Analyse Progressive de Données

Progressive Data Analysis

Jean-Daniel Fekete

English Abstract—Our civilization is collecting data at a pace never seen before. While data analysis has made tremendous progresses in scalability in the last decade, this progress has only benefited "confirmatory" analysis or model-based computation; progress in data exploration has lagged behind. The main reason is that, to maintain their efficiency during exploration, humans need a rapid feedback loop of about 10 seconds. However, when data becomes larger or algorithms more complex, bounding the latency while preserving accuracy is not possible with existing computation paradigms. To address this problem, I propose a new computation paradigm: Progressive Data Analysis, tying together the traditional imperative paradigm, stream processing, and sample-based approaches.

1 Introduction

Existing data analysis systems do not support data exploration at scale because, for large amounts of data or for expensive computations, their latency is not controllable: computations can take minutes, hours, even days and months. Miller [6] has shown that humans' cognitive capabilities degrade when latency increases. He points out that the feedback of a system should remain below 10 seconds to maintain the user's attention. Therefore, to try to limit the latency, analysts currently resort to complex, inefficient, and unsatisfactory strategies, such as sampling without being able to know if a sample is representative.

To address the scalability problem under controlled latency, I propose a new computation paradigm: *progressive data analysis*. Instead of performing each computation in one long step—forcing the analyst to wait for an unbounded amount of time—a progressive data analysis system generates estimates of the results and updates the analyst continuously at a bounded pace. The process continues until the computation is complete, or earlier if the analyst considers that the quality of the estimates is sufficient to make a decision. During the process, progressive data analysis allows users to monitor and steer the computation, with data visualizations and interactions.

1.1 Semantics of Progressive Data Analysis

A data analysis function F can be specified as: $F(S, P, D) \mapsto (S', R, t, M)$ where t is the time it runs, M the amount of memory it uses, S the starting machine state, and S' the ending state. Also, in most data analysis functions, we distinguish between the parameters P of the function from a set of input *data tables* $D = d_1, \ldots, d_m$ that the function takes, and the ones it returns as output $R = r_1, \ldots, r_o$.

• *Jean-Daniel Fekete: Inria*
E-mail: Jean-Daniel.Fekete@inria.fr.

Existing data analysis systems rely on sequential computations (or *eager* semantics) to deliver their results, *i.e.*, if a function F operates on the results of a function G, as in $F \circ G$, it has to wait for G to finish before starting to run. Therefore, eager functions cannot control their latency.

An equivalent progressive data analysis is made of multiple calls to a progressive function F_p [4]:

$$\begin{array}{rcl} F_p(S_1, q, P, D_1) & \mapsto & (S_2, R_1, t_1, m_1) \\ & \cdots & \\ F_p(S_z, q, P, D_z) & \mapsto & (S_{z+1}, R_z, t_z, m_z) \end{array}$$

where D_i converges to D, and the partial results returned R_i converge to R. Furthermore, if q is the desired latency allowed to the progressive data analysis (we call it *quantum*), $t_i \leq q$.

1.2 Algorithm Transformation and Computation Strategies

Progressive computation needs progressive algorithms and novel computation strategies. For example: data analysis often requires a pre-processing stage where data is normalized, such as *min-max scaling*: rescaling the values contained in a data vector in the interval $[0, 1]$. For a progressive computation, the minimum and maximum values of the vector will be computed progressively so their values will converge step by step along the computation. A change will trigger a recomputation of the whole normalized vector if it has to stay accurate, and thus of all the computations relying on the normalized vector. Progressive data analysis needs to explore possible strategies to delay these changes, and analyze the impact of these delays in terms of performance and accuracy.

Besides strategies, traditional data analysis algorithms have to be translated into some progressive counterpart. The good news is that, according to our studies, progressive counterparts can be implemented for most of the families of data analysis algorithms: classification, regression, clustering, dimensionality reduction, and pre-processing methods like mean removal and variance scaling. Several research groups are currently working on adapting traditional algorithms to become progressive [2], [7], [10], [11], and some algorithms are natively progressive [8], [9]. Yet, a lot of work remains to be done to actually implement the most useful algorithms.

1.3 Visualization

Data visualization has so far been limited in scalability to 10,000–1,000,000 data items shown. Scaling to higher sizes needs data summarization techniques: either sampling, aggregation, or projection methods. Aggregation and projection cannot be computed in real-time; some systems pre-compute them to allow interaction, but it takes a long time, sometimes hours [5]. With progressive computation, summarization techniques can be used at an interactive rate [7], but visualization techniques need adaptations.

There have been several articles on visualizing aggregated data [1], [3], but the problem remains hard and algorithms to address it are computationally expensive. When connected to progressive computations, the visualization evolves with time. Analysts are able to monitor the behavior of algorithms, generate early hypotheses, and make early decisions, but new issues arise: analysts need to track the changes, and to navigate back and forth in the history of the visualization to link an early hypothesis to a later stage of the progression for confirmation or rejection.

The user interface, linked to the visualizations, should also be adapted to provide information about how the algorithm progresses, with multiple indicators: how much of the whole dataset has been processed? How is the algorithm improving/converging?

