Proc. 1st International Conference on
Machine Learning and Data Engineering (iCMLDE2017)
20-22 Nov 2017, Sydney, Australia
ISBN: 978-0-6480147-3-7

Time Series Anomaly Detection in Online Social Network: Challenges & Solutions

Md Rafiqul Islam[1], Rafiqul Islam[2] and Abu Raihan M Kamal[3]

[1, 3] Islamic University of Technology (IUT), Dhaka, Bangladesh
[2] Charles Sturt University, NSW, Australia
Corresponding author's E-mail: rafiqulislam.cse24@gmail.com

Abstract

In the field of data mining, the social network is one of the complex systems that poses significant challenges in this area. Time series anomaly detection is one of the critical applications. Recent developments in the quantitative analysis of social networks, based largely on graph theory, have been successfully used in various types of time series data. In this survey paper, we review the studies on graph theory to investigate and analyze time series social networks data including different efficient and scalable experimental modalities. We provide some applications, challenging issues and existing methods for time series anomaly detection.

Keywords: Social networks, Time Series Analysis, Anomaly Detection.

1. INTRODUCTION

Social networks have been an upcoming research field over the last few years, and they have gained an established position globally. Social networks provide online hang out space for everybody and using this technology anybody can communicate with their interested friend and share their information, photos, and videos. But this prominent technology also opens the door for unlawful activities. These illicit activities are also alluded to as anomalies. Anomalies emerge in online social networks as a result of specific people, or gatherings of people, rolling out sudden improvements in their examples of connection or communicating in a way that extraordinarily contrasts from their companions. The effects of this strange conduct can be seen in the subsequent network structure. In the field of mathematics, graph theory is a major area to model relations between objects and to represent a connected network structure. From the past decade, researchers are using graph theory to quantify aspects such as similarity, hierarchy and network efficiency of complex network structure in many other fields.

Anomalies are typically defined in terms of deviation from some expected behavior. In general, the definitions of anomalies as "patterns in data that do not conform to a well-defined notion of normal behavior". A record of phenomenon irregularly varying with time is referred to as time series. The analysis of time series is critical for many different scientific areas as often it contains some anomalies. However, it is necessary to carefully examine graphs of the data as a first step of the time series analysis. For example, to experiments the time series analysis of social networks, S. Asur and B. A. Huberman [1] Collected the quantity of tweets about a specific 'point from twitter. Also in their trials, each question is characterized by the Twitter handle of a prevalent artist recorded on the site reverbnation.com, and they recover all tweets that straightforwardly address this craftsman.

Over the past decade, many of the researchers are working to improve their understanding of how the social network work and how they can detect the anomalies. A large number of prominent techniques have been applied to detect time series anomalies in social network. Time series anomaly detection contains many challenging issues. Predetermined number of tasks has been done on applying time arrangement oddity recognition systems to online social networks because of the issues, for example, precision, computational multifaceted nature, and protection, absence of marked datasets and absence of adequate data. Different methods (distance-based, distribution-based, and clustering based.) have been applied to find anomalies in time series data. But these methods do not work well.

The major contributions of this survey paper are as follows:

We aim to familiarize time series anomaly detection in social network. We describe how the anomalies can be identified and how time series anomalies can be detected from social network.

We discuss many of the relevant works that have shown how various anomaly detection techniques of social network have already been used to detect anomalies.

Finally, with these ideas established we propose one technique that can help ameliorate the future directions of social network research.

The remainder of the paper is organized as follows: Section 2 presents the related work of time series anomaly detection in social network. While the analysis of time series anomaly detection in social networks in Section 3. The discussion is presented in Section 4. Finally, the conclusion is provided in Section 5.

2. RELATED WORK

A Social network is an online platform which provides an accessible space for everyone, especially young adults. They use this technology to socialize with absorbed accompany and acquaintances, and to allotment photos, information, and videos. This controlling phenomenon, which captures the structure and dynamics of person-to-technology and person-to-person interaction, is being used for numerous purposes such as education, business, medical, telemarketing, ball and adulterous activities. Social networks have been upcoming research areas over the last few years and they have increased a recognized position globally. In general, Social networks are self-organizing, emergent, and complex, such that a world-wide articular pattern appears from the local alternation of the elements that make up the system. These patterns become added credible as network size increases. Some of the most popular social network websites are Facebook, Google+, LinkedIn, Twitter, Messenger, Viber, WhatsApp, YouTube, WeChat etc. Sometimes this popular social network also opens the door for anomalies. Anomalies in online social networks can announce irregular and generally illegal behavior. Detection of such anomalies has been acclimated to analyze awful individuals, including spammers, animal predators, and online fraudsters [2, 3]. It is truly difficult to take care of an anomaly detection problem in a general frame. In this manner, an anomaly detection method should be developed and modified for a particular application by embracing thoughts from various teaches, for example, insights, machine learning, and data mining.

