Report from Dagstuhl Seminar 19282

Data Series Management

Edited by

Anthony Bagnall[1], Richard L. Cole[2], Themis Palpanas[3], and Kostas Zoumpatianos[4]

1 **University of East Anglia – Norwich, GB,** anthony.bagnall@uea.ac.uk
2 **Tableau Software – Palo Alto, US,** ricole@tableau.com
3 **University of Paris, FR,** themis@mi.parisdescartes.fr
4 **Harvard University – Cambridge, US,** kostas@seas.harvard.edu

Abstract

We now witness a very strong interest by users across different domains on data series (a.k.a. time series) management. It is not unusual for industrial applications that produce data series to involve numbers of sequences (or subsequences) in the order of billions (i.e., multiple TBs). As a result, analysts are unable to handle the vast amounts of data series that they have to manage and process. The goal of this seminar is to enable researchers and practitioners to exchange ideas and foster collaborations in the topic of data series management and identify the corresponding open research directions. The main questions answered are the following: i) What are the data series management needs across various domains and what are the shortcomings of current systems, ii) How can we use machine learning to optimize our current data systems, and how can these systems help in machine learning pipelines? iii) How can visual analytics assist the process of analyzing big data series collections? The seminar focuses on the following key topics related to data series management: 1)Data series storage and access paterns, 2) Query optimization, 3) Machine learning and data mining for data serie, 4) Visualization for data series exploration, 5) Applications in multiple domains.

Seminar July 7–12, 2019 – http://www.dagstuhl.de/19282
2012 ACM Subject Classification Information systems → Data management systems
Keywords and phrases data series; time series; sequences; management; indexing; analytics; machine learning; mining; visualization
Digital Object Identifier 10.4230/DagRep.9.7.24

1 Executive Summary

Anthony Bagnall (University of East Anglia, GB)
Richard L. Cole (Tableau Software, US)
Themis Palpanas (Paris Descartes University, FR)
Kostas Zoumpatianos (Harvard University, US)

We now witness a very strong interest by users across different domains on data series[1] (a.k.a. time series) management systems. It is not unusual for industrial applications that produce data series to involve numbers of sequences (or subsequences) in the order of billions. As

[1] A data series, or data sequence, is an ordered set of data points.

Data Series Management, *Dagstuhl Reports*, Vol. 9, Issue 7, pp. 24–39
Editors: Anthony Bagnall, Richard L. Cole, Themis Palpanas, and Konstantinos Zoumpatianos
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

a result, analysts are unable to handle the vast amounts of data series that they have to filter and process. Consider for instance that in the health industry, for several of their analysis tasks, neuroscientists are reducing each of their 3,000 point long sequences to just the global average, because they cannot handle the size of the full sequences. Moreover, in the quest towards personalized medicine, scientists are expected to collect around 2-40 ExaBytes of DNA sequence data by 2025. In engineering, there is an abundance of sequential data. Consider for example that each engine of a Boeing Jet generates 10 TeraBytes of data every 30 minutes, while domains such as energy (i.e., wind turbine monitoring, etc.), data center, and network monitoring continuously produce measurements, forcing organizations to develop their custom solutions (i.e., Facebook Gorilla).

The goal of this seminar was to enable researchers and practitioners to exchange ideas in the topic of data series management, towards the definition of the principles necessary for the design of a big sequence management system, and the corresponding open research directions.

The seminar focused on the following key topics related to data series management:
Applications in multiple domains: We examined applications and requirements originating from various fields, including astrophysics, neuroscience, engineering, and operations management. The goal was to allow scientists and practitioners to exchange ideas, foster collaborations, and develop a common terminology.
Data series storage and access patterns: We described some of the existing (academic and commercial) systems for managing data series, examined their differences, and commented on their evolution over time. We identified their shortcomings, debated on the best ways to lay out data series on disk and in memory in order to optimize data series queries, and examined how to integrate domain specific summarizations/indexes and compression schemes in existing systems.
Query optimization: One of the most important open problems in data series management is that of query optimization. However, there has been no work on estimating the hardness/selectivity of data series similarity search queries. This is of paramount importance for effective access path selection. During the seminar we discussed the current work in the topic, and identified promising future research directions.
Machine learning and data mining for data series: Recent developments in deep neural network architectures have also caused an intense interest in examining the interactions between machine learning algorithms and data series management. We discussed machine learning from two perspectives. First, how machine learning techniques can be applied for data series analysis tasks, as well as for tuning data series management systems. Second, we how data series management systems can contribute towards the scalability of machine learning pipelines.
Visualization for data series exploration: There are several research problems in the intersection of visualization and data series management. Existing data series visualization and human interaction techniques only consider very small datasets, yet, they can play a significant role in the tasks of similarity search, analysis, and exploration of very large data series collections. We discussed open research problems along these directions, related to both the frontend and the backend.

