Example-based Exploration: Exploring Knowledge through Examples

Matteo Lissandrini
Aalborg University
matteo@cs.aau.dk

Davide Mottin
Aarhus University
davide@cs.au.dk

Themis Palpanas
Unversity of Paris
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 to 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 or SPARQL) 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.

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 [49]. The need for new and effective exploratory search methods is particularly relevant given the current abundance and richness of today's large datasets (e.g., Linked Open Datasets). In common exploratory settings, the user progressively acquires the knowledge by issuing a sequence of generic queries to gather intelligence about the data. The existing body of work in data analysis assumes the user is willing to pose several well defined or structured queries to the underlying database to progressively gather the required information. This assumption stems from the intuition that the user is accustomed to data analysis techniques. Yet, this assumption is not always true.

Recently, *examples* became a popular proxy for data exploration. Examples avoid the need for complex query languages (e.g., SQL or SPARQL). One of the earliest attempts to bring examples as a query method is query-by-example (QBE) [54]. The main idea was to help the user in the query formulation, allowing them to specify the results in terms of templates for tuples. This idea, though, was particularly bounded to the relational model. 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 QBE, and are successfully applied not only to relational data [11, 48], but also to textual [7, 51], and graph [9, 16, 28] data as well, with several applications to semantic web data.

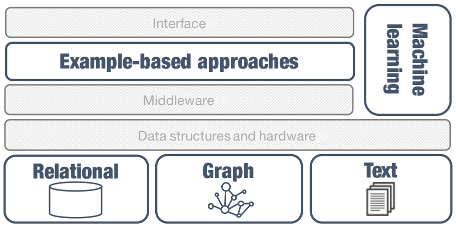

Figure 1: A view of example-based data exploration in relation to the broader scope of data-management techniques.

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 search systems with a complementary tool: whenever a query is too complex to be expressed with a detailed set of conditions, examples represent a natural alternative. In this respect (see 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 [21]. We describe how example-based methods can be employed as an expressive and powerful method for exploratory search systems that naturally enhance user capabilities in accessing information from web documents, knowledge graphs, and other datasets.

2 OBJECTIVES

We survey the main approaches for exploratory queries, highlighting the main differences among data models, and presenting in-depth insights into the current status of research in this area. The

ESWC '20, May 31– June 4, 2020, Heraklion, Crete, Greece

final goal is to provide a comprehensive overview of novel data-management techniques that can empower advanced exploratory search systems. In particular, we will highlight the existing example-based methods that have been already studied to improve knowledge graph search, SPARQL query formulation, and data exploration of RDF data. Moreover, we aim to present techniques that have been studied in other research areas and that could be successfully applied in the Semantic Web domain.

The first and second part of the tutorial introduce the broad topic of data exploration, highlighting the hardness of query languages for simple users and advocating the need for different query methods. We will introduce the example-based methods as flexible delegates for more complex search tasks that would otherwise need to be expressed through very complex traditional queries. In this part, we will discuss various cases where queries cannot be expressed in declarative languages without requiring complex constructs. We will also present an expressive formulation of example-based approaches as seeking a similarity among objects. This part will also introduce the concept of *query reverse engineering* and the original solution for reverse engineering of SQL queries from examples and more advanced techniques. These concepts will provide the necessary background to understand similar solutions for SPARQL queries that will be presented later.

The third part of the tutorial discusses the current main techniques for textual and graph data to provide a complete picture of the power of the approach. In this part, we will present the algorithms, show how they work, and demonstrate their ability to (conceptually) solve complex search tasks (e.g., goal-oriented search, focused community retrieval, graph search) from simple examples. We will also highlight the differences among data models, focusing on the scalability perspective, presenting the motivations and drawing parallels among methods for different data types.

The techniques for *text exploration* include search approaches based on documents used as representatives for the set of results [51], and serendipitous search based on the current visited pages [7]. These approaches focus on documents as examples for retrieving related information and are well versed to be expanded with additional information (e.g., from graphs and ontologies in the Semantic Web). Recently, examples have been successfully employed in entity extraction [14, 41], in which the user provides either mentions of entities in a text [14], or tuples and similarities among attributes [41], 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 [9, 16, 20, 28], while the second focuses on the vertices of the graph (usually entities [26]), which are used for making the selections [17, 34]. Among the existing approaches Exemplar Queries [27, 28] and Graph Query by Example (GQBE) [16] 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 [9] in which the input is a set of positive and negative entity mentions in an RDF dataset. Examples can also be employed for targeted analysis of networks, to discover communities [17], dense regions [12], or subspaces along with outliers [34].

