Interactive Data Discovery in Data Lakes

Andra Ionescu
supervised by Geert-Jan Houben and Asterios Katsifodimos
Delft University of Technology
Delft, Netherlands
a.ionescu-3@tudelft.nl

ABSTRACT

As data is produced at an unprecedented rate, the need and expectation to make it easily available for the end-users is growing. Dataset Discovery has become an important subject in the data management community, as it represents the means of providing the data to the user and fulfilling an information need. Since the end-user is the one that needs the information and knows what type of information to look for, little has been done to involve the user in the discovery process.

This PhD project addresses the topic of *interactive data discovery*, where the user's interests are modelled through interactions and used as a context for the discovery process. We aim to develop a system that addresses the problem of minimising the trade-off between efficiency and effectiveness, thus providing accurate results in an interactive fashion. The innovative part of the system consists of extracting the user's interests and data needs through interactions and using them to enrich the data context and provide tailored results to the user. We describe the steps to create models and methods that would be used in designing the prototype and we relate to previous systems and neighbouring communities for optimising the system.

1 INTRODUCTION

It is often said that "data is the new oil". We have the means to produce and collect data at a fast pace and new repositories emerge such as the data lakes. Having this vast amount of data, both expert and non-expert users expect to create accurate insights much easier and faster [8]. To support the users, we need to extract knowledge out of the data, and thus much work has been done in all the areas concerning data processing. One of the first steps in the pipeline is data discovery, which gives meaningful input for data integration tasks [21].

Data discovery is the process of finding relevant data among thousands of disparate heterogeneous datasets. The relevance is defined by the users, it is based on their needs and it often means finding joinable or unionable datasets [12]. However, data discovery is a tedious process. In some cases, the user inspects the data manually in order to find the relevant information, while in other cases, the data discovery algorithms provide the most accurate results, although not relevant for the user. Both scenarios have one common aspect: the user involvement in the process. In previous works, the user involvement is limited. The users can perform a keyword search [7, 22] or provide a base table as a starting point of the discovery process [5, 10, 20, 25, 27]. More sophisticated systems offer the user a broader space of interactions such as using a specific query language [12] or data science notebooks [26].

Interactive Data Exploration is another research area where the users actively engage with the system. Interactive data exploration is the process of extracting knowledge from data when the users explore the data space without a specific target. Thus, the users engage in an iterative process where they pose queries, review the result, adjust the query and repeat the cycle until they decide to terminate the process [11].

The main difference between the two processes, the discovery and the exploration, is the user's data need. In the discovery process, the users have an information need, but are unaware of the location of the data. In the exploration process, they are unaware of both the kind of information that resides in the repository and the kind of information that they need.

We focus on data discovery in data lakes and therefore we assume that the users have an information need. Providing them with relevant datasets implies that the data lake contains sufficient information about the datasets such as comments and documentation about the entities, their semantic types, and the relations among datasets. However, data lakes only provide minimum metadata about the datasets, thus they lack the contextual information, which aids in finding related datasets. In the rest of the paper we use *context* to refer to the contextual information, which means additional data about the datasets. The works focused on the web tables use as context the text surrounding the tables to infer the topic and determine which tables match [22, 24]. To address this problem in data lakes, we plan to leverage the interactivity of exploration systems to provide context for data discovery. Thus, we propose a fusion between the areas and we introduce the concept of Interactive Data Discovery (IDD). By interacting with the system, the users can provide valuable information which we can leverage to enhance the discovery process. Through user interactions, we enrich the data with more context and teach the system about the user preferences. As search engines use past searches and users' interests to tailor the results, an interactive data discovery system uses the user interactions to tailor the recommendations.

However, involving the user raises a new set of challenges. Firstly, we have to take into account the interaction time and the fact that the users lose their patience and interest after one second [14, 16]. Secondly, we face with the trade-off between the efficiency and effectiveness [14], where improving one metric results in compromising the other one. Finally, the interactions must not exhaust the users [3, 17], but they should help them in reaching their goal, therefore a careful design must be employed.

Proceedings of the VLDB 2021 PhD Workshop, August 16th, 2021. Copenhagen, Denmark.