With colleagues, we have investigated user interface elements required to use progressive data analysis efficiently from an analyst's point of view [1] (see Figure 1). We have shown that by adding well designed user interface elements on top of each visualizations, users can assess the progression and the current quality of the computed results and decide whether to monitor the evolution or make decisions early on the results, validating our hypothesis on the effectiveness and usability of progressive systems with a suitable user interface. Yet, more work is required *e.g.*, to understand the human support needed to avoid biases and manage the exploration of competing hypotheses.

2 Implementation

ProgressiVis is a work-in-progress implementation of the progressive data analysis paradigm in the Python language: all the constructs of the language are progressive so data analysis pipelines are progressive by design. We chose Python because it is among the most popular languages for Data Science, providing high-performance libraries for a wide variety of general and domain-specific analytics. We initially hoped to rely on the well-known "stack" of core libraries, such as *NumPy*, *SciPy*, *Pandas*, and scikit-learn for machine-learning. Unfortunately, the requirements of the progressive data analysis paradigm are incompatible with many assumptions of these libraries, in particular that the data structures are mostly pre-allocated in memory; we had to re-implement most of the stack.

ProgressiVis relies on a kernel that defines the base *Module* class, a *scheduler*, and several supporting classes and mechanisms. A web server embedded in ProgressiVis is used to deliver visualization and manage interaction through a web browser; see https://github.com/jdfekete/progressivis for the code and [4] for more details.

3 Conclusion

We believe progressive data analysis will become a leading paradigm for exploratory data analysis at scale in the future. We provide our experimental system ProgressiVis to help understand the issues raised by this paradigm, and hope it will steer interesting development in the visual analytics community, as well as in others such as databases, machine-learning, and HPC to name a few.

References

[1] S. K. Badam, N. Elmqvist, and J.-D. Fekete. Steering the Craft: UI Elements and Visualizations for Supporting Progressive Visual Analytics. *Computer Graphics Forum*, 36(3):12, June 2017.

[2] U. Brandes and C. Pich. Eigensolver methods for progressive multidimensional scaling of large data. In M. Kaufmann and D. Wagner, editors, *14th International Symposium on Graph Drawing*, pages 42–53. Springer, 2007.

[3] N. Elmqvist and J.-D. Fekete. Hierarchical Aggregation for Information Visualization: Overview, Techniques, and Design Guidelines. *IEEE Trans. Vis. Comput. Graphics*, 16(3):439–454, 2010.

[4] J.-D. Fekete and R. Primet. Progressive analytics: A computation paradigm for exploratory data analysis. *CoRR*, abs/1607.05162, 2016.

[5] L. Lins, J. T. Klosowski, and C. Scheidegger. Nanocubes for real-time exploration of spatiotemporal datasets. *IEEE Trans. Vis. Comput. Graphics*, 19(12):2456–2465, Dec 2013.

[6] R. B. Miller. Response time in man-computer conversational transactions. In *Proc. of the Fall Joint Computer Conference, Part I*, pages 267–277. ACM, 1968.

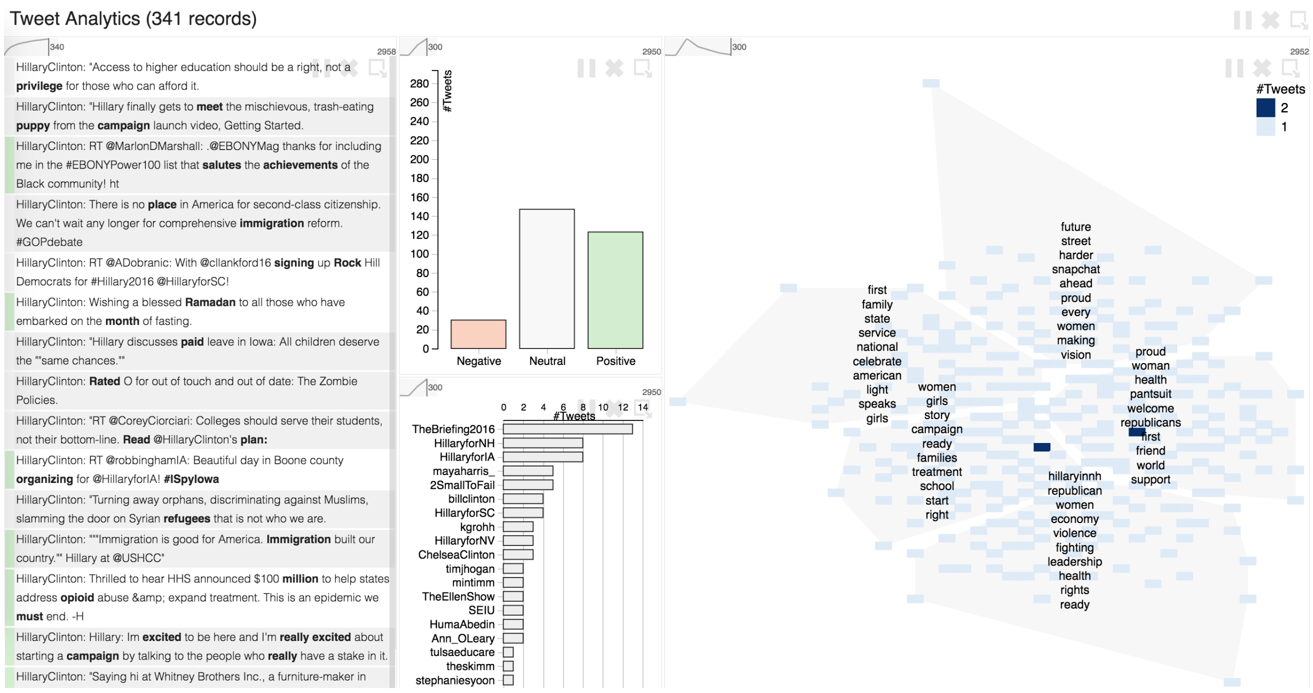
[7] N. Pezzotti, B. P. F. Lelieveldt, L. van der Maaten, T. Höllt, E. Eisemann, and A. Vilanova. Approximated and user steerable tsne for progressive visual analytics. *IEEE Trans. Vis. Comput. Graphics*, PP(99):1–1, 2016.

