M4: A Visualization-Oriented Time Series Data Aggregation

Uwe Jugel, Zbigniew Jerzak,
Gregor Hackenbroich
SAP AG
Chemnitzer Str. 48, 01187 Dresden, Germany

Volker Markl
Technische Universitat Berlin
Strae des 17. Juni 135
10623 Berlin, Germany
volker.markl@tu-berlin.de

I. Introduction

Visualization of large scale time series data is a crucial need of modern exploratory bigdata analysis [1]. But the huge size of the data is a barrier to visualization [2], [3], [4]. To address this challenge of bigdata different data reduction and sampling strategies are used to overcome the barrier [5], [6]. But for preserving the semantics of trend line of time series data these sampling strategies show huge limitations [7].

In this review paper we present a review of the paper [7] which address this issue of preserving the semantic of time series data and present some related works in the line. The paper appeared in the Proceedings of the VLDB Endowment, 2014.

The authors present M4, an aggregation based time series data reduction strategy that guarantees error free visualization of time series data as line chart as well as higher rate of data reduction. The approach is generic to any visualization system as long as the visualization systems uses RDBMS as data source.

II. Controbutions of the paper

The authors of the paper rewrite visualization queries Q using data reduction operator M_R such that the visualization of the original data from query Q and the visualization from the query $Q_R = M_R(Q)$ are similar and error free. As

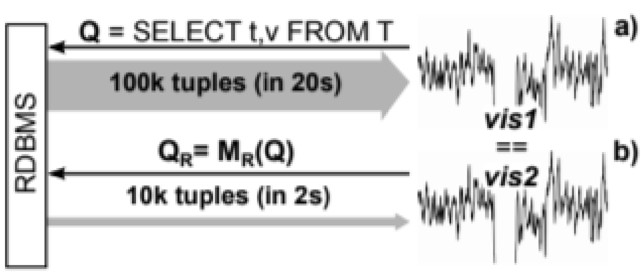

Figure 1. Time series visualization: a) based on original data; b) Using data reduction operator;

shown in Figure 1 Q_R produced the same visualization as Q with almost 10 times less tuples and 10 times reduced time. The main contributions of the paper are following:

- Proposed a visualization driven query rewriting technique relying on relational operators and parameterized with width and height of the desired visualization
- Focusing on the detailed semantics of the line charts, they propose a visualization driven aggregation strategy that only select necessary points needed for visualization. For visualization, in every time interval which corresponds to a pixel column in the visualization they select four tuples. The starting tuple, ending tuple, max tuple and the min tuple.

A. Query Rewriting

Most queries for time series visualization are of the form **SELECT time, value FROM SERIES WHERE** $time > t_1$ **AND** $time < t_2$. In addition to the query the visualization parameters like width and height are also passed for query rewriting. The rewritten query Q_R contains the following subqueries:

1) Original Query Q
2) A cardinal query Q_C on Q
3) a cardinality check (conditional execution)
4) to either use the result of Q directly or to execute an additional reduction Q_D on Q.

$M4$ system composes all those subqueries into single SQL query to avoid bandwidth consumption.

B. M4 Aggregation

M4 is a value preserving aggregation strategy for time series data. It divides the entire time series dataset into w equal groups and thus each pixel column in the visualization takes only one group. For each group M4 select the aggregates $min(v)$, $max(v)$, $min(t)$,$max(t)$ and that is why it is called M4 aggregation and then it joins the aggregated data to the time series and add missing timestamps t_{bottom},t_{top} and missing values v_{first}, v_{last}. In Figure 2 an example M4 query and the corresponding visualization is shown.

Complexity of M4: The grouping and computation of aggregated values can be done in $O(n)$ time where n is the number of tuples in the original query Q. Then the subsequent joining of the $4.w$ aggregated tuples with Q requires $O(n + 4.w)$ using hash join.

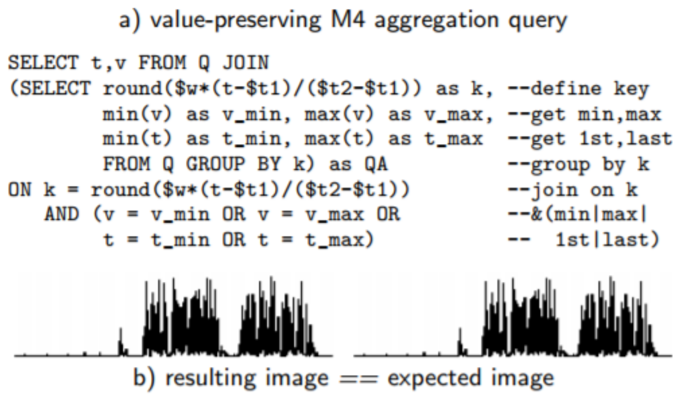

Figure 2. M4 query and visualization

C. *Time Series Data Reduction*

III. Evalaution

IV. Related Works

TODO

V. Our Proposal

TODO

VI. Conclusion

The conclusion goes here. this is more of the conclusion

References

[1] T.-c. Fu, "A review on time series data mining," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 1, pp. 164–181, 2011.

[2] A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with big data," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 2032–2033, 2012.

[3] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *National science review*, vol. 1, no. 2, pp. 293–314, 2014.

[4] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Information Sciences*, vol. 275, pp. 314–347, 2014.

[5] G. Cormode and N. Duffield, "Sampling for big data," 2014.

[6] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[7] U. Jugel, Z. Jerzak, G. Hackenbroich, and V. Markl, "M4: a visualization-oriented time series data aggregation," *Proceedings of the VLDB Endowment*, vol. 7, no. 10, pp. 797–808, 2014.