Particular focus will be given on *how example-based search can exploit knowledge graphs* to provide semantic search capabilities for documents as well as on example-based methods for *entity and structure search within knowledge graphs*.

The fourth part of the tutorial focuses on the latest developments of machine learning to progressively discover user intention. We will introduce the general area of online learning, some early methods based on relevance feedback [15], and show some recent applications of multi-armed bandits theories that include active search [24, 43].

Challenges and open research questions. The last part of the tutorial is dedicated to the challenges and open research questions. Exploratory search based on examples is rapidly attracting attention and getting traction, though, the support for such techniques in modern search and data management systems is lagging behind. The need for understanding the semantics behind the user query is also of paramount importance. Some challenges have already been discussed in recent vision papers [4, 47, 50]. Finally, we will conclude the tutorial with remarks about the current state of affairs, and engage the audience in a discussion about their experiences with needs, tools, and challenges in this area.

Tutorial outline:

I. **Introduction, motivation, and formulation**
   - Why example-based approaches are important
     - Usefulness of exploratory analysis
     - Main characteristics of exploratory analysis
     - Example-based methods for exploratory analysis
     - Use cases of failing keyword and declarative queries
     - Applications in current data management
   - Connection to data exploration
   - Problem formulation as similarity discovery

II. **The origin: Example-based approaches for structured data**
   - Query-by-example: [54]
   - Example methods in relational databases:
     - Reverse engineering of SQL queries [18, 32, 35, 40, 44, 45, 48, 52];
     - Schema mapping [1, 6, 13];
     - Exploratory Analytics [8, 38, 39].

III. **Example-based approaches for semi-structured and unstructured data**
   - Example methods in textual data:
     - Exploring Web documents as examples [7, 53];
     - Example based Entity and Relation extraction [14, 41];
     - Web table search and augmentation [51];
     - Goal oriented content discovery [33];
   - Example methods in graphs:
     - Cluster and Community exploration by Example Nodes [12, 17, 34, 37];
     - Semantic Entity Search [26, 42];
     - Reverse Engineering Path Queries [5] and SPARQL [2, 9] from Examples;
     - Example-based Knowledge Graph search [16, 20, 23, 28].

IV. **Learning methods based on examples**
   - Passive similarity learning: MindReader [15]
   - Active learning:

  - Multi-armed bandits and the Upper Confindence Bound algorithm [3]
  - Gaussian processes and GP-Select [46]
  - Relevance feedback learning [11] and for graphs [24, 43]

V. **Challenges and Discussion**

- Can we *interactively* assist the user toward the retrieval of the correct answer?
- Can we provide *explanations* for the query results?
- How can machine learning help in exploratory analysis?
- Can we build a Personal Knowledge Assistant?

3 RELEVANCE

The topic of exploration has been of interest in many communities related to information retrieval, data management, and semantic web for many years now [9, 26, 49]. Exploratory search involves the study of information retrieval paradigms that move the process beyond predictable fact retrieval [25]. The large availability of knowledge graphs and open data provide both abundant resources and unique challenges to users aiming to find relevant information on the web. This tutorial represents a research bridge across data-management, information-retrieval, and semantic web techniques. In particular, this tutorial will show how to combine results from research areas that are already prominent in the semantic web community (e.g., search and retrieval, knowledge graphs, and machine learning) to novel techniques based on *example driven* query paradigms from the data management world to the benefit of enabling user-friendly knowledge exploration. Past tutorials that cover relevant topics are for instance, "Utilizing Knowledge Graphs in Text-centric Information Retrieval"[10] by Dietz et al., presented at SIGIR 2018 (and earlier at WSDM 2017); "Graph Exploration: Let me Show what is Relevant in your Graph" [31] by Mottin and Müller at KDD 2018; and "Information Discovery in E-commerce" [36] by Ren et al. at SIGIR 2018. Yet, none of them focuses on the topic of exploratory search in general, nor they cover example-driven query paradigms in detail.

In contrast, this tutorial builds upon the earlier "New Trends on Exploratory Methods for Data Analytics" presented at VLDB 2017 [29] that has been expanded with the material from the book "Data Exploration using Example-based Methods" [19] for SIGMOD 2019 [30], as well as on the material of the tutorial on Exploratory Search presented at SIGIR 2019 [22] and will introduce the audience to these novel methods to empower data exploration for and with semantic web resources and knowledge graphs.

REFERENCES

[1] Bogdan Alexe, Balder Ten Cate, Phokion G Kolaitis, and Wang-Chiew Tan. 2011. Designing and refining schema mappings via data examples. In *SIGMOD*.

[2] Marcelo Arenas, Gonzalo I Diaz, and Egor V Kostylev. 2016. Reverse engineering SPARQL queries. In *WWW*.