Enos, et al. [4] has proposed social network analysis to comprehend the interoperability related with the DoD arrangement of frameworks. It applied a few centrality measurements to a system of Major Defence Acquisition Programs (MDAPs) to evaluate the interoperability of individual frameworks inside the arrangement of frameworks. In particular, it inspects the contrasts between the degree, closeness, and an eigenvector centrality measurement to distinguish which metric best speaks to the interoperability of individual frameworks. Gardounis et al.[5] has purposed a SNA applied system to dissect venture forms and their changeability as an outcome of BIM implementation. The proposed point of view features applied endeavours for upgrading and connecting holes apparent in the current methodologies. Such holes prompt a failure to oblige the mind boggling frameworks surfacing among partaking groups and their refined needs, likewise influencing the incorporation of various mechanical interfaces and the coordination of the viable trade of huge scale data.

P. Kazienko and N. Chawla. [6] has presented the idea of the social network classifier that can autonomously order an on-screen character from the relationship among performers. They have tried the likelihood of utilizing the social network classifier as a supporting classifier. It can be utilized to enhance the certainty level of different classifiers paying little heed to their temperament. Himaja, N., and G. Murali [7] has suggested a general straggler-aware execution technique, SAE, to help the assessment transporter inside the cloud. It introduces a novel computational disintegration strategy that causes straggling capacity extraction strategies into more magnificent grained sub systems that are then distributed over groups of PCs for parallel execution. Test comes about demonstrate that SAE can pace up the assessment by method for as much as 1.77 occasions in correlation with best in class arrangements.

Awasthi, Abhishek [8] has designed a Genetic Clustering Algorithm which can be utilized for a) recognizing every one of the bunches in a chart and b) the group of only one given hub. The Genetic algorithm utilized for the grouping issue additionally consolidates the Island demonstrate with movement making the calculation much proficient than a solitary populace Genetic algorithm. The execution of the island show with movement can be upgraded substantially more by parallel programming. Doostmohammadian et al. [9] has built up the vital conditions for circulated discernibleness of social networks modelled as LSI frameworks. They have a portray essential perceptions, specialist order, and system network that empower every operator to deduce any social marvels advancing over a given social digraph. Specifically, they demonstrated that the circulated recognisability requires no a bigger number of perceptions than the brought together case; nonetheless, it requires certain characterization and availability necessities on the specialists watching those states. Cutillo et al. [10] has brook down the connection between the social network graph topology and the achievable security. They have watched three measurements, in particular degree dispersion, bunching coefficient and blending time, and demonstrate that they give key bits of knowledge on the protection level of the OSN.

Patil, Neha An. Et al [11] has actualized and enhanced run time post or message channel for a social network application. Because of enhanced manage set it is demonstrated that different sorts of assaults as for private information of client and malignant social exercises are avoided. They have made a constant application on to break down framework execution and it is discovered its execution is expanded essentially when contrasted with other online networking like Facebook, Twitter, LinkedIn and so forth.; Mathematicians Andrew Beveridge and Jie Shan published Network of Thrones in Math Horizon Magazine where they analyzed graph theory and then applied social network analysis algorithms to the network to find the most important characters in the network and a community detection algorithm to find clusters of characters [12].

K. Saha et al.,[13] has pointed of this investigation were to (1) build a record that deliberate the consciousness of various statistic bunches around schizophrenia-related data on Facebook; (2) consider how this list contrasted crosswise over statistic gatherings and how it corresponded with correlative Web-based (Google Trends) and non– Web-based factors about populace prosperity (emotional well-being markers and framework), and (3) look at the relationship of Facebook determined schizophrenia list with different sorts of online movement and in addition disconnected health and mental health and indicators. K. Garimella et al.[14] has exactly inspected the part of a recently presented Twitter highlight, 'cite retweets' (or 'quote RTs') in political talk, particularly whether it has prompted enhanced, common, and adjusted trade. Quote RTs enable clients to cite the tweet they retweet, while including a short remark. They investigation utilizing substance, system and group marked information demonstrates that the element has expanded political talk and its dispersion, contrasted with existing highlights. They talk about the ramifications of our discoveries in comprehension and lessening on the web polarization.