2 Table of Contents

Executive Summary

Overview of Talks

3 Overview of Talks

3.1 Interaction Metaphors for Time Series Analysis

Azza Abouzied (New York University – Abu Dhabi, AE)

Through Qetch, I describe how a simple canvas metaphor can afford an intuitive and powerful querying language by allowing users to sketch patterns of interest, annotate them, as well as apply regular expression operations to search for repeated patterns or anomalies. The canvas metaphor also affords powerful multi-series querying functionality through the relative positioning of sketches. Through revisiting fundamental interaction metaphors, we can uncover elegant mechanisms for other complex time series analysis tasks.

3.2 Mini Tutorial on Time Series Data Mining Top of Form

Anthony Bagnall (University of East Anglia – Norwich, GB)

TSDM is a research are that involves developing algorithms for tasks relating to time series. These can be grouped into two families of tasks:

1. Specializations of generic machine learning tasks: classification, regression, clustering, rule discovery and query problems, and all variants thereof, such as semi-supervised/active learning, attribute selection, reinforcement learning, etc.
2. Time series specific tasks:
   a. Forecasting/panel forecasting;
   b. Time to event modelling/survival analysis;
   c. Annotation, such as segmentation, anomaly detection, motif discovery, discretization, imputation.

Problems can move from one task to another through a reduction strategy. For example, a regression task can be transformed to a classification task through discretizing the response variable, and forecasting can be reduced to regression through applying a sliding window. The challenges for TSDM include promoting reproducibility through open source code and improving evaluation strategies through better use of public data repositories and dealing with the challenges of large data so that algorithms can balance scalability vs accuracy. This becomes hugely important when dealing with streaming data, in particular with IoT applications involving widespread sensor nets where decisions need to be made about what data to store.

3.3 Visualizing Large Time Series (a brief overview)

Anastasia Bezerianos (INRIA Saclay – Orsay, FR)

Visually representing in a meaningful way large timeseries remains a research challenge for the visualization community. We present examples of existing approaches that attack the problem using different solutions, such as representing visual aggregations, illustrating representative patterns in the data, or creating novel compact visual representations. One key aspect in deciding what to visualize and how, is to understand why the timeseries needs to be visualized – i.e., what tasks the viewer needs to perform. This influences both what type of visual representation is more appropriate to use, but also what interactions need to be supported to help visual analysis. We conclude with general challenges (and new directions) in visualizing and iteratively interacting with large amounts of data in real time.

3.4 Anomaly Detection in Large Data Series

Paul Boniol (Paris Descartes University, FR)

Main reference Paul Boniol, Michele Linardi, Federico Roncallo, Themis Palpanas: “Automated Anomaly Detection in Large Sequences”, ICDE, 2020.

Subsequence anomaly (or outlier) detection in long sequences is an important problem with applications in a wide range of domains. However, the approaches that have been proposed so far in the literature have limitations: they either require prior domain knowledge, or become cumbersome and expensive to use in situations with recurrent anomalies of the same type. We briefly discuss these problems in this talk.

3.5 Data Series Management and Query Processing in Tableau

Richard L. Cole (Tableau Software – Palo Alto, US)

Tableau supports operations on time series, such as formatting, filters, calcs, date parts, date parse, and time zones. This talk is about Tableau’s aspirations to support query processing of exceptionally large data series, including complex data mining analytics, such as similarity search. Query processing may be divided into query compilation and query execution. New query compiler language elements and query execution operators will be needed. Additionally, support for data series data sources, i.e., time focused database systems, and federated query processing for data series in general will be desirable.

