Report on the Third International Workshop on Exploratory Search in Databases and the Web (ExploreDB 2016)

Senjuti Basu Roy
New Jersey Institute of Technology
senjutib@njit.edu

Kostas Stefanidis
University of Tampere
kostas.stefanidis@uta.fi

Georgia Koutrika
HP Labs, Palo Alto
koutrika@hp.com

Laks V.S. Lakshmanan
University of British Columbia
laks@cs.ubc.ca

Mirek Riedewald
Northeastern University, Boston
mirek@ccs.neu.edu

1. INTRODUCTION

The traditional way of interaction between a user and a database system is through queries, for which the correctness and completeness of their answers are key challenges. Structured query languages, such as SQL, XQuery, and SPARQL, allow users to submit queries that may precisely identify their information needs, but often require users to be familiar with the structure of data, the content of the database, and also have a clear understanding of their needs. As databases get larger and accessible to a more diverse audience, new forms of data exploration and interaction become increasingly more attractive to aid users navigate through the information space and overcome the challenges of information overload [6, 5].

The Web represents the largest and most complex repository of content. Users seek information through two predominant modes: by browsing or by searching. In the first mode, the interaction between the user and the data repository is driven directly by the user's needs interpretation. In the latter mode, a search engine typically mediates the user-data interactions and the process starts with the user entering query-terms that act as surrogates for the user information goals. Commonly, independently from data models and query languages, the query results are presented to the user as a ranked list.

Clearly, there is a need to develop novel paradigms for exploratory user-data interactions that emphasize user context [13] and interactivity with the goal of facilitating exploration, retrieval, and assimilation of information. A huge number of applications need an exploratory form of querying. Ranked retrieval techniques is a first step in this direction [1, 3]. Recently, several new aspects for exploratory search, such as preferences [12], diversity [14], novelty [9], surprise [10] and serendipity [4], are gaining increasing importance. From a different perspective, recommender systems tend to anticipate user needs by suggesting the most appropriate to the users information [11], while a new line of research in the area of exploratory search is fueled by the growth of online social interactions within social networks and Web communities [2]. Overall, the query-answering task needs to be further enhanced to capture the intent that the user may have in mind during querying. Exploratory search techniques are of great assistance that facilitates and guides users to focus on the relevant aspects of their search results.

To sum up, the field of data exploration is diverse in terms of research directions and potential user base. Hence, the ExploreDB workshop intends to bring together researchers and practitioners from different fields, ranging from data management and information retrieval to data visualization and human computer interaction. Its goal is to study the emerging needs and objectives for data exploration, as well as the challenges and problems that need to be tackled, and to nourish interdisciplinary synergies. We summarize the outcomes of the third workshop instance held in conjunction with ACM SIGMOD 2016 in San Francisco, USA.[1]

2. WORKSHOP OUTLINE

The workshop program consisted of two keynote

[1]For a summary of the first and second instances of ExploreDB, please refer to [8] and [7], respectively.

talks and six research papers.

2.1 Invited Talks

The first keynote talk titled *"Unifying Data Exploration and Curation"* was given by Shan Shan Huang from LogicBlox.

Shan Shan pointed out that recent years have seen a surge in "self-service" business intelligence tools. These tools primarily focus on supporting decision-making by non-technical "end users", through data exploration – the querying of data and inspection of results.

Exploration, however, is only part of the story. Curation is its complement. As Shan Shan discussed, curation is the ability to organize data into structures that are meaningful for a particular problem domain and convenient for building further explorations upon. Curation is also the ability to modify data, as well as creating new data through rules and constraints, in order to support what-if's, forecasting, and planning for the future. Exploration and curation often need to interleave in the decision-making process of an end-user.

Shan Shan presented the LogicBlox Modeler, namely a unifying environment that provides support for both exploration and curation. She explained the need for a unifying environment through applications in government, major financial institutions, and large global retailers. She also discussed the employed language – in its visual and textual representation – that supports not only querying, but also the creation and modification of schema and data. Finally, Shan Shan expounded the challenges imposed on the database runtime by the use cases of exploration and curation at scale and aspects of the LogicBlox database designed to meet these challenges.

In the second keynote, titled *"Why would you recommend me THAT!?"*, Aish Fenton from Netflix focused on problems in the area of recommender systems.