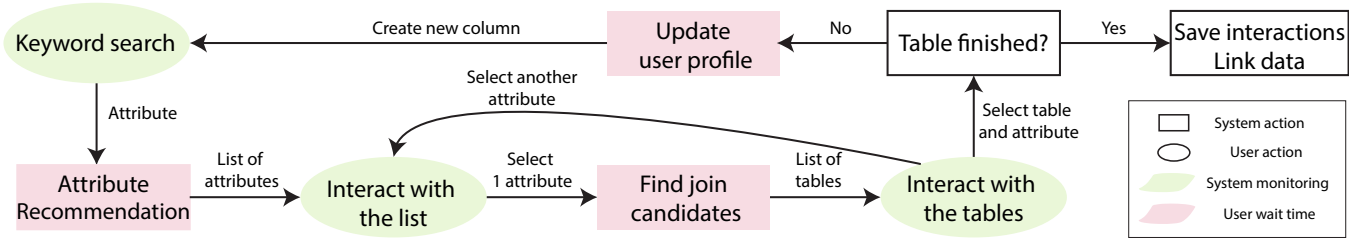

Figure 1: User-system workflow

With this PhD project, we aim at tackling the above challenges by making the following contributions:

- We aim to devise a strategy to create easy, meaningful and sufficient interactions to support both the user and the system in the discovery process.
- We will research ranking mechanisms based on a rich graph structure containing the key-foreign key constraints, similarities between attributes and datasets, the local relevance of a column (being part of a table) as well as the user interactions.
- We will research a model to link the data in order to save the connections created during a session for the future users.
- We aim to achieve interactive speeds (execution time per iteration smaller than 1000ms [19, 23]) without significant losses in accuracy.

Following, we present the research goal in Section 2 and the research overview in Section 3. We describe the challenges we meet to achieve the goal and the shortcoming of existing methods in Section 4 and conclude in Section 5.

2 RESEARCH GOAL

The goal of this PhD research is to improve the usability and performance of the data discovery process. On one hand, we aim to improve the performance by achieving high accuracy while decreasing the latency and the user effort. On the other hand, we aim to improve the usability by providing the exact result to the users, instead of a list with the top-k results as it is often done in the literature.

Currently, the data discovery systems are faced with the trade-off between effectiveness and efficiency. Moreover, they return uniform impersonal results, disregarding the users' needs and offering a limited involvement in the process. These systems focus on automating the process and minimising the user input as much as possible [10]. Additionally, the users can not influence the automated discovery process, thus their needs are not captured. Often times the only input consists of a target table used to find the most relevant datasets [5, 10, 20, 24, 27]. Moreover, the process finishes with a list of top-k results that can be integrated with the target table, thus the result is not tailored to the users' data needs [7, 20, 24]. On the other hand, the interactive data exploration systems involve the user in the process. However, these systems achieve high accuracy by increasing the user effort [11].

We propose an interactive system, where the users become first-class citizens. They are in control of the discovery process and by interacting with the system they provide meaningful information about their interests. This information is further modelled and leveraged by the system in order to retrieve the most relevant datasets. We aim to combine these datasets through user interactions and to return to the user one single relevant table.

3 RESEARCH OVERVIEW

We envision a system that helps the users create tables with N attributes based on keywords. The exploration is clearly defined and it is tailored to assist the user in the process in an iterative manner. The system uses the existing information from the data lake to create the desired result and presents the most relevant datasets to the user based on a number of specific interactions. Through these interactions, the users find and combine the attributes in a single table. We propose an iterative process, where the user and the system engage in a conversation as illustrated in Figure 1.

The user starts the exploration with a keyword and waits for the systems to provide similar attribute names existing in the repository. Next, the user selects one attribute from the list of recommendations and the system starts searching for join candidates. The user receives a list of join candidates and selects a table together with the attributes that might help in the next search. Finally, the process continues with the same steps until the user explicitly stops it. Once the process terminates, the system links the data and saves the interactions for future users.

During this process, the focus is on extracting information through interactions, illustrated in Figure 1 by the system monitoring the user. Furthermore, we focus on modelling this information such that the system recommends meaningful attributes and datasets to the user. Next, we want to capture the interactions and use mechanism for short-term memory to understand the user's interests and long-term memory to help other users with similar interests [4]. We need to understand the users' interests in order to help them find the right information in minimum N iterations, where N represents the number of attributes. Finally, we aim to minimise the user wait time, which represents the moments when the system should perform at interactive speeds, as illustrated in Figure 1.