3.6 Location Intelligence

Michele Dallachiesa (Minodes GmbH – Berlin, DE)

For the first time in human's history, the position of more than three-quarters of the world's population is recorded at a fine-grained spatiotemporal resolution. This massive data source provides a unique view for infrastructure planning, retail development, and demographic research. In this talk, I overview two important localization strategies based on cellular and WiFi networks. In addition, the correct handling of missing or imprecise data points is presented as one of the major challenges in providing actionable insights with quality guarantees.

3.7 Data Series Similarity Search: Where Do We Stand Today? And Where Are We Headed?

Karima Echihabi (ENSIAS-Mohammed V University – Rabat, MA)

Main reference Karima Echihabi, Kostas Zoumpatianos, Themis Palpanas, Houda Benbrahim: "The Lernaean Hydra of Data Series Similarity Search: An Experimental Evaluation of the State of the Art", Proc. VLDB Endow., Vol. 12(2), pp. 112–127, VLDB Endowment, 2018.
URL https://doi.org/10.14778/3282495.3282498
Main reference Karima Echihabi, Kostas Zoumpatianos, Themis Palpanas, Houda Benbrahim: "Return of the Lernaean Hydra: Experimental Evaluation of Data Series Approximate Similarity Search", PVLDB, Vol. 13(3), pp. 403–420, 2019.
URL http://dx.doi.org/10.14778/3368289.3368303

Increasingly large data series collections are becoming commonplace across many different domains and applications. A key operation in the analysis of data series collections is similarity search, which has attracted lots of attention and effort over the past two decades. We presented the results of two extensive experimental evaluations. The three main lessons learned are as follows: 1) choosing the best approach is an optimization problem that depends on several factors (hardware, data characteristics, summarization quality and clustering efficacy); 2) exact search is slow; 3) approximate search can be fast and accurate (our extensions to exact techniques outperform the state-of-the-art on disk). We also outlined our future research directions: 1) building a new index that outperforms the state-of-the-art in-memory and on-disk; and 2) exploring query optimization for data series.

3.8 Progressive PCA for Time-Series Visualization

Jean-Daniel Fekete (INRIA Saclay – Orsay, FR)

With EDF, we are interested in the visual sensitivity analysis for ensemble simulation. EDF uses simulation software for forecasting the evolution of rivers and sea levels in the next 100 years. Their simulation system produces a large number of results, called "ensemble simulations", that are plausible evolutions for a river, such as the level every month for the next 100 years. These time series are then analyzed to find out if the results cluster around one value, or spread over multiple possible “regimes” or “modes”. This analysis is usually performed by clustering of the results but should be supervised and interpreted by analysts. Therefore, we use a dimensionality reduction algorithm to project the resulting time-series using Principal Component Analysis, explore the results, cluster it, and allow experts to write reports on the possible outcomes of the simulation. We are adapting PCA to cope with a large number of time series. First, PCA has seen recently a surge of new results related to its use online for out-of-core datasets. Second, iterative PCA computations have also been recently improved recently to boost its convergence using momentum. We are exploring these new algorithms as well as multi-resolution computation to reach interactive rates for computing PCA over a large number of time-series.

3.9 Deep Learning for Time Series Classification, and Applications in Surgical Data Science

Germain Forestier (University of Mulhouse, FR)

In recent years, deep learning approaches have demonstrated a tremendous success in multiple domains like image processing, computer vision or speech recognition. In this talk, I reviewed recent advances in deep learning for univariate and multivariate time series classification. I presented experimental results obtained with the principal architectures proposed in the literature. I also discussed the main challenges linked with the use of deep learning like transfer learning, data augmentation, ensembling and adversarial attacks. Moreover, I presented some applications in the field of Surgical Data Science which is an emerging field with the objective of improving the quality of interventional healthcare through capturing, organizing, analyzing and modeling of data. Finally, I discussed an application of the above in Surgical Data Science. The need for automatic surgical skills assessment is increasing, especially because manual feedback from senior surgeons observing junior surgeons is prone to subjectivity and time consuming. Thus, automating surgical skills evaluation is a very important step towards improving surgical practice. I presented how we used a Convolutional Neural Network (CNN) to evaluate surgeon skills by extracting patterns in the surgeon motions performed in robotic surgery. The proposed method has been validated on the JIGSAWS dataset and achieved very competitive results with 100% accuracy on the suturing and needle passing tasks. While we leveraged from the CNNs efficiency, we also managed to mitigate its black-box effect using class activation map. This feature allows our method to automatically highlight which parts of the surgical task influenced the skill prediction and can be used to explain the classification and to provide personalized feedback to the trainee.