Specifically, his talk focused on the complexities and nuances of a real world recommendation problem: With so many advances in machine learning recently, why recommendations are not yet perfect?

Aish's talk started with a brief overview of recommender systems. After that, he provided a walkthrough of the open problems in the area of recommender systems, especially as they apply to Netflixs personalization and recommender algorithms. He described several challenging aspects of obtaining real world feedback from the users - in particular, he illustrated the difference between the implicit and the explicit feedback and how they are being used in the matrix factorization model inside Netflix. Aish also summarized the use of "latent" vs "explicit" users and item features inside the recommendation model. Aish captured several critical issues in presenting recommended items in the user interface many of which lend themselves to challenging HCI design and research problems. Last but not the least, his talk focused on the scalability challenges, as the Netflix user base contains millions of users and items giving rise to a gigantic yet very sparse user-item matrix on which the matrix factorization algorithm needs to run. Finally, for many of these aforementioned challenges, he sketched out some tentative solutions and future directions.

2.2 Paper Presentations

The six talks of the technical program covered a variety of issues related to different perspectives of exploratory data analysis.

In *"Towards Large-Scale Data Discovery"*, Raul Castro Fernandez, Ziawasch Abedjan, Samuel Madden and Michael Stonebraker presented their vision towards making a data discovery system that facilitates locating relevant data among thousands of data sources. The proposed work represents data sources succinctly through signatures, and then creates search paths that permit quick execution of a set of data discovery primitives used for finding relevant data. Authors have built a prototype that is being used to solve data discovery challenges of two big organizations, namely the MIT data warehouse team and a big pharma company.

Zhan Li, Olga Papaemmanouil and Georgia Koutrika focused in the course selection decision making problem in the work *"CourseNavigator: Interactive Learning Path Exploration"*. Specifically, they introduced CourseNavigator, which is a new course exploration service. The service identifies all possible course selection options for a given academic period, referred to as learning paths, that can meet the students customized goals and constraints. CourseNavigator offers a suite of learning path generation algorithms designed to meet a range of course exploration end-goals, such as learning paths for a given period and desired degree, as well as the highest ranked paths based on user-defined ranking functions.

In *"Space Odyssey - Efficient Exploration of Scientific Data"*, Mirjana Pavlovic, Eleni Tzirita Zacharatou, Darius Sidlauskas, Thomas Heinis and Anastasia Ailamaki presented Space Odyssey, a novel approach enabling scientists to efficiently explore multiple spatial datasets of massive size. Without any prior information, Space Odyssey in-

crementally indexes the datasets and optimizes the access to datasets frequently queried together. The experimental evaluation, showed, through incrementally indexing and changing the data layout on disk, that Space Odyssey accelerates exploratory analysis of spatial data by substantially reducing query-to-insight time compared to the state of the art.

Hisham Benotman, Lois Delcambre and David Maier noticed, in *"Multiple Diagram Navigation (MDN)"*, that navigation systems with rich user interfaces could go beyond search and browse facilities by providing overviews and exploration features. Specifically, authors presented MDN to assist domain novices by providing multiple overviews of the content matter. MDN superimposes any type of diagram or map over a collection of information resources, allowing content providers to reveal interesting perspectives of their content. Users can navigate through the content in an exploratory way using three different types of browsing. The authors also discussed their vision for using heuristics about diagram structures to help rank results returned by MDN queries.

In *"Collection, Exploration and Analysis of Crowdfunding Social Networks"*, Miao Cheng, Anand Sriramulu, Sudarshan Muralidhar, Boon Thau Loo, Laura Huang and Po-Ling Loh presented their initial results at understanding the phenomenon of crowdfunding using an exploratory data-driven approach. They developed a big data platform for collecting and managing data from multiple sources, including company profiles (CrunchBase and AngelList) and social networks (Facebook and Twitter). Using Spark, they studied the impact of social engagement on startup fund raising success. Finally, they explored visualization techniques that allow visualizing communities of investors that make decisions in a close-knit fashion vs. looser communities where investors largely make independent decisions.

