On the Trend and Problems of IoT Data Anomaly Detection

Shuai Li[1], Lejie Li[2(✉)], Kaining Xu[2], Jiafeng Yang[2], and Siying Qu[2]

[1] School of EE, University of Jinan, Jinan 250022, China
[2] Jinan Lingsheng Info Tech. Co. Ltd., Huaiyin District, Jinan 250000, China
lejie.li@dmml.stream

Abstract. With the rapid development of Internet technology, the Internet of Things is also constantly developing and progressing. More and more areas are starting to see connected devices, and more and more data is being generated by them. Effective data analysis and detection can prevent network intrusions and predict future trends. In recent years, with the breakthrough of computer technology, machine learning has shown good results in anomaly detection. Therefore, the research on anomaly detection of Internet of Things data has gradually increased and deepened. This work analyzes and summarizes the research trends in this field. First, we use keyword search to export articles in this field. Then we use the tool bibliometrix to generate statistical charts and trend charts for exported articles. At last, we analyze and summarize the generated two graphs. In the process of analysis, we have a detailed description of the phenomenon and a cause analysis. Finally, the future research direction in this field is derived.

Keywords: Internet of Things · Anomaly detection · Trend

1 Introduction

With the rapid development of modern society, Internet of Things technology has penetrated into everyone's daily life [1]. Network devices are constantly increasing, and numerous devices record large amounts of data. Because it is recorded in time, this data is called time series data. The analysis and detection of recorded data can play an important role in many fields such as transportation, aerospace, communications, and military.

In recent years, many effective detection algorithms have emerged in the field of artificial intelligence, and the precision and accuracy of detection are also constantly improving [2]. The accurate detection and analysis of data can provide good guidance for human production and life, which is of great significance.

This work is supported by National Key R&D Program of China 2018AAA0101703, Shandong Key Technology R&D Program 2019JZZY021005 and Natural Science Foundation of Shandong ZR2020MF067.

B. Wang et al. (Eds.): ICMTEL 2023, LNICST 535, pp. 358–368, 2024.
https://doi.org/10.1007/978-3-031-50580-5_31

2 Background

At the moment, data anomaly detection in the Internet of Things has been deeply studied and explored in many fields. In the field of transportation, Liu et al. [3] proposed a two-step anomaly data processing method for millimeter-wave radar in traffic flow detection applications, studied the reasonable range of distance, used an appropriate threshold to reduce the sample of single-parameter anomalous obviousness, and used the nearest neighbor method for analysis. Sun et al. [4] improved an algorithm based on Mahalanobis distance to estimate traffic fluctuations and detect accidents by processing differential data. In terms of network intrusion, Kim et al. [5] proposed a new detection method for intrusion detection based on RNN algorithm. Aleksieva et al. [6] used a genetic algorithm to analyze and process the packet lines received by the host, and protected the system by detecting whether there was a spoofing attack. In order to improve the detection ability of malicious traffic and malware, Sun et al. [7] considered examining the characteristics of the session dataset and fusing the appropriate one into a dataset. In the industrial field, Aygün et al. [8] designed an anomaly detection model combining autoencoder and de-autoencoder, and compared with other non-mixed models, the model performed well. Ferriyan et al. [9] made improvements in feature selection, using genetic algorithms to design systems. The parameters of the genetic algorithm are optimized. Shen Yan et al. [10] developed a hybrid robust convolutional autoencoder (HRCAE) for unsupervised anomaly detection of machine tools under noise.

Anomalous data from IoT is valuable [11]. The research and analysis in the field of Internet of Things anomaly detection has changed greatly in recent years, so this paper collects articles about this field, and uses the Bibliometrix tool to analyze and summarize the collected articles.