[8] C. Qin and F. Rusu. Speculative approximations for terascale distributed gradient descent optimization. In *Proceedings of the Fourth Workshop on Data analytics in the Cloud, DanaC 2015, Melbourne, VIC, Australia, May 31 - June 4, 2015*, pages 1:1–1:10, 2015.

[9] D. Sculley. Web-scale k-means clustering. In *Proc. of the 19th International Conference on World Wide Web*, WWW '10, pages 1177–1178. ACM, 2010.

[10] C. D. Stolper, A. Perer, and D. Gotz. Progressive visual analytics: User-driven visual exploration of in-progress analytics. *IEEE Trans. Vis. Comput. Graphics*, 20(12):1653–1662, Dec 2014.

[11] E. Zgraggen, A. Galakatos, A. Crotty, J.-D. Fekete, and T. Kraska. How progressive visualizations affect exploratory analysis. *IEEE Trans. Vis. Comput. Graphics*, page to appear, 2016.


Tweet Analytics (341 records)

HillaryClinton: "Access to higher education should be a right, not a **privilege** for those who can afford it.

HillaryClinton: "Hillary finally gets to **meet** the mischievous, trash-eating **puppy** from the **campaign** launch video, Getting Started.

HillaryClinton: RT @MarlonDMarshall: .@EBONYMag thanks for including me in the #EBONYPower100 list that **salutes** the **achievements** of the Black community! ht

HillaryClinton: There is no **place** in America for second-class citizenship. We can't wait any longer for comprehensive **immigration** reform. #GOPdebate

HillaryClinton: RT @ADobranic: With @cllankford16 **signing** up **Rock** Hill Democrats for #Hillary2016 @HillaryforSC!

HillaryClinton: Wishing a blessed **Ramadan** to all those who have embarked on the **month** of fasting.

HillaryClinton: "Hillary discusses **paid** leave in Iowa: All children deserve the ""same chances.""

HillaryClinton: **Rated** O for out of touch and out of date: The Zombie Policies.

HillaryClinton: "RT @CoreyCiorciari: Colleges should serve their students, not their bottom-line. **Read** @HillaryClinton's **plan:**

HillaryClinton: RT @robbinghamIA: Beautiful day in Boone county **organizing** for @HillaryforIA! **#ISpyIowa**

HillaryClinton: "Turning away orphans, discriminating against Muslims, slamming the door on Syrian **refugees** that is not who we are.

HillaryClinton: ""Immigration is good for America. **Immigration** built our country."" Hillary at @USHCC"

HillaryClinton: Thrilled to hear HHS announced $100 **million** to help states address **opioid** abuse &amp; expand treatment. This is an epidemic we **must** end. -H

HillaryClinton: Hillary: Im **excited** to be here and I'm **really excited** about starting a **campaign** by talking to the people who **really** have a stake in it.

HillaryClinton: "Saying hi at Whitney Brothers Inc., a furniture-maker in

340
2958
300
2950
300
2952
#Tweets
280
260
240
220
200
180
160
140
120
100
80
60
40
20
0
Negative
Neutral
Positive
0 2 4 6 8 10 12 14
#Tweets
TheBriefing2016
HillaryforNH
HillaryforIA
mayaharris_
2SmallToFail
billclinton
HillaryforSC
kgrohh
HillaryforNV
ChelseaClinton
timjhogan
mintimm
TheEllenShow
SEIU
HumaAbedin
Ann_OLeary
tulsaeducare
theskimm
stephaniesyoon
#Tweets
2
1
future street harder snapchat ahead proud every women making vision
first family state service national celebrate american light speaks girls
women girls story campaign ready families treatment school start right
proud woman health pantsuit welcome republicans first friend world support
hillaryinnh republican women economy violence fighting leadership health rights ready

Fig. 1. The ProgressiveInsights system with specific user interface elements for monitoring and control [1].