Matrix Profile Evolution: An Initial Overview

Bin Sun, Liyao Ma(✉), Renkang Geng, and Yuan Xu

School of Electrical Engineering, University of Jinan, Jinan 250022, China
cse_maly@ujn.edu.cn

Abstract. Time series data have been investigated for decades in different domains. Recent fast development of wireless networks and cheaper price of small electronic monitoring devices, especially cheap IoT (internet of things) devices start to providing a lot time series data. However, those time series data are mixed with different patterns across lifetime. The patterns should be distinguished so the data can be separated and sent to corresponding process. There are different ways to tackle this challenge, for example by traditional pattern discovery or classification/clustering machine learning algorithms. The matrix profile (MP) method provides a way to handle this problem which can be used individually or together with other methods as an indicator variable or feature. This work aims to take an initial overview of MP method history and evolution from bibliometric aspect.

Keywords: Pattern recognition · Matrix profile · Machine learning · Time series

1 Introduction

Nowadays, with the fast application and deployment of 5G telecommunication technology, internet of things (IoT) devices start to cover bigger daily life domains as well as their enormous volume of data. The data from those devices come with timestamps, thus time series data. IoT data are now much more useful for data analysis in many domains such as positioning [14] and e-health [13]. Equipment and even humans are now monitored by IoT devices and the data are used for different purposes such as incident detection and requirement forecasting.

At the same time, the need of mining those time series to find useful information such as health monitoring and alerting, short-term weather and traffic prediction and so on. However, the time series data from IoT devices are mixed with different patterns across lifetime.

W. Fu et al. (Eds.): ICMTEL 2021, LNICST 387, pp. 492–501, 2021.
https://doi.org/10.1007/978-3-030-82562-1_48

2 Related Papers and Sources

From the work of Keogh E., we select the time-series related ones and their citations in the Web of Knowledge core collection databases. This leads to 4195 documents and more than half of them are journal articles. They are published from 1997 until 2021 from 2065 sources (including journals, books, proceedings etc.) and contributed by 9597 authors. The papers' information is imported into an R library bibliometrix [1] to generate and analyze bibliometrics.

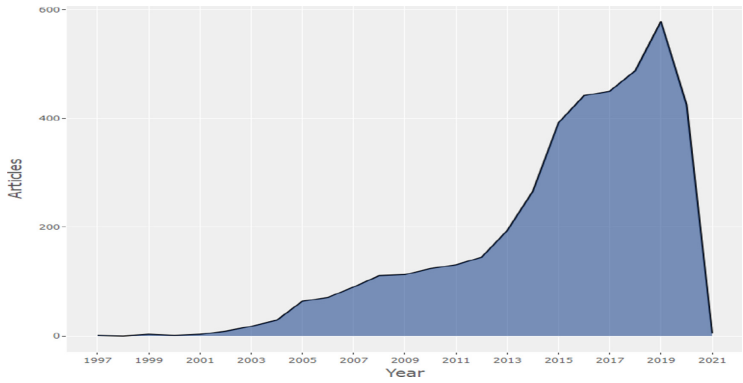

Fig. 1. Annual scientific production

Figure 1 shows that the time-series related research is still climbing a steep hogback even after decades of continuous research. It reflects the fact that the modern life brings more time series which are in more urgent need of fast and effective processing methods.

Figure 2 shows the sources which produce most in our database according to Bradford's Law [10]. Bradford's Law orders sources by publication numbers and then divides and groups sources from top to down and each group has the same amount of publications. In most cases, Bradford's Law follows Pareto distribution.

However, the higher amount does not mean better quality. If we compare with the list of sources ordered by impact (H-index) as shown in Fig. 3, we can see that producing more papers does not mean the papers are in good quality.

