Interactive Time Series Analytics Powered by ONEX

Rodica Neamtu, Ramoza Ahsan, Charles Lovering, Cuong Nguyen,
Elke A. Rundensteiner, and Gabor Sarkozy
Worcester Polytechnic Institute, Worcester, MA 01609
rneamtu | rahsan | cjlovering | ctnguyendinh | rundenst | gsarkozy @wpi.edu

ABSTRACT

Modern applications in this digital age collect a staggering amount of time series data from economic growth rates to electrical household consumption habits. To make sense of it, domain analysts interactively sift through these time series collections in search of critical relationships between and recurring patterns within these time series. The ONEX (Online Exploration of Time Series) system supports effective exploratory analysis of time series collections composed of heterogeneous, variable-length and misaligned time series using robust alignment dynamic time warping (DTW) methods. To assure real-time responsiveness even for these complex and compute-intensive analytics, ONEX precomputes and then encodes time series relationships based on the inexpensive-to-compute Euclidean distance into the ONEX base. Thereafter, based on a solid formal foundation, ONEX uses DTW-enhanced analytics to correctly extract relevant time series matches on this Euclidean-prepared ONEX base. Our live interactive demonstration shows how our ONEX exploratory tool, supported by a rich array of visual interactions and expressive visualizations, enables efficient mining and interpretation of the MATTERS real data collection composed of economic, social, and education data trends across the fifty American states.

1. INTRODUCTION

Motivation and Background. In the era when time series are prevalent in most application domains, we extract important insights from time series datasets. Let's consider the use case of finding and leveraging time series similarities for decision making. In 2013 in Massachusetts, organizations set out to repeal the Sales and Use Tax on computer and software services, perceived as potentially having a negative impact on the economic health of the state. Vast amounts of data were analyzed to show similarities between tax rates and fluctuations of social and economic factors, all modeled as *time series*, obtained from a large spectrum of public governmental websites such as the Tax Policy Center[1], the Census Bureau[2], and the Bureau of Economic Analysis [3]. Finding and interpreting the similarities between diverse economic indicators represented as time series were central to this process.

(1) Analysts had to answer complex questions that were not always based on traditional similarity searches. For example, they had to look for recurring similarity patterns in the growth or unemployment rate of a state over a few years. (2) The presence of data from different domains reported over specific intervals required comparisons of time series of different lengths and alignments, as the impact of a tax change might play out with different time durations.

(3) Data from diverse domains required using different parameter settings such as similarity thresholds, leading to repeated and redundant computations for each parameter value. For example, the most suitable thresholds for studying similarity of demographic data are different than those used for growth rates.

Challenges. To enable interactive time series analytics, many challenges must be tackled:

1. High data cardinality leading to decreased responsiveness. Time series datasets such as census, financial and tax data tend to be huge. To answer questions involving varying temporal granularities one needs to consider the whole time series as well as all shorter subsequences. Given the huge number of such subsequences, performing similarity comparisons among them is impractical.

2. Sequences of different lengths and alignments. To perform meaningful comparisons between sequences of different lengths and alignments, complex elastic distances such as Dynamic Time Warping (DTW) [2] must be used. Its popularity among robust alignment tools is only overshadowed by its high computational cost [6]. The exploration of the huge number of pairwise similarity comparisons between sequences using DTW leads inherently to long response times. Many systems settle either on increased latency or on decreased accuracy using cheaper-to-compute distances.

3. Offering rich classes of exploratory queries. As discussed above, to gain insights into data more operations are needed, including finding repeating patterns and best matches to discover similarity patterns, all performed using varying parameters such as similarity thresholds.

State of the Art. Most systems face the trade-off between accuracy and time response especially when dealing

SIGMOD'17, May 14-19, 2017, Chicago, IL, USA
© 2017 ACM. ISBN 978-1-4503-4197-4/17/05...$15.00
DOI: http://dx.doi.org/10.1145/3035918.3058729

[1] http://www.taxpolicycenter.org/
[2] www.census.gov
[3] www.bea.gov/

with high volumes of data. Some provide an exact or a highly accurate solution [7] at the expense of responsiveness. Others [1] use preprocessing steps to improve the timely responsiveness, but their requirement for setting many different parameters limits their efficiency [6]. Clearly, a dilemma exists between choosing complex similarity distances and time responsiveness, namely, while shorter time responses are guaranteed by the use of fast-to-compute distances like the Euclidean Distance [4], such distances cannot handle sequences with different time alignments. Meaningful comparisons of such sequences require the use of time-warping distances like DTW, whose computational complexity [3] leads to slow responsiveness and poor scaling as data grows.