3.10 Seismic Time Series: Introduction and Applications

Pierre Gaillard (CEA de Saclay – Gif-sur-Yvette, FR)

Seismometers, also called seismic stations, are sensitive instruments located all over the world, that allow to record continuously the smallest displacements of the ground. The given data take the form of discrete time series that are the basis of various studies: seismic risk analysis, seismic wave propagation, tomography of the Earth and seismic monitoring. This presentation is focused on seismic monitoring, and we present a standard pipeline dedicated to detect and characterize seismic event. To perform this task automatically, fast and reliable processing is required to extract as much information as possible from all the available time series. Such processing includes quality control, detection of event, measurement of features (amplitude, direction of arrival, polarity...), clustering or classification (e.g. anthropic versus natural events). All this information is then used by seismologists and are controlled, improved, shared or stored. This user intervention is usually performed through interactive software that need to manage and display large collection of time-series, as well as the associated data (detections, events, features...). Due to the increase of data available to perform seismic monitoring, we emphasize in the conclusion of this presentation, the need of new management system as well as new processing techniques based on machine learning in order to improve the analysis pipeline.

3.11 Progressive Similarity Search in Large Data Series Collections

Anna Gogolou (INRIA Saclay – Orsay, FR)

Main reference Anna Gogolou, Theophanis Tsandilas, Themis Palpanas, Anastasia Bezerianos: "Progressive Similarity Search on Time Series Data", in Proc. of the Workshops of the EDBT/ICDT 2019 Joint Conference, EDBT/ICDT 2019, Lisbon, Portugal, March 26, 2019., CEUR Workshop Proceedings, Vol. 2322, CEUR-WS.org, 2019.
URL http://ceur-ws.org/Vol-2322/BigVis_5.pdf

Time series data are increasing at a dramatic rate, yet their analysis remains highly relevant in a wide range of human activities. Due to their volume, existing systems dealing with time series data cannot guarantee interactive response times, even for fundamental tasks such as similarity search. Therefore, in this talk, we present our vision to develop analytic approaches that support exploration and decision making by providing progressive results, before the final and exact ones have been computed. We demonstrate through experiments that providing first approximate and then progressive answers is useful (and necessary) for similarity search queries on very large time series data. Our findings indicate that there is a gap between the time the most similar answer is found and the time when the search algorithm terminates, resulting in inflated waiting times without any improvement. We present preliminary ideas on computing probabilistic estimates of the final results that could help users decide when to stop the search process, i.e., deciding when improvement in the final answer is unlikely, thus eliminating waiting time. Finally, we discuss two additional challenges: how to compute efficiently these probabilistic estimates, and how to communicate them to users.

3.12 Model-Based Management of Correlated Dimensional Time Series

Søren Kejser Jensen (Aalborg University, DK)

Owners and manufacturers of wind turbines would like to collect and store large quantities of high-quality sensor data. However, the amount of storage required makes this infeasible and only simple aggregates are stored. This removes outliers and hides fluctuations that could indicate problems with the wind turbines. As a remedy, these high-quality regular time series can instead be stored as models which reduces the amount of storage required by approximating the time series within a user-defined error bound (possibly 0%). As time series change over time, each time series should be represented using multiple different model types, and as a data set often contains multiple similar time series, correlation should be exploited to further reduce the amount of storage required. ModelarDB is a time series management system that stores time series as models and takes all the above factors into account. There are still open-questions related to model-based storage of time series. How do we assist users with selecting a good set of model types to use for a particular data set? Can similarity search be performed directly on models instead of on data points reconstructed from the models? Can models be fitted at the turbines without significantly increasing the latency and/or the amount of data being transferred? And can the error bound be inferred from the user's query workload?

