Example-based Search: a New Frontier for Exploratory Search

Matteo Lissandrini
Aalborg University
matteo@cs.aau.dk

Davide Mottin
Aarhus University
davide@cs.au.dk

Themis Palpanas
Paris Descartes University
themis@mi.parisdescartes.fr

Yannis Velegrakis
Utrecht University
i.velegrakis@uu.nl

ABSTRACT

Exploration is one of the primordial ways to accrue knowledge about the world and its nature. As we accumulate, mostly automatically, data at unprecedented volumes and speed, our datasets have become complex and hard to understand. In this context, *exploratory search* provides a handy tool for progressively gather the necessary knowledge by starting from a tentative query that can provide cues about the next queries to issue. An exploratory query should be simple enough to avoid complicate declarative languages (such as SQL) and convoluted mechanism, and at the same time retain the flexibility and expressiveness required to express complex information needs. Recently, we have witnessed a rediscovery of the so called *example-based methods*, in which the user, or the analyst circumvent query languages by using examples as input. This shift in semantics has led to a number of methods receiving as query a set of example members of the answer set. The search system then infers the entire answer set based on the given examples and any additional information provided by the underlying database. In this tutorial, we present an excursus over the main example-based methods for exploratory analysis. We show how different data types require different techniques, and present algorithms that are specifically designed for relational, textual, and graph data. We conclude by providing a unifying view of this query-paradigm and identify new exciting research directions.

ACM Reference Format:
Matteo Lissandrini, Davide Mottin, Themis Palpanas, and Yannis Velegrakis. 2019. Example-based Search: a New Frontier for Exploratory Search. In *42nd Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, Article 2, 2 pages. https://doi.org/10.1145/3331184.3331387

1 MOTIVATION

Exploratory search includes methods to efficiently extract knowledge from data repositories, even if we do not know what exactly we are looking for, nor how to precisely describe our needs [24]. The need for new and effective exploratory search methods is particularly relevant given the current abundance and richness of today's large datasets. In common exploratory settings, the user progressively acquires the knowledge by issuing a sequence of generic queries to gather intelligence about the data. However, the existing body of work in data analysis, data visualization, and predictive models, assumes the user is willing to pose several well defined or structured queries to the underlying database in order to progressively gather the required information. This assumption stems from the intuition that the user is accustomed to data analysis techniques. Yet, very often, this assumption is not true.

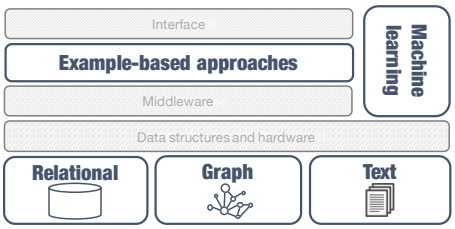

Figure 1: A view of example-based data exploration.

Recently, *examples* became a popular proxy for data exploration. Examples avoid the need for complex query languages. One of the earliest attempts to bring examples as a query method is query-by-example [26]. The main idea was to help the user in the query formulation, allowing them to specify the results in terms of templates for tuples. Nowadays, examples are not anymore a mere template for relational queries, but rather the representative of the intended results the user would like to have. These example-based approaches are fundamentally different from the initial query-by-example idea, and are successfully applied not only to relational data [5, 23], but also to textual [2, 25], and graph [10, 16] data as well.

We note that the flexibility examples provide does not compromise the richness of the results, yet, it can overcome the ambiguity of generic keyword searches, which are frequently found in information retrieval. On the other hand, while data exploration techniques assume the user is willing to pose several exploratory queries, the use of examples allows the searcher to provide more information with less effort, making example-based methods a more palatable choice for novice users, as well as for practitioners. This new functionality can empower existing information retrieval systems with a complementary tool: whenever a query is too complex to be expressed with detailed set of conditions, examples represent a natural alternative. In this respect (cfr. Figure 1) example-based exploration is a middle ground between the user interface, and the data-management layer, enabling new functionalities for the former and allowing more natural exploitation of the latter. Moreover, the use of examples has been demonstrated to be very effective in visual query interfaces [13]. Here we demonstrate how example-based methods can be employed as an expressive and powerful method for exploratory search systems.