The most influencing source is the "Data Mining and Knowledge Discovery" journal which publishes some key papers from Keogh's team. In total, it publishes 76 papers in the collected database including the survey and empirical demonstration regarding time series data mining benchmarks with 391 Web Of Science (WOS) citations (1511 in Google scholar), the SAX algorithm paper "Experiencing SAX: a novel symbolic representation of time series" with 605 WOS citations (1383 in Google scholar), and "Experimental comparison of representation methods and distance measures for time series data" with 376 WOS

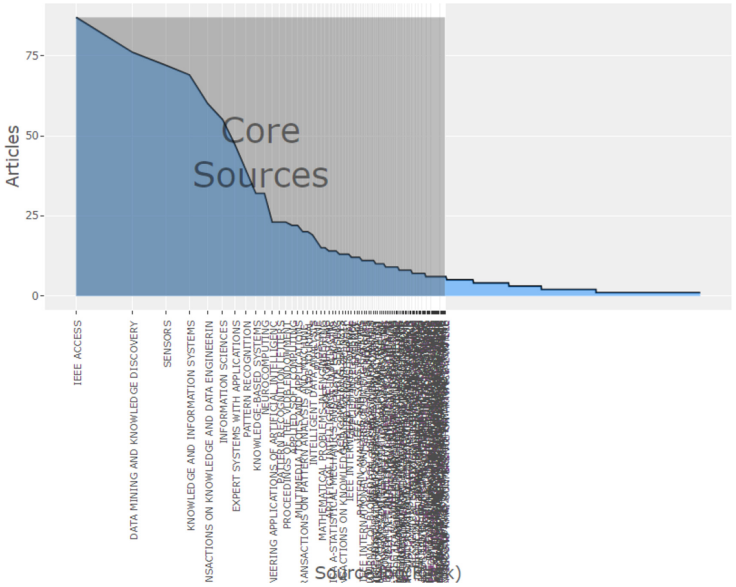

Fig. 2. Most publishing sources clustering with Bradford's law

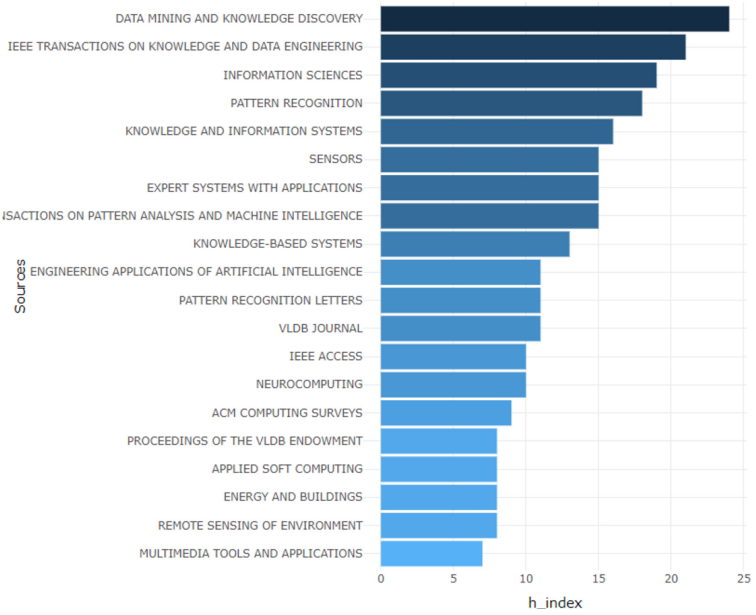

Fig. 3. Source impact (Source H-index)

citations (748 in Google scholar) among many other with more than three-digit citations. The second influencing source is the "IEEE Transactions on Knowledge and Data Engineering" journal. Within our collected database, it publishes four papers with about two hundred WOS citations.

3 Citations and References

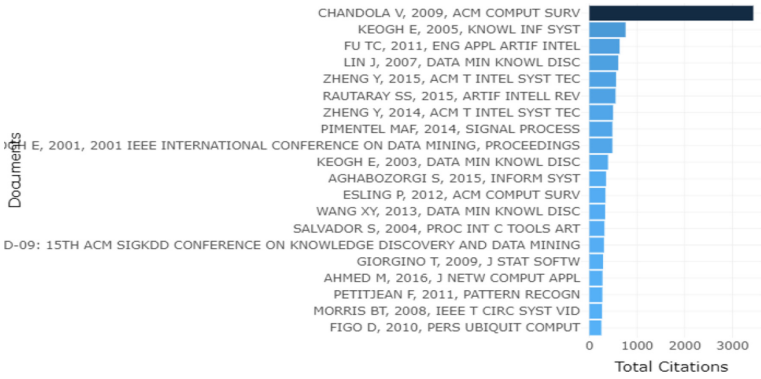

Fig. 4. Most cited documents

According to most cited document list (Fig. 4), 4 out of top 10 papers are from Keogh team as follows. The 2nd position is "Clustering of time-series subsequences is meaningless: implications for previous and future research" [6] with 759 citations in WOS. The 4th paper "Experiencing SAX: a novel symbolic representation of time series" [9] has 605 citations. The 9th paper introducs an online algorithm for segmenting time series [5] with 483 citations. The 10th paper is a survey and empirical demonstration which focuses on the need for time series data mining benchmarks [8] with 391 citations.