3 Related Papers and Sources

We selected papers related to anomaly detection of IoT data and their citations in the Web of Knowledge core collection database. A total of 334 files were exported based on keyword searches. Most of these articles come from magazines and journals, with a total of 110 sources. The articles were published between 2014 and 2022, with a total of 1289 authors. We use the Bibliometrix tool to analyze these articles, and we can obtain different kinds of trend charts and statistical charts, and then we analyze and summarize these graphs.

Figure 1 presents the overall trend of traffic time series anomaly detection research since 2014. As can be seen from Fig. 1, the overall research showed an upward trend, and it has gone through four stages, namely slow development period (2014–2016), stable development period (2016–2018, 2021–2022), and rapid development period (2020–2021). After 2016, the number of articles began to grow rapidly, reaching a maximum value in 2022, with 100 articles. Generally speaking, the past two years have been in a stage of prosperity and development. According to the data, from 2014 to 2022, the average annual growth rate of articles published in this field is 22.03%, showing an overall trend of continuous rapid development.

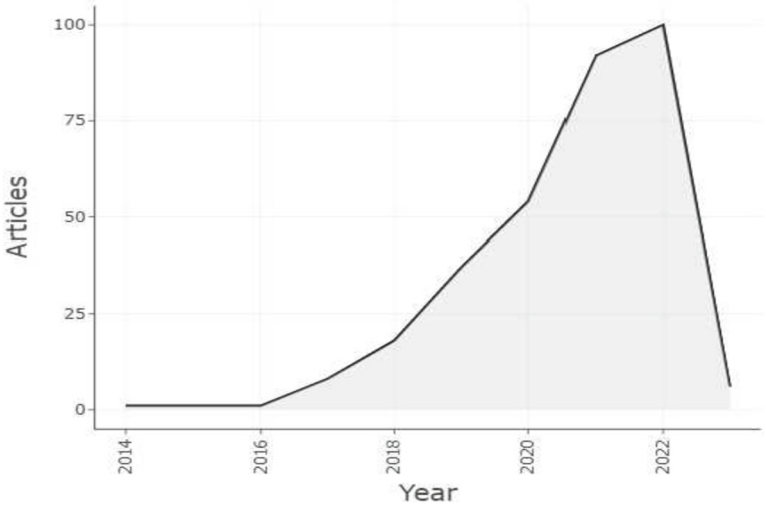

Fig. 1. Annual scientific production

Figure 2 shows the growth trend of the top 5 journals in our field for research and the amount of relevant literature produced by these journals. Among them, IEEE ACCESS has the fastest growth rate and the largest number of publications, bringing together 46 related articles in 2022. Figure 3 is based on Bradford's law, also known as the law of literature dispersion, so we can analyze the degree of dispersion of research in a certain field according to Fig. 3. The graph shows the 4 major journals published in our field and their distribution, and the rest of the journals are evenly distributed, thus indicating that research in our field is largely fragmented.

However, a high number of publications in a journal does not mean good quality of publications. As shown in Fig. 4, we used Total citation to rank and compare all sources. The most influential source is the IEEE INTERNET OF THINGS JOURNAL, which published a total of 76 papers in the collected database, including a large number of research based on machine learning algorithms for IoT security detection.

4 Citations and Reference

According to Fig. 5, among the exported literature, the most cited document in the world is "Data Mining for the Internet of Things: Literature Review and Challenges", which was cited 312 times in WOS, and the second place was cited 251 times. The third place was cited 244 times. In fourth place is a review on cybersecurity, which has been cited 238 times. It can be observed that the second, third and fourth digits are cited an equal number of times.