ONEX: Our Approach for Time Series Analytics. ONEX offers a viable compromise between real time responsiveness and the high computational complexity needs of the elastic distances. We counter the prohibitively expensive use of DTW by employing a "marriage" of two distances: we use the computationally inexpensive Euclidean Distance to construct compact similarity groups for specific lengths, and then we support the exploration of these groups using the robust time-warping method DTW [2]. This unique combination yields very accurate results at much reduced response time rates, as DTW is successfully applied over the compact ONEX base instead of the raw data [5]. Our theoretical foundation rests on the proof of a triangle inequality between ED and DTW that builds a conceptual bridge between the offline construction of the ONEX base and its online exploration. Offering a novel answer to the trade-off dilemma between the use of complex time warped distances and timely responsiveness, ONEX has been shown to be several times faster than the fastest known method [6], while still delivering up to 19% more accurate results [5]. Our web-based visual analytics interface enable analysts to explore and directly interact with time series data sets in an intuitive manner using rich classes of operations. Further, diverse visualizations capture the matched and/or clustered times series data and their similarity relationships to enrich the interpretation of the matches between sequences.

Our live interactive demonstration on real world economic and social data sets illustrates how our ONEX visual web interface enables analysts to interactively explore similarity and its recurring patterns.

2. THE ONEX FRAMEWORK

The ONEX infrastructure is depicted in Fig. 1. The pre-processing step (top) encodes similarity relationships between time series into "ONEX similarity groups" using Euclidean Distance. Exploration of these groups is achieved by DTW-empowered operations via the query processor (middle). The middle of the figure indicates the diverse exploratory operations performed by ONEX which include retrieving of the best match for a given sample sequence, finding repeating patterns of specific time series or seasonal similarity, and showing the changes in the similarity between sequences for varying parameters. Lastly, the visual analystics interface, depicted at the bottom, enables the user to view, analyze and interpret information for interactive time series exploration.

3. KEY INNOVATIONS

The theoretical foundation of ONEX guarantees efficient exploration of time series datasets by first creating a compact ONEX base using ED and then enabling time-warped comparisons between any sample sequence and this compact set of representative sequences in the ONEX base.

