Nested Intellectual Data Grouping and Clusterization for the Interactive Visual Explorer

Maria Grigorieva[1,3,⋆], *Mikhail* Titov[1,3], *Timofei* Galkin[2], and *Igal* Milman[2]

[1] *Lomonosov Moscow State University, Moscow, Russia*
[2] *National Research Nuclear University "MEPhI", Moscow, Russia*
[3] *Plekhanov Russian University of Economics, Moscow, Russia*

Abstract. The Interactive Visual Explorer (InVEx) application is designed as a visual analytics tool for Big Data analysis. Visual analytics is an integral approach to data analysis, combining methods of intellectual data analysis with advanced interactive visualization. One of the main objectives of InVEx is to process large data samples by decreasing their level of detail (LoD). The proposed approach includes clustering as well as flexible grouping by different parameters, providing the exploration of data from the lowest to the highest level of details. The results of grouping and clusterization are visualized using interactive 3D scene and parallel coordinates, allowing the user to gain insight into data, to explore hidden correlations and trends of parameters.

1 Introduction

The Interactive Visual Explorer (InVEx) is developed as a generic interactive visual analytics tool for the analysis and exploration of big volumes of multidimensional data [1]. InVEx is based on the combined usage of intellectual data analysis methods and advanced interactive visualization techniques. It should be noted that data analysis and data exploration are not the same. In data analysis, the user knows in advance what he/she is looking for, and in data exploration, the user does not. Data analysis implies deep understanding of the structure of the data, while data exploration is aimed at uncovering the general structure of the data. This paper is mostly focused on data exploration issues related to big data volumes.

At the first stage of the exploration of huge data samples, an analyst is usually interested in searching for groups of similar objects, representing the aggregated overview of data. The aggregated data objects can be visualized and explored together to determine some specific considerations about the structure of the initial data. The Level-of-Detail (LoD) method was implemented specially for this task. It allows grouping or clustering the initial data and representing groups as aggregated objects for further visualization as spheres on a 3D scene. Thereby, the application provides the ability of nested groupings, so aggregated data objects created from the initial data sample (or the previous step of grouping) could be grouped again while giving the possibility to explore objects within the group. If an analyst finds an interesting object, the next stage of the exploration is to dig into this object, which, however, may be also huge. Then, the LoD method may be applied to this object again. This process

[⋆] e-mail: maria@srcc.msu.ru

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

is repeated as many times as needed until the group of data objects is small enough to be visualized and explored.

2 HEP computing metadata as a test ground for InVEx

High Energy Physics (HEP) computing metadata [2] is one of the most representative examples of high-dimensional large data volumes, having all statistical types of features (continuous, nominal, ordinal, range). HEP experiments have a unique global computing infrastructure, complex data management and processing systems that handle exabytes of data, execute billions of physics analysis and processing jobs. In InVEx we use jobs metadata: each record represents a single computing job having up to 200 different parameters of different types, including type of jobs, execution time, utilized resources, volumes of processed data, error logs, and many others. One of the crucial tasks of the distributed computing is to ensure the stability and efficiency of the execution of the jobs, which implement massive data processing in a diverse distributed environment. This requires a lot of operational efforts. However, these efforts can be reduced by applying intelligent methods of data analysis. The HEP computing metadata have been accumulated for more than a decade. The knowledge of the functioning of the distributed infrastructure under different circumstances for such a long period may help analyse and forecast its future behavior. But most of the intelligent methods are treated as "black boxes" and must be validated and interpreted by a human. That's where visual analytics may help by combining intellectual data analysis and interactive visualization.

3 InVEx basics

InVEx was designed based on open source web technologies. All heavy data processing and analysis algorithms are implemented on a server, and interactive visualization is achieved in a web-application that the user works with. The server-side of the application provides integration with external sources of data for real-time data analysis and exploration. The intellectual data analysis, clusterization and dimensionality reduction are implied. Currently, InVEx is focused on clustering and data grouping methods as one of the most important stages of real-time data exploration. It provides highly interactive GUI to facilitate the interpretation of the results of intellectual data analysis. It uses methods of scientific visualization such as 3D visualization and parallel coordinates [3]. And finally, InVEx provides the ability to explore data at different levels of detail.

4 The Level-of-Detail generator with nested grouping and clusterization

The LoD Generator for InVEx provided by our group allows to choose the degree of interest into data analysis and exploration. Its structure is presented in Figure 1. The starting point is a data sample uploaded to the server. If the initial data sample is small enough, it can be visualized and explored immediately in the web-application. Otherwise, if the data sample is large (has more than 10K objects), it may be problematic to visualize all that data objects in the browser. However, we assume that the first stage of data exploration usually requires just an overview of the initial data to uncover its general structure. The LoD Generator module provides several options to split the initial large data sample into groups or clusters for further analysis using visual and intelligent methods: K-Means clusterization (MiniBatch K-Means method from python's Scikit-learn library), grouping of data by categorical or by continuous numerical features. In the case of K-Means clusterization, the user may choose the level of detail such as the number of clusters, and a set of features used for cluster definition. The initial data sample will be split into the specified number of clusters, and values of all features in clusters will be aggregated. In the case of grouping by categorical feature(s), the initial data sample is split into some groups by following the number of categories. A third option is to group data by a numerical continuous feature. Then, its value range (maximum – minimum) is split into a specified number of intervals. After that, all data are grouped by these intervals. As a result, the LoD Generator returns an aggregated data sample with generated groups of data, and saves these data into the internal storage.