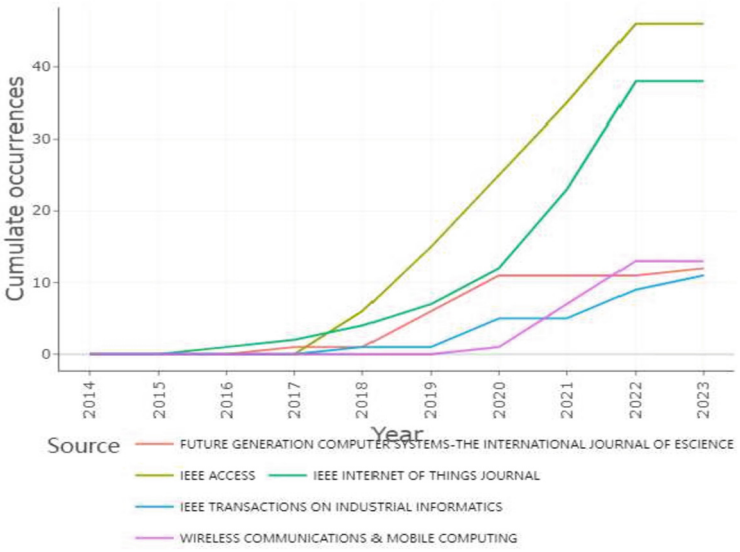

Fig. 2. Source dynamics

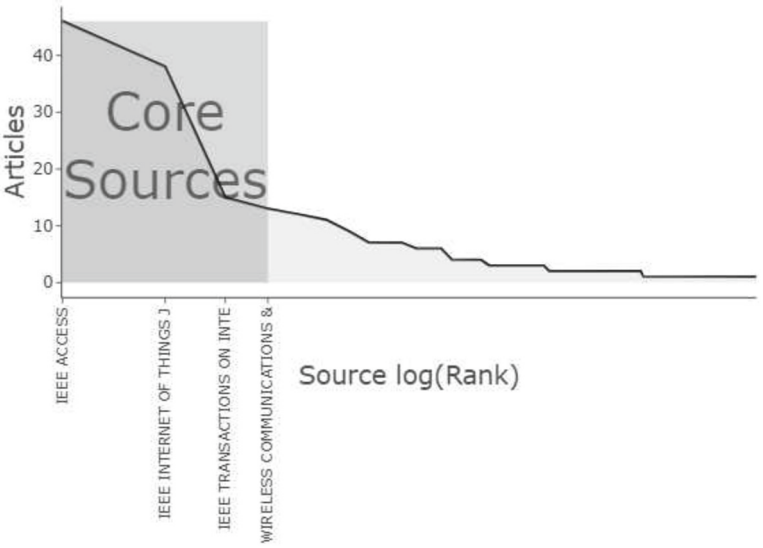

Fig. 3. Most published sources clustering with Bradford's law

We can also find that three of the top 15 papers were by Shafiq Muhammad and were cited 128, 99 and 77 times, respectively. Among them, the seventh paper introduces a new feature selection algorithm CorrAUC for network malicious traffic detection, the ninth paper proposes a method for selecting the corre-

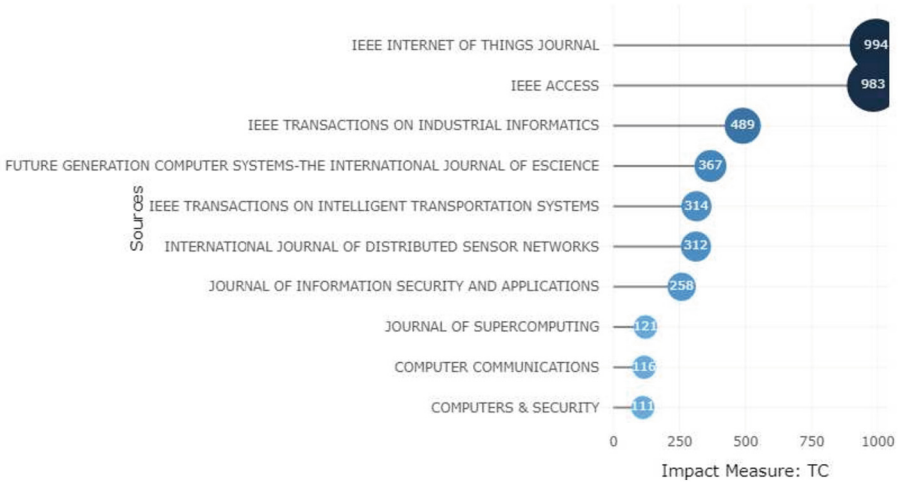

Fig. 4. Source impact (Total citation)