[3] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47, 2-3 (2002).

[4] Krisztian Balog and Tom Kenter. 2019. Personal Knowledge Graphs: A Research Agenda. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '19)*. ACM, New York, NY, USA, 217–220.

[5] Angela Bonifati, Radu Ciucanu, and Aurélien Lemay. 2015. Learning path queries on graph databases. In *EDBT*.

[6] Angela Bonifati, Ugo Comignani, Emmanuel Coquery, and Romuald Thion. 2017. Interactive Mapping Specification with Exemplar Tuples. In *SIGMOD*.

[7] 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*.

[8] D. Deutch and A. Gilad. 2016. QPlain: Query by explanation. In *ICDE*.

[9] Gonzalo Diaz, Marcelo Arenas, and Michael Benedikt. 2016. SPARQLByE: Querying RDF data by example. *Proceedings of the VLDB Endowment* 9, 13 (2016).

[10] Laura Dietz, Alexander Kotov, and Edgar Meij. 2018. Utilizing Knowledge Graphs for Text-Centric Information Retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM.

[11] Kyriaki Dimitriadou, Olga Papaemmanouil, and Yanlei Diao. 2014. Explore-by-example: An automatic query steering framework for interactive data exploration. In *SIGMOD*. ACM.

[12] Aristides Gionis, Michael Mathioudakis, and Antti Ukkonen. 2015. Bump hunting in the dark: Local discrepancy maximization on graphs. In *ICDE*. 1155–1166.

[13] Georg Gottlob and Pierre Senellart. 2010. Schema mapping discovery from data instances. *JACM* 57, 2 (2010).

[14] Maeda F Hanafi, Azza Abouzied, Laura Chiticariu, and Yunyao Li. 2017. Synthesizing Extraction Rules from User Examples with SEER. In *SIGMOD*.

[15] Yoshiharu Ishikawa, Ravishankar Subramanya, and Christos Faloutsos. 1998. MindReader: Querying databases through multiple examples. In *VLDB*.

[16] Nandish Jayaram, Arijit Khan, Chengkai Li, Xifeng Yan, and Ramez Elmasri. 2015. Querying knowledge graphs by example entity tuples. *TKDE* 27, 10 (2015).

[17] Isabel M Kloumann and Jon M Kleinberg. 2014. Community membership identification from small seed sets. In *KDD*.

[18] Hao Li, Chee-Yong Chan, and David Maier. 2015. Query from examples: An iterative, data-driven approach to query construction. *PVLDB* 8, 13 (2015).

[19] Matteo Lissandrini, Davide Mottin, Themis Palpanas, and Yannis Velegrakis. 2018. *Data Exploration Using Example-Based Methods*. Synthesis Lectures on Data Management, Vol. 10. Morgan & Claypool Publishers.

[20] Matteo Lissandrini, Davide Mottin, Themis Palpanas, and Yannis Velegrakis. 2018. Multi-Example Search in Rich Information Graphs. In *ICDE*.

[21] Matteo Lissandrini, Davide Mottin, Themis Palpanas, and Yannis Velegrakis. 2018. X2Q: Your Personal Example-based Graph Explorer. In *PVLDB*. ACM, 901–904.

[22] Matteo Lissandrini, Davide Mottin, Themis Palpanas, and Yannis Velegrakis. 2019. Example-based Search: A New Frontier for Exploratory Search *(SIGIR'19)*. ACM, New York, NY, USA, 1411–1412.

[23] Matteo Lissandrini, Davide Mottin, Themis Palpanas, and Yannis Velegrakis. 2020. Graph-Query Suggestions for Knowledge Graph Exploration. In *World Wide Web Conference (WWW 2020)*.

[24] Yifei Ma, Tzu-Kuo Huang, and Jeff G Schneider. 2015. Active Search and Bandits on Graphs using Sigma-Optimality. In *UAI*. 542–551.

[25] Gary Marchionini. 2006. Exploratory Search: From Finding to Understanding. *Commun. ACM* 49, 4 (April 2006), 41–46.

[26] Steffen Metzger, Ralf Schenkel, and Marcin Sydow. 2013. QBEES: query by entity examples. In *CIKM*. 1829–1832.

[27] Davide Mottin, Matteo Lissandrini, Yannis Velegrakis, and Themis Palpanas. 2014. Searching with xq: the exemplar query search engine. In *SIGMOD*. ACM.

[28] Davide Mottin, Matteo Lissandrini, Yannis Velegrakis, and Themis Palpanas. 2016. Exemplar queries: a new way of searching. *VLDB J.* (2016).

