Under One Sky: The IAU Centenary Symposium
Proceedings IAU Symposium No. ,
C. Sterken, J. Hearnshaw & D. Valls-Gabaud, eds.
doi:10.1017/xxxxx

Easy, Accurate, and Fast Machine Learning on Very Large Data Series Collections: Similarity Search and Subsequence Anomaly Detection

Themis Palpanas

LIPADE, Université Paris Cité

Keywords. data series, time series, similarity search, template matching, anomaly detection

1. Extended Abstract

There is an increasingly pressing need, by several applications in diverse domains, for developing techniques able to manage and analyze very large collections of sequences, or data series. Examples of such applications come from various monitoring applications, including in power utility companies, where we need to apply machine learning techniques for knowledge extraction. It is not unusual for these applications to involve numbers of data series in the order of hundreds of millions to billions, which are often times not analyzed in their full detail due to their sheer size. However, no existing data management solution can offer native support for sequences and the corresponding operators necessary for complex analytics.

In this talk, we describe our efforts in designing techniques for indexing and analyzing truly massive collections of data series that enable scientists to run complex analytics on their data. These techniques are orders of magnitude faster than the state of the art. We also present our recent work on (essentially, parameter-free) subsequence anomaly detection and explanation, which is both more accurate and faster than competing approaches.

In the following, we summarize the main points along these two areas of research.

[Similarity Search] Similarity search in high-dimensional data spaces was a relevant and challenging data management problem in the early 1970s Bentley (1975), when the first solutions to this problem were proposed. Today, fifty years later, we can safely say that the exact same problem is more relevant (from Time Series Management Systems to Vector Databases) and challenging than ever. This is true, not because the research community has been idle; on the contrary, the literature on this topic is very large and diverse Bentley (1975); Guttman (1984); Lin et al. (1994); Berchtold et al. (1996); Ciaccia et al. (1997); Berchtold et al. (1998); Palpanas (2020); Wang et al. (2023), demonstrating both the interest in this problem, as well as the wide range of ideas that have been applied to it and led to impressive advances. This is true, rather because very large amounts of high-dimensional data are now omnipresent (ranging from traditional multidimensional data to time series and deep embeddings) Palpanas and Beckmann (2019); Echihabi et al. (2020), and the performance requirements (i.e., response-time and accuracy) of a variety of applications that need to process and analyze these data have become very stringent and demanding. In these past fifty years, high-dimensional similarity search has been studied in its many flavors. Similarity search algorithms for exact and approximate, one-off and progressive query answering Echihabi et al. (2018, 2019, 2023). In this talk, we review the state of the art solutions for data series similarity search Peng et al. (2021b); Echihabi et al. (2022); Peng et al. (2020, 2021a); Chatzakis et al. (2023); Fatourou et al. (2023); Schäfer et al.

(2025); Azizi et al. (2023), we observe that the data series solutions are the methods of choice for several of the similarity search problem flavors even for high-dimensional vector search, and we point to interesting open research problems, including the role of machine learning in this space Wang et al. (2025); Chatzakis et al. (2026).

[Anomaly Detection] Subsequence anomaly detection in long data series is an important problem with applications in a wide range of domains Boniol et al. (2024); Darban et al. (2025). However, several of the approaches that have been proposed in the literature have limitations: they require prior domain knowledge that is used to design the anomaly discovery algorithms, they need training with labeled examples that are hard/expensive to produce, or have scalability problems. In this talk, we present recent unsupervised methods suitable for domain agnostic subsequence anomaly detection. We discuss two possible way to represent the normal behavior of a long data series that lead to fast and accurate identification of abnormal subsequences. These normal representations are either based on subsequences (using a data structure called the normal model) Boniol et al. (2021a), or on graphs, by taking advantage of graph properties to encode the normal transitions between neighboring subsequences of a long series Boniol and Palpanas (2020). We describe their properties, and explain how they can be extended to handle streaming time series Boniol et al. (2021b), or adapted to address other problems, such as interpretable time series clustering Boniol et al. (2025).

References

Ilias Azizi, Karima Echihabi, and Themis Palpanas. Elpis: Graph-based similarity search for scalable data science. *Proc. VLDB Endow.*, 16(6):1548–1559, 2023. doi: 10.14778/3583140.3583166. URL https://www.vldb.org/pvldb/vol16/p1548-azizi.pdf.

Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, 1975. doi: 10.1145/361002.361007. URL https://doi.org/10.1145/361002.361007.