sponding algorithm according to the feature characteristics and then mixing it to achieve efficient anomaly detection of IoT data, and the thirteenth paper introduces a dual-target feature selection algorithm CorrACC for detecting network intrusion. It can be seen that author Shafiq Muhammad's main research on the Internet of Things network traffic intrusion detection problem is about feature extraction and feature selection algorithms. The author has made a considerable contribution in this field.

Figure 6 depicts the year distribution of the references of the papers we exported, while Fig. 7 shows the top ten most cited references of all references. According to Fig. 6 and Fig. 7, most of the references cited in the paper are distributed from 2017 to 2019, which may be due to the fact that most of the foundations relied on in this field were published in these two years. The broader period is from 2002 to 2022, during which the number of references experienced slowly increase, rapid increase, and rapid decline.

We use a three-paragraph diagram to show the relationship between the three aspects of the selected paper, as shown in Fig. 8. From the connection chart of countries, authors, and keywords, we can get the main publishing countries, main publishing authors, and the keywords mentioned in the collection. Key keywords are: Internet of Things, anomaly detection and deep learning. The lead authors are: Srivastava G., Moustafa N., Garg S., etc. Among them, China and Australia have contributed to various research directions in the field of Internet of Things data anomaly detection.

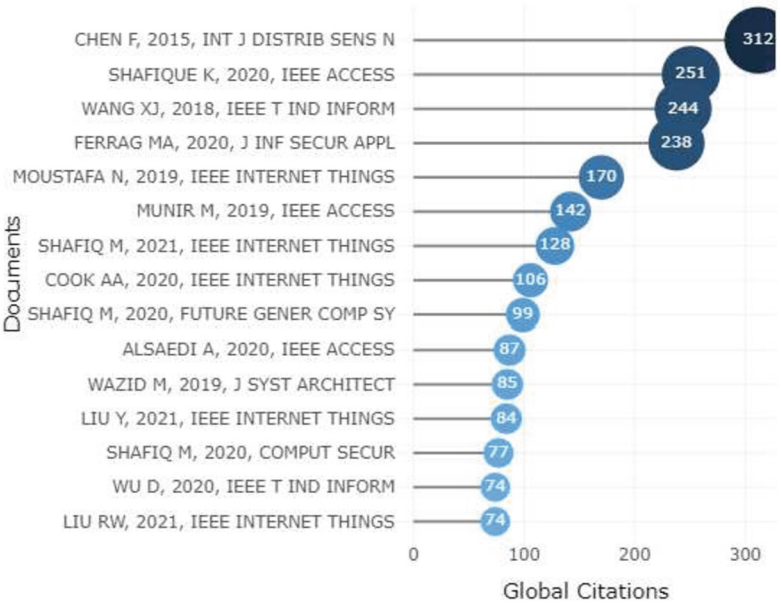

Fig. 5. Most global cited documents

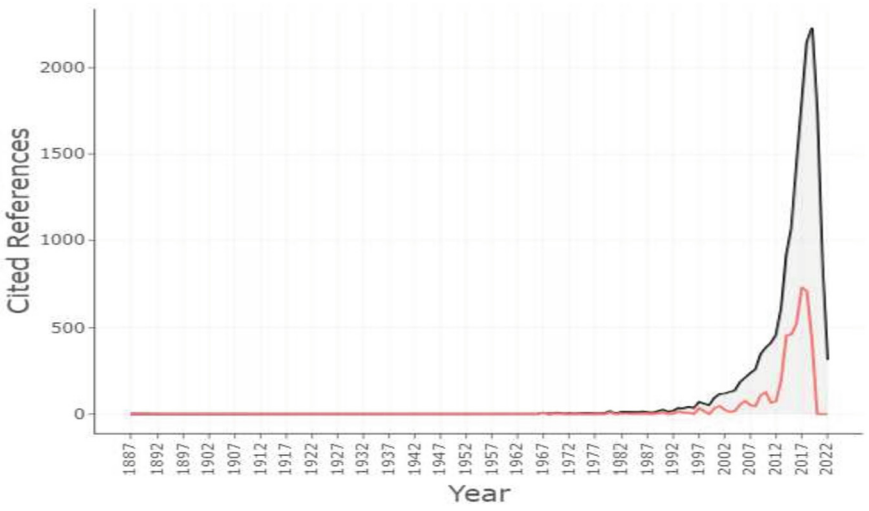

Fig. 6. Reference publication year spectroscopy (RPYS))

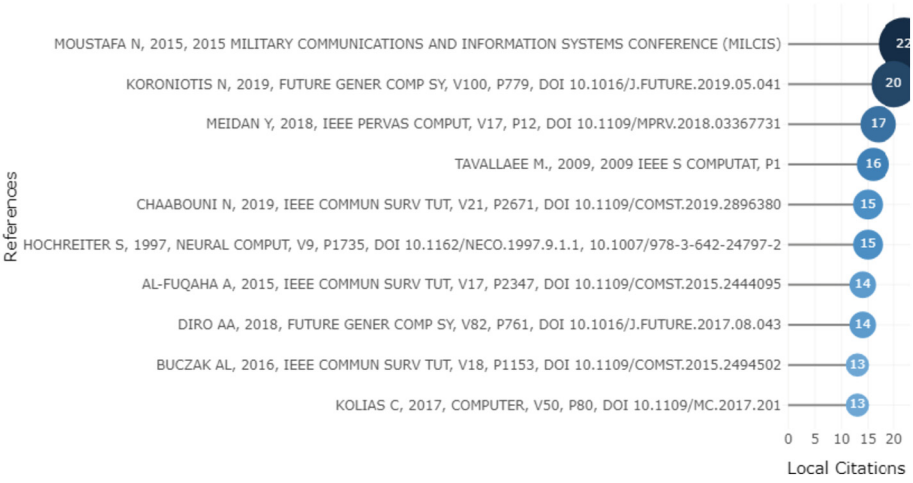

