International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Visual Analysis using Modified Ramer-Douglas–Peucker Algorithm on Time Series Data

Saksham Lakhera[1], Praveena T[2]

[1]*Department of Computer Science and Engineering, R.V. College of Engineering, Bengaluru, Karnataka, India*
[2]*Department of Computer Science and Engineering, R.V. College of Engineering, Bengaluru, Karnataka, India*

Abstract - *Analytical engines are used to conduct analytics. Data is required to identify patterns in current trends. When this data is stored at any other location then it becomes time consuming to retrieve such a large amount of data. It is more complicated when the user needs to do a real-time visual analysis of the data, which is when they want to visually track the data. This research paper aims to establish a solution for handling these data for real-time visual analysis using the modified Ramer-Douglas – Peucker algorithm (RDP Algorithm) without forfeiting the statistical value of data such as mean and standard deviation*

***Key Words*: RDP Algorithm, indexing, visual data analysis, Ramer-Douglas – Peucker algorithm, time series data.**

1. INTRODUCTION

Data analysis has become crucial for the discovery of patterns in this modern-day environment. As computation is usually done far from data, the user gets data from the source and this data is large enough to have thousands to millions of data points. Our objective is to enable a fast-visual analysis of this data by compressing it without losing its statistical significance, such as the mean and standard deviation [8]. This paper proposes a method for sending data to the user more easily for analysis. The data is a continuous time series. Data is collected at a particular time period and displayed in the form of a graph on the user's device [1]. It generates thousands to millions of points, but even after plotting a few points on the screen, the new graph and the original graph are visually indistinguishable [10].

This paper exploits this principle by reducing the number of points depending on the size of the data without losing its mathematical significance. A variety of algorithms have been developed for compressing data points. These algorithms attempt to reduce the size of the curve while preserving the accuracy [4],[5]. Few of these many compression algorithms are Douglas-peucker algorithm, Bellman's Algorithm, STTrace Algorithm, Opening Window Algorithms [3], [7]. According to [4], [6] the Douglas-Peucker (DP) algorithm is considered to be one of the most reliable and efficient methods of compressing curve by the number of data points while maintaining only significant positions but this algorithm is unable to preserve the statistical quality of the data and the fixed value of epsilon is given for compression. Due to this fact, the RDP algorithm is not dynamic in nature. To avoid this, a modified approach of Ramer-Douglas-Peucker algorithm is taken into consideration that takes into account the size of the window and number of points within the given range. The proposed method of visual analysis is as follows and the same is depicted in Fig. 1.

1. The user asks for data within a time interval.
2. This request is sent to the data source.
3. Data source calculates the difference between the time interval and calculate number of data points in between the time interval.
4. Data is then extracted and compressed, keeping in mind the number of points between the intervals.
5. Data is sent back to the user.
6.

Advantages of using modified RDP for visual analysis are as follows:

1. Use less Bandwidth
2. Visually indistinguishable from original graph
3. Maintains mathematical importance of data points such as mean, standard deviation
4. Easier to process data as number of datapoints are less
5. Cost Effective solution
6. Real time analysis