Finally, Anna Gogolou, Marialena Kyriakidi and Yannis Ioannidis, in *"Data Exploration: A Roll Call of All User-Data Interaction Functionality"*, pointed out that data exploration begins when a user is given a set of data and ends when the user extracts all information and knowledge hidden in the data. Although a plethora of systems have been developed to tackle different data exploration aspects, there is no framework devoted to it as a whole, and several interaction types and data functionalities, such as search, data analysis, curation, constraint satisfaction, data mining and visualization, are kept out of sight. In this work, authors claimed that any user-data interaction is essential for data exploration and sketch a prototype with both automated and user-induced functionality.

3. WORKSHOP CONCLUSIONS

Several themes emerged in the discussions.

- The presented papers cover a variety of domains - scientific data, spatial data, structured and unstructured data or a combination thereof - in all of these domains data exploration is an important as well as necessary operation.
- The papers presented in the workshop employ a variety of interesting technical solutions - discrete and continuous optimization problems, innovative data structures, and novel algorithmic solutions.
- Data exploration is an active area of research, as it involves a handful challenging sub-problems that span across data analysis, curation, constraint satisfaction, visualization, mining, and most importantly scale.
- The audience acknowledges and appreciates the necessity of data exploration in a variety of domains in the context of pure academic research as well as solving a real world industry scale business problem.
- Going forward, data exploration research is likely to make new strides due to the variety of data, its scale and velocity, as well as due to the emergence of new applications.

This third instance of ExploreDB made clear that a lot of research work still needs to be done in the general area of data exploration and discovery. Given the growing interest in industry and academia, we are looking forward to the next instance of this workshop.

4. REFERENCES

[1] S. Agrawal, S. Chaudhuri, G. Das, and A. Gionis. Automated ranking of database query results. In *CIDR*, 2003.

[2] S. Amer-Yahia, L. V. S. Lakshmanan, and C. Yu. Socialscope: Enabling information discovery on social content sites. In *CIDR 2009, Fourth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2009, Online Proceedings*, 2009.

[3] S. Chaudhuri, G. Das, V. Hristidis, and G. Weikum. Probabilistic information retrieval approach for ranking of database query results. *ACM Trans. Database Syst.*, 31(3):1134–1168, 2006.

[4] A. Foster and N. Ford. Serendipity and information seeking: an empirical study. *Journal of Documentation*, 59(3):321–340, 2003.

[5] H. Garcia-Molina, G. Koutrika, and A. G. Parameswaran. Information seeking: convergence of search, recommendations, and advertising. *Commun. ACM*, 54(11):121–130, 2011.

[6] M. L. Kersten, S. Idreos, S. Manegold, and E. Liarou. The researcher's guide to the data deluge: Querying a scientific database in just a few seconds. *PVLDB*, 4(12):1474–1477, 2011.

[7] G. Koutrika, L. V. S. Lakshmanan, M. Riedewald, M. A. Sharaf, and K. Stefanidis. Report on the second international workshop on exploratory search in databases and the web (exploredb 2015). *SIGMOD Record*, 44(4):49–52, 2015.

[8] G. Koutrika, L. V. S. Lakshmanan, M. Riedewald, and K. Stefanidis. Report on the first international workshop on exploratory search in databases and the web (exploredb 2014). *SIGMOD Record*, 43(2):49–52, 2014.

[9] A. Labrinidis and N. Roussopoulos. Exploring the tradeoff between performance and data freshness in database-driven web servers. *VLDB J.*, 13(3):240–255, 2004.

[10] N. Sarkas, N. Bansal, G. Das, and N. Koudas. Measure-driven keyword-query expansion. *PVLDB*, 2(1):121–132, 2009.

[11] Y. Shi, M. Larson, and A. Hanjalic. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Comput. Surv.*, 47(1):3:1–3:45, 2014.

[12] K. Stefanidis, G. Koutrika, and E. Pitoura. A survey on representation, composition and application of preferences in database systems. *ACM Trans. Database Syst.*, 36(3):19, 2011.

[13] K. Stefanidis, E. Pitoura, and P. Vassiliadis. Adding context to preferences. In *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007*, pages 846–855, 2007.

[14] C. Yu, L. V. S. Lakshmanan, and S. Amer-Yahia. It takes variety to make a world: diversification in recommender systems. In *EDBT 2009, 12th International Conference on Extending Database Technology, Saint Petersburg, Russia, March 24-26, 2009, Proceedings*, pages 368–378, 2009.