Big Data Research 36 (2024) 100445

Contents lists available at ScienceDirect

Big Data Research

journal homepage: www.elsevier.com/locate/bdr

Interactive big data visualization and analytics

One of the major challenges of the Big Data era is the need to support the analysis of a great amount and variety of massive datasets by non-corporate data analysts with little or no support and expertise on data processing such as research scientists, data journalists, policy makers, SMEs and individuals. These datasets are made accessible in a raw format (e.g., plain text, json) and are not being loaded or indexed in a database. Furthermore, they are dynamic, noisy, and heterogeneous in nature. The level of difficulty in transforming a data-curious user into someone who can access and analyze this data is even more burdensome. The purpose of visual data exploration is to facilitate information perception and manipulation, knowledge extraction and inference by non-expert users. The visualization techniques, used in a variety of modern systems, provide users with intuitive means to interactively explore the content of the data, identify interesting patterns, infer correlations and causalities, and support sense-making activities that are not always possible with traditional data traditional data analysis techniques.

In the Big Data era, several challenges arise in the field of data visualization and analytics. First, the modern exploration and visualization systems should offer scalable data management techniques in order to efficiently handle billion objects datasets, limiting the system response in a few milliseconds. Besides, nowadays systems must address the challenge of on-the-fly scalable visualizations over large and dynamic sets of volatile raw data, offering efficient interactive exploration techniques, as well as mechanisms for information abstraction, sampling and summarization for addressing problems related to visual information overplotting. Further, they must encourage user comprehension by offering customization capabilities to different user-defined exploration scenarios and preferences according to the analysis needs. Overall, the challenge is to enable users to gain value and insights out of the data as rapidly as possible, minimizing the role of IT-expert in the loop.

This special issue aimed to publish work on multidisciplinary research areas spanning from Data Management and Mining to Information Visualization and Human-Computer Interaction. In addition to the normal submissions, this special issue considered to invite some of the best papers from the 3rd International Workshop on Big Data Visual Exploration and Analytics (BigVis 2020), held in conjunction with the 23rd International Conference on Extending Database Technology & 23rd International Conference on Database Theory (EDBT/ICDT 2020).

In response to the call for papers, 22 submissions on different applications of visual data analysis techniques were received. From these, 9 articles were accepted after a two-stage review process supported by a reviewer board of internationally renowned experts in the field.

The article " OL-HeatMap: Effective Density Visualization of Multiple Overlapping Rectangles", by Niloy Eric Costa, Tilemachos Pechlivanoglou, and Manos Papagelis introduces a novel heat-map visualization, termed OL-HeatMap, designed for identifying and displaying the precise density of overlapping rectangles. The proposed approach not only reveals valuable insights, such as the actual position and size of the formed overlapping rectangle. The authors exploit a state-of-the-art computational geometry method based on the sweep-line paradigm. The adopted sweep-line algorithm enables the implementation of fast and exact density-based visualization Furthermore, the authors present two evaluation metrics that quantitatively assess the accuracy of grid-based overlap visualizations. An extensive evaluation, incorporating both synthetic and real datasets, showcases the accuracy and efficiency of the proposed methods.

In "ViewSeeker: An Interactive View Recommendation Framework", Xiaozhong Zhang, Xiaoyu Ge, Panos K. Chrysanthis, and Mohamed A. Sharaf explore the problem of context-aware view recommendation by introducing an Interactive View Recommendation framework named ViewSeeker. The framework engages in a human-in-the-loop approach, interacting with the user to identify the most suitable utility function for the current analysis context. ViewSeeker provides two forms of adaptation: utility function tuning, constituting its initial and optional phase, and utility function integration, constituting its subsequent phase. In the utility function tuning phase, ViewSeeker collaborates with the user to fine-tune the parameters of a utility function. This ensures that the parameters accurately capture the data analysis context. The utility function integration phase employs active-learning techniques to select informative example views for labeling and to predict the contribution of each utility function within a multi-objective utility function. Extensive experiments utilizing two real-world datasets have been conducted, demonstrating the effectiveness and efficiency of the adaptation methods.

In “ExplorerTree: A Focus+Context Exploration Approach for 2D Embeddings”, Wilson E. Marcílio-Jr, Danilo M. Eler, Fernando V. Paulovich, José F. Rodrigues-Jr, and Almir O. Artero Artero introduces a hierarchical exploration approach known as ExplorerTree. This methodology tackles visual scalability challenges encountered when presenting dimensionality reduction results through scatter plots. The approach adopts a hierarchical exploration paradigm enriched with Focus+context interaction. To mitigate visual clutter, the construction of the hierarchy incorporates a sampling selection algorithm and encoding strategies. The application of this approach in exploring images of convolutional filters, along with a user evaluation, illustrates the effectiveness of the proposed methodology.