2. MODIFIED RDP ALGORITHM

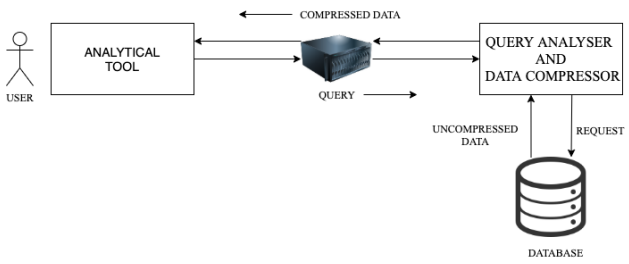


Fig-1: Architecture of the system

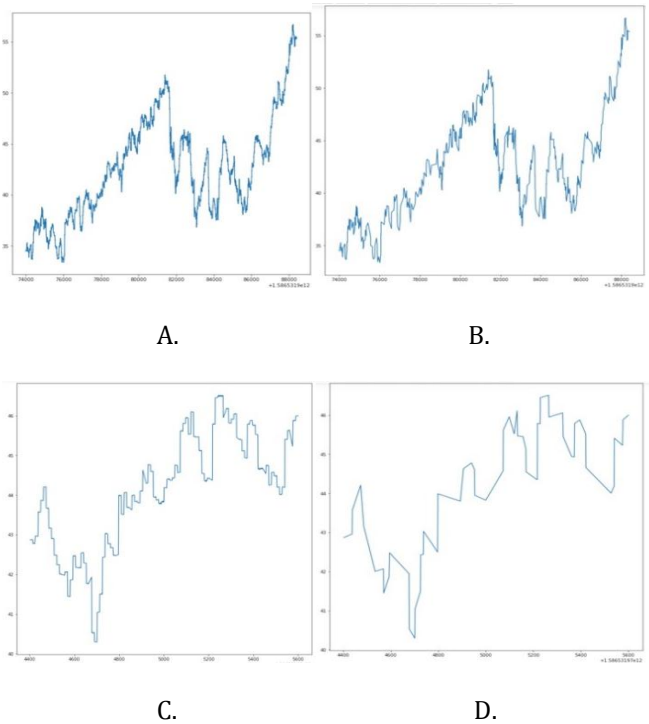
As depicted in Fig.1, The user sends a query which includes a time range to the data source, this data may be old data or real-time data. In this time range there may be thousands to millions of data and sending all these points to the user may

not be the best choice since even a smaller subset of data points generates almost identical graphs.

The Ramer-Douglas-Peucker algorithm is an algorithm that can drastically reduce points to 99 percent without affecting the graph's visual appearance [5], [9]. It handles the outlier points very well, it includes outlier in the dataset as it is the point which greatly contributes to the forming of the graph. [2], [9] The Ramer-Douglas – Peucker algorithm is an algorithm designed to reduce the number of points in a curve approximated by a sequence of dots. It does so by "thinking" of a line in a series of points forming the curve between the first and the last point. It checks which point is farthest from the line in between. If the point is closer than a given 'epsilon' distance, it eliminates all of these intermediate points. On the other hand, if this 'outlier point' is more distant than epsilon from our imaginary line, the curve is split into two parts:

1. From the first point up to and including the outlier.
2. The outlier and the remaining points.

On both the resulting curves, the function is recursively called up, and the two reduced curve forms are put back together.



A. B.

C. D.

Fig-2: A) Dataset with 14400 points. **B)** 14400 data points reduced to 504 data points using RDP. **C)** Dataset with 1200 points. **D)** 1200 data points reduced to 66 data point using RDP.

In Fig.2, Graph 'A' represents a dataset with 14400 points, and when the RDP algorithm is applied to it, Graph 'B' is generated with 504 points, these two graphs are visually identical, but when the dataset is less as shown in graph 'C' with only 1200 points, the graph produced by the RDP algorithm is very vague and looks completely different from the original graph.

The justification for the RDP algorithm to dramatically decrease the points in graph 'C' is that the value of epsilon is constant in both cases, due to which the high value of epsilon resulted in high reduction of data points and Such a dramatic reduction is not necessary in graph 'C' and thus a lower value of epsilon is needed for this graph [1].

To avoid the problem of specifying different values of the epsilon for different graphs, a modified approach of the Ramer-Douglas – Peucker algorithm is taken into account in compressing dataset.

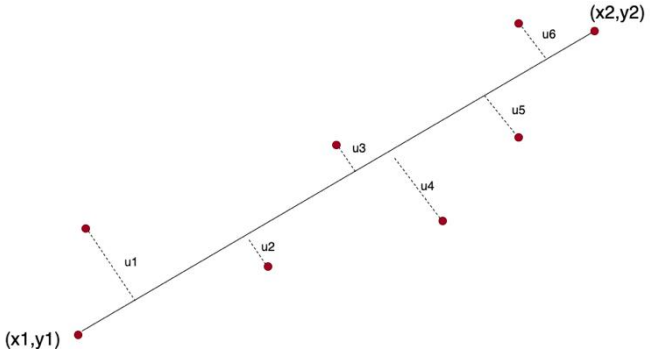
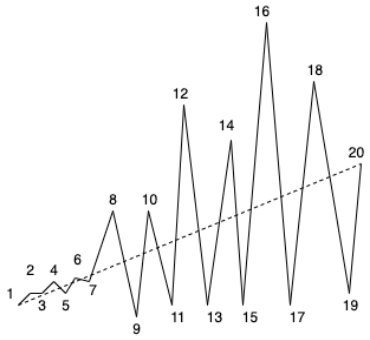


Fig-3: perpendicular distance of all the points on the line formed by start and end point of the window. Epsilon = (ΣU_i) * Time interval/ (x_2-x_1), number of points = (x_2-x_1)/ Time interval

The epsilon value of the modified Ramer-Douglas-Peucker algorithm is determined automatically making it dynamic in nature. For this method, a moving window is taken over the entire dataset and the algorithm is applied over the window, the entire dataset is split into subsets by the window and RDP is applied over it. For each window, the value of the epsilon is calculated by taking the mean of the perpendicular distance of all points along the line created by the start and end data point of the window as shown in Fig.3, due to which data points are not taken into subsets with very little variance.

