ISSN No. 0976-5697

Volume 8, No. 1, Jan-Feb 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Threats Detection using Big Data Analytics

Savita Kumari Sheoran
Associate Professor & Chairperson
Department of Computer Science & Applications
Chaudhary Ranbir Singh University, Jind (Haryana) – India

Pratibha Yadav
Ph.D. Research Scholar
Department of Computer Science & Applications
Indira Gandhi University Meerpur, Rewari (Haryana) - India

Abstract: Social networking sites (SNS) are being rapidly increased in recent years, which provides platform to connect people all over the world and share their interests. The ubiquitous use of social media has generated unparalleled amounts of social data. Data may be in the form of text, audio, video, images. It is necessary to analyze this massive amount of data and extracting useful information from it. Information can be spread across social networks quickly and effectively, hence have now become prone to different types of undesired and malicious spammer/hacker actions. Spam's can be in the form of images, text, videos, audio etc. However, Social Networking Sites is providing opportunities for cybercrime activities. Therefore, there is a pivotal need for security in social media and industry. The aim of this paper is to detect threats in a social media networks using Big Data analytics.

Keywords: Threat classification, Security Threats, Big data Analytics

1.INTRODUCTION

With the development of Information and Communications Technologies and increasing accessibility to the Internet, organizations become vulnerable to both insiders and outsiders threats. A threat is a description of a potential undesirable incident. Information systems are constantly exposed to various types of threats, and these threats can cause different types of damages, which might lead to significant financial losses. Sizes of these damages can range from small errors, which only harm the integrity of databases, to those that destroy whole computer centers . The social media data is generated in the form of text, audio, video, images, numbers or facts that are computable by a computer. A particular data is absolutely useless until and unless it converted into some useful information. Therefore it is necessary to analyse this massive amount of data and extracting useful information from it. Information can be spread across social networks quickly and effectively, therefore become susceptible to different types of undesired and malicious spammer and hacker actions. Social Networking Sites are providing opportunities for cybercrime activities [1], hence there is an essential need for security in social media networks and industry [2]. Data are now woven into every sector of industry and function in the worldwide economy. These are generated from emails, online transactions, search queries, audios, videos, images, click streams, logs, health records, posts, social networking communications, sensors and science data, mobile phones and their applications.. By the end of year 2020, 50 billion hosts will be connected to networks and the internet [3]. With the fast development of information digitization, huge amounts of unstructured, semi-structured and structured data are generated quickly. This data is known as "Big Data" due to its volume, velocity and the variety.

There are three characteristics of Big Data [4], called "3V":

- **Volume** (the data volumes are huge which cannot be processed by traditional methods),
- **Velocity** (the data is created with great velocity and must be captured and processed quickly) and
- **Variety** (variety of data types: structured, semi-structured, and unstructured).
- Based on data quality, IBM has added a fourth V called: **Veracity.**
- However, Oracle has added a fifth V called: **Value**, highlighting the added value of Big Data.

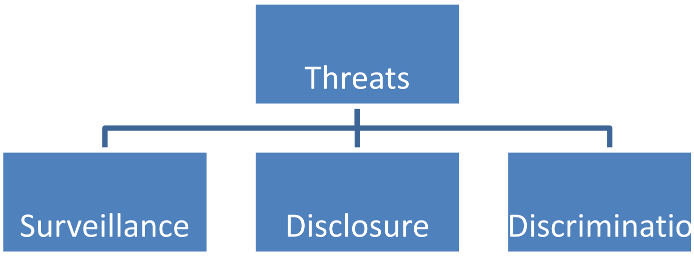
For instance a total 5 exabytes of data were produced by human until 2003. Today this amount of information is generated and collected in just two days. In year 2012, digital data was extended to 2.72 zettabytes. It is predicted that this data will be double in every two years, and reaching about 8 zettabytes of by 2015. According to IBM, 2.5 exabytes of data generated daily and also 90% of the data produced in last two years.

2. THREATS

A threat is the opponent's goal, or what an opponent might try to do to a system. In computer security a threat is a possible danger that might exploit a vulnerability to breach security and cause possible harm. Big Data is not useful without sharing data between the users [5]. But, sharing of data faces the challenge of data privacy and security .These issues have little attention till now. Only researchers pointed out that due to its big volumes Big Data creates new threats. Traditional security policies are less adequate to protect the Big Data from the threat to secure sensitive data. Big Data involve both the width of sources as well as depth of the information needed for programs to specify risks correctly, to defend against illegitimate activity and advanced cyber threats [6].Financial institution and healthcare provider are more effected if attacker attacks on their data repository because the data volume is high and government regulations are exists . Still we are lacking the proper policy to secure the data therefore hackers can call any time. It is also a big challenge by research community. Threats can occur from outside or from the inside of an organization [7]. A threat can be internal, external or both external and internal entities. Outside threats are the attacks on system by someone from outside the organization. Hackers are outside attackers who break into computer systems and cause destruction within an organization. Ecological threats can be either internal, due to natural processes or external, due to natural process that begin outside the system boundaries . There are hackers who start diffusion viruses, and these viruses cause enormous harm to the files in various computer systems. By nature of action a threat can be malicious or non-malicious. The goal of attackers on a system can be malicious or non-malicious.