Fig. 7. Most Local Cited References

5 Trending Topics and Theme Evolution

Changing trends in topics provide us with more information. Figure 9 shows trends in key topics over time. Figure 10 depicts the evolution of the theme in two stages. After analysis, we can get that no topics have received sustained attention during the time period in the figure, but there are still more topics that continue to receive attention after 2020.

After 2020, the research topics in this field have undergone major changes. Among them, related problems such as traffic analysis have been solved to a high degree, and neural networks and attacks have received more and more attention. The frequency of use of the Internet topic peaked. The Internet and anomaly detection have become mainstream trends, evolving from a variety of themes, and the theme of big data as a whole has not changed much.

The theme evolution diagram uses another form, represented in the form of four quadrants of horizontal and vertical coordinates. As shown in Fig. 11 and Fig. 12, the abscissa represents the degree of relevance to the central topic, and the abscissa represents the intensity of development. The first quadrant has the highest density and central relevance, indicating that themes in this region are both central and highly developed, and are more important than the other quadrants. The themes of the second quadrant are niche themes, and although they are highly developed, they are off-center and therefore not very specialized. The themes of the third quadrant are marginal themes, themes that are in a state of rise or decline, usually peripheral or underdeveloped. The fourth quadrant is general topics, which are usually more basic, conceptual, and although more relevant to the center, they are not better researched and developed.