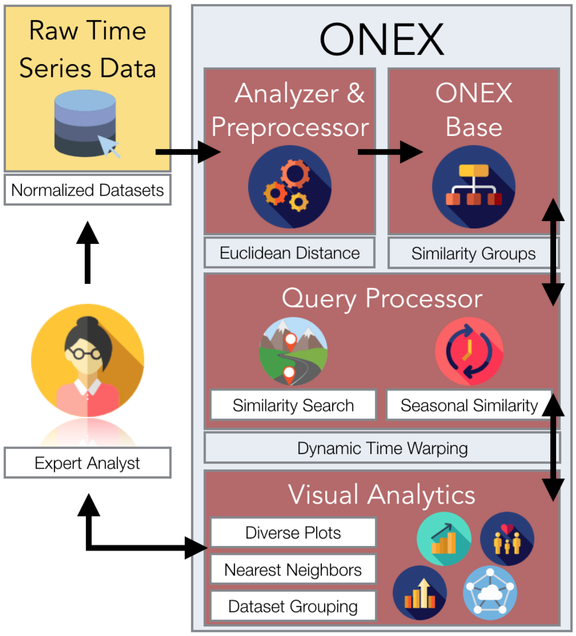


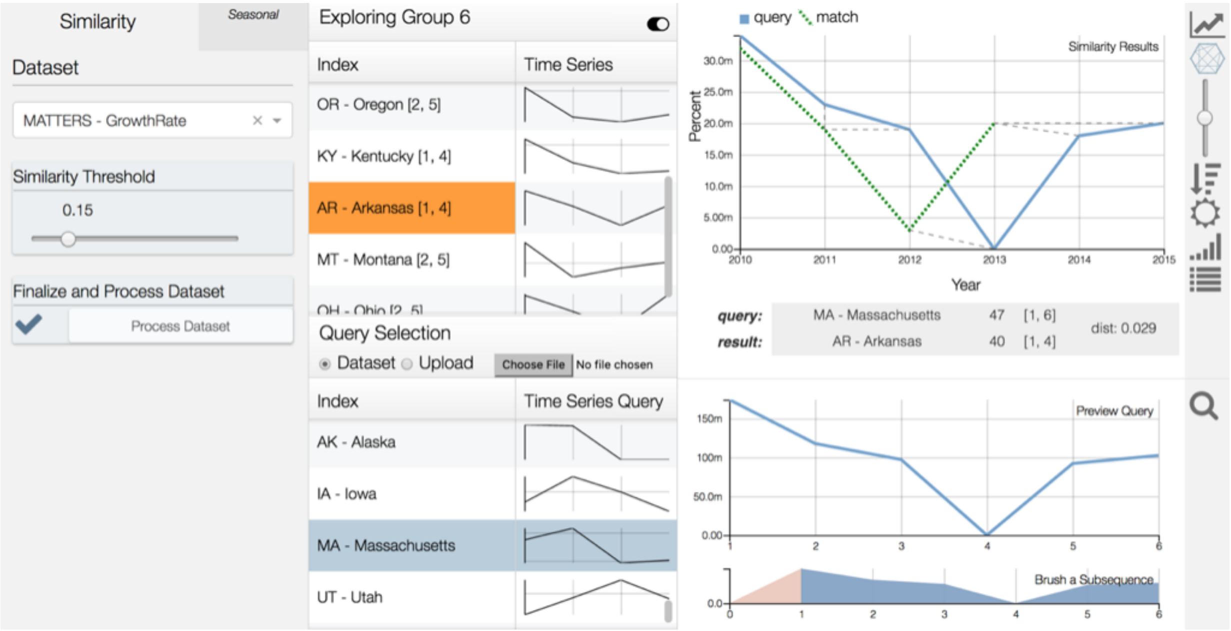
Figure 1: ONEX Framework

3.1 ONEX Base: Using Clustering for Data Reduction

Our ONEX base preserves and compacts similarity relationships between subsequences of the time series. We first group subsequences of the same length that are similar using the ubiquitous and inexpensive Euclidean Distance into so called "ONEX similarity groups". We then summarize these groups by their centroid, or the average of all sequences in each group, to be the representative of that group. Our construction methodology insures that theses similarity groups contain sequences that are similar to each other within the similarity threshold ST, while each sequence is similar to the representative within half of the similarity threshold. The use of the compact ONEX base instead of the entire dataset for exploration using DTW guarantees speed-up while assuring highly accurate results.

3.2 Time-Warped Exploration Model Based on Fundamental Similarity Mapping

Our ONEX time-warped retrieval framework rests upon the proven insight of a triangle inequality between ED and DTW. Namely, the DTW similarity between a sample sequence *seq* provided by the analyst and a match found in form of a representative of one of the ONEX similarity groups can be proven to hold true for all subsequences in that respective group. This empowers ONEX to perform time-warped comparisons of the sample sequence over the compact ONEX base instead of the entire dataset. Namely, it guarantees that the best match to a sample sequence *seq* is found in the group with the "best match representative" and the DTW between *seq* and its best match is always within the similarity threshold ST.



MATTERS - GrowthRate

Similarity Threshold

0.15

Index	Time Series
OR - Oregon [2, 5]	
KY - Kentucky [1, 4]	
AR - Arkansas [1, 4]	
MT - Montana [2, 5]	
OH - Ohio [2, 5]	

Query Selection

● Dataset ○ Upload Choose File No file chosen

Index	Time Series Query
AK - Alaska	
IA - Iowa	
MA - Massachusetts	
UT - Utah	

query: MA - Massachusetts 47 [1, 6]
result: AR - Arkansas 40 [1, 4]
dist: 0.029

Figure 2: Similarity View with *Overview Pane*, *Query Selection Pane*, *Query Preview Pane*, and *Similarity Results Pane* showed respectively in counter-clockwise order.