3. TIME SERIES ANOMALY DETECTION IN ONLINE SOCIAL NETWORK

Time-series data is a type of temporal data which is naturally high dimensional and large in data size. Time-series data are of interest due to their ubiquity in various areas ranging from science, engineering, business, finance, economics, healthcare, to government. While each time-series is consisting of a large number of data points it can also be seen as a single object. Genshiro Kitagawa, [15] described the definition of time series data in his book "Introduction to Time Series Modeling" as follows: A record of phenomenon irregularly varying with time is called time series. He also provided the examples of time series data. Typical time series examples are economic data like stock prices; meteorological data like temperature or rainfall and also medical data.

Anomaly detection is an important problem that has been studied within various research areas including online social networks, fraud, industrial damage, critical systems, image processing insurance, health care, military and bank. Anomaly detection, also named as outlier detection, refers to detecting patterns which do not accede with accepted behaviors. Besides especially anomaly detection is coming to play an increasing great role in social network. Detecting time series anomalies is an important issue in an area as it could be either a sign of a cogent botheration or carrying advantageous advice for the analyzer.

3.1. Application of time series anomaly detection

The analysis of time series anomaly detection on scientific applications can be examined of the diverse fields. Kitagawa [15] points out that it is also essential to carefully examine graphs of the data as a first step of the time series analysis. So it is easier to identify the next step of the analysis and find appropriate strategies for statistical modelling. Some of the important applications of time series anomaly detections are as follows.

	Applications	References
Time series Anomaly Detection	Social network	[16,17]
	Flight sequences data from aircrafts	[18,19]
	Shape of medical data	[20,21]
	Complex brain network	[22,23]
	Outlier light curves of periodic stars	[24,25]
	Eco-system	[26,27]
	Attack detection in recommender systems	28
	Heart beat pulses	29

3.2. Challenges of time series anomaly detection

There are some major challenges related with anomaly detection for time series are:

- There are many different means in which an anomaly occurring in a time series may be identified. An event within a time series may be anomalous; a subsequence within a time series may be anomalous; or an entire time series may be anomalous with respect to a set of normal time series.
- The training and test time series can be of different lengths.
- For detecting anomalous subsequence, the real length of the subsequence is generally unknown.
- Best similarity/distance measures which can be used for different types of time series is not easy to determine. Simple measures like Euclidean distance do not consistently perform well as they are awful acute to outliers and they also cannot be used if the time alternation are of different lengths.
- Analyzing performances of many anomaly detection algorithms are awful affected to noise in the time series data, as detect appropriate anomalies from noise is an arduous task.
- Time series in absolute applications are usually continued and as the length increases the computational complication also increases.

3.3. Solutions of time series anomaly detection

A large number of approaches have been developed to study time series anomaly detection of social networks and each of these technologies aims to aid in examine and assessing the extent of social network. Although these techniques are able to detect anomalies from social network, sometimes they are not able to provide the clear results. They have not the capacity to cover the entire social network rather they provide only information. Moreover, many of the anomaly detection algorithms apprehend multiple time series to be at a comparable scale in consequence while for a lot of the data it is not true.

Method	**Description**	**References**
Principal Component Analysis (PCA)	Used for identifing unusual time series in a large collections of time series.	[30]
RLPSVDD (Relaxed of linear programming Support Vector Data Description)	RLPSVDD solves a linear programming problem to provide a flexible data description for time series anomaly detection.	[31]
Decision tree classifier	Automatically extracting and summarizing reports	[32]
Latent Dirichlet Allocation (LDA)	Uses an abnormality estimation scheme based on probabilistic topic modeling and seasonal-trend decomposition to find and examine relevant message subsets.	[33]
Two-stage approach	Used for anomaly detection in large dynamic networks, in a context where in principle any type of anomaly should be detected.	[34]
EFS (Evolving fuzzy ststem)	Automatically analyze the user profiles in real times of a specific community of users. This analysis includes the detection of outliers, the clustering of profiles and their classification.	[35]
Statistical modeling	Used for the early and accurate anomaly detection from the time series of the negative tweets.	[36]
Regression Model	Describe the behavior	[37]
Multiple Linear Regression model	Describe the behavior of random variables	[37]
Autoregressive model	Ability to predict	[38]
Multilevel Regression Model	Examine independent and interactive effects of variables	[39-42]
Gaussian Mixed Model	Support categorical and continues values	[43]
ARIMA	Able to cluster seasonality patterns	[44]
Markov Chain	Multiple variable support	[45]
Hidden Markov Chain	Able to capture the dependencies between variables	[46]

4. DISCUSSION

Nowadays, Time series anomaly detection in the social network is a novel research field globally. Many researchers already engaged with this research field still there are some challenging issues need to be identified. In this paper we have surveyed a number of methods for detecting time series anomalies in social networks. We found that these different methods can be usefully categorized based on characterization of anomalies as being static or dynamic and labelled or unlabeled. Depending on this characterization, different features of the network may be examined, and for this we are proposing Model based clustering algorithm is one of the best technique. A future aspect could be to apply this unique algorithm or framework to further and may be larger datasets to get better insights. Based on these new datasets and algorithm it would also be a good idea to extend the social network properties. As a result of this it would also be possible to compare the determined results better and have more possibilities to find anomalies.

