BIGDATA: F: DKA: Usable Multiple-Scale Big-Data Analytics through Interactive Visualization

National Science Foundation under Grant No. 1447416; Project duration 9/2014 – 8/2018
PIs: Chris North, Leanna House, Scotland Leman, and Wu Feng (Virginia Tech)

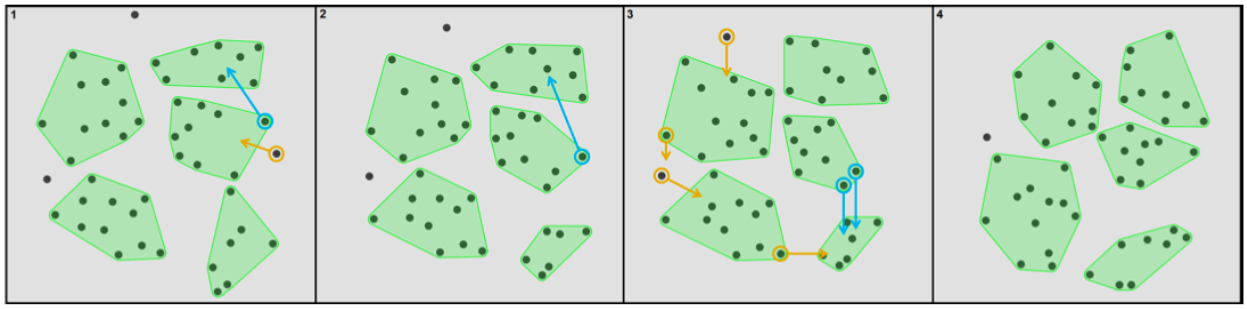

Figure 1: Example of combining Dimension Reduction and Clustering models from Table 1.

Project Challenges: Gaining big insight from big data requires big analytics, which poses big usability problems. Analyses of big data often rely on several computational and statistical models that operate on multiple levels of data scale to discover and characterize latent data structure. The models work jointly or in sequence to filter, group, summarize, and visualize big data so that analysts may assess the data. As a simple example in big text analytics, massive text is first sampled for relevant or representative words, then further reduced by topic modeling, then visualized by applying a dimension reduction algorithm. As the size of data increases, so does the number of models and, likewise, the need for human interaction in the analytical process. By interacting, humans inject expert judgment into the analytical process, and efficiently explore and make sense of big data from varying perspectives. However, because of complex low-level parameters and enforced premature formality, interacting with any individual model is difficult, and now, there is a need to interact with a growing number of models. In this proposal, current human-computer-interaction research is merged with complex statistical methods and fast computation to make big data analytics usable and accessible to professional and student users.

Project Goals: The proposed solution is to scale up Visual to Parametric Interaction (V2PI) to a new framework called Multi-scale V2PI (MV2PI). V2PI currently supports usable small-data analytics, and enables users to adjust model parameters by interacting directly with data in a visualization. That is, V2PI interprets visual interactions quantitatively to update parameters and produce new visualizations. MV2PI is a new interactive framework that links together multiple models operating at multiple levels of data scale in a unified interactive space. Model results are combined into a common visual representation. Directly manipulating the small-scale visual representation propagates to larger scale models by inverting the models to update their parameters, ultimately producing a new output result. In the text analytics example, if the user drags several data points together to hypothesize a cluster, the inverted dimensionality reduction model computes updated dimension weights, queries relevant new hits at the large scale, identifies changed topics, and updates the layout to show big-data support for the new cluster. This approach enables users to interactively explore large-scale data and complex inter-relationships between models in real time, and in a usable fashion that directly supports their natural cognitive sensemaking process.

Intellectual Merit: Intellectual merits are the fundamentally novel approach to interactively combining multiple statistical data models across levels of data scale to enable usable big-data analytics. This research will (1) create the conceptual MV2PI pipeline, and identify alternatives for communication flow between models, visualization, and interaction, including possible shared parameters; (2) establish several new useful models, covering different levels of scale, that support the V2PI model inversion approach to machine learning and can operate within the new pipeline; (3) develop new computational methods for

high-performance updates to inverted models in support of real-time interaction with MV2PI; and (4) evaluate the usability of MV2PI and measure its impact on human sensemaking in big data analytics.

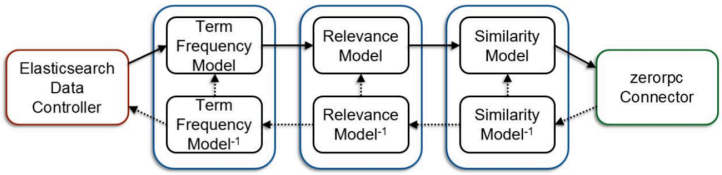

Figure 2: Example of software architecture for combining Dimension Reduction and Information Retrieval models from Table 1.

Broader Impacts: Broader impacts stem from bringing attention to the critical role of usability in big data analytics. The outcomes of this research include (1) clear impacts of making big-data analytics accessible to end users who are experts in various data domains, but not in advanced statistical data models and algorithms; (2) development of educational programs in support of pedagogy for exploratory analytical thinking in the context of big data; (3) establishing a workshop focused on usability in big-data analytics to increase awareness and promote collaboration between computational and usability researchers; (4) outreach to government agencies with needs in big text analytics, through our involvement in DHS VACCINE and the national laboratories; and (5) involvement of diverse student populations in the research project.

Table 1: Example of a three-level MV2PI with 3 types of models at 3 levels of scale. The models, visualizations, and interactions shown are examples of a possible MV2PI instantiation and should not be considered an exhaustive list.

Scale of Interaction	***Small***	***Medium***	***Large***
Model purpose	Spatially project small scale data points onto the display, e.g. based on similarity	Aggregate medium scale data into smaller scale, e.g. based on similarity	Extract useful data from large scale, e.g. based on relevance or coverage
Usage Description	System lays out displayed data, according to user's spatial organization feedback	System groups larger clusters of data in the layout, according to user's grouping feedback	System uses layout to query very large data and retrieve additional relevant data
Data scale of operation	Millions (e.g. display scale)	Billions (e.g. database scale)	Trillions (e.g. cloud scale)
Model operations	Project	Aggregate	Select
Model algorithms	Dimensionality reduction to 2D, spatialization	Clustering, classification, topic modeling	Information retrieval, entity extraction, sampling
Model parameters	Dimension weights	Dimension weights, centroids, count	Dimension weights, object weights
Model metrics	Similarity metric	Aggregation metric	Relevance metric
Visual representations	Similarity mapped to visual proximity	Aggregates mapped to visual groups, containment	Relevance mapped to visual salience, 3rd dimension
User interactions	Re-organize proximities, highlight	Re-group, annotate, landmarks, zoom	Open/close, like/dislike, search
Interactive feedback loop for machine learning	Change proximity -> update similarity -> update dimension weights -> update proximity -> update other models	Change membership -> update centroids -> update dimension weights -> update aggregates-> update other models	Change salience -> update relevance -> update dimension weights -> update extraction-> update other models

Keywords: Usability, visual analytics, interactive analysis, high-performance computing.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.