Report on the First International Workshop on Exploratory Search in Databases and the Web (ExploreDB 2014)

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

Laks V.S. Lakshmanan
Department of Computer Science, University of British Columbia
laks@cs.ubc.ca

Mirek Riedewald
College of Computer and Information Science, Northeastern University, Boston
mirek@ccs.neu.edu

Kostas Stefanidis
ICS-FORTH, Heraklion
kstef@ics.forth.gr

1. INTRODUCTION

Databases are well-organized collections of data. Structured query languages, such as SQL, XQuery, and SPARQL, enable users to formulate precise queries over the data stored in a database. To be successful, users need to be familiar with the query language and the underlying data organization. They also need to understand, to some extent, the data stored, and have a fairly clear idea of what they are looking for.

In contrast, the World Wide Web represents the largest and arguably the most complex repository of content, where the above assumptions do not hold. Structured and unstructured data, files and records, multimedia and text, scientific and user-generated data co-exist peacefully on the Web. Free-text queries provide an easy way for users to express their information seeking needs without having to worry about the underlying data organization. Search engines typically act as mediators for user-data interactions on the Web. Given a query, results are most commonly presented as a ranked list. Users subsequently peruse the list to satisfy their information needs through browsing the links and by issuing further queries. This information seeking paradigm has prevailed on the Web for many years as witnessed by the success of search engines, such as Google and Bing.

Despite the popularity and success of querying combined with browsing of data and results, it is worth exploring if this paradigm is still suitable and sufficient in the age of Big Data.

Consider for example `www.data.gov`: this site alone provides access to more than 100,000 different datasets, making it difficult for users to determine which of them could be relevant for their analysis. At the same time, increasingly more applications no longer rely on queries specified by experts. Instead, queries are generated by a diverse and not necessarily programming-aware audience. For instance, consider the Sloan Digital Sky Survey, accessed by domain scientists for a variety of analysis purposes, or the Amazon catalog, where users search over a huge number of products. In these settings, knowing what to look for or how to find it is not easy. With seemingly infinite options, long and repetitive query-browse sessions frustrate users and shake their confidence in the results: "Did I find what I was really looking for?" "Are these good results?" "Did I explore all my options?"

Information on the Web gets rapidly diversified both in terms of its complexity as well as the media through which it is encoded, spanning from large amounts of unstructured and semi-structured data to semantically rich knowledge. Many useful facts about entities (e.g., people, locations, organizations, and products) and their relationships can be found in a multitude of semi-structured and structured data sources, such as Wikipedia, Linked Data cloud[1], Freebase[2], and many others. Increasing demands for sophisticated discovery capabilities are now being imposed by numerous applications in various domains, such as social media, healthcare, e-commerce and web analytics, telecommunications, business intelligence, and cyber-security. Yet, many of these data are hidden behind barriers of language constraints, data heterogeneity, ambiguity, and the lack of proper interfaces.

[1] `http://linkeddata.org`
[2] `http://freebase.com`

Furthermore, the complexity and heterogeneity of the information implies that the associated semantics is often user-dependent and emergent. Individual aspects, such as age, gender, profession, or experience, are often not taken into account, for example, the difference in searching between children and adults. In addition, most systems still assume that the user has a static information need, which remains unchanged during the seeking process. Hence, they are strongly optimized for lookup searches, expecting that the user is only interested in facts and not in complex problem solving.

Consequently, there is a need to develop novel paradigms for exploratory user-data interactions that emphasize user context and interactivity with the goal of facilitating exploration, interpretation, retrieval, and assimilation of information. Ranked retrieval techniques for relational, XML, RDF and graph databases, text, multimedia, scientific and statistical databases, social networks and many others, comprise a first step towards this direction. From a different perspective, recommendation applications tend to anticipate user needs by automatically suggesting the information which is most appropriate to the users and their current context. Recently, new aspects of exploratory search, such as preferences, diversity, novelty, and surprise, are gaining increasing importance. Also, 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.

The purpose of the ExploreDB workshop is to bring together researchers and practitioners that approach data exploration from different angles, ranging from data management and information retrieval to data visualization and human computer interaction, in order 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 first workshop instance held in conjunction with EDBT/ICDT 2014 in Athens, Greece.