Stefan Berchtold, Daniel A. Keim, and Hans-Peter Kriegel. The x-tree : An index structure for high-dimensional data. In *VLDB'96, Proceedings of 22th International Conference on Very Large Data Bases, September 3-6, 1996, Mumbai (Bombay), India*, pages 28–39. Morgan Kaufmann, 1996. URL http://www.vldb.org/conf/1996/P028.PDF.

Stefan Berchtold, Christian Böhm, and Hans-Peter Kriegel. The pyramid-technique: Towards breaking the curse of dimensionality. In *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA*, pages 142–153. ACM Press, 1998. doi: 10.1145/276304.276318. URL https://doi.org/10.1145/276304.276318.

Paul Boniol and Themis Palpanas. Series2graph: Graph-based subsequence anomaly detection for time series. *Proc. VLDB Endow.*, 13(11):1821–1834, 2020. URL http://www.vldb.org/pvldb/vol13/p1821-boniol.pdf.

Paul Boniol, Michele Linardi, Federico Roncallo, Themis Palpanas, Mohammed Meftah, and Emmanuel Remy. Unsupervised and scalable subsequence anomaly detection in large data series. *VLDB J.*, 30(6):909–931, 2021a. doi: 10.1007/S00778-021-00655-8. URL https://doi.org/10.1007/s00778-021-00655-8.

Paul Boniol, John Paparrizos, Themis Palpanas, and Michael J. Franklin. SAND: streaming subsequence anomaly detection. *Proc. VLDB Endow.*, 14(10):1717–1729, 2021b. doi: 10.14778/3467861.3467863. URL http://www.vldb.org/pvldb/vol14/p1717-boniol.pdf.

Paul Boniol, Qinghua Liu, Mingyi Huang, Themis Palpanas, and John Paparrizos. Dive into time-series anomaly detection: A decade review, 2024. URL https://arxiv.org/abs/2412.20512.

Paul Boniol, Donato Tiano, Angela Bonifati, and Themis Palpanas. -graph: A graph embedding for interpretable time series clustering. *IEEE Trans. Knowl. Data Eng.*, 37(5):2680–2694, 2025. doi: 10.1109/TKDE.2025.3543946. URL https://doi.org/10.1109/TKDE.2025.3543946.

Manos Chatzakis, Panagiota Fatourou, Eleftherios Kosmas, Themis Palpanas, and Botao Peng. Odyssey: A journey in the land of distributed data series similarity search. *Proc. VLDB Endow.*, 16(5):1140–1153, 2023. doi: 10.14778/3579075.3579087. URL https://www.vldb.org/pvldb/vol16/p1140-chatzakis.pdf.

Manos Chatzakis, Yannis Papakonstantinou, and Themis Palpanas. Darth: Declarative recall through early termination for approximate nearest neighbor search. *Proceedings of the ACM on Management of Data*, 2026.

Paolo Ciaccia, Marco Patella, and Pavel Zezula. M-tree: An efficient access method for similarity search in metric spaces. In *VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases, August 25-29, 1997, Athens, Greece*, pages 426–435. Morgan Kaufmann, 1997. URL http://www.vldb.org/conf/1997/P426.PDF.

Zahra Zamanzadeh Darban, Geoffrey I. Webb, Shirui Pan, Charu Aggarwal, and Mahsa Salehi. Deep learning for time series anomaly detection: A survey. *ACM Comput. Surv.*, 57(1):15:1–15:42, 2025. doi: 10.1145/3691338. URL https://doi.org/10.1145/3691338.

Karima Echihabi, Kostas Zoumpatianos, Themis Palpanas, and Houda Benbrahim. The lernaean hydra of data series similarity search: An experimental evaluation of the state of the art. *Proc. VLDB Endow.*, 12 (2):112–127, 2018. doi: 10.14778/3282495.3282498. URL http://www.vldb.org/pvldb/vol12/p112-echihabi.pdf.

Karima Echihabi, Kostas Zoumpatianos, Themis Palpanas, and Houda Benbrahim. Return of the lernaean hydra: Experimental evaluation of data series approximate similarity search. *Proc. VLDB Endow.*, 13 (3):403–420, 2019. doi: 10.14778/3368289.3368303. URL http://www.vldb.org/pvldb/vol13/p403-echihabi.pdf.

Karima Echihabi, Kostas Zoumpatianos, and Themis Palpanas. Scalable machine learning on high-dimensional vectors: From data series to deep network embeddings. In *WIMS 2020: The 10th International Conference on Web Intelligence, Mining and Semantics, Biarritz, France, June 30 - July 3, 2020*, pages 1–6. ACM, 2020. doi: 10.1145/3405962.3405989. URL https://doi.org/10.1145/3405962.3405989.