In "Intelligent Narrative Summaries: From Indicative to Informative Summarization", Samira Ghodratnama, Amin Beheshti, Mehrdad Zakershahrak, and Fariborz Sobhanmanesh address introduce a generic

https://doi.org/10.1016/j.bdr.2024.100445

Available online 14 February 2024

hierarchical personalized summarization framework named Narrative Summaries, empowering users to express preferences, resulting in the generation of user-specific hierarchical summaries. The framework incorporates two models: a semi-structured summarization approach and a fully-structured summarization approach. In the semi-structured summarization, the basic unit of representation is a sentence. The model employs objective functions to cluster sentences coherently and in a logical order. Furthermore, an objective function is applied during the summarization phase to ensure that cluster summaries for a given hierarchy level are logically distinct and also fit within a user-defined budget size. In the fully-structured summarization model, the basic unit of representation is a concept. The model employs an objective function to hierarchically cluster concepts at different levels of detail. The proposed methods are assessed in comparison to state-of-the-art approaches using a news articles dataset. The results demonstrate that the generated summaries enhance users' comprehension of the topic.

In "Scaling the Growing Neural Gas for Visual Cluster Analysis", Elio Ventocilla, Rafael M. Martins, Fernando Paulovich, and Maria Riveiro present two methods that enhance growing neural gas algorithms within the context of visualizing clusters in large-scale, high-dimensional datasets. The first method tackles overplotting and clutter issues by avoiding connections that result in high-dimensional graphs. Consequently, the generated visualization becomes more accurate and meaningful. The second method addresses performance concerns related to the time required to generate the results. This method parallelizes the process by modeling and merging different parts of a dataset using the MapReduce model. The proposed methods are evaluated across nine datasets, utilizing various quantitative and qualitative metrics. The experiments reveal that the first method generates lower-dimensional graphs with reduced overplotting and clutter. The second method preserves visual quality while requiring less execution time.

In "Visual Exploration of Anomalies in Cyclic Time Series Data with Matrix and Glyph Representations", Josef Suschniggab, Belgin Mutluac, Georgios Koutroulisa, Vedran Sabolc, Stefan Thalmannd, and Tobias Schrecke present an interactive glyph-based visual analytics approach for anomaly detection in multivariate time series. This approach considers sensor data collected by durability tests performed in industrial settings (e.g. automotive sector) and displays the iterations of such tests as a collection of color-encoded cycle glyphs, which enable users to visually identify conspicuous data through a matrix representation for drill-down and further comparative analysis. The article provides results and discussion of a pair analytics evaluation, which has been conducted in collaboration with the target user group and preliminary results on a visual interactive labelling concept for anomaly classification.

In "Dependency Visualization in Data Stream Profiling", Bernardo Breve, Loredana Caruccio, Stefano Cirillo, Vincenzo Deufemia, and Giuseppe Polese provide methods for the automatic detection and visualization of functional dependencies (FDs) in streaming data and more specifically functional dependencies and their extensions, relaxed functional dependencies. Although many algorithms for discovering FDs from static datasets, the article contributes to methods that continuously monitor FDs that evolve in data streams. It presents a tool that allows users to visualize FDs, explore and compare results based on their types, and employ quantitative measures to monitor how discovery results evolve. A user study has been conducted to assess the effectiveness of the proposed visualization tool.

In "Graph Waves", James Abello and Daniel Nakhimovich address problems related to the visualization of very large and complex networks. They present efficient algorithmic mechanisms to partition very large graphs into subgraphs for generating intuitive and interactive visualizations of the meta-structure of complex networks. To achieve interactivity, a graph is processed into graph layers, called fixed points, which are further decomposed into visual representations called graph waves and fragments. This decomposition is used to create spanning views of fixed points in a graph. They provide illustrative examples on publicly available data sets including social, web, and citation networks.

In "An Analytic Graph Data Model and Query Language for Exploring the Evolution of Science", Ke Li, Hubert Naacke and Bernd Amann introduce a data model (called pivot topic graph) and a query language for the visualization and exploration of topic evolution networks representing the research progress in scientific document archives. The model is independent of a particular topic extraction and alignment method and proposes a set of semantic and structural metrics for characterizing and filtering meaningful topic evolution patterns. These metrics are particularly useful for the visualization and the exploration of large topic evolution graphs. They also present a first implementation of their model on top of Apache Spark and experimental results obtained from four real-world document archives.

In closing, we would like to thank all authors who have submitted their work to this special issue, and the authors of the accepted papers for their timely revisions. We sincerely thank the Editors-in-Chief of the Journal of Big Data Research, Themis Palpanas and Zhaohui Wu, for hosting this special issue and supporting us during all stages. Additionally, we are very grateful to all reviewers for their time and valuable constructive comments.

David Auber[a], Nikos Bikakis[b,*], Panos K. Chrysanthis[c], George Papastefanatos[d], Mohamed Sharaf[e]
[a] *University Bordeaux, France*
[b] *Hellenic Mediterranean University, Greece*
[c] *University of Pittsburgh, USA*
[d] *ATHENA Research Center, Greece*
[e] *United Arab Emirates University, UAE*

[*] Corresponding author.
E-mail address: bikakis@athenarc.gr (N. Bikakis).