2.1 BIG DATA THREATS

Three broad categories of big data threats: surveillance, disclosure, and discrimination.



Surveillance means the feeling of being watched, which can result from the collection, aggregation, and/or use of one’s information. The feeling of being surveilled might be an intrinsic problem, similar to emotional distress. It might also be a problem because such a feeling can affect how people behave, if people start to think twice about the things they do, read, or search for.

Disclosure of data outside of the context in which it was collected. One disclosure threat might be the nosy employee who looks up people he knows in a corporate database. Another might be an identity thief who successfully hacks into a database. Problems of insecurity are in this sense problems of disclosure. Less maliciously, information might be revealed to people who happen to be nearby and see the ads on another person’s computer.

Discrimination, that is, treating people differently on the basis of information collected about them. There are many different kinds of discrimination threats. The most obvious might be trying to predict membership in some protected class, such as race or religion, and then discriminating on that basis. Some might further object to any discrimination that is correlated with a protected characteristic, whether or not it forms the explicit basis for the targeting.

3. REVIEW OF LITERATURE

During the last decade, many researchers have been contributed in the areas of big data analytics. However a little works has done in the area of security and privacy. Here we are presenting the works of various authors on big data analytics..Xindong Wuet al [8] presented a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. They analyzed the challenging issues in the data-driven model and also in the Big Data revolution. Zhe Yao et al [9] described and analyzed an approach to anomaly detection using proximity graphs and the PageRank algorithm. They showed that PageRank produces point-wise consistent probability density estimates for the data points in an asymptotic sense, and with much less computational effort.
Oliver Brdiczka et al [10] proposed an approach that combines Structural Anomaly Detection (SA) from social and information networks and Psychological Profiling (PP) of individuals. SA uses technologies including graph analysis, dynamic tracking, and machine learning to detect structural anomalies in large-scale information network data, while PP constructs dynamic psychological profiles from behavioral patterns. Threats are finally identified through a fusion and ranking of outcomes from SA and PP..Kevin M. Carter et al [11] presented a method for detecting malicious activity within networks of interest. They leverage prior community detection work by propagating threat probabilities across graph nodes, given an initial set of known malicious nodes. Pi- Zubair Ahmad et al [12] presented scenarios related to identity theft, unlawful information gathering and tracking. They showed the main issue of lack of platform trust in platforms involve in federated systems and discussed the consequences of respective threats on them. Michael Fire et al [13] presented the different security and privacy risks[14], which threaten the well-being of online social network (OSN) users in general, and children in particular. In addition, we present an overview of existing solutions that can provide better protection, security, and privacy for OSN users. Arash Golibagh Mahyari et al [15] introduced an aspect of threat detection, which is identifying abrupt changes in edges’ weights over time. Wavelet decomposition method is used to separate the transient activity from the stationary activity in the edges. Amit Kumar Bhardwaj etal [16] proposed a strong and viable solution to overcome different threats, network security using data mining approach and techniques through visual graphical representation. They demonstrates two new visualization schemes called as: Grid and Platter for visualize threats. Leman Akoglu et al [17] proposed a comprehensive overview of graph-based techniques for anomaly, event, and fraud detection, as well as their use for post-analysis and sense-making in explaining the detected abnormalities.

4. PROPOSED WORK

For a solution to be viable, it must be highly scalable and support multiple heterogeneous data sources. Current state-of-the-art solutions do not scale well and preserve accuracy. Furthermore, by utilizing the data acquisition techniques, we are able to easily integrate and align heterogeneous data. Thus, our approach will create a scalable solution in a dynamic environment. Presently, no existing threat detection tools offer this level of scalability and interoperability. We will use novel data mining techniques to create an efficient threat detection solution. Particularly, in our approach, we will use primary data from social media data and/or server logs or secondary data as a dataset. After loading, data will be refined to reduce the redundancy and stop words. Database will be created from raw data collected from social networks or server logs. Our solution will pull data from multiple sources and then data feature selection and extraction techniques will be applied to the database. Database may be large to process, so after feature selection, reduction of data to relevant data will be done. An optimization algorithm will be applied for this reduction of data. After that, data classification techniques will be used to classify the dataset into malware or legitimate data. On the basis of classification techniques, we will be able to predict the anomalies or threats found on social media data and servers logs. Finally a comparison of classification techniques will be presented on the basis of result obtained in classification step. Hence security can be enhanced by detecting the anomalies on server's logs and malicious activities on social media data.