3.3 ONEX Exploratory Model and Processing

The ONEX conceptual framework provides rich classes of exploratory operations as will be demonstrated using our economic analytics use case. These include *similarity queries* that retrieve the best match for a given sequence. Using these, referring back to our motivating example, an analyst can find the state that has the most similar economic growth rate with that of MA. *Seasonal similarity queries* find repeated patterns within a given time series. Thus one could find if a specific growth or decline in the economical growth of MA has previously been experienced in this state. *Threshold recommendations* help analysts to select appropriate parameter settings in a data-driven fashion. This is important as the similarity in growth rate percentages may require very small thresholds, whereas similarity between unemployment figures is expressed in tens of thousands of people uses higher thresholds. To efficiently support this diversity of time-warped queries, our ONEX query processor does not only leverage the compact ONEX base, but it also applies several optimization strategies ranging from indexing of time series using bounding envelopes to early pruning of unpromising candidates.

3.4 ONEX Visual Analytics

An array of complementary visualization techniques from stacked lines charts to connected scatter plots compose our ONEX web interface. Visualizations are critical for interactive time series analytics as they allow for intuitive interactions by analysts with large data sets. For the robust alignment of sequences of different lengths or alignments, a *warping path* constructed by ONEX corresponds to a a set of indices marking which points are most similar, including possible multiple matchings, contrary to the use of pointwise distances. Our ONEX views uniquely display these "warped" points to highlight the shape matching. Our *Multiple Lines Charts* display dotted lines between corresponding points of the sequences highlighting the role of the time-warped matching. *Radial Plots* compact the time series to a radial display that allows analysts to evaluate how close the shapes are aligned. Similarly, the *Connected Scatter Plots* showcase the ordering of a sequence by connecting consecutive points.

4. ONEX DEMONSTRATION

In our demonstration, the audience will be able to directly interact with ONEX via an intuitive visual web interface to understand how it assists analysts in addressing complex societal and economical questions, such as the ones described in our motivating examples. We use real datasets from diverse domains such as economic, census and tax datasets from MATTERS[4] and a power usage dataset Electricity-Load[5]. Our interface empowers analysts to draw insights with ease from these collections, as described below.

Data Loading into ONEX. With a click of a button, analysts can load new data sets into ONEX. Loading a new dataset, such as the MATTERS GrowthRate, triggers the preprocessing of this data at the server side and its loading into the respective ONEX Base. Thereafter, the ONEX server provides near real-time responsiveness to the analyst exploring the data via a client-server architecture.

Making Sense of Overall Time Series Trends. To offer an overview of the data, the *Overview Pane* (Fig. 2 left top) displays the representatives of the similarity groups, color-coded such that the color intensity increases proportional with the cardinality of sequences in the group. This gives a quick sense of the typical patterns within the data set as well as the overall data distribution. Each representative is shown as a small graph that captures the general shape of the group. This supports analysts in finding the

[4] http://matters.mhtc.org/

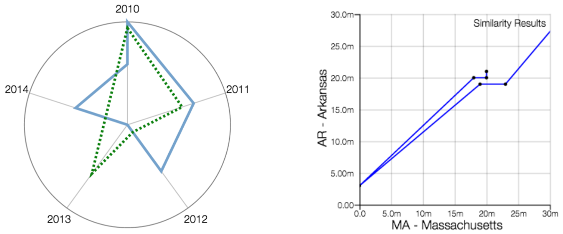
[5] *www.cs.ucr.edu/ eamonn/time_series_data/*

states with similar growth rates. Upon drill-down, the audience can scroll through the states in the *Query Selection Pane*, each visualized by its name and a small line graph displaying the growth rate over the last 6 years. In Fig. 2 left bottom, MA is selected from this group.

Honing in On Specific Temporal Trends. The *Similarity View* shown in Fig. 2 on the right assists the analyst in finding states with a similar economic growth rate to that of MA, while the *Seasonal View* in Fig. 4 enhances the understanding of a specific time series by highlighting repeating patterns. The *Query Preview Pane* displays the chosen sample query in more detail. Brushing the second half of the graph will focus the attention on the recent trends in MA. As the first preview graph is brushed, the upper chart is updated to show the selected subsequence in more detail.

