Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Closed-loop Big Data Analysis with Visualization and Scalable Computing ☆

Guangchen Ruan [a], Hui Zhang [b]

[a] *Indiana University, Bloomington, IN, United States*
[b] *University of Louisville, Louisville, KY, United States*

ARTICLE INFO

Article history:
Received 17 April 2016
Received in revised form 12 November 2016
Accepted 2 January 2017
Available online xxxx

ABSTRACT

Many scientific investigations require data-intensive research where big data are collected and analyzed. To get big insights from big data, we need to first develop our initial hypotheses from the data and then test and validate our hypotheses about the data. Visualization is often considered a good means to suggest hypotheses from a given dataset. Computational algorithms, coupled with scalable computing, can perform hypothesis testing with big data. Furthermore, interactive visual interfaces can allow domain experts to directly interact with data and participate in the loop to refine their research questions and redirect their research directions. In this paper we discuss a framework that integrates information visualization, scalable computing, and user interfaces to explore large-scale multi-modal data streams. Discovering new knowledge from the data requires the means to exploratively analyze datasets of this scale—allowing us to freely “wander” around the data, and make discoveries by combining bottom-up pattern discovery and top-down human knowledge to leverage the power of the human perceptual system. We start with a novel interactive temporal data mining method that allows us to discover reliable sequential patterns and precise timing information of multivariate time series. We then proceed to a parallelized solution that can fulfill the task of extracting reliable patterns from large-scale time series using iterative MapReduce tasks. Our work exploits visual-based information technologies to allow scientists to interactively explore, visualize and make sense of their data. For example, the parallel mining algorithm running on HPC is accessible to users through asynchronous web service. In this way, scientists can compare the intermediate data to extract and propose new rounds of analysis for more scientifically meaningful and statistically reliable patterns, and therefore statistical computing and visualization can bootstrap each another. Furthermore, visual interfaces in the framework allows scientists to directly participate in the loop and can redirect the analysis direction. All these combine to reveal an effective and efficient way to perform closed-loop big data analysis with visualization and scalable computing.

1. Introduction

A recent trend in many scientific investigations is to conduct data-intensive research by collecting a large amount of high-density high-quality data [14,18,38,40,36]. These data, such as text, video, audio, images, RFID, and motion tracking, are usually multi-faceted, dynamic, and extremely large in size, and likely to be substantially publically accessible for the purposes of continued and deeper data analysis. Indeed, data-driven discovery has already happened in various research fields, such as earth sciences, medical sciences, biology and physics, to name a few.

Top-down human knowledge plays an important role in knowledge discovery. Getting big insights from big data is no exception. We need to first develop our initial hypotheses from the data, and visualization is considered a good means to suggest initial hypotheses from a given dataset—very often in small size. We next need to test and validate our hypotheses about the data, and make discoveries by combining bottom-up pattern discovery and top-down human knowledge to leverage the power of the human perceptual system. Computational algorithms, coupled with scalable computing, can perform hypothesis testing with much larger data sets. Discovering new knowledge from the data also requires the means to allow us to freely “wander” around the data, refine and retest our hypotheses. Interactive visual interfaces can allow us to directly interact with data and participate in the analysis loop to refine their research questions and redirect their research directions in multiple iterations.

☆ This article belongs to HPC Tutorial for Big Data.
E-mail address: h0zhan22@louisville.edu (H. Zhang).

http://dx.doi.org/10.1016/j.bdr.2017.01.002
2214-5796/Published by Elsevier Inc.

Please cite this article in press as: G. Ruan, H. Zhang, Closed-loop Big Data Analysis with Visualization and Scalable Computing, Big Data Res. (2017), http://dx.doi.org/10.1016/j.bdr.2017.01.002