5. CONCLUSION

Time series anomaly detection in social networks have been played an important rule globally. In this survey paper we describe some review of studies and identify some challenging issues of time series analysis. Besides, detecting these challenging problems we propose one novel technique. We hope this technique will be played an increasingly important rule in the evolvement of social network in near future.

REFERENCES

1. S. Asur and B. A. Huberman, "Predicting the future with social media," in Web Intelligence and Intelligent
2. Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, 2010, pp. 492-499.
3. B. A. Huberman, D. M. Romero, and F. Wu, "Social networks that matter: Twitter under the microscope," 2008.
4. J. M. J. M. Gottman, Time-series analysisa comprehensive introduction for social scientists, 1981.
5. J. R. Enos, M. Mansouri, and R. Nilchiani, "Applying social network analysis to systems of systems: The case of the 2016 Puerto Rico power outage," in System of Systems Engineering Conference (SoSE), 2017 12th, 2017, pp. 1-6.
6. F. Gardounis, H.-Y. Chong, and X. Wang, "A conceptual framework for Social Network Analysis of Building Information Modelling in construction projects," in Research and Innovation in Information Systems (ICRIIS), 2017 International Conference on, 2017, pp. 1-6.
7. P. Kazienko and N. Chawla, Applications of Social Media and Social Network Analysis: Springer, 2015.
8. N. Himaja and G. Murali, "Social network analysis for efficient cloud data access," in Communication and Electronics Systems (ICCES), International Conference on, 2016, pp. 1-4.
9. A. Awasthi, "Clustering algorithms for anti-money laundering using graph theory and social network analysis," 2012.
10. M. Doostmohammadian and U. A. Khan, "Graph-theoretic distributed inference in social networks," IEEE Journal of Selected Topics in Signal Processing, vol. 8, pp. 613-623, 2014.
11. L. A. Cutillo, R. Molva, and M. Onen, "Analysis of privacy in online social networks from the graph theory perspective," in Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE, 2011, pp. 1-5.
12. N. A. Patil and A. S. Manekar, "A Novel Approach to Prevent Personal Data on a Social Network Using Graph Theory," in Computing Communication Control and Automation (ICCUBEA), 2015 International Conference on, 2015, pp. 186-189.
13. A. Beveridge and J. Shan, "Network of thrones," Math Horizons, vol. 23, pp. 18-22, 2016.

