KD-Box: Line-segment-based KD-tree for Interactive Exploration of Large-scale Time-Series Data

– Supplementary Material –

1 Comparative Evaluation in Real Datasets

In this section, we provide detailed information of some real world data sets in the comparative evaluation. This includes the source of the data and the corresponding benchmark.

Weather. In this case, we explore an environmental sensor dataset obtained from the Applied Climate Information System (ACIS)Web Services 4, which records the daily temperature of 6187 public weather stations in the United States in 2019. The range of temperature is -35°F to 120°F.
https://www.rcc-acis.org/docs_webservices.html

Table 1: Benchmark of weather data

Method	Recall (%)	Precision (%)	RNL time (ms)	Timebox time (ms)	Angular time (ms)	Build time (s)
Ours	97.1	99.8	9.128	5.321	3.112	1.412
Sequential Search	100	100	27.317	25.842	61.432	0
KD-tree	72.3	100	0.051	1.428	0.393	5.313
R-tree	100	73.7	1.212	5.729	0.132	1.827

Stock. This data shows the price of 4440 stocks from year 2005 to 2018, which price range from 0 to 100 dollars.
https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs

Table 2: Benchmark of stock data

Method	Recall (%)	Precision (%)	RNL time (ms)	Timebox time (ms)	Angular time (ms)	Build time (s)
Ours	97.0	99.9	9.017	3.521	1.937	1.231
Sequential Search	100	100	15.732	17.217	30.328	0
KD-tree	72.0	100	0.023	1.183	0.438	3.012
R-tree	100	74.9	0.821	3.312	0.083	1.032

Hard Drives. This data shows the temperature (SMART 194) of hard drives in Backblaze since 2013. Backblaze is a cloud storage provider, who publishes detailed statstics about the hard drives in their data centers every quarter of the year. The dataset is sampled to 92K time series to make it loadable for browser. The range of temperature is 10°C to 58°C.
https://www.backblaze.com/b2/hard-drive-test-data.html

Table 3: Benchmark of hard drives data

Method	Recall (%)	Precision (%)	RNL time (ms)	Timebox time (ms)	Angular time (ms)	Build time (s)
Ours	98.4	99.9	225.428	87.437	62.432	21.4
Sequential Search	100	100	462.318	404.128	627.127	0
KD-tree	83.2	100	26.127	17.328	1.239	263.9
R-tree	100	89.7	31.238	129.483	0.283	36.2

Beat Basics. This data comes from the paper "Capture & Analysis of Active Reading Behaviors for Interactive Articles on the Web"[1]. The dataset has 4430 time series, each representing the reading behavior of a user in a session and recording their scroll position from opening to closing the article.
https://s3-us-west-2.amazonaws.com/interactive-analytics-eurovis/data.zip

Web Traffic. This is the web traffic data of Wikipedia in 2016. The dataset is sampled to 100K time series to make it loadable for browser. The range of traffic flow is 0 to 67M.
https://www.kaggle.com/c/web-traffic-time-series-forecasting/data?select=train_2.csv.zip

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

Table 4: Benchmark of beat basics data

Method	Recall (%)	Precision (%)	RNL time (ms)	Timebox time (ms)	Angular time (ms)	Build time (s)
Ours	97.4	99.7	10.100	3.834	2.269	1.149
Sequential Search	100	100	17.243	18.915	33.329	0
KD-tree	68.3	100	0.058	1.204	0.636	2.235
R-tree	100	84.2	0.934	3.546	0.130	1.052

Table 5: Benchmark of web traffic data

Method	Recall (%)	Precision (%)	RNL time (ms)	Timebox time (ms)	Angular time (ms)	Build time (s)
Ours	99.3	99.9	248.028	96.157	68.775	32.4
Sequential Search	100	100	508.537	444.658	739.289	0
KD-tree	87.3	100	28.838	19.083	1.524	293.1
R-tree	100	92.1	34.372	142.402	0.331	39.5

ECG. This data comes from the paper "Components of a New Research Resource for Complex Physiologic Signals"[2]. The dataset has 5000 time series, each representing the electrocardiogram (ECG) signal of a person. An ECG is a simple test that can be used to check ones heart's rhythm and electrical activity.
`https://timeseriesclassification.com/description.php?Dataset=ECG5000`

Table 6: Benchmark of ECG data

Method	Recall (%)	Precision (%)	RNL time (ms)	Timebox time (ms)	Angular time (ms)	Build time (s)
Ours	97.1	99.6	11.230	4.827	2.535	1.303
Sequential Search	100	100	18.960	20.745	36.781	0
KD-tree	73.2	100	0.167	1.328	0.682	2.520
R-tree	100	79.6	1.018	3.901	0.268	1.144

Electricity. This includes 16K one day electricity consumption records in U.K. The intention of this dataset was to collect behavioural data about how consumers use electricity within the home to help reduce the UK's carbon footprint.
`https://timeseriesclassification.com/description.php?Dataset=ElectricDevices`

Table 7: Benchmark of electricity data

Method	Recall (%)	Precision (%)	RNL time (ms)	Timebox time (ms)	Angular time (ms)	Build time (s)
Ours	90.4	99.9	20.258	7.669	4.509	3.913
Sequential Search	100	100	34.231	37.396	66.208	0
KD-tree	78.2	100	0.430	2.387	1.302	23.84
R-tree	100	89.8	1.830	7.003	0.664	6.829

REFERENCES

[1] M. Conlen, A. Kale, and J. Heer. Capture & Analysis of Active Reading Behaviors for Interactive Articles on the Web. *Computer Graphics Forum*, 38:687–698, June 2019. doi: 10.1111/cgf.13720

[2] A. Goldberger, L. Amaral, L. Glass, S. Havlin, J. Hausdorg, P. Ivanov, R. Mark, J. Mietus, G. Moody, C.-K. Peng, H. Stanley, and P. Physiobank. Components of a new research resource for complex physiologic signals. *PhysioNet*, 101, 01 2000.