4 CHALLENGES

In this section, we present the strategy to achieve our goal and relate to current solutions from the literature. We present their limitations, the challenges to build our system and propose solutions to overcome them.

4.1 User Interactions

Most data discovery systems use similar user interactions, such as providing a target table to find the relevant datasets [5, 10, 20, 24, 27], providing a key attribute or indicating the number of expected results [24, 27]. We envision a process where the user reuses the information from a data lake, without indicating a source table or a key column. Therefore, the users should use keywords in an iterative fashion to describe the attributes they wish to find. Furthermore, by interacting with the results, we aim to infer the interests and provide more relevant datsets in the next iterations. Alternative interactive systems use Jupyter Notebooks [26] or active learning to increase the relevance of results [6], while Aurum [12] offers a query language that permits the user to pose various queries. Octopus [7] offers specific actions to enrich a dataset such as asking for more context, extending a table with more rows and concatenating multiple datasets.

Although some of these data discovery systems involve the user in the process, they do not use the interactions to develop a user profile. We plan on extracting information from the user interactions and develop a user profile to model the users behaviour and interests. With this, we aim to provide more accurate and informative results for the user. However, the human-in-the-loop approach presents a number of challenges:

- What is the minimum number of iterations that offers sufficient information for the system while keeping the user engaged in the process?
- What kind of information can we extract from the interactions in order to model the user's interests?

To address these challenges we plan on experimenting with techniques from information retrieval regarding the user search behaviour [1, 2] as well as the ones from human-computer interaction related to the design of user interfaces and modelling a user profile [18]. Some of the interactions that can provide information about the user interests are click and hover, the click-through rate, the mouse movement. Another factor to consider in modelling the behaviour is the time spent on each of these interactions.

4.2 Modelling Interactions

The effectiveness of an algorithm can be summarised as the approach used to retrieve the most relevant datasets. One such popular approach is the top-k ranking. A ranking function replaces the need of a threshold and eases the discovery process, as the users do not need prior knowledge to indicate the number of results. In data discovery, top-k ranking is used in combination with intersection estimation [27], graph matching [20] and LSH indexes [5]. Besides top-k ranking, other methods used to retrieve joinable or unionable datasets are sorted lists [7], graph traversal [12] or machine learning techniques [10].

These approaches use the similarities between the datasets under different signals in order to retrieve joinable or unionable datasets. However, the similarity is highly dependent on the quality of the data, as shown in our previous work [15], where each algorithm shows high effectiveness given a certain data configuration. Thus, we must understand how the algorithms perceive the data, as it might appear very dissimilar to algorithms, but quite similar to the human eye.

The systems using web tables overcome this problem by giving more context to the data, such as the text surrounding the tables, the title of the web page, the URL [7, 22]. In data lakes, the context is minimum and it is based on the metadata usually extracted after the ingestion phase [5, 13]. As the search systems use the user profiles to enhance the search process, we propose (i) to employ the user in order to generate the context and (ii) to create a ranking function not only based on the similarities between the datasets or the explicit relations, but also on the user interactions. However, such an approach poses the next challenges:

- What type of data structure is most suitable to model the user interactions as the context for the datasets in a data lake?
- What is the aggregation function that allows us to combine several signals in a single ranking function in order to increases the accuracy of the system?

Our hypothesis is that a graph structure will allow us to model both the context and the signals as depicted in Figure 2. By context we refer to the user's interests, while by signals we relate to the data profiles and similarities between the datasets. We will employ several machine learning algorithms both linear and based on decision trees to achieve a ranking function with a high accuracy based on the the context and signals.

4.3 Linking the Data

The purpose of data discovery is to facilitate the access to the data for the integration tasks, such as schema matching, merging, mapping and query reformulation [21]. The data discovery systems often combine both the access to the data and the integration, by modelling and developing algorithms for matching the datasets.

We envision one more step, where the semantic relations are captured, thus working on creating a mapping. As depicted in Figure 1, the last step in the workflow consists of saving all the interactions and linking the data. The challenge consists of creating the proper signals such that we enrich the data structure created in Section 4.2. This new information should help the next users in their discovery process when the system lacks information about their interests (in the beginning of the process, before building the user profile).

Solving the challenges from Section 4.1 and Section 4.2 should point us in the right direction for extracting and modelling the signals to create a mapping and integrate it in the data model used for ranking.