In the first thematic map, the distribution of themes up to 2020 in the four quadrants can be seen. First of all, the topics in the first quadrant are safety and

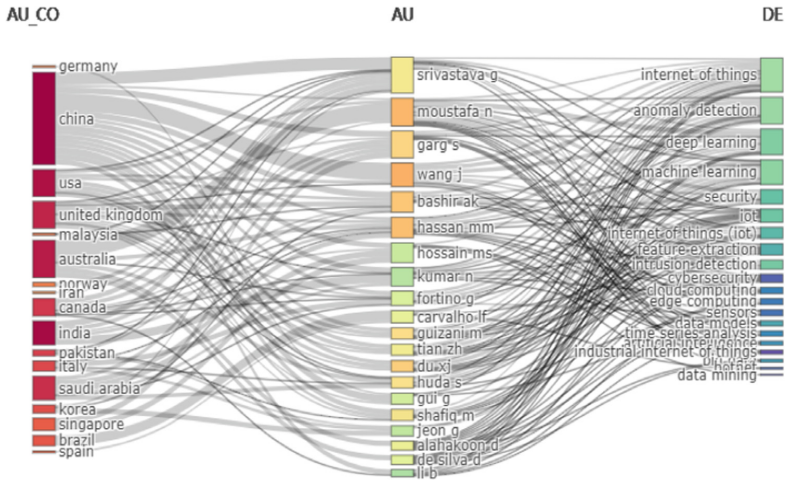

Fig. 8. Three field plot

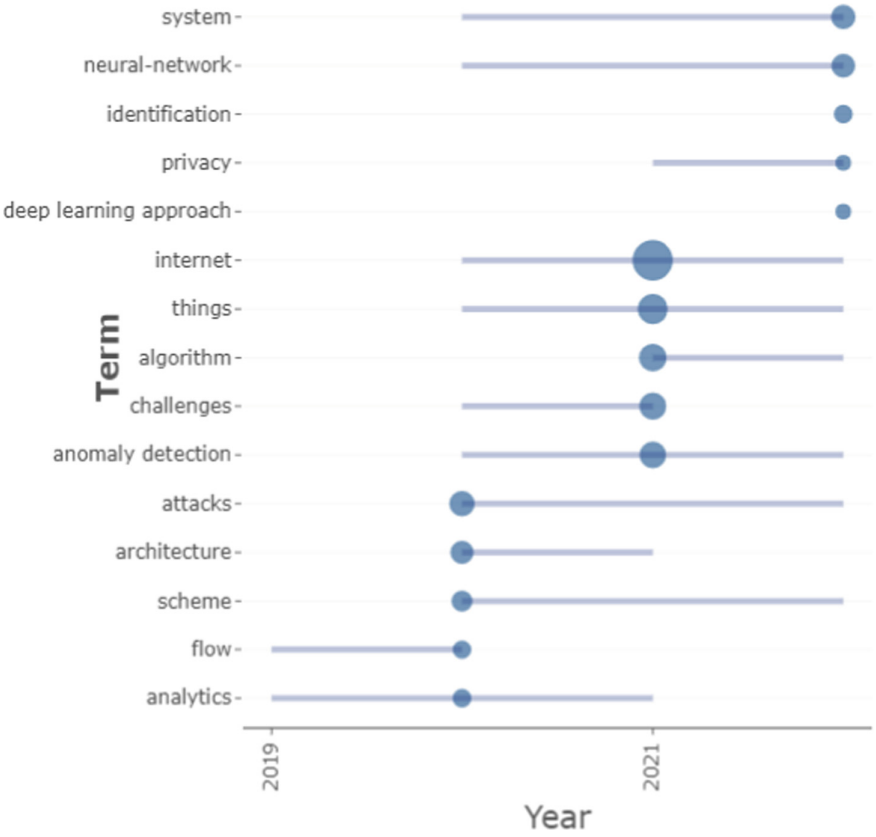

Fig. 9. Topic trend (Frequency)