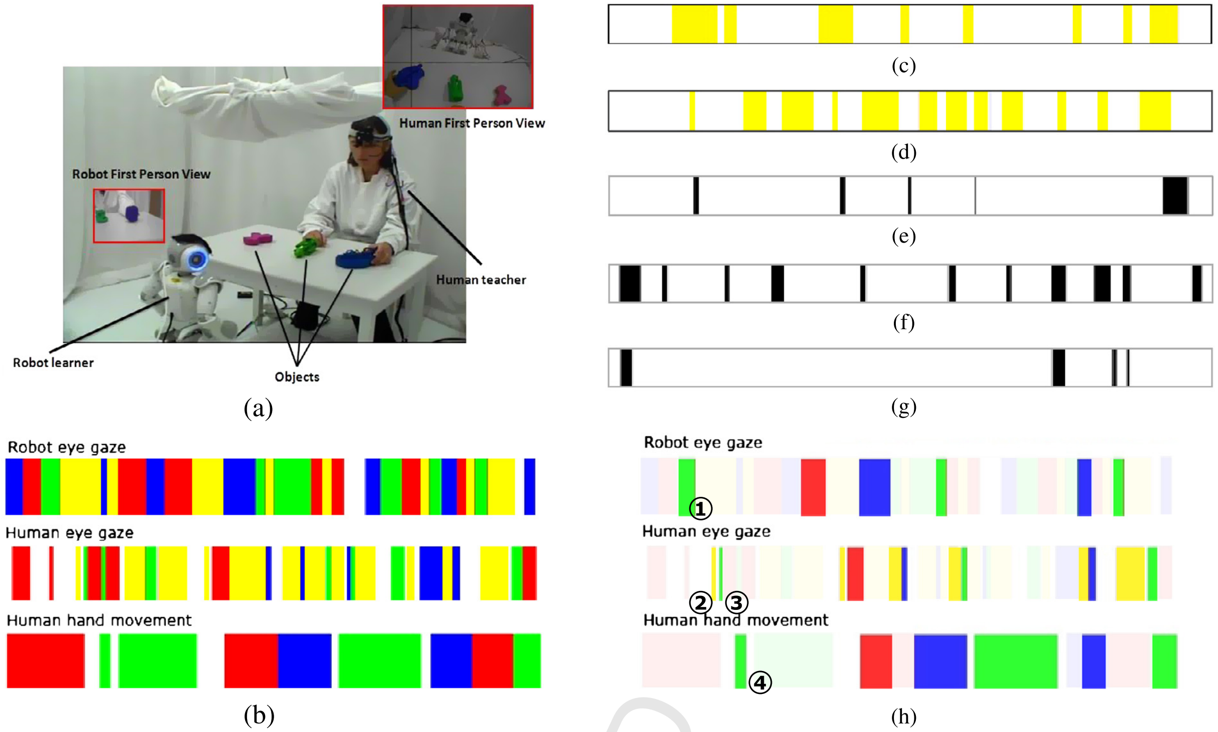

Fig. 1. (a)→(b): Using multi-modal sensing systems to collect and analyze fine-grained behavioral data including motion tracking data, eye tracking data, video and audio data. A family of studies using this research paradigm are conducted to collect multi-stream multi-modal data and convert them into multi-streaming time series for further data analysis and knowledge discovery. (c)–(g): Integrating over ROI event streams by overlaying pictures that record individual ROI streams. (c) Robot face-looking event. (d) Human face-looking event. (e) Face-to-face coordination: (Robot face-looking) *XOR* (Human face-looking). (f) Human eye–hand-coordination: (Human eye gaze) *XOR* (Human hand movement). (g) Human–Robot joint attention: (Human eye–hand-coordination) *XOR* (face-to-face coordination). (h) Six instances of momentary interactive behavior highlighted in the AOI streams. This sequential pattern starts with the situation that the robot agent and the human teacher attend to different objects (①), and then the human teacher checks the robot agent's gaze (②) and follows the robot agent attention to the same object (③) and finally reach to that object (④).

Thus one goal in this research is to develop, use and also share tools that enable researchers to find new patterns and gain new knowledge from such data. But how can we discover new and meaningful patterns if we do not know what we are looking for? Although standard statistics or data mining algorithms may find a subset of these meaningful patterns, they may miss a great deal more. We suggest here that visualization and visual mining is a powerful approach for new pattern discovery. Discovering new knowledge requires the ability to detect unknown, surprising, novel, and unexpected patterns. To achieve this goal, our proposed solution is to use visualization system that allows us to easily spot interesting patterns through both our visual perception systems and our domain knowledge. Consequently, the exploratory process should be highly iterative and interactive—visualizing not only raw data, but also the intermediate results of current statistical computations for further analysis. In this way, computational algorithm and visualization can bootstrap each other—informative visualization based on new results leads to the discovery of more complex patterns which can in turn be visualized, leading to more findings. Human experts play a critical role in this human-in-the-loop knowledge discovery by applying statistical techniques to the data, examining visualization results, and deciding/directing the research focus based on their theoretical knowledge. In this way, domain knowledge, computational power, and information visualization techniques can be integrated together to understand massive datasets.

2. A motivating use case—developing top-down knowledge hypotheses from visual analysis of multi-modal data streams

Analyzing fine-grained behavioral data in the format of multiple time-series (see e.g., Fig. 1(a)) is the motivating use case in our work. Interacting embodied agents, be the groups of people engaged in a coordinated task, autonomous robots acting in an environment, or an avatar on a computer screen interacting with a human user, must seamlessly coordinate their actions to achieve a collaborative goal. The pursuit of a shared goal requires mutual recognition of the goal, appropriate sequencing and coordination of each agent's behavior with others, and making predictions from and about the likely behavior of others. Such interaction is multimodal as we interact with each other and with intelligent artificial agents through multiple communication channels, including looking, speaking, touching, feeling, and pointing.

To gain insight of how interaction happens across multiple channels, we need to first find a way to "look" at the data. Information visualization can be effectively exploited to explore *moments-of-interest* in multi-modal data streams (see e.g., [37]). Computational algorithms such as data mining and sequential pattern mining (see e.g., [21,25,28,29,24,39,16] are other variants that have requirement similar to ours. While we of course exploit many techniques of information visualization and interactive data mining that have been widely used in other interfaces, we found that many of the problems we encounter have been fairly unique, and thus require customized hybrid approaches.

Color-based representation of temporal events. Multi-modal data streams are first converted to multi-streaming temporal events with categorical type values. We represent an event *e* as a rectangular bar by assigning the distinct color key to the event type (i.e., *e.t*), with the length corresponding to the event's duration (i.e., *e.d*). The visual display of temporal events themselves in a sequence allows us to examine how frequently each temporal event happens over time, how long each instance of an event takes, how one event relates to other events, and whether an event appears

Please cite this article in press as: G. Ruan, H. Zhang, Closed-loop Big Data Analysis with Visualization and Scalable Computing, Big Data Res. (2017), http://dx.doi.org/10.1016/j.bdr.2017.01.002