2 TUTORIAL OUTLINE

In this tutorial, we provide a detailed overview of the new area of example-based methods for exploratory search, surveying the relevant state-of-the-art techniques. Moreover, we present future directions discussing various machine learning techniques used to

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07.
https://doi.org/10.1145/3331184.3331387

infer user preferences in an online fashion. All the materials are freely available at https://data-exploration.ml.

2.1 Example-based approaches

We survey the main approaches for exploratory queries, highlighting the main differences among data models, and presenting in-depth insights of the current status of research in this area. We first introduce query-by-example [26] as a first attempt to simplify query formulation. In query-by-example the user, instead of explicitly typing a query, specifies the shape of the results in a tabular fashion. We present the main body of work within relational, textual, and graph data.

For *relational data* we provide an overview of techniques that solve various tasks using examples. We show how from examples we can infer fully specified SQL queries through reverse engineering [18, 23]. This very active area has reached maturity discovering both approximate and exact queries with different expressiveness and SQL operators. The use of examples is also beneficial in more complex tasks, such as data integration via schema mapping [1]. More recently, example-based approaches have been used for data cleaning by finding duplicate entities [21] or cleaning rules [8]. Last, we present prototype systems that build upon examples, such as Bleau [19] and QPlain [3].

For *textual data* the techniques include search approaches based on documents used as representatives for the set of results [25], and serendipitous search based on the current visited pages [2]. These approaches focus on documents as examples for retrieving related information. Recently, examples has been successfully employed in entity extraction [7, 20], in which the user provides either mentions of entities in a text [7], or tuples and similarities among attributes [20], and the system automatically returns extraction rules that can be applied to the given dataset.

For *graph data* there are two prominent approaches: the first use subgraphs, or partially specified structures as input examples [4, 10, 12, 16], while the second focuses on the vertices of the graph, which are used for making the selections [11, 17]. Structures convey a more precise information and therefore can be used to quickly prune the search space. Among the existing approaches Exemplar Queries [15, 16] and Graph Query by Example (GQBE) [10] use subgraph isomorphism or structural similarities to identify structures related to the one the user provided. A different approach is the reverse engineering of SPARQL queries [4] in which the input is a set of positive and negative entity mentions in a RDF dataset. Examples can also be employed for targeted analysis of networks, in order to discover communities [11], dense regions [6], or subspaces along outliers [17].

2.2 Machine learning with examples

Current techniques use ad-hoc notions of similarity to retrieve results that are likely to be part of the solution of an unknown query. The current development in machine learning and active search [14, 22] present a different perspective: user preferences can be learned from user interactions instead of manually crafted in the system. Current hardware capabilities allow to process large amount of data, and at the same time dynamically change the internal preference model. One of the earliest work in this direction is MindReader [9] in which the user specifies a set of tuples and optional relevance scores and the system infers a distance function on the objects in the database. The exploration of such *relevance learning* or *metric learning* approaches form the basis of interactive exploratory systems. Moreover, the study of Gaussian Processes as a mean of interactively learning any function given a set of points from the user has recently found applications in graphs [14]. Therefore, we will present a body of work that takes the machine learning perspective into account. The research in this area is still at its infancy and forms a fertile ground for a new generation of data management systems.

