M2 Internship: Interactive Data Exploration at Scale

Yanlei Diao (École Polytechnique)

Duration: 5-6 months, the starting date is flexible (ideally March 1st, 2016)
Location: École Polytechnique, Palaiseau, France
Keywords: Big Data Analytics, **Databases**

Background: Interactive Data Exploration

Traditional data management systems are suited for applications in which the structure, meaning and contents of the database, as well as the questions (queries) to be asked, are all well-understood. However, this is no longer true when the volume and diversity of data grow at an unprecedented rate, as we are witnessing in scientific computing, social network analysis, and business data analysis. At the same time, the human ability to comprehend data remains as limited as before. To address the increasing disparity in the "big data - same humans" problem, this project explores a new approach of system-aided exploration of big data space and automatic learning of the user interest in order to retrieve all objects that match the user interest – we call this new service "interactive data exploration", which complements the traditional querying interface for large databases.

Challenges: Interactive Data Exploration at Scale

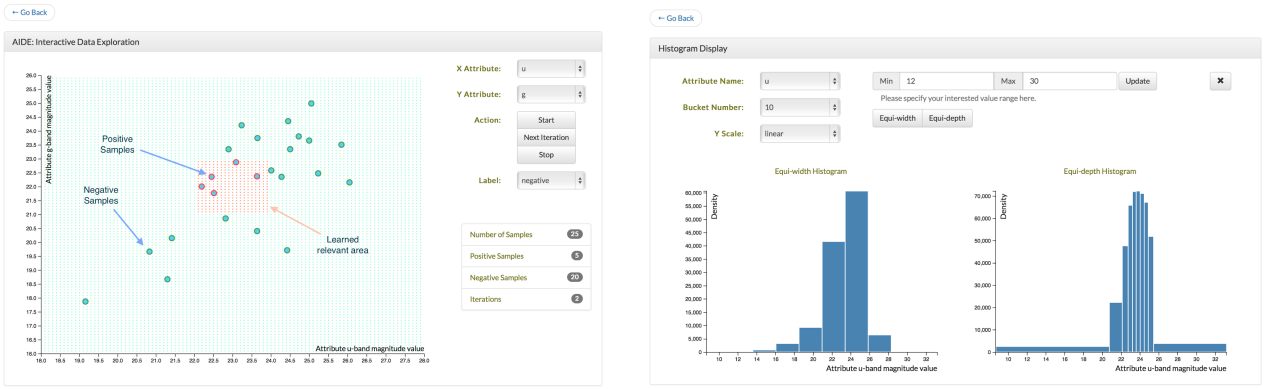
We aim to design a new system that obtains user feedback on database samples as relevant or irrelevant to the user interest. Such feedback is incrementally incorporated into a machine learning model that captures the user interest. The system further decides which subarea of the data space and which samples in the subarea to present to the user for further feedback (i.e., *exploration strategies*). Figure 1(a) illustrates the distribution of relevant and irrelevant database samples in a two dimensional space, and a current model of the user interest (which may not be accurate yet). Figure 1(b) illustrates the data distribution along one of the two dimensions, which may provide insight for developing new exploration strategies. Existing work has developed exploration strategies based on the insights from Decision Tree and Support Vector Machine (SVM) learning algorithms, and focused on how to make the best exploration strategies over a database such that the user interest model converges to the true model with minimum user feedback [2, 3]. Interactive data exploration at scale, however, poses two additional challenges:

1. *Interactive performance*: The above data exploration activities are organized in a series of iterations: in each iteration, the system collects user feedback on a few data samples, improves the user interest model, and makes an exploration decision to present the user with a new subarea of the data space and samples from that subarea for feedback. Each iteration must complete with interactive performance, e.g., within seconds, as the user may be waiting for new results online. Hence, guaranteeing interactive performance when making good exploration decisions over a large database becomes another technical challenge.

2. *Scalability:* As the volume and diversity of data continue to grow, data exploration techniques are expected to scale in both the database size, e.g. terabytes of data, as well as in the number of dimensions, e.g., hundreds to thousands of the attributes in the database.

Internship Goal:

The goal of this internship is to develop the key data exploration techniques and optimizations to enable scalable, interactive performance.



(a) A snapshot of interactive data exploration.

(b) Histogram visualization for exploration attributes.

Figure 1: Visualization of labeled database samples and data distriutions in interactive data exploration.

Scaling in Dimensionality: Initial results reveal that existing exploration strategies such as decision tree-based and SVM-based techniques suffer from slow convergence when the data space involves more than 4 or 5 dimensions [2, 3]. One issue is how to present a relatively small, relevant data space for the user to explore. A large database often contains dozens of tables and hundreds of attribute per table, while the user may be interested in a data space of a dozen attributes (dimensions). Presenting such a conceptual data space for each user, including recognizing all relevant objects and mapping their locations in the space, requires effective views constructed from the underlying physical database. In addition, even for a data space of a dozen attributes, if the true user interest lies in a lower-dimensional space, it is crucial to also explore dimensionality reduction techniques to construct lower-dimensional subspaces, either pre-computed or computed on-the-fly, to help expedite the convergence of the user interest model.

Scaling in Database Size. We also seek to scale data exploration to multi-terabyte databases. Such scaling requires fundamental changes of the database architecture from a single sever to a compute cluster, while ensuring interactive performance. Recent projects like BlinkDB [1] aim to achieve scalability and interactive performance by performing sampling and providing probabililstic bounds on the query answers. Such sampling, however, is designed for traditional database queries but not data access patterns in interactive data exploration. To ensure fast convergence of data exploration, our system must work with the new sampling and indexing methods designed specially for data exploration and scale them across a distributed storage with interactive performance.

This project will involve the design of new algorithms for dimensionality reduction in data exploration, and database optimization such as sampling and indexing to deal with large databases and provide interactive performance. The proposed techniques and optimizations will be evaluated using real-world datasets, including housing datasets, the IMDB database, and digital sky surveys.

Contact

- Yanlei Diao (yanlei.diao@polytechnique.edu), http://www.lix.polytechnique.fr/~yanlei.diao/

References

[1] Sameer Agarwal, Henry Milner, Ariel Kleiner, Ameet Talwalkar, Michael Jordan, Samuel Madden, Barzan Mozafari, and Ion Stoica. Knowing when you're wrong: Building fast and reliable approximate query processing systems. In *SIGMOD Conference*, pages 481–492, 2014. 2

[2] Yanlei Diao, Kyriaki Dimitriadou, Zhan Li, Wenzhao Liu, Olga Papaemmanouil, Kemi Peng, and Liping Peng. AIDE: an automatic user navigation system for interactive data exploration. *PVLDB*, 8(12):1964–1967, 2015. 1, 2

[3] Kyriaki Dimitriadou, Olga Papaemmanouil, and Yanlei Diao. Explore-by-example: an automatic query steering framework for interactive data exploration. In *SIGMOD Conference*, pages 517–528, 2014. 1, 2