The reason to accept epsilon's value as an average perpendicular distance is that it informs us how many points have deviated from the line. When the variance is big enough, the average will be high and the data that are less deviated will not be included in the data set but when the variance is even lesser than that, it is expensive to delete less deviated points because they play a significant role in defining the curve. An example can be used to explain why this approach works.



Points	1	2	3	4	5	6	7	8	9	10
Dist.	0	0.05	0.01	0.07	0.05	0.05	0.08	1.2	1.1	1

Points	11	12	13	14	15	16	17	18	19	20
Dist.	1.3	2.4	1.1	1.0	1.2	31.	2.2	2.1	1.8	0

Fig-4: A high variance graph with 20 points and perpendicular distance of all the points from line joining point 1 and point 20

Fig.4 represent a high variance graph with 20 points and their perpendicular distance from the virtual line joining the 1st and 20th point. According to the modified RDP algorithm, points with perpendicular distance lesser than epsilon are removed. In Fig.3 the value of epsilon is measured as 0.97. The end points and the points excluding 2, 3, 4, 5, 6 are included in the dataset. Point 7 is also lesser than the epsilon value, but it is kept because it is the outlier with respect to point 8 as the graph indicates a sudden jump.

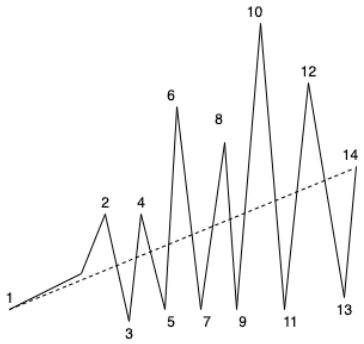


Fig-5: 20 points graph reduced to 14 points graph using RDP

The RDP algorithm preserves the outlier points.since these are the points that determine the shape of the graph. Point 7 is outlier with regard to point 8 and point 8 is outlier with regard to point 7, thus both points are retained. Points 2,3,4,5,6 which have not contributed much to the dataset are removed and a reduced graph is generated. The Fig.5 represents the Reduced graph.

The size of the window plays an important role in determining the compressed size of the data. The more the size of the window, the more the compression and lesser the number of windows, the greater the compression. For datasets that do not show high fluctuating data, the number of windows can vary from 60 to 70 for best performance. Here, in this project we are dealing with data that does not display a strong fluctuation, so the window size of 60-70 is ideally suited to the most reliable result.

Time	open	high	low	close	volume	Name
1586531974344	34.39	34.66	34.29	34.41	10237828	ABT
1586531974345	34.42	34.49	34.24	34.26	7928236	ABT
1586531974346	34.27	34.5	34.21	34.3	7070536	ABT
1586531974347	34.29	34.58	34.25	34.46	6688100	ABT
1586531974348	34.28	34.75	34.28	34.7	8561425	ABT
1586531974349	34.84	35.1	34.775	35.08	10016820	ABT
1586531974350	35.18	35.29	34.75	34.82	9584941	ABT

Fig-6: ABT Stock Dataset

ABT stock dataset is used in this paper to evaluate data points separated by a time interval of 1 millisecond as shown in Fig 6. The estimated number of compressed data points of around 360 to 540 is sufficient enough to make the graph visually indistinguishable from the original graph. This can be seen mathematically as follows, if the user uses a 1080pixel display and every point is represented by 2 to 3 pixels, because it is difficult to see a single pixel, there would be no visual difference if 1 data point is used per 2-3 pixel, the number of points would range from 360 (1080/3) to 540 (1080/2).

3. RESULT AND DISCUSSION

The key factor in the RDP algorithm is the value of epsilon as it specifies the number of points in the compressed dataset [2], [3], [4]. After performing modified RDP on ABT stock datasets of different sizes, it was found that 60-70 numbers of windows offer the best result for datasets with less fluctuation.

As shown in Table 1, dataset of 11331 points takes about 143.88 KB of space and 451 points data takes about 5.69 KB of space. As all data points are reduced between 300 and 600 due to which the size of the reduced data points is always around 4 KB to 7 KB range, which is far lesser than the 145kB of space taken by 11331, Reduced data points can be sent very easily across the network to the user.