Other papers come from different affiliations. The top paper is a survey about anomaly detection [2] with more than three thousand citations. The third is a review on time series data mining [3]. The fifth and seventh are overviews about trajectory data mining [15] and urban computing [16] from Microsoft. The sixth is a survey regards vision based gesture recognition [12]. The eighth is a review of novelty detection [11].

From the statistics of references showing in Fig. 5 and Fig. 6, we can see that the top three papers from Keogh team [6,7,9] regarding similarity measurement and fast search are influencing the latter papers most. The reason could be the fact that similarity measurement for time series data is always a hard question and suffers from many difficulties especially the well-known curse of dimensionality.

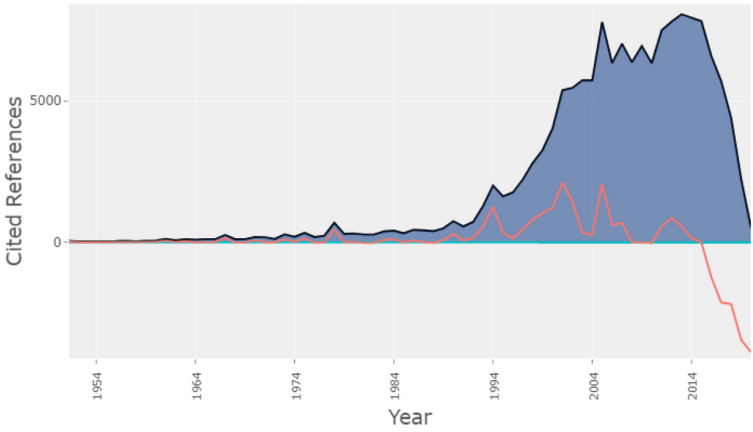

Fig. 5. Reference publication year spectroscopy (RPYS)

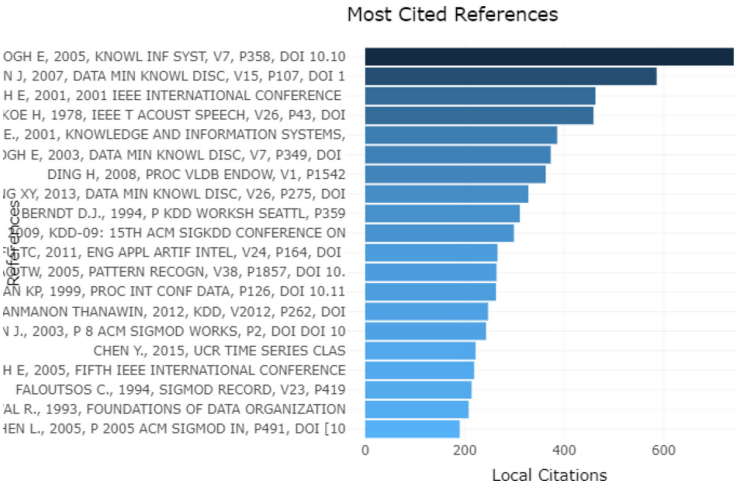

Fig. 6. Most cited references

The three fields plot is used to show relations among three aspects of collected publications. From the three fields plot by papers, authors and keywords as shown in Fig. 7, we can see the structure of the most important papers, authors and keywords which forms the main research trends and topics. The impact or significance of authors is represented by the amount of relations. We can see that the key authors are Keogh E., Palpanas T., Li H., Anh D. among others.