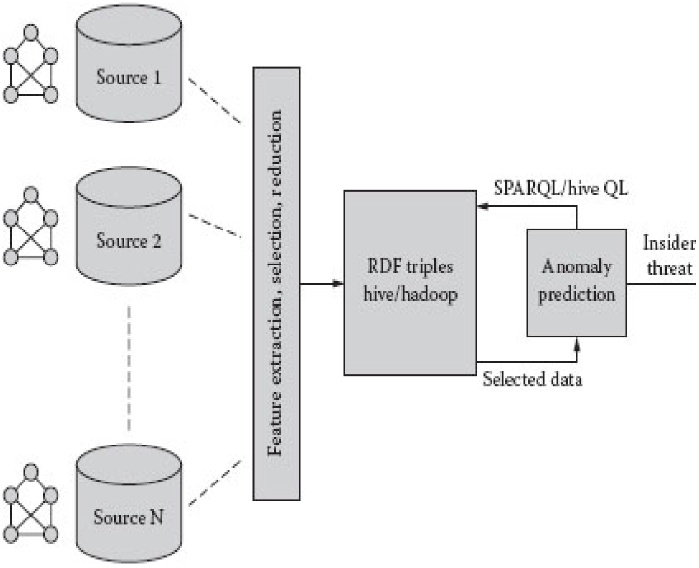


Fig1:Threat Detection Using Big Data Analytics

5. CONCLUSION

In this paper data mining techniques with big data are used to detect threast. A large graph as input then apply data mining techniques for feature extraction ,selection and reduction.Ddata is stored on Hadoop in RDF triplet format. Apply SPARQL/hive query language on selected data for anomaly detection. Big data analytics is more efficient to detect threats from social media data. We can provide security to detect malicious activity from social media data.

6. REFERNCES

1. Nandhini B.Sri, Sheeba J.I., "Online Social Network Bullying Detection Using Intelligence Techniques", International Conference on Advanced Computing Technologies and Applications, Elsevier, pp.485- 492, 2015.
2. S. Multani Harshal, Sinh Marod Amrita, Pillai Vinita, Gaware Vishal , "Spam Detection in Social Media Networks: A Data Mining Approach", International Journal of Computer Applications , Volume 115 – No. 9, pp.1-4, April 2015.
3. Giroglu Serefsa and Sinanc duygu ," Big Data: A Review" ,IEEE International Conference, pp.42-47, 2013.
4. Alguliyev Rasim, Imamverdiyev Yadigar, "Big Data: Big Promises for Information Security", IEEE International Conference, pp.1-4, 2014.
5. Beckwith Richard and Mainwaring Scott, " Privacy: Personal Information, Threats, and Technologies" , IEEE Conference, pp.9-16, 2005.
6. Nikolaos E. Petroulakis, Ioannis G. Askoxylakis, Theo Tryfonas, " Life-logging in Smart Environments: Challenges and Security Threats", IEEE Conference, pp.5680-5684, 2012.
7. Bhatt Parth, Yano Edgar Toshiro, Jose Sao, M.Gustavsson Dr.Per, "Towards a Framework to Detect Multi-Stage Advanced Persistent Threats Attacks", IEEE International Symposium on service Oriented System Engineering, pp.390-395, 2014.
8. Wu Xindong ,Zhu Xingquan ,Wu Gong-Qing and Ding Wei, "Data Mining with Big Data", IEEE Transactions on Knowledge and data Engineering,vol 26, no. 1,pp.97-107, January 2014.
9. Yao Zhe ,Mark Philip and Rabbat Michael , " Anomaly Detection Using Proximity Graph and PageRank Algorithm", IEEE Transactions on Information Forensics and Security, Vol. 7, No. 4, pp.1288-1300 ,August 2012.
10. Mahyari, Arash Golibagh, Aviyente Selin, " A multi-scale energy detector for anomaly detection in dynamic networks", IEEE Conference, pp.962-965,2013.
11. Kevin M. Carter, Nwokedi Idika, and W.William Streilein, "Probabilistic threat propagation for malicious activity detection", IEEE International Conference, pp.2940-2944, 2013.
12. Ahmad Zubair and Ab Manan Jamalul-Lail, Sulaiman Suziah, " User Requirement Model for Federated Identities Threats", IEEE International Conference on Advanced Computer Theory and Engineering(ICACTE), pp.317-321,2010.
13. Mahyari, Arash Golibagh, Aviyente Selin, "A multi-scale energy detector for anomaly detection in dynamic networks", IEEE Conference, pp.962-965,2013.
14. Raissi-Dehkordi Majid and Carr David, "A Multi-Perspective Approach to Insider Threat Detection", IEEE Conference, pp.1164-1169, 2011.
15. Miller Benjamin, S. Beard Michelle and T. Bliss Nadya, " Eigen space Analysis for Threat Detection in Social Networks", IEEE International Conference, pp.1-7, 2011.
16. Bhardwaj Amit Kumar, Singh Maninder, "Data mining-based integrated network traffic visualization framework for threat detection", Springer, pp.117–130, 2015.
17. Akoglu Leman,Tong Hanghang, Koutra Danai, "Graph based anomaly detection and description: a survey", Springer, pp.626–688, 2015.