Table -1: Reduced data points using modified RDP

Number of Data points	630	1260	2518	6295	8813	11331
Reduced data Points	319	381	357	568	541	451
Original size on machine (KB)	7.78	15.91	31.97	79.93	111.91	143.88
Reduced size on machine (KB)	4.02	4.81	4.50	7.17	6.83	5.69
Reduction % of data points	49.36	69.76	85.82	90.96	93.86	96.01

Keeping the number of the window as 63 for the experiment, the TABLE 1 reveals that as the number of data points increases, the reduced data points remain approximately the same, which is true for datasets that do not fluctuate. For the data set with 11331 points, the reduced data points are only 451 points, which is lesser than 8813 data points, with 541 reduced points, this concludes that size of reduced dataset is not increasing along when there is increase in size of original dataset. When this technique is applied on dataset with less fluctuation a reduced data set is produced whose size is between 300 and 600 even for large dataset.

It is evident from TABLE 1, as the number of data points is rising with the reduction rate, it is difficult for larger datasets to maintain statistical value because they lose a lot of data, but even in dataset with 11331 points, there was not much difference between the mean of reduced dataset and the mean of the original dataset. The key explanation is that the algorithm just eliminates points in the sequence that does not add a lot to the entire structure of the curve. As a result, the points that lead to its structure and its mathematical significance are preserved in reduced dataset.

Case I: Fig.7 represents the ABT stock dataset with 2518 points separated by a time span of 1 millisecond and Fig.8 represents the ABT stock dataset, which is reduced to 357 point, i.e. 85.82 percent reduction in dataset. The graph in Fig.3 and Fig.4 is visually and mathematically the same. The mean of data points in Fig.7 is 45.64 and in Fig.8 is 45.75. Taking into account how dramatically the data points have been reduced this small difference is acceptable. The standard deviation has been reduced from 4.81 in the original to 4.67 in the reduced dataset, indicating that the curve is much smoother than the original curve.

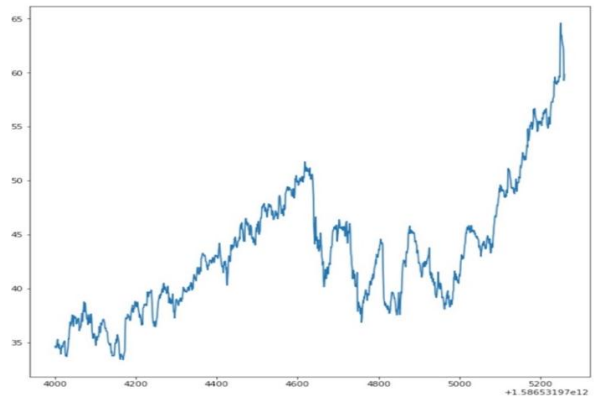


Fig-7: Original graph with 2518 points

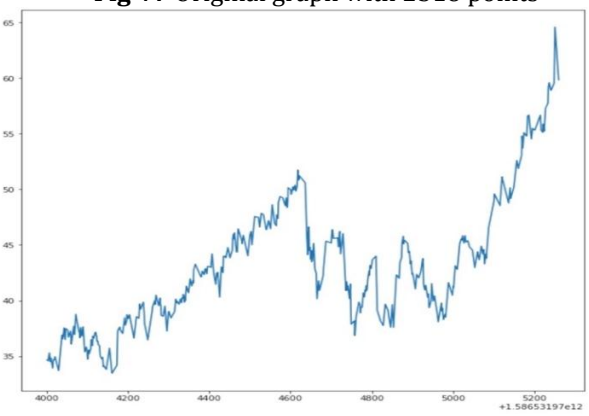


Fig-8: Reduced data points with 357 points

Case II: Fig.9 represents a portion of the dataset with 6295 data points and after reduction the data point is reduced to just 568 points, the graph of which can be seen in Fig.10. The mean of the original dataset is 42.90 and that of the reduced dataset is 42.81 and there is a decrease in the standard deviation from 3.671 to 3.66, suggesting that the curve is smoother than the original dataset.