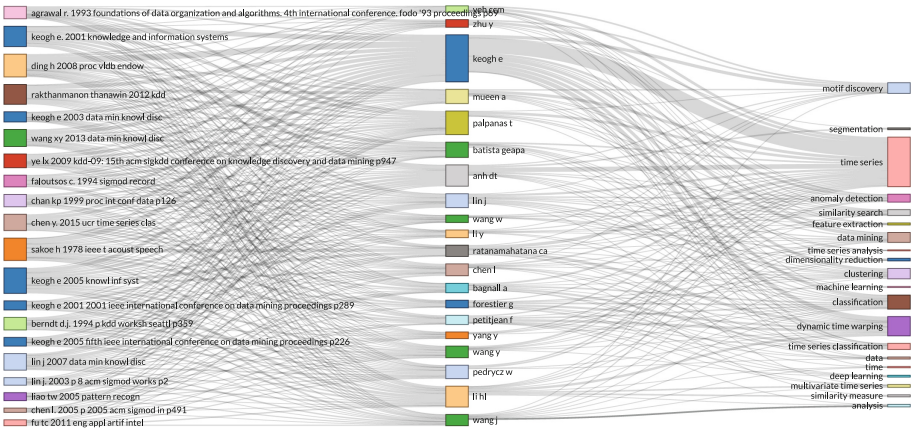

Fig. 7. Three fields plot by papers, authors and keywords

4 Thematic Evolution and Trending Topics

Some more trend details can be seen from the yearly topic frequencies as shown in Fig. 8. It reveals topic trends from WOS keyword-plus. Keyword-plus field from WOS is digested from corresponding paper's title and content. It is a semi-automated and normalized field and it is able to show deeper content and wider variety. In this analysis, we can see that visualization gets continuous attention and is on a peak. Recognition of different types of patterns provides the highest frequency. High frequently mentioned topics.

The thematic evolution of WOS keyword-plus with one cut point after 2016 shows the change of research focus. Classification related problems have been partly solved while models and patterns are now getting more attention and replaces "search". The percentage of focus on time-series has been doubled. The "dynamics" topic is replaced by "variability". The identification (of motif, outliers etc.) gets a bit higher sharing.

The thematic map reveals also the keyword plus field but from another aspect which separate and visualizes trending topics into four different types as shown in Fig. 10 and 11 in four quadrants [4]. The first quadrant contains motor themes with high centrality and high density which constructs the main structure and drives the work of this MP research field. The second quadrant contains niche themes (low centrality, high density) that are well developed but very specialised. The third quadrant contains peripheral themes (low centrality, low density) that are either emerging or declining and are often under developed and marginal. The fourth quadrant contains basic themes (high centrality, low density) which are transversal and general.

From the first thematic map representing keywords until 2016 as shown in Fig. 9, it can be seen that the upper right (first) quadrant with motor themes (high density, high centrality) that presents topics made up of time series, algorithms, and systems. The topics within the lower right (fourth) quadrant with