[29] Davide Mottin, Matteo Lissandrini, Yannis Velegrakis, and Themis Palpanas. 2017. New trends on exploratory methods for data analytics. *PVLDB* 10, 12 (2017).

[30] Davide Mottin, Matteo Lissandrini, Yannis Velegrakis, and Themis Palpanas. 2019. Exploring the Data Wilderness Through Examples *(SIGMOD '19)*. ACM, New York, NY, USA, 2031–2035.

[31] Davide Mottin and Emmanuel Müller. 2018. Graph Exploration: Let me Show what is Relevant in your Graph *(KDD '18)*.

[32] Kiril Panev and Sebastian Michel. 2016. Reverse Engineering Top-k Database Queries with PALEO.. In *EDBT*.

[33] Dimitra Papadimitriou, Georgia Koutrika, Yannis Velegrakis, and John Mylopoulos. 2017. Finding related forum posts through content similarity over intention-based segmentation. *TKDE* 29, 9 (2017).

[34] Bryan Perozzi, Leman Akoglu, Patricia Iglesias Sánchez, and Emmanuel Müller. 2014. Focused clustering and outlier detection in large attributed graphs. In *KDD*. 1346–1355.

[35] Fotis Psallidas, Bolin Ding, Kaushik Chakrabarti, and Surajit Chaudhuri. 2015. S4: Top-k Spreadsheet-Style Search for Query Discovery. In *SIGMOD*. 2001–2016.

[36] Zhaochun Ren, Xiangnan He, Dawei Yin, and Maarten de Rijke. 2018. Information Discovery in E-commerce: Half-day SIGIR 2018 Tutorial. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*. 1379–1382. https://doi.org/10.1145/3209978.3210185

[37] Natali Ruchansky, Francesco Bonchi, David García-Soriano, Francesco Gullo, and Nicolas Kourtellis. 2015. The Minimum Wiener Connector Problem. In *SIGMOD*. 1587–1602.

[38] Thibault Sellam and Martin Kersten. 2016. Cluster-driven navigation of the query space. *TKDE* 28, 5 (2016).

[39] Thibault Sellam and Martin Kersten. 2016. Ziggy: Characterizing query results for data explorers. *PVLDB* 9, 13 (2016).

[40] Yanyan Shen, Kaushik Chakrabarti, Surajit Chaudhuri, Bolin Ding, and Lev Novik. 2014. Discovering Queries Based on Example Tuples. In *SIGMOD*. 493–504.

[41] Rishabh Singh. 2016. Blinkfill: Semi-supervised programming by example for syntactic string transformations. *PVLDB* 9, 10 (2016).

[42] Grzegorz Sobczak, Mateusz Chochół, Ralf Schenkel, and Marcin Sydow. 2015. iQbees: Towards Interactive Semantic Entity Search Based on Maximal Aspects. In *Foundations of Intelligent Systems*. Springer.

[43] 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*.

[44] Quoc Trung Tran, Chee-Yong Chan, and Srinivasan Parthasarathy. 2009. Query by output. In *SIGMOD*.

[45] Quoc Trung Tran, Chee-Yong Chan, and Srinivasan Parthasarathy. 2014. Query reverse engineering. *VLDB J.* 23, 5 (2014).

[46] Hastagiri P. Vanchinathan, Andreas Marfurt, Charles-Antoine Robelin, Donald Kossmann, and Andreas Krause. 2015. Discovering Valuable Items from Massive Data. In *KDD*. 1195–1204.

[47] Abdul Wasay, Manos Athanassoulis, and Stratos Idreos. 2015. Queriosity: Automated Data Exploration. In *Proceedings of the IEEE International Congress on Big Data*.

[48] Yaacov Y Weiss and Sara Cohen. 2017. Reverse Engineering SPJ-Queries from Examples. In *SIGMOD*.

[49] 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).

[50] Eugene Wu, Leilani Battle, and Samuel R. Madden. 2014. The Case for Data Visualization Management Systems: Vision Paper. *Proc. VLDB Endow.* 7, 10 (June 2014), 903–906.

[51] Mohamed Yakout, Kris Ganjam, Kaushik Chakrabarti, and Surajit Chaudhuri. 2012. InfoGather: Entity Augmentation and Attribute Discovery by Holistic Matching with Web Tables. In *SIGMOD*.

[52] Meihui Zhang, Hazem Elmeleegy, Cecilia M Procopiuc, and Divesh Srivastava. 2013. Reverse engineering complex join queries. In *SIGMOD*.

[53] Mingzhu Zhu and Yi-Fang Brook Wu. 2014. Search by Multiple Examples. In *WSDM*. 667–672.

[54] Moshé M. Zloof. 1975. Query by Example. In *AFIPS NCC*. 431–438.