4.4 Interactive Speed

In the most basic approach, the data discovery is an $O(N^2)$ problem, where every candidate is matched against another. However, the execution time of the data discovery algorithms becomes a concern in a large repository that contains thousands or millions of tuples. Therefore, various approaches have been proposed to minimise the sample space. Indexing is one of the preferred methods, such as using LSH [5, 12, 20] or creating particular indexing solutions based on Elasticsearch[1] [24, 26]. Other approaches leverage information retrieval algorithms [27], active learning sampling techniques [11], approximate query processing [9], search engines in combination

[1] https://www.elastic.co/elasticsearch

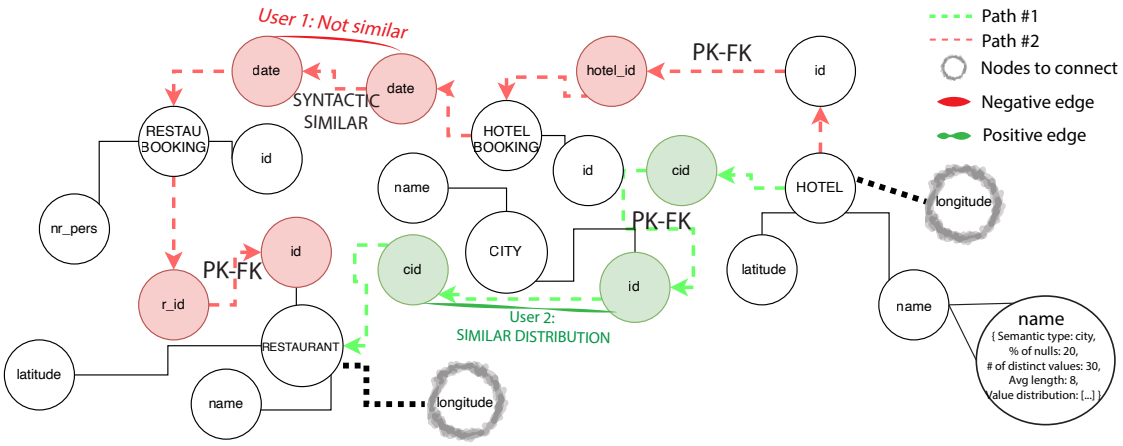

Figure 2: Graph example showing different signals such as similarities between attributes, data profiles and the context captured from user input.

with probabilities [7] or coreset construction [10]. Each method presents features that help finding the information faster, some compromising the accuracy while some increasing the amount of interactions. Therefore the challenge is maintaining the accuracy, while improving the interactivity without increasing the user effort.

We will follow an empirical methodology to assess what method or which combination of methods helps us create an interactive and effective system. Moreover, we will take into account the power of distributed systems and we envision a hybrid approach consisting of sample space minimisation and distributed algorithms.

5 CONCLUSION

In conclusion, this paper presents the plan for a PhD project that aims to improve the data discovery process by using the user interactions. We introduce the concept of Interactive Data Discovery and propose a system to model this process. We address several challenges that concern the user interactions and the methodology to achieve high accuracy and low latency. Finally, we are interested in minimising the trade-off between the efficiency and effectiveness by leveraging the user interactions without a significant increase in the user effort.

ACKNOWLEDGMENTS

This work has been partially funded by the H2020 project Opertus-Mundi No. 870228.