Karima Echihabi, Panagiota Fatourou, Kostas Zoumpatianos, Themis Palpanas, and Houda Benbrahim. Hercules against data series similarity search. *Proc. VLDB Endow.*, 15(10):2005–2018, 2022. doi: 10.14778/3547305.3547308. URL https://www.vldb.org/pvldb/vol15/p2005-echihabi.pdf.

Karima Echihabi, Theophanis Tsandilas, Anna Gogolou, Anastasia Bezerianos, and Themis Palpanas. Pros: data series progressive k-nn similarity search and classification with probabilistic quality guarantees. *VLDB J.*, 32(4):763–789, 2023. doi: 10.1007/S00778-022-00771-Z. URL https://doi.org/10.1007/s00778-022-00771-z.

Panagiota Fatourou, Eleftherios Kosmas, Themis Palpanas, and George Paterakis. Fresh: A lock-free data series index. In *42nd International Symposium on Reliable Distributed Systems, SRDS 2023, Marrakesh, Morocco, September 25-29, 2023*, pages 209–220. IEEE, 2023. doi: 10.1109/SRDS60354.2023.00029. URL https://doi.org/10.1109/SRDS60354.2023.00029.

Antonin Guttman. R-trees: A dynamic index structure for spatial searching. In Beatrice Yormark, editor, *SIGMOD'84, Proceedings of Annual Meeting, Boston, Massachusetts, USA, June 18-21, 1984*, pages 47–57. ACM Press, 1984. doi: 10.1145/602259.602266. URL https://doi.org/10.1145/602259.602266.

King-Ip Lin, H. V. Jagadish, and Christos Faloutsos. The tv-tree: An index structure for high-dimensional data. *VLDB J.*, 3(4):517–542, 1994. URL http://www.vldb.org/journal/VLDBJ3/P517.pdf.

Themis Palpanas. Evolution of a Data Series Index - The iSAX Family of Data Series Indexes. In *Communications in Computer and Information Science (CCIS)*, volume 1197, 2020.

Themis Palpanas and Volker Beckmann. Report on the first and second interdisciplinary time series analysis workshop (ITISA). *SIGMOD Rec.*, 48(3):36–40, 2019. doi: 10.1145/3377391.3377400. URL https://doi.org/10.1145/3377391.3377400.

Botao Peng, Panagiota Fatourou, and Themis Palpanas. MESSI: in-memory data series indexing. In *36th IEEE International Conference on Data Engineering, ICDE 2020, Dallas, TX, USA, April 20-24, 2020*, pages 337–348. IEEE, 2020. doi: 10.1109/ICDE48307.2020.00036. URL https://doi.org/10.1109/ICDE48307.2020.00036.

Botao Peng, Panagiota Fatourou, and Themis Palpanas. SING: sequence indexing using gpus. In *37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19-22, 2021*, pages 1883–1888. IEEE, 2021a. doi: 10.1109/ICDE51399.2021.00171. URL https://doi.org/10.1109/ICDE51399.2021.00171.

Botao Peng, Panagiota Fatourou, and Themis Palpanas. Paris+: Data series indexing on multi-core architectures. *IEEE Trans. Knowl. Data Eng.*, 33(5):2151–2164, 2021b. doi: 10.1109/TKDE.2020.2975180. URL https://doi.org/10.1109/TKDE.2020.2975180.

Patrick Schäfer, Jakob Brand, Ulf Leser, Botao Peng, and Themis Palpanas. Fast and exact similarity search in less than a blink of an eye. In *41st IEEE International Conference on Data Engineering, ICDE 2025, Hong Kong, May 19-23, 2025*, pages 2464–2477. IEEE, 2025. doi: 10.1109/ICDE65448.2025.00186. URL https://doi.org/10.1109/ICDE65448.2025.00186.

Qitong Wang, Ioana Ileana, and Themis Palpanas. Leafi: Data series indexes on steroids with learned filters. *Proc. ACM Manag. Data*, 3(1):51:1–51:27, 2025. doi: 10.1145/3709701. URL https://doi.org/10.1145/3709701.

Zeyu Wang, Peng Wang, Themis Palpanas, and Wei Wang. Graph- and tree-based indexes for high-dimensional vector similarity search: Analyses, comparisons, and future directions. *IEEE Data Eng. Bull.*, 47(3):3–21, 2023. URL http://sites.computer.org/debull/A23sept/p3.pdf.