3.13 Time Series Recovery

Mourad Khayati (University of Fribourg, CH)

Recording sensor data is seldom a perfect process. Missing values often occur as blocks in time series data due to multiple reasons, e.g., sensor failure, server transmission, etc. In my talk, I introduced the problem of missing values in real-world time series data. Then, I introduced our solution to recover missing blocks in time series with mixed correlation. Finally, I summarized the main open research problems in the field.

3.14 Adaptive and fractal time series analysis: methodology and applications

Alessandro Longo (University of Rome III, IT)

A methodology for adaptive and fractal time series analysis, based on Empirical Mode Decomposition, time-varying filter EMD and Detrended Fluctuation Analysis has been applied to characterize time series data from different physical systems. It has been applied to seismometer data from sensors monitoring the Virgo interferometer and to data of activity concentration of cosmogenic beryllium-7, sampled worldwide by the International Monitoring

System of the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO). In the first case, seismometer was recording during an acoustic noise injection performed for detector characterization purposes. Using adaptive and fractal algorithms, the seismic perturbation due to acoustic noise performed in the room can be separated from the underlying nonlinear nonstationary noise affecting the seismometer. Furthermore, applying Hilbert Spectral analysis provided with a high-resolution time frequency representation, even though the data length is short due to the low sampling frequency. In the second case, extracting the yearly oscillatory mode of beryllium-7, sampled by a worldwide distributed network, allowed to characterize its shift in time of occurrence in term of patterns of large scale atmospheric dynamics, namely in term of the seasonal shift of the Hadley cell.

3.15 Helicopters Time Series Management & Analysis

Ammar Mechouche (Airbus Helicopters – Marignane, FR)

Massive time series data are collected within the aerospace domain, making its management and analysis a challenging task. This presentation offers a return of experience regarding AIRBUS Helicopter flight data management and analysis. It illustrates first the big amount of time series data collected at every flight. Then, it shows how this data is managed and exploited using latest big data technologies. After that, illustrative examples that show the benefits from the analysis of this data are provided. Finally, some challenging use cases are presented, highlighting some limitations of existing tools and analysis methods.

3.16 Socio-temporal Data Mining

Abdullah Mueen (University of New Mexico, US)

Modern social media produce timestamped events that can be considered as a data series to mine for patterns. We consider a large number of tweets from millions of Twitter users in a streaming manner for several years and mine clusters, motifs and cluster-dynamics. Emerged patterns represent automated user behavior, host curation events, and political events. We also show an application of pattern mining to extract hidden seismic events.

3.17 Data Series Mining and Applications

Rodica Neamtu (Worcester Polytechnic Institute, US)

My research interests are at the crossroads of theoretical computer science and Big Data analytics. In this light, my work reveals that domain-specific distances preferred by analysts for exploring similarities among time series tend to be “point-to-point” distances. Unfortunately, this point-wise nature limits their ability to perform meaningful comparisons between sequences of different lengths and with temporal mis-alignments. Analysts instead need "elastic" alignment tools such as Dynamic Time Warping (DTW) to perform such flexible comparisons. However, the existing alignment tools are limited in that they do not incorporate diverse distances. To address this shortcoming, our work introduces the first conceptual framework called Generalized Dynamic Time Warping (GDTW) that supports now alignment (warping) of a large array of domain-specific distances in a uniform manner. We further use these warped distances to explore data in diverse application domains including neuroscience, finance, healthcare and to facilitate communication for people with disabilities. My talk discusses briefly three projects in these areas and highlights the common denominator represented by the omnipresence of data series. First I showcase our work incorporating machine learning to support an Augmentative Alternative Communication app for people with several verbal and motor-skill challenges. Our LIVOX application incorporates artificial intelligence algorithms to reduce the so-called ”reciprocity gap” that acts as a communication barrier between disabled people and their interlocutors, thus enabling people with disabilities, especially children, to participate in daily social and educational activities. Integrating them into the existing social structures is central to making the world a more inclusive place. Then I discuss our use of generalized warping distances to explore data for neuroadaptive technology. We show that our exploratory tool can use different similarity distances for robust identification of similar patterns in the brain data during complex tasks. This builds a foundation for interactive systems that are capable of identifying cognitive states and adapting system behavior to better support users. Lastly, I discuss a tool for automatic website content classification for the financial technology (Fintech) domain that facilitates the identification of promising startup fintec companies.