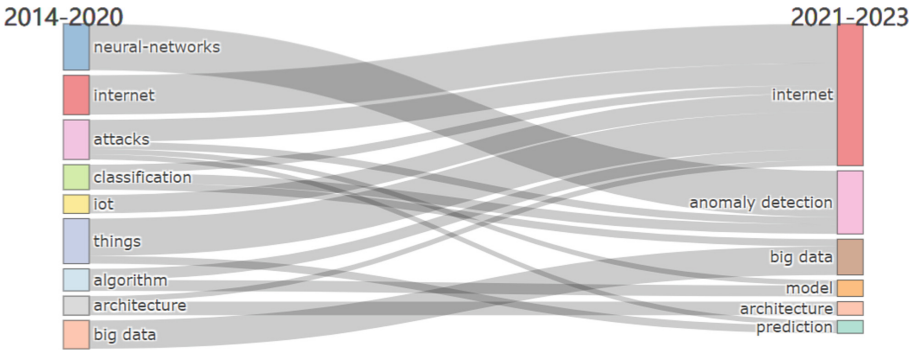

Fig. 10. Thematic evolution

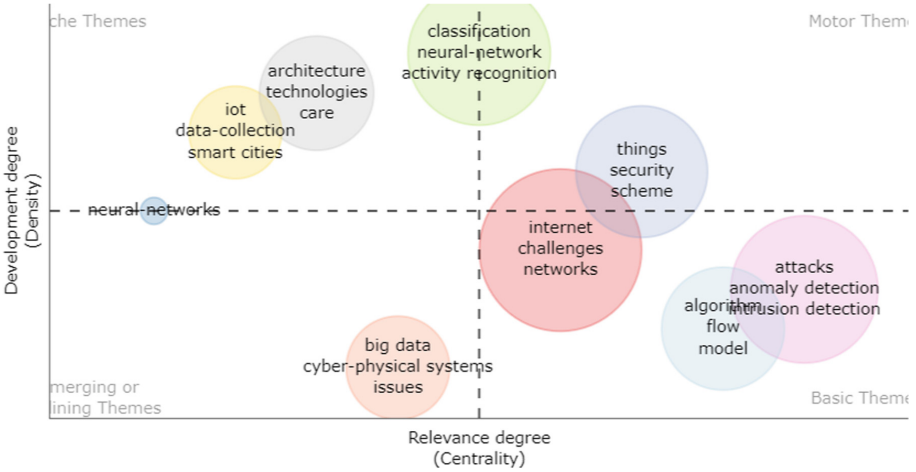

Fig. 11. Thematic map until 2020

planning. The topic of the third quadrant is big data, which can be analyzed as being in its infancy. The fourth quadrant is mainly the Internet of Things, anomaly detection, attacks, etc., which are more related to the central theme, but they were less studied at that time.

However, after 2020, the Thematic map underwent important changes. Thematic of anomaly detection has moved from the fourth quadrant to the first quadrant, proving that the research of anomaly detection has made great progress in recent years. Neural networks have slipped from the junction of the second and third quadrants to the third quadrant, which shows that the study of neural networks is becoming more and more important in our field.