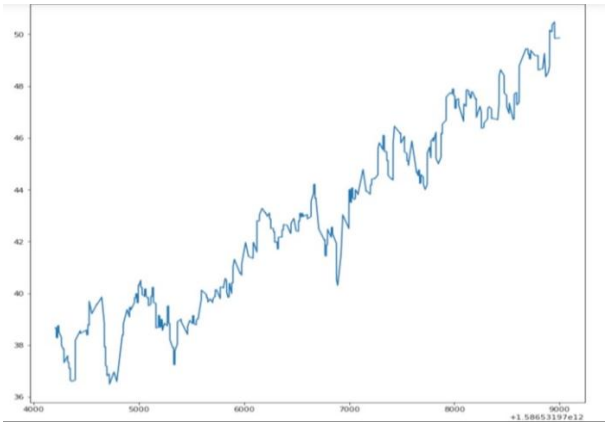


Fig-9: Original graph with 6295 point

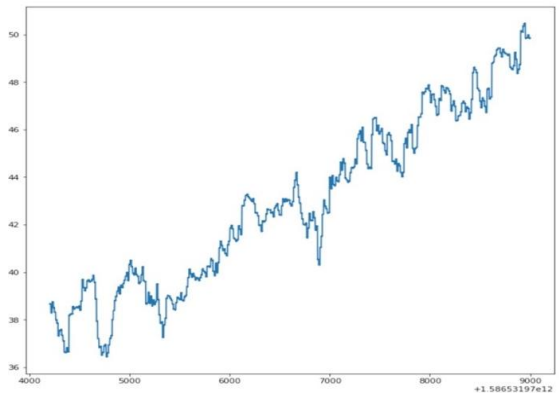


Fig-10: Reduced data points with 568 points

The structure of graph remains the same, only difference between original and reduced graph in both the case is that the reduced graph form is much lighter and smoother as compared to original graph. As in original graph, multiple points are represented by a pixel which is why it looks thicker. From the above two cases we can say that the statistical value of the data points is maintained in reduced dataset.

4. CONCLUSION

The modified RDP algorithm is very useful for data with less fluctuation and it is able to reduce data to a range of 300 to 600 points without losing its structural significance and mathematical importance. There is no need for us to determine the value of epsilon in this algorithm as it is determined by the algorithm itself and due to reduced dataset, it takes less bandwidth and can be sent to user with very high speed. When this approach is used for highly fluctuated data, the number of points is usually higher than what we would ideally like to achieve, because the main goal of the paper is to retain statistical significance hence most of the points are included in the data set. For faster processing, caching and indexing techniques can be applied to the database so that it does not waste more time compressing the data

REFERENCES

[1] Systems, K., 2020. Dynamically Shrinking Big Data Using Time-Series Database Kdb+ – White Papers – Q And Kdb+ Documentation. [online] Code.kx.com. Available at: <https://code.kx.com/q/wp/ts-shrink/> [Accessed 14 April 2020].

[2] Ehret, U. and Neuper, M., 2020. Applying The Ramer-Douglas-PeuckerAlgorithm To Compress And Characterize Time-Series And Spatial Fields Of Precipitation. [online] NASA/ADS. Available at: <https://ui.adsabs.harvard.edu/abs/2014EGUGA..1613537E/abstract> [Accessed 15 April 2020].

[3] F. Goz, A. Mutlu and O. Akbulut, "Analysis of Ramer-Douglas-Peucker algorithm as a discretization method," 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1-4.

[4] M. Mozumdar, A. Shahbazian and N. Ton, "A big data correlation orchestrator for Internet of Things," 2014 IEEE World Forum on Internet of Things (WF-IoT), Seoul, 2014, pp. 304-308.

[5] Muckell, J. and Hwang, J., 2010. Algorithms For Compressing GPS Trajectory Data | Proceedings Of The 18Th SIGSPATIAL International Conference On Advances In Geographic Information Systems. [online] Dl.acm.org. Available at: <https://dl.acm.org/doi/abs/10.1145/1869790.1869847> [Accessed 14 April 2020].

[6] L. Li and W. Jiang, "An improved Douglas-Peucker algorithm for fast curve approximation," 2010 3rd International Congress on Image and Signal Processing, Yantai, 2010, pp. 1797-1802.

[7] H. Ratschek, J. Rokne & M. Leriger (2001) . Robustness In GIS Algorithm Implementation With Application To Line Simplification. [online] Taylor & Francis. Available at: <https://www.tandfonline.com/doi/abs/10.1080/13658810110053107>

[8] McMaster, R. B. (1986). A statistical analysis of mathematical measures for linear simplification. The American Cartographer 13(2), 113-116.

[9] Douglas, D. and Peucker, T, 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. The International Journal for Geographic Information and Geovisualization, 10(2), pp.112-122.

[10] Ramer, U. (1972). An iterative procedure for the polygonal approximation of plane curves. Computer Graphics and Image Processing 1(3), 244-256.