this model are observed. Finally, they did an autoregressive integrated moving average (ARIMA) model to forecast a user's future traffic based on data stationarity of the time and parameter estimation where the main purpose of this stage is finding the optimal value for the parameters by the use of autocorrelation functions and partial autocorrelation function for the completion of the prediction process. The authors believe that this model helps satisfying the requirements of the user without affecting the QoS being provided.

In [7], the authors created a system based on graphical representation using Neo4j on the data set extracted from the CDR which is an embedded, disk based, completely transaction Java engine for saving structured data graphs instead of tables. Neo4j has been proven to have lager scaling capabilities with an easy to apply API. They assumed that any call has 4 aspects connected together including the caller, the receiver, location of both, the duration. The tables of the collected data are then converted into graphs then the records are analyzed to find the possible suspect through. The graphical representation of this data helps making the whole processes easier that's why the authors decided to only focus on the people involved in the call instead of focusing on the people, duration, and location. They created an uninterrupted relation between people involved in a call called "KNOWS".

A CDR model for criminal investigations based graphical representation of the attributes of the archived record details proposed in [1]. Their motive was to help investigation parties in the tracking of criminals through the data of the mobile service providers. They built the framework using database of previous criminal cases and anti-social elements which they believe would help in solving many criminal cases and they are using the graphical representation for easier displaying of the attributes involved in the CDR. The attributes used involve the caller number, the receiver's number, the location, the duration of the call and start and end time, along with the IMEI number which is a unique number for each cell phone and its actually an important attribute. The first step of the framework is the analysis of the CDR meta-data of previous criminal cases along with the anti-social elements and analyzing the associations between them. The second part of their model is investigation using CDR where they analyze the data with the purpose of finding a specific individual based on their involvement in the crime or previous crimes after that they start to identify the potential suspects using simple search method that answer specific questions about the behavior or pattern of a specific caller ID. After that they use the search results to identify the network suspects and merge it with the CDR database in an attempt to find association between the caller Ids resulting from the search query and old cases. After the suspects are vaguely selected they start doing the visual representation and analysis of the network of suspects they reached to find the specific criminal they are looking for.

Another framework for anomaly detection using big data analytics was proposed by karatepe and Zeydan [5]. They presented a rule based anomaly detection method called CADM. The main idea was to help the mobile service providers to enhance the systems consistency and minimize the time and effort for finding location based anomalies. They believe that the use of big data analytics would create better accuracy results and at the same time would be a flexible, reliable, and easy way to get results. They used big data techniques like MapReduce and Hadoop for better results besides being cost beneficial. The first step happens when a communication is started the system records the time and location then the mediation starts to gather the data of the CDR from the nodes on the network then send the collected data to their specific department. Then the CADM starts to gather the previously collected and distributed data for the detection of any irregularities and the output of this phase is the new input of the Mediation. The authors believe that the framework provides a more flexible and sturdy way of anomaly detection for cell operators with the help of higher coordinating their configurations in the course of the provider lifecycle to meet their policy level needs.

4. CONCLUSION

Mobiles are used nowadays by almost everyone and it has many uses starting from connecting people to emergency calls to meetings in a nutshell. So they are used by everyone all the time. This created a way to help study the behavior of people using the mobile services and telecommunication. A very useful use of the huge amount of data created from these connections is trying to capture crime through analyzing that data. It could also be used for marketing reasons and much more. The Call Details Records (CDR) is considered as a big data source and the analysis of such huge amount of data is a big challenge. In this paper, we first discussed the importance of CDR and its uses as well as the techniques and technologies applied in the process. We then reviewed the different proposed frameworks that employed graphical representation and big data techniques for analyzing the CDR. We believe that further research is required to increase the effectiveness and efficiency of the analysis of that huge amount of data.

5. REFERENCES

[1] Kumar, D., Hanumanthappa, D., & Kumar, D. V. (2016). Crime Investigation and Criminal Network Analysis Using Archive Call Detail Records. International Conference on Advanced Computing (ICoAC) (pp. 46-50). IEEE

[2] S. B. Elagib, A. -H. A. Hashim and R. F. Olanrewaju (2015).CDR Analysis using Big Data Technology . International Conference on Computing, Control, Networking, Electronics and Embedded Systems Engineering (pp. 467-471). IEEE.

[3] Abuhamoud, N., & Geepalla, E. (2019). Analysis CDR for Crime Investigation using graph- based method (Neo4j). International Conference on Technical Sciences. (ICST2019).

[4] Iglesias, J. A., Ledezma, A., Sanchis, A., & Angelov, P. (2017). Real-Time Recognition of Calling Pattern and Behaviour of Mobile Phone Users through Anomaly Detection and Dynamically-Evolving Clustering. Applied Science MDPI, 7 (798), 1-14.

[5] Karatepe, I. A., & Zeydan, E. (2014). Anomaly Detection In Cellular Network Data Using Big Data Analytics. European Wireless 2014; 20th European Wireless Conference. Spain: VDE.

[6] Sultan, K., Ali, H., & Zhang, Z. (2018). Call Detail Records Driven Anomaly Detection and Traffic Prediction in Mobile Cellular Networks. IEEE ACCESS JOURNAL (July).

[7] Georgen, D., Mendiratta, V., State, R., & Engel, T. (2014). Analysis of large Call Data Records with Big Data. IPTComm.