14. K. Saha, I. Weber, M. L. Birnbaum, and M. De Choudhury, "Characterizing Awareness of Schizophrenia Among Facebook Users by Leveraging Facebook Advertisement Estimates," Journal of medical Internet research, vol. 19, 2017.
15. K. Garimella, I. Weber, and M. De Choudhury, "Quote RTs on Twitter: usage of the new feature for political discourse," in Proceedings of the 8th ACM Conference on Web Science, 2016, pp. 200-204.
16. G. Kitagawa, "Introducing to Time Series Modeling, Chapman & Hall," ed: USA, CRC Press, 2010.
17. M. C. Chuah and F. Fu, "ECG anomaly detection via time series analysis," in International Symposium on Parallel and Distributed Processing and Applications, 2007, pp. 123-135.
18. S. Akhavan and G. Calva, "‘Automatic Anomaly Detection in ECG Signal by Fuzzy Decision Making," in Proceedings of 6th International Conference on Fuzzy Theory and Technology, 1998, pp. 96-98.
19. N. A. Heard, D. J. Weston, K. Platanioti, and D. J. Hand, "Bayesian anomaly detection methods for social networks," The Annals of Applied Statistics, vol. 4, pp. 645-662, 2010.
20. D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang, "Anomaly detection in online social networks," Social Networks, vol. 39, pp. 62-70, 2014.
21. V. Rajagopalan and A. Ray, "Symbolic time series analysis via wavelet-based partitioning," Signal Processing, vol. 86, pp. 3309-3320, 2006.
22. D. K. Tolani, M. Yasar, A. Ray, and V. Yang, "Anomaly Detection in Aircraft Gas Turbine Engines," JACIC, vol. 3, pp. 44-51, 2006.
23. M. Bicego and V. Murino, "Investigating hidden Markov models' capabilities in 2D shape classification," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, pp. 281-286, 2004.
24. Z. Liu, J. X. Yu, L. Chen, and D. Wu, "Detection of shape anomalies: A probabilistic approach using hidden markov models," in Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on, 2008, pp. 1325-1327.
25. L. Wei, E. Keogh, and X. Xi, "Saxually explicit images: Finding unusual shapes," in Data Mining, 2006. ICDM'06. Sixth International Conference on, 2006, pp. 711-720.
26. P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long short term memory networks for anomaly detection in time series," in Proceedings, 2015, p. 89.
27. E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," Nature reviews. Neuroscience, vol. 10, p. 186, 2009.
28. P. Protopapas, J. Giammarco, L. Faccioli, M. Struble, R. Dave, and C. Alcock, "Finding outlier light curves in catalogues of periodic variable stars," Monthly Notices of the Royal Astronomical Society, vol. 369, pp. 677-696, 2006.
29. U. Rebbapragada, P. Protopapas, C. E. Brodley, and C. Alcock, "Finding anomalous periodic time series," Machine learning, vol. 74, pp. 281-313, 2009.
30. H. Cheng, P.-N. Tan, C. Potter, and S. Klooster, "Detection and characterization of anomalies in multivariate time series," in Proceedings of the 2009 SIAM International Conference on Data Mining, 2009, pp. 413-424.
31. S. Zhang, A. Chakrabarti, J. Ford, and F. Makedon, "Attack detection in time series for recommender systems," in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, pp. 809-814.
32. R. J. Hyndman, E. Wang, and N. Laptev, "Large-scale unusual time series detection," in Data Mining Workshop (ICDMW), 2015 IEEE International Conference on, 2015, pp. 1616-1619.
33. C. Huang, G. Min, Y. Wu, Y. Ying, K. Pei, and Z. Xiang, "Time Series Anomaly Detection for Trustworthy Services in Cloud Computing Systems," IEEE Transactions on Big Data, 2017.
34. J. Krumm and E. Horvitz, "Eyewitness: Identifying local events via space-time signals in twitter feeds," in Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2015, p. 20.
35. J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, et al., "Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend

decomposition," in Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on, 2012, pp. 143-152.

36. J. A. Iglesias, A. García-Cuerva, A. Ledezma, and A. Sanchis, "Social network analysis: Evolving Twitter mining," in Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on, 2016, pp. 001809-001814.
37. K. Ikeda, G. Hattori, C. Ono, H. Asoh, and T. Higashino, "Early detection method of service quality reduction based on linguistic and time series analysis of twitter," in Advanced Information Networking and Applications Workshops (WAINA), 2013 27th International Conference on, 2013, pp. 825-830.
38. T. Nguyen, D. Phung, B. Dao, S. Venkatesh, and M. Berk, "Affective and content analysis of online depression communities," IEEE Transactions on Affective Computing, vol. 5, pp. 217-226, 2014.
39. M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in Proceedings of the 5th Annual ACM Web Science Conference, 2013, pp. 47-56.
40. O. L. Haimson, N. Andalibi, M. De Choudhury, and G. R. Hayes, "Relationship breakup disclosures and media ideologies on Facebook," New Media & Society, p. 1461444817711402, 2017.
41. K. Saha, I. Weber, M. L. Birnbaum, and M. De Choudhury, "Characterizing Awareness of Schizophrenia Among Facebook Users by Leveraging Facebook Advertisement Estimates," Journal of medical Internet research, vol. 19, 2017.
42. K. Garimella, I. Weber, and M. De Choudhury, "Quote RTs on Twitter: usage of the new feature for political discourse," in Proceedings of the 8th ACM Conference on Web Science, 2016, pp. 200-204.
43. C. Biernacki, G. Celeux, and G. Govaert, "Assessing a mixture model for clustering with the integrated completed likelihood," IEEE transactions on pattern analysis and machine intelligence, vol. 22, pp. 719-725, 2000.
44. M. Corduas and D. Piccolo, "Time series clustering and classification by the autoregressive metric," Computational Statistics & Data Analysis, vol. 52, pp. 1860-1872, 2008.
45. M. Ramoni, P. Sebastiani, and P. Cohen, "Multivariate clustering by dynamics," in AAAI/IAAI, 2000, pp. 633-638.
46. M. Bicego, V. Murino, and M. A. Figueiredo, "Similarity-based clustering of sequences using hidden Markov models," in International Workshop on Machine Learning and Data Mining in Pattern Recognition, 2003, pp. 86-95.