Thus, instead of visualizing all data objects from a large data sample, InVEx visualizes only an aggregated overview allowing to represent the general structure of the initial data. Interactive visualization allows to select groups of interest and investigate all data objects belonging to these groups separately, but in the same way. All interim stages of data exploration are stored at the back-end, providing the navigation through these stages and data exploration at different levels of detail.

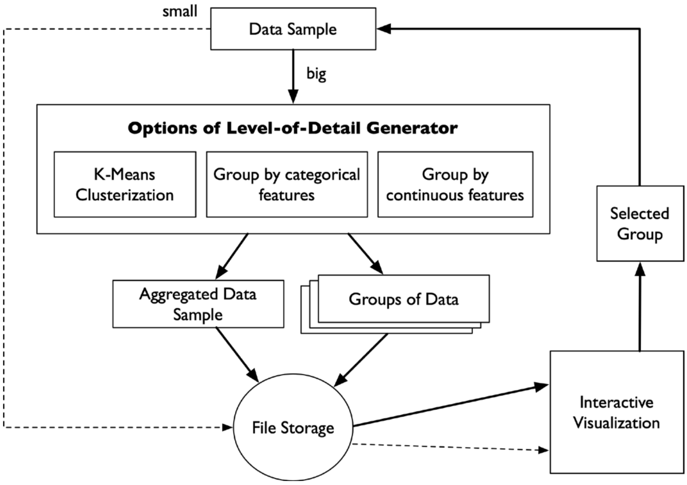
small
Data Sample
big
Options of Level-of-Detail Generator
K-Means Clusterization
Group by categorical features
Group by continuous features
Aggregated Data Sample
Groups of Data
File Storage
Interactive Visualization
Selected Group

Figure 1. Structure of the Level-of-Detail Generator

5 An instance of InVEx data exploration

The current implementation of InVEx was tested on a random portion of HEP computing metadata. The initial data sample was a number of computing jobs (160K jobs). At the first stage of data exploration, the LoD Generator was activated and all data objects (computing jobs) were grouped by the categorical feature “computingSite”. As a result, 94 groups of data were obtained. Figure 2 represents the 3D visualization of these aggregated data objects in projections: ReadRate, WallTime, CPU Consumption. The radius of each sphere depends on the size of the corresponding group or cluster. The user may select a group of interest (on the left picture), and explore it separately. Then all data objects (jobs) belonging to the selected computing site are visualized in a new window. The right part of the figure shows the visualization of all jobs for different computing sites. The user can interactively change the projections. InVEx allows to apply various clustering algorithms (KMeans, DBSCAN, Hierarchical, K-Prototypes) to these data. The spheres are highlighted with different colors, representing the results of clusterization. Additionally, InVEx provides the interactive parallel coordinates to analyze trends of features for all clusters. The combination of these two visualization methods,

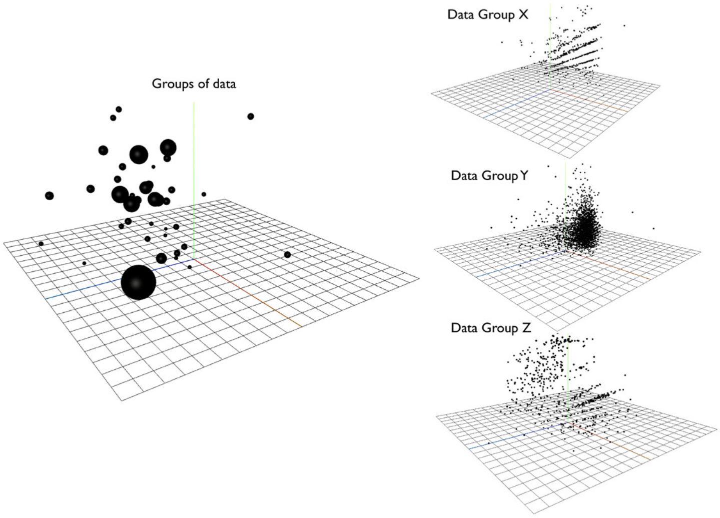

Figure 2. Aggregated data objects exploration with InVEx

3D visualization and parallel coordinates, allows to investigate the structure of data and to analyze correlations between features, search for anomalies, and make a comparative analysis.

6 Conclusion

InVEx is an actively developed visual analytics tool. The near-term plans are to implement more clustering algorithms, including clusterization of strings and mixed values (categorical and numerical). The interactive parallel coordinates are to be linked with tables to make the process of data exploration more efficient. The file storage is going to be replaced with a database to ensure the scalability. As InVEx is aimed at the exploration of Big Data, the next step is to integrate data processing algorithms into supercomputer infrastructures, which will allow to implement more complicated algorithms of intellectual data analysis at a reasonable time, providing the user the ability to substantially accelerate the visualization and interpretation of the results of intellectual data analysis.

Acknowledgements

We gratefully acknowledge the financial support from the Russian Science Foundation grant No.18-71-10003.

References

[1] T. Galkin, M. Grigoryeva, A. Klimentov, T. Korchuganova, I. Milman, V. Pilyugin, M. Titov, Scientific Visualization **10**, 5, 32–44 (2018)

[2] Albrecht et al. HEP Software Foundation, *A Roadmap for HEP Software and Computing R&D for the 2020s* (2017)

[3] https://www.d3-graph-gallery.com/parallel.html [online, visited November 4, 2019]