REFERENCES

[1] Eugene Agichtein, Eric Brill, Susan Dumais, and Robert Ragno. 2006. Learning user interaction models for predicting web search result preferences. In *SIGIR*. 3–10.
[2] Leif Azzopardi. 2016. Simulation of interaction: A tutorial on modelling and simulating user interaction and search behaviour. In *SIGIR*. 1227–1230.
[3] Nigel Bevana, Jurek Kirakowskib, and Jonathan Maissela. 1991. What is usability. In *HCI*.
[4] Daniel Billsus and Michael J Pazzani. 2014. A Hybrid User Model for News Story Classification. In *CISM UM99*. 99.
[5] Alex Bogatu, Alvaro A.A. Fernandes, Norman W. Paton, and Nikolaos Konstantinou. 2020. Dataset discovery in data lakes. *ICDE*, 709–720.
[6] Angela Bonifati, Radu Ciucanu, and Slawomir Staworko. 2014. Interactive join query inference with JIM. *PVLDB* (2014), 1541–1544.
[7] Michael J. Cafarella, Alon Halevy, and Nodira Khoussainova. 2009. Data integration for the relational web. *PVLDB* (2009), 1090–1101.
[8] Ugur Cetintemel, Mitch Cherniack, Justin Debrabant, Yanlei Diao, Kyriaki Dimitriadou, Alex Kalinin, Olga Papaemmanouil, and Stan Zdonik. 2013. Query Steering for Interactive Data Exploration. In *CIDR*.
[9] Surajit Chaudhuri, Bolin Ding, and Srikanth Kandula. 2017. Approximate query processing: No silver bullet. In *SIGMOD*. 511–519.
[10] Nadiia Chepurko, Ryan Marcus, Emanuel Zgraggen, Raul Castro Fernandez, Tim Kraska, and David Karger. 2020. ARDA: automatic relational data augmentation for machine learning. *PVLDB* (2020), 1373–1387.
[11] Kyriaki Dimitriadou, Olga Papaemmanouil, and Yanlei Diao. 2016. AIDE: an active learning-based approach for interactive data exploration. *TKDE* (2016), 2842–2856.
[12] Raul Castro Fernandez, Ziawasch Abedjan, Famien Koko, Gina Yuan, Samuel Madden, and Michael Stonebraker. 2018. Aurum: A data discovery system. In *ICDE*. 1001–1012.
[13] Rihan Hai, Sandra Geisler, and Christoph Quix. 2016. Constance: An intelligent data lake system. *SIGMOD* 26-June-20, 2097–2100.
[14] Niranjan Kamat, Prasanth Jayachandran, Karthik Tunga, and Arnab Nandi. 2014. Distributed and interactive cube exploration. In *ICDE*. 472–483.
[15] Christos Koutras, George Siachamis, Andra Ionescu, Kyriakos Psarakis, Jerry Brons, Marios Fragkoulis, Christoph Lofi, Angela Bonifati, and Asterios Katsifodimos. 2021. Valentine: Evaluating Matching Techniques for Dataset Discovery. , to appear pages.
[16] Zhicheng Liu and Jeffrey Heer. 2014. The effects of interactive latency on exploratory visual analysis. *IEEE TVCG* (2014), 2122–2131.
[17] Hao Ma, Xueqing Liu, and Zhihong Shen. 2016. User fatigue in online news recommendation. In *WWW*. 1363–1372.
[18] Aaron Marcus and Elizabeth Rosenzweig. 2020. *DESIGN, USER EXPERIENCE, AND USABILITY. INTERACTION DESIGN: 9th International*. Vol. 12200.
[19] Robert B Miller. 1968. Response Time in Man-Computer Conversational Transactions. In *MRK*. 267–267.
[20] Fatemeh Nargesian, Erkang Zhu, Ken Q. Pu, and Rene´e J Miller. 2018. Table union search on open data. *PVLDB*, 813–825.
[21] M Tamer Özsu and Patrick Valduriez. 1999. *Principles of distributed database systems*. Vol. 2.
[22] Rakesh Pimplikar and Sunita Sarawagi. 2012. Answering table queries on the web using column keywords. *PVLDB* (2012), 908–919.
[23] Ben Shneiderman. 1984. Response time and display rate in human performance with computers. *CSUR* (1984), 265–285.
[24] Mohamed Yakout, Kris Ganjam, Kaushik Chakrabarti, and Surajit Chaudhuri. 2012. InfoGather: Entity augmentation and attribute discovery by holistic matching with web tables. *SIGMOD*, 97–108.
[25] Ying Zhang, Yao Yi Chiang, Pedro Szekely, and Craig A. Knoblock. 2013. A semantic approach to retrieving, linking, and integrating heterogeneous geospatial data. *AIIP/Semantic Cities@ IJCAI*, 31–37.
[26] Yi Zhang and Zachary G Ives. 2020. Finding Related Tables in Data Lakes for Interactive Data Science. *SIGMOD*, 1951–1966.
[27] Erkang Zhu, Dong Deng, Fatemeh Nargesian, and Renée J Miller. 2019. Josie: Overlap set similarity search for finding joinable tables in data lakes. In *SIGMOD*. 847–864.