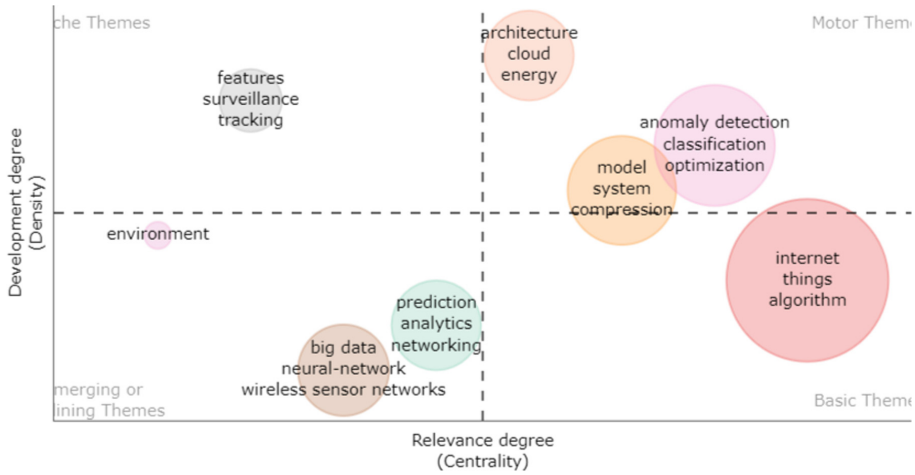

Fig. 12. Thematic map after 2020

6 Conclusion

According to the above analysis, we can see that the research on anomaly detection of Internet of Things data is obviously on the rise in recent years, and the research methods are constantly updated and optimized [12]. With the gradual rise of deep neural networks, anomaly detection based on neural networks has shown a great trend, which has also brought important opportunities and challenging to our research. Therefore, the main research direction in the next few years and even decades will be more and more inclined to deep learning.

References

1. Sun, B., Geng, R., Xu, Y., Shen, T.: Prediction of emergency mobility under diverse IoT availability. EAI Endorsed Trans. Pervasive Health Technol. **8**(4), e2 (2022)
2. Sun, B., Geng, R., Shen, T., Xu, Y., Bi, S.: Dynamic emergency transit forecasting with IoT sequential data. Mobile Networks Appl. 1–15 (2022)
3. Liu, H., Teng, K., Rai, L., Zhang, Y., Wang, S.: A two-step abnormal data analysis and processing method for millimetre-wave radar in traffic flow detection applications. IET Intel. Transport Syst. **15**(5), 671–682 (2021)
4. Sun, B., Cheng, W., Bai, G., Goswami, P.: Correcting and complementing freeway traffic accident data using Mahalanobis distance based outlier detection. Tech. Gazette **24**(5), 1597–1607 (2017)
5. Kim, J., Kim, J., Thu, H.L.T., Kim, H.: Long short term memory recurrent neural network classifier for intrusion detection. In: 2016 International Conference on Platform Technology and Service (PlatCon), pp. 1–5. IEEE (2016)
6. Aleksieva, Y., Valchanov, H., Aleksieva, V.: An approach for host based botnet detection system. In: 2019 16th Conference on Electrical Machines, Drives and Power Systems (ELMA), pp. 1–4. IEEE (2019)

7. Sun, B., Geng, R., Zhang, L., Li, S., Shen, T., Ma, L.: Securing 6G-enabled IoT/IoV networks by machine learning and data fusion. EURASIP J. Wirel. Commun. Netw. **2022**(1), 1–17 (2022)
8. Aygün, R.C., Yavuz, A.G.: A stochastic data discrimination based autoencoder approach for network anomaly detection. In: 2017 25th Signal Processing and Communications Applications Conference (SIU), pp. 1–4. IEEE (2017)
9. Ferriyan, A., Thamrin, A.H., Takeda, K., Murai, J.: Feature selection using genetic algorithm to improve classification in network intrusion detection system. In: 2017 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC), pp. 46–49. IEEE (2017)
10. Yan, S., Shao, H., Xiao, Y., Liu, B., Wan, J.: Hybrid robust convolutional autoencoder for unsupervised anomaly detection of machine tools under noises. Robot. Comput.-Integr. Manuf. **79**, 102441 (2023)
11. Sun, B., Ma, L., Shen, T., Geng, R., Zhou, Y., Tian, Y.: A robust data-driven method for multiseasonality and heteroscedasticity in time series preprocessing. Wirel. Commun. Mob. Comput. **2021**, 6692390:1–6692390:11 (2021)
12. Chen, M., Shao, H., Dou, H., Li, W., Liu, B.: Data augmentation and intelligent fault diagnosis of planetary gearbox using ILoFGAN under extremely limited samples. IEEE Trans. Reliab. **72**, 1029–1037 (2022)