REFERENCES

[1] Angela Bonifati, Ugo Comignani, Emmanuel Coquery, and Romuald Thion. 2017. Interactive Mapping Specification with Exemplar Tuples. In *SIGMOD*.
[2] Ilaria Bordino, Gianmarco De Francisci Morales, Ingmar Weber, and Francesco Bonchi. 2013. From machu_picchu to rafting the urubamba river: anticipating information needs via the entity-query graph. In *WSDM*.
[3] D. Deutch and A. Gilad. 2016. QPlain: Query by explanation. In *ICDE*.
[4] Gonzalo Diaz, Marcelo Arenas, and Michael Benedikt. 2016. SPARQLByE: Querying RDF data by example. *Proceedings of the VLDB Endowment* 9, 13 (2016).
[5] Kyriaki Dimitriadou, Olga Papaemmanouil, and Yanlei Diao. 2014. Explore-by-example: An automatic query steering framework for interactive data exploration. In *SIGMOD*. ACM.
[6] Aristides Gionis, Michael Mathioudakis, and Antti Ukkonen. 2015. Bump hunting in the dark: Local discrepancy maximization on graphs. In *ICDE*. 1155–1166.
[7] Maeda F Hanafi, Azza Abouzied, Laura Chiticariu, and Yunyao Li. 2017. Synthesizing Extraction Rules from User Examples with SEER. In *SIGMOD*.
[8] Jian He, Enzo Veltri, Donatello Santoro, Guoliang Li, Giansalvatore Mecca, Paolo Papotti, and Nan Tang. 2016. Interactive and deterministic data cleaning. In *SIGMOD*.
[9] Yoshiharu Ishikawa, Ravishankar Subramanya, and Christos Faloutsos. 1998. MindReader: Querying databases through multiple examples. In *VLDB*.
[10] Nandish Jayaram, Arijit Khan, Chengkai Li, Xifeng Yan, and Ramez Elmasri. 2015. Querying knowledge graphs by example entity tuples. *TKDE* 27, 10 (2015).
[11] Isabel M Kloumann and Jon M Kleinberg. 2014. Community membership identification from small seed sets. In *KDD*.
[12] Matteo Lissandrini, Davide Mottin, Themis Palpanas, and Yannis Velegrakis. 2018. Multi-Example Search in Rich Information Graphs. In *ICDE*.
[13] Matteo Lissandrini, Davide Mottin, Themis Palpanas, and Yannis Velegrakis. 2018. X2Q: Your Personal Example-based Graph Explorer. In *PVLDB*. ACM, 901–904.
[14] Yifei Ma, Tzu-Kuo Huang, and Jeff G Schneider. 2015. Active Search and Bandits on Graphs using Sigma-Optimality. In *UAI*. 542–551.
[15] Davide Mottin, Matteo Lissandrini, Yannis Velegrakis, and Themis Palpanas. 2014. Searching with xq: the exemplar query search engine. In *SIGMOD*. ACM.
[16] Davide Mottin, Matteo Lissandrini, Yannis Velegrakis, and Themis Palpanas. 2016. Exemplar queries: a new way of searching. *VLDB J.* (2016).
[17] Bryan Perozzi, Leman Akoglu, Patricia Iglesias Sánchez, and Emmanuel Müller. 2014. Focused clustering and outlier detection in large attributed graphs. In *KDD*. 1346–1355.
[18] Fotis Psallidas, Bolin Ding, Kaushik Chakrabarti, and Surajit Chaudhuri. 2015. S4: Top-k Spreadsheet-Style Search for Query Discovery. In *SIGMOD*. 2001–2016.
[19] Thibault Sellam and Martin Kersten. 2016. Cluster-driven navigation of the query space. *TKDE* 28, 5 (2016).
[20] Rishabh Singh. 2016. Blinkfill: Semi-supervised programming by example for syntactic string transformations. *PVLDB* 9, 10 (2016).
[21] Rohit Singh, Venkata Vamsikrishna Meduri, Ahmed Elmagarmid, Samuel Madden, Paolo Papotti, Jorge-Arnulfo Quiané-Ruiz, Armando Solar-Lezama, and Nan Tang. 2017. Synthesizing entity matching rules by examples. *PVLDB* 11, 2 (2017).
[22] Yu Su, Shengqi Yang, Huan Sun, Mudhakar Srivatsa, Sue Kase, Michelle Vanni, and Xifeng Yan. 2015. Exploiting relevance feedback in knowledge graph search. In *KDD*.
[23] Yaacov Y Weiss and Sara Cohen. 2017. Reverse Engineering SPJ-Queries from Examples. In *SIGMOD*.
[24] Ryen W White and Resa A Roth. 2009. Exploratory search: Beyond the query-response paradigm. *Synthesis lectures on information concepts, retrieval, and services* 1, 1 (2009).
[25] Mohamed Yakout, Kris Ganjam, Kaushik Chakrabarti, and Surajit Chaudhuri. 2012. InfoGather: Entity Augmentation and Attribute Discovery by Holistic Matching with Web Tables. In *SIGMOD*.
[26] Moshé M. Zloof. 1975. Query by Example. In *AFIPS NCC*. 431–438.