2. WORKSHOP OUTLINE

The workshop program included a keynote talk, six research papers, and a panel, which examined data exploration from the standpoints of data visualization, information retrieval, Web search, data mining, and database queries.

2.1 Invited Talk

The keynote talk was given by Prof. Daniel A. Keim from the University of Konstanz, Germany, and was entitled "*Exploring Big Data using Visual Analytics.*" With ever-growing volumes of data, Prof. Keim highlighted that effective large-data exploration has to include the human in the process. Specifically, it is important to combine the flexibility, creativity, and general knowledge of the human with the enormous storage capacity and the computational power of today's computers.

Visual analytics naturally integrate the human in the data analysis process and enable her to apply her perceptual abilities to large data sets. Presenting data in an interactive, graphical form often fosters new insights and encourages the formation and validation of new hypotheses for better problem solving and gaining deeper domain knowledge. Visual analytics techniques have proven to be of high value both for the first steps of the data exploration process, namely understanding the data and generating hypotheses about the data, as well as for the actual knowledge discovery by guiding the search using visual feedback.

However, in putting visual analysis to work on big data, it is not obvious where the boundary lies between what can be done by automated analysis and what should be done by interactive visual methods. In dealing with massive data, the use of automated methods is mandatory—and for some problems it may be sufficient. On the other hand, there is a wide range of problems where the use of interactive visual methods is necessary. The keynote speaker discussed when it is useful to combine visualization and analytics techniques, as well as the options on how to combine techniques from both areas. He provided examples from a wide range of application areas that illustrated the benefits of visual analytics techniques.

2.2 Paper Presentations

Sean Chester, Michael Lind Mortensen, and Ira Assent in their paper entitled "*On the Suitability of Skyline Queries for Data Exploration*" studied the data exploration problem based on a sequence of incrementally changing queries to the data. They focused on the skyline operator as a tool of exploratory querying both analytically and empirically. They showed how the results evolve as users modify their queries, and suggested different ways to guide users in formulating reasonable queries.

From a different perspective, George Valkanas, Apostolos N. Papadopoulos, and Dimitrios Gunopulos in their paper "*Skyline Ranking a la IR*" studied a quality-based ranking technique of the results of a skyline query. They described a novel IR-style ranking mechanism for generating such results,

based on the renowned tf-idf weighting scheme, and efficient algorithms to compute the quality of a result and induce a total ordering of the skyline set.

Panagiotis Papadakos and Yannis Tzitzikas in their paper "*Preference-enriched Faceted Exploration*" presented Hippalus, a system for exploratory search enriched with preferences. Hippalus supports the popular interaction model of Faceted and Dynamic Taxonomies, enriched with user actions. These allow users to define preferences, while offering automatic conflict resolution. Preferences can be expressed over attributes (facets), whose values can be hierarchically valued and/or multi-valued.

Marina Drosou and Evaggelia Pitoura in their paper "*The DisC Diversity Model*" considered that diversification can be used as a means for exploration, and they described the notion of DisC diversity. A DisC diverse subset of a query result contains objects, such that each object in the result is represented by a similar object in the diverse subset and the objects in the diverse subset are dissimilar to each other. Locating a minimum DisC diverse subset is an NP-hard problem, hence, they provided heuristics for its approximation.

Haridimos Kondylakis and Dimitris Plexousakis in their paper titled "*Exploring RDF/S Evolution using Provenance Queries*" discussed how to reduce the human effort spent on understanding ontology evolution. They presented a module, named ProvenanceTracker, which receives as input the list of changes between two or more RDF/S ontology versions and can answer fine-grained provenance queries about ontology resources. The module can identify when and how a resource was created, as well as the sequence of changes that led to the creation of that specific resource.

Finally, Steven Simske, Igor Boyko, and Georgia Koutrika in their paper "*Multi-Engine Search and Language Translation*" summarized approaches that focus on improving the quality and accuracy of the search and language translation tasks in the process of interaction between a user and a database. Specifically, multi-engine and related meta-algorithmic approaches are shown to be promising means of improving the performance of both search and translation. They also described a vision of how these tasks can be combined to create a more robust overall text mining project.