3.18 Fulfilling the Need for Big Sequence Analytics

Themis Palpanas (Paris Descartes University, FR)

Main reference Themis Palpanas: "Data Series Management: The Road to Big Sequence Analytics", SIGMOD Record, Vol. 44(2), pp. 47–52, 2015.
URL https://doi.org/10.1145/2814710.2814719

Massive data sequence collections exist in virtually every scientific and social domain, and have to be analyzed to extract useful knowledge. However, no existing data management solution (such as relational databases, column stores, array databases, and time series management systems) can offer native support for sequences and the corresponding operators necessary for complex analytics. We argue for the need to study the theory and foundations for sequence management of big data sequences, and to build corresponding systems that will enable scalable management and analysis of very large sequence collections. To this effect, we need to develop novel techniques to efficiently support a wide range of sequence queries and mining operations, while leveraging modern hardware. The overall goal is to allow analysts across domains to tap in the goldmine of the massive and ever-growing sequence collections they (already) have.

3.19 Accelerating IoT Data Analytics through Time-Series Representation Learning

John Paparrizos (University of Chicago, US)

Main reference John Paparrizos, Michael J. Franklin: "GRAIL: Efficient Time-series Representation Learning", Proc. VLDB Endow., Vol. 12(11), pp. 1762–1777, VLDB Endowment, 2019.
URL http://dx.doi.org/10.14778/3342263.3342648
Main reference John Paparrizos, Luis Gravano: "Fast and Accurate Time-Series Clustering", ACM Trans. Database Syst., Vol. 42(2), pp. 8:1–8:49, 2017.
URL https://doi.org/10.1145/3044711

The analysis of time series is becoming increasingly prevalent across scientific disciplines and industrial applications. The effectiveness and the scalability of time-series mining techniques critically depend on design choices for three components: (i) representation; (ii) comparison; and (iii) indexing. Unfortunately, these components have to date been investigated and developed independently, resulting in mutually incompatible methods. The lack of a unified approach has hindered progress towards fast and accurate analytics over massive time-series collections. To address this major drawback, we present GRAIL, a generic framework to learn in linear time and space compact time-series representations that preserve the properties of a user-specified comparison function. Given the comparison function, GRAIL (i) extracts landmark time series using clustering; (ii) optimizes necessary parameters; and (iii) exploits approximations for kernel methods to construct representations by expressing time series as linear combination of the landmark time series. We build GRAIL on top of Apache Spark to facilitate analytics over large-scale settings and we extensively evaluate GRAIL's representations for querying, classification, clustering, sampling, and visualization of time series. For these tasks, methods leveraging GRAIL's compact representations are significantly faster and at least as accurate as state-of-the-art methods operating over the raw high-dimensional time series. GRAIL shows promise as a new primitive for highly accurate, yet scalable, time-series analysis.

3.20 Contradictory Goals of Classification, Accuracy, Scalability and Earliness

Patrick Schäfer (HU Berlin, DE)

Time series classification (TSC) tries to mimic the human understanding of similarity. Classification approaches can be divided into 6 areas: whole series, Shapelets, Dictionary and Interval, Ensembles and Deep Learning. Our research focusses on three contradictory goals of TSC, namely accuracy, scalability and earliness. Much research has gone into improving the accuracy of TSC. When it comes to long or larger time series datasets, these state-of-the-art classifiers reach their limits because of high training or prediction times. To improve scalability, a classifier has to sacrifice on accuracy. In contrast, early time series classification (eTSC) is the problem of classifying a time series after seeing as few measurements as possible with the highest possible accuracy. The most critical issue is to decide when enough data of a time series has been seen to take a decision: Waiting for more data points usually makes the classification problem easier but delays the time in which a classification is made.

3.21 More Reliable Machine Learning through Refusals

Dennis Shasha (New York University, US)

SafePredict is a meta-algorithm that sits on top of one or more machine learning algorithms. It takes each prediction from these algorithms (which may be weighted) and decides whether to accept or refuse to accept that prediction. Suppose that a user sets an error threshold E. SafePredict will endeavor to guarantee that among all accepted predictions the fraction of errors doesn't exceed E. Under very general assumptions, SafePredict can guarantee this. When the data points are i.i.d. (independent and identically distributed), SafePredict does even better.