Highlighting Time-Warped Shape Matching. When the analyst performs a similarity search, the best match sequence in the dataset is displayed along with the sample query subsequence in the *Results Pane* (Fig. 2 top right). The default "multiple lines" chart displays both time series on a single graph. The "matched points" are connected with dotted lines helping the analyst get a better intuition of how similar the time series shapes are and their relative warping.

Contrasting Trends Across Multiple Linked Perspectives. Different visualizations illustrate different aspects of similarity. For example, to get a richer understanding of the similarity between MA and ARK, the analyst can switch to different visuals by selecting the mode via the right menu bar of the *Results Pane*. The same pair of time series can now be viewed in a compacted *Radial Chart* (Fig. 3a). This view allows a consistent compression of the data, providing the analyst with alternative views to compare sequences. Further, in the *Connected Scatter Plot* (Fig. 3b), the shape is close to a 45 degree angle. This indicates that the match is extremely close – when a point in such plot lies on the diagonal, it has the exact same value in both series. This observation coupled with the fact that all values are very close in range indicates that the subsequences are a close match.



(a) Radial Chart (b) Connect Scatter Plot

Figure 3: Visualizing MATTERS Tech Employment

Exploring Re-occurrence of Motives Within Time Series. Our ONEX exploratory tool can work with data from diverse domains. We showcase now its use in exploring electrical usage data using the ElectricityLoad collection. The *similarity view* provides a wealth of information about repeated patterns in electricity usage. Focusing on the electrical consumption of a single household, Fig. 4 shows a single time series across one year in Portugal and finds repeated patterns within it. The alternating blue and green coloration are used to clarify instances of consecutive segments. The top graph displays a monthly pattern indicating that this household tends to use electricity in a consistent manner throughout the summer months. The bottom pattern shows that winter months too have similar trends, empowering the analyst to determine that a few small habit changes could have a large savings impact.

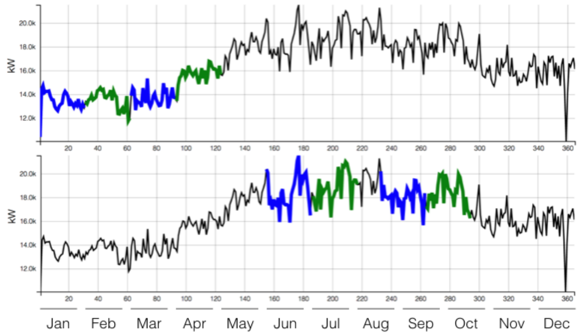


Figure 4: Seasonal View displaying Patterns in the Power Usage Dataset

5. CONCLUSION

ONEX is a truly interactive time series exploration tool that enables efficient exploration of time series datasets based on the combination of two similarity distances. This leads to shorter time responses compared to the fastest known state-of-the-art method. Complemented by novel visual analytics, ONEX offers actionable insights into similarity through rich classes of operations.

Acknowledgments *We thank the Computer Science Dept. at WPI and Fulbright for financial support. We thank Dr. David Brown and Dr. Lane Harrison from WPI for good feedback and suggestions. We thank former undergraduate students Samantha Swartz and Michael Andrews from WPI for help with the implementation.*

6. REFERENCES

[1] V. Athitsos et al. Approximate embedding-based subsequence matching of time series. In *ACM SIGMOD*, pages 365–378. ACM, 2008.

[2] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.

[3] S. Chu, E. Keogh, et al. Iterative deepening dynamic time warping for time series. In *SDM*, pages 195–212. SIAM, 2002.

[4] C. Faloutsos, R. M, et al. *Fast subsequence matching in time-series databases*. ACM, 1994.

[5] R. Neamtu et al. Interactive time series exploration powered by the marriage of similarity distances. *VLDB Endowment*, 10(3):169–180, 2016.

[6] T. Rakthanmanon et al. Searching and mining trillions of time series subsequences under dynamic time warping. In *ACM SIGKDD*, pages 262–270. ACM, 2012.

[7] Y. Sakurai et al. Stream monitoring under the time warping distance. In *ICDE*, pages 1046–1055. IEEE, 2007.