2.3 Panel

The panel's theme was "*Exploratory search in databases and the Web—New name for an old hat?*" The moderator, Georgia Koutrika, challenged five panelists on the novelty and future of data exploration and its connection to databases and the Web: Amelie Marian (Rutgers University, USA), Melanie Herschel (Université Paris Sud 11, France), Daniel Keim (University of Konstanz, Germany), Yannis Tzitzikas (University of Crete, Greece), and K. Selçuk Candan (Arizona State University, USA).

Amelie Marian viewed exploratory search from the personalization perspective, and she pointed out the importance of exploring the past when aiming at personalizing the user experience. Such process may include exploration based on personal data, e.g., data from e-mails, Skype, calendars, browser history and file systems, as well as exploration based on social data, e.g., data from Facebook, Twitter, and Foursquare.

Melanie Herschel took the OLAP angle, arguing that learning and investigation are important components of exploratory search. Aspects of both, such as knowledge acquisition, comparison (through the global view of data), aggregation (data warehouses), analysis (OLAP, data mining) and synthesis (provided by reports), have already been solved. However, traditional OLAP querying is constrained by hierarchical dimensions and cube operations, and the limited capabilities for changing and evolving information needs. Intuitively, it is about harvesting what has been planted. On the other hand, exploratory search is about the unknown, creating opportunities for future work, including the *search-refine-expand* paradigm, the freedom to adapt to changing user questions and the fact that learning is an iterative process.

Daniel Keim talked about visual analytics in exploratory search. He described the general goals of data visualization as presentation (visualization of data), confirmatory analysis (visualization of data that allows confirmation or rejection of input hypotheses), and exploratory analysis (visualization of data that provides hypotheses about data). He claimed that exploratory analysis is an open topic with many applications, such as visually exploring complex, semantically ambiguous, dynamic, and uncertain data.

Yannis Tzitzikas spoke from the Web search perspective. He pointed out that Web searching has mainly focused on ranking, but ranking alone is not adequate for exploration. The integration of interaction models for exploring structured and unstructured data, faceted browsing of search results and gradual restrictions for different types of queries and different domains are important topics that have not been (adequately) covered yet. He also highlighted the issue of the applicability of this interac-

tion mode to distributed and heterogeneous sources of varying structural complexity, and the need for better support of the decision making process with user-provided and user-controllable preferences.

Finally, K. Selçuk Candan spoke about multimedia data exploration. Challenges in multimedia exploration arise from special characteristics of such data, including imprecision, sparsity, volume, velocity, variety, high-dimensionality, and multimodality. Selçuk pointed out that there are several directions for future work not (fully) covered so far, including media summarization and dimensionality reduction, multi-modal and richly structured/linked data exploration, dynamic/evolving multimedia exploration, and bridging the semantic gap in media exploration.

3. WORKSHOP CONCLUSIONS

One important message was made clear by the workshop presentations and the participants: given the proliferation of data and applications, there is a need to view data exploration at various levels and from different perspectives. A number of key observations and research directions emerged that we summarize below.

- In databases, much work has been done on generating precise answers for precise queries. In contrast, exploratory search being about the unknown opens up the door to novel information seeking paradigms that focus on the *user-data interaction*, and the need to support an *iterative learning process that adapts to evolving user objectives*.

- In the Web, ranking alone is not adequate for exploration. For developing novel paradigms for exploratory interactions between users and data, *user context and interactivity* need to be emphasized. The integration of interaction models for exploring structured and unstructured data, faceted browsing of search results and gradual restrictions for different types of queries and different domains are important topics not adequately covered yet.

- Exploratory analysis is an open topic with many applications, such as exploring complex, semantically ambiguous, dynamic, and uncertain data. *Different types of data* bring different research challenges at the table.

- Challenges in multimedia exploration arise from special characteristics of such data, including imprecision, sparsity, and multimodality. *Media summarization and dimensionality reduction, multi-modal and richly structured/linked data exploration, dynamic/evolving multimedia exploration* are some critical challenges in multimedia data exploration.

- As yet another example, personal data can be leveraged for *exploring the past and personalizing the user experience*. Personal data exploration can take into account psychological and behavioral patterns to build novel exploration paradigms. For example, in an *associative learning paradigm*, where exploration becomes a learning experience that allows the individual to learn and remember the relationship between unrelated items such as the name of someone in an article and the name of a company in a personal email.