3.22 Systems and Tools for Time Series Analytics

Nesime Tatbul (Intel Labs & MIT – Cambridge, US)

Main reference Metronome Project
URL http://metronome.cs.brown.edu/

From autonomous driving to industrial IoT, the age of billions of intelligent devices generating time-varying data is here. There is a growing need to ingest and analyze time series data accurately and efficiently to look for interesting patterns at scale. Our key goal in the Metronome Project is to build novel data management, machine learning, and interactive visualization techniques for supporting the development and deployment of predictive time series analytics applications, such as anomaly detection [1]. In this talk, I give three example tools that we have recently built for time series anomaly detection: (i) a customizable scoring model for evaluating accuracy, which extends the classical precision/recall model to range-based data; (ii) a zero-positive learning paradigm, which enables training anomaly detectors in absence of labeled datasets; and (iii) a visual tool for interactively analyzing time series anomalies.

3.23 Data Series Similarity Search

Peng Wang (Fudan University – Shanghai, CN)

Similarity search is a fundamental task for data series mining. In this talk, I introduce our works for both whole matching problem and subsequence matching problem, DSTree and KV-match. Also, some ongoing works and open problems are discussed.

3.24 Tableau for Data Series

Richard Wesley (Tableau Software – Seattle, US)

Tableau is an interface for converting visual specifications into queries. To enable this, it uses a unified data model that interfaces to a large number of query engines. This model has many advantages for simplifying the user experience and integrating data, but it also leads to a lowest common denominator approach that restricts analysis to a small number of data types and makes it hard to integrate complex data types like data series and the associated query operations.

3.25 Managing and Mining Large Data Series Collections

Konstantinos Zoumpatianos (Harvard University – Cambridge, US)

Main reference Kostas Zoumpatianos, Themis Palpanas: "Data Series Management: Fulfilling the Need for Big Sequence Analytics", in Proc. of the 34th IEEE International Conference on Data Engineering, ICDE 2018, Paris, France, April 16-19, 2018, pp. 1677–1678, IEEE Computer Society, 2018.
URL https://doi.org/10.1109/ICDE.2018.00211

Data series management has recently gathered a great amount of attention. This is mainly driven by the large amount of sequential information that analysts both in science as well as in industry need to be able to monitor and analyze. In this talk we will look at how we can manage large collections of data series, the types of data analysis tasks that are commonly performed, and how they can be efficiently performed from a data systems perspective. Specifically, we will look at an overview of the most commonly found query templates and data mining tasks (clustering, classification, deviation detection, frequent pattern mining), specialized index structures for efficiently answering such queries, as well as pinpoint the open problems and research directions.

Participants

- Azza Abouzied
New York University –
Abu Dhabi, AE
- Anthony Bagnall
University of East Anglia –
Norwich, GB
- Anastasia Bezerianos
INRIA Saclay – Orsay, FR
- Paul Boniol
Paris Descartes University, FR
- Richard L. Cole
Tableau Software – Palo Alto, US
- Michele Dallachiesa
Minodes GmbH – Berlin, DE
- Karima Echihabi
ENSIAS-Mohammed V
University – Rabat, MA
- Jean-Daniel Fekete
INRIA Saclay – Orsay, FR
- Germain Forestier
University of Mulhouse, FR
- Pierre Gaillard
CEA de Saclay –
Gif-sur-Yvette, FR
- Anna Gogolou
INRIA Saclay – Orsay, FR
- Søren Kejser Jensen
Aalborg University, DK
- Mourad Khayati
University of Fribourg, CH
- Alessandro Longo
University of Rome III, IT
- Ammar Mechouche
Airbus Helicopters –
Marignane, FR
- Abdullah Mueen
University of New Mexico, US
- Rodica Neamtu
Worcester Polytechnic
Institute, US
- Themis Palpanas
Paris Descartes University, FR
- John Paparrizos
University of Chicago, US
- Patrick Schäfer
HU Berlin, DE
- Dennis Shasha
New York University, US
- Nesime Tatbul
Intel Labs & MIT –
Cambridge, US
- Peng Wang
Fudan University – Shanghai, CN
- Richard Wesley
Tableau Software – Seattle, US
- Konstantinos Zoumpatianos
Harvard University –
Cambridge, US