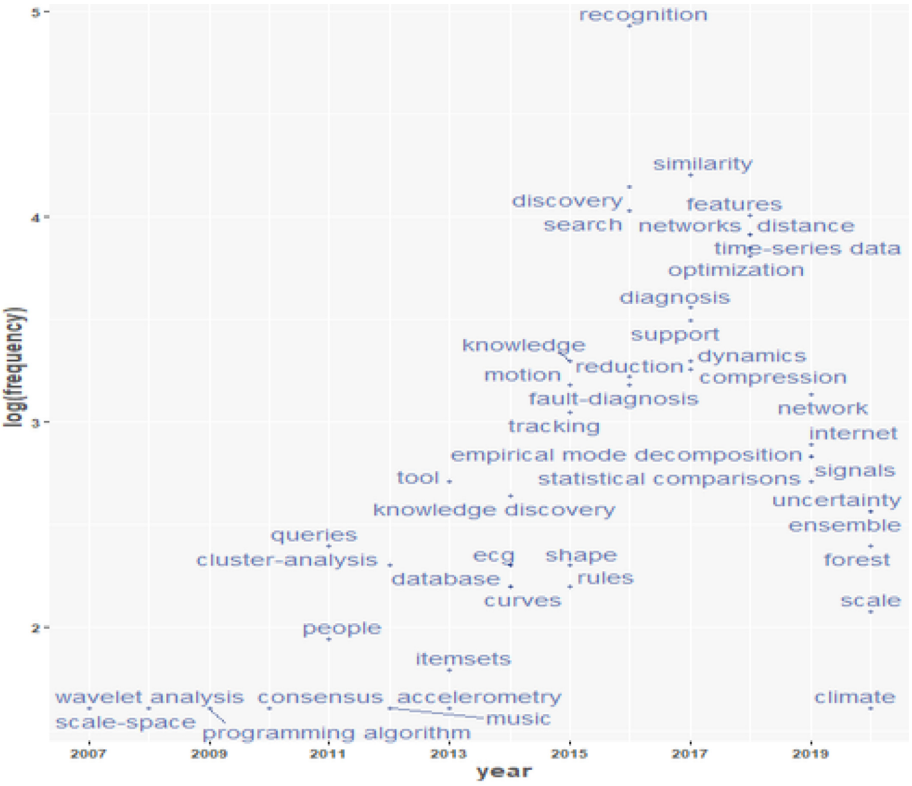

Fig. 8. Topic trend (Frequency)

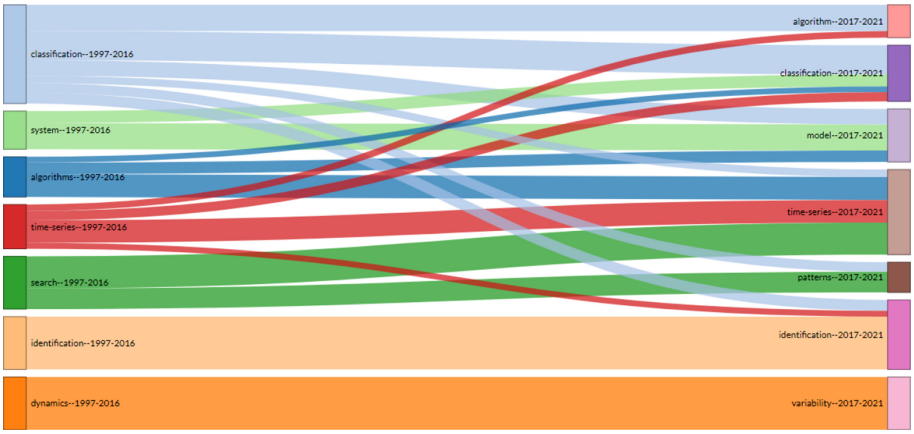

Fig. 9. Thematic evolution

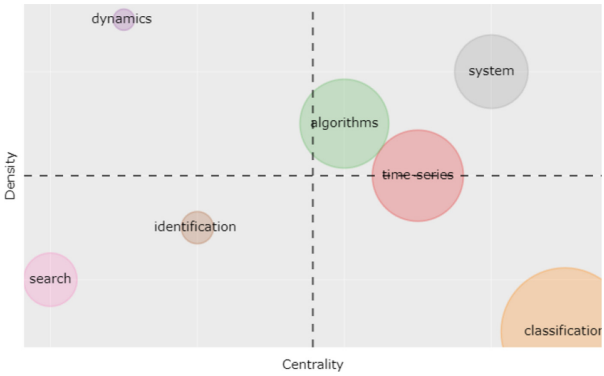

Fig. 10. Thematic map until 2016

basic themes is the classification of time series. Before 2016, dynamics was the one specialized research direction of the upper left quadrant.

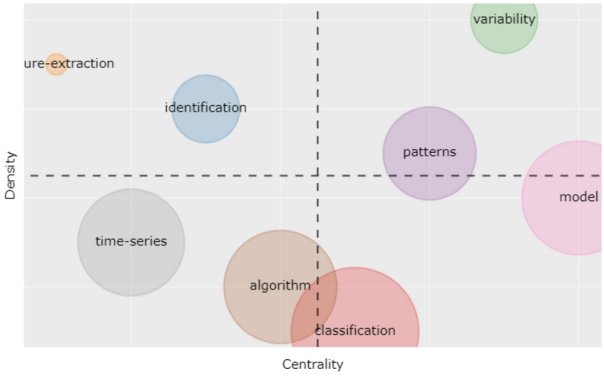

Fig. 11. Thematic map after 2016

However, after 2016, great changes have taken place in the thematic map. Keywords and distribution have produced great changes, the upper right corner appeared variability and patterns. The upper left corner is composed of identification and pure-extraction. The lower left corner contains algorithms of time series. The lower right area mainly consists of classification models, indicating the development of different models for classification is still an important task though it is sliding to the third quadrant.

5 Conclusion and Future Work

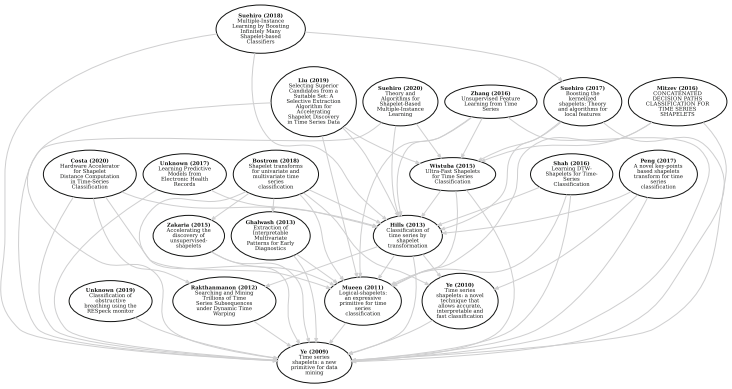

Fig. 12. Citation graph

From the influential citations as shown in Fig. 12 we can see that the matrix profile has showing some promising results in many domains such as computer science and health care. However, few transportation work has been found for urban transportation IoT/AIoT application. Thus, we tend to investigate the related area in the near future.

References

1. Aria, M., Cuccurullo, C.: Bibliometrix: an r-tool for comprehensive science mapping analysis. J. Inf. **11**(4), 959–975 (2017)
2. Chandola, V.: Anomaly Detection for Symbolic Sequences and Time Series Data. Ph.D. thesis, University of Minnesota (2009)
3. Fu, T.C.: A review on time series data mining. Eng. Appl. Artif. Intell. **24**(1), 164–181 (2011)
4. Fusco, F., Marsilio, M., Guglielmetti, C.: Co-production in health policy and management: A comprehensive bibliometric review. BMC Health Serv. Res. **20**(1), 504 (2020)
5. Keogh, E., Chu, S., Hart, D., Pazzani, M.: An online algorithm for segmenting time series. In: Proceedings 2001 IEEE International Conference on Data Mining, pp. 289–296 (November 2001)
6. Keogh, E., Lin, J.: Clustering of time-series subsequences is meaningless: implications for previous and future research. Knowl. Inf. Syst. **8**(2), 154–177 (2005)
7. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. Knowl. Inf. Syst. **3**(3), 263–286 (2001)
8. Keogh, E., Kasetty, S.: On the need for time series data mining benchmarks: a survey and empirical demonstration. Data Min. Knowl. Disc. **7**(4), 349–371 (2003)

9. Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing SAX: a novel symbolic representation of time series. Data Min. Knowl. Disc. **15**(2), 107–144 (2007)
10. Nisonger, T.E.: The 80/20 rule and core journals. Serials Librarian **55**(1–2), 62–84 (2008)
11. Pimentel, M.A.F., Clifton, D.A., Clifton, L., Tarassenko, L.: A review of novelty detection. Sign. Process. **99**, 215–249 (2014)
12. Rautaray, S.S., Agrawal, A.: Vision based hand gesture recognition for human computer interaction: a survey. Artif. Intell. Rev. **43**(1), 1–54 (2015)
13. Sun, M., et al.: Methods to characterize the real-world use of rollators using inertial sensors–a feasibility study. IEEE Access **7**, 71387–71397 (2019)
14. Xu, Y., Shmaliy, Y.S., Ahn, C.K., Shen, T., Zhuang, Y.: Tightly-coupled integration of INS and UWB using fixed-lag extended UFIR smoothing for quadrotor localization. IEEE Internet Things J. **8**(3), 1716-1727 (2020)
15. Zheng, Y.: Trajectory data mining. ACM Trans. Intell. Syst. Technol. (TIST) **6**(3), 1-41 (2015)
16. Zheng, Y., Capra, L., Wolfson, O., Yang, H.: Urban Computing. ACM Trans. Intell. Syst. Technol. (TIST) **5**(3), 1–55 (2014)