Enhancing Natural Language-Based Data Exploration with Analysis Pipeline Illustration

Yi Guo
Tongji University

Nan Cao *
Tongji University

Xiaoyu Qi
Tongji University

Haoyang Li
Renmin University of China

Danqing Shi
Finnish Center for Artificial Intelligence

Jing Zhang
Renmin University of China

Qing Chen
Tongji University

Daniel Weiskopf
University of Stuttgart

Abstract

Exploratory Data Analysis (EDA) is a necessary yet laborious task when examining new datasets. To facilitate it, natural language interfaces (NLIs) can help users explore data through questions. However, existing NLIs lack explanations and visualizations of the analysis process used to uncover the answer. To address this, we introduce Urania, an interactive system that visualizes data analysis pipelines for resolving input questions. It first leverages a novel algorithm to break down questions into analysis pipelines, and then presents pipelines as datamations, with animated operations and data changes. Our experiments show that our algorithm outperforms existing methods in terms of accuracy and that Urania can help people explore datasets better.

Index Terms: Natural language interfaces

1 Introduction

Exploratory Data Analysis (EDA) is an important analytics process for knowledge discovery [3] in which analysts interactively explore data by performing a series of analysis operations (e.g., filter, aggregate). However, performing EDA requires a great deal of time and profound analytical skills [3]. Consequently, recent years have witnessed a proliferation of natural language interfaces (NLIs) designed to facilitate intuitive data exploration for users [2]. Existing NLIs primarily focus on improving question comprehension and delivering more precise responses, neglecting to explain or present the data analysis pipelines behind these answers. The absence of such explanations and presentations in NLIs undermines the interpretability and reliability of the answers, while hindering users' comprehension of the analysis process and their ability to extract insights.

To address this gap, we present Urania, an interactive system that combines natural language processing and visualization techniques to answer data-related questions and demonstrate the underlying data analysis pipelines. Our approach involves decomposing user queries using a data-aware algorithm to extract the pipeline required to obtain the answer. This generated pipeline then guides the creation of an action-oriented unit visualization, known as a datamation [4], which is an efficient way to interpret the results of analysis tasks through step-by-step animation of a detailed analysis pipeline. The resulting datamation is displayed in a carefully designed user interface, allowing users to refine it and explore the dataset interactively. To validate our approach, we conducted two quantitative experiments to evaluate the performance of the data-aware question decomposition algorithm. Furthermore, we interviewed three domain experts to assess the usability of Urania, comparing it with Tableau.

In summary, we present Urania, an interactive system that combines a natural language interface (NLI) with a question decomposition algorithm. The system visualizes the data analysis pipeline for each user question and provides datamations [4] as answers. Users can interact with the datamations to improve the analysis steps or correct any algorithmic errors.

*Nan Cao is the corresponding author. E-mail:nan.cao@tongji.edu.cn

2 System Overview

Figure 1 details three primary modules: (a) the *Preprocessing Module*, (b) the *Decomposition Algorithm*, and (c) the *Natural Language Interface*. A user-inputted tabular dataset, X, and a natural language question, q_x, trigger the *Preprocessing Module* (Fig. 1(a)), which extracts and converts the schema, s, into a token-included word sequence, s_x, for future computations.

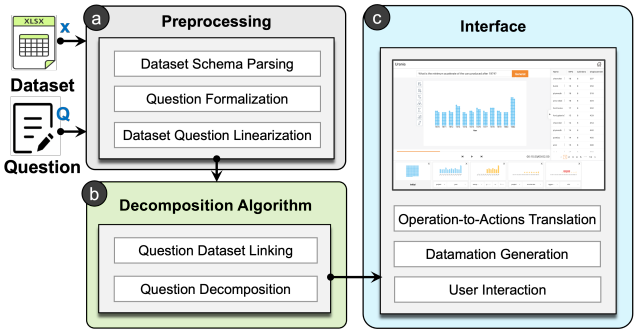

Figure 1: The architecture of Urania system consists of three major modules: (a) preprocessing module, (b) decomposition algorithm, (c) natural language interface.

Subsequently, the *Decomposition Algorithm*, as depicted in Fig. 1(b), processes the question q_x in the context of the schema s_x. It breaks down the question into a sequence of QDMR operations [5]:

$$S_{op} = [op_1, op_2, ..., op_n] \leftarrow Decompose(q_x, s_x) \quad (1)$$

Each operation, op_i, conducts an analysis of the dataset. Executing these operations sequentially provides the answer to q_x, forming a data analysis pipeline, S_{op}.

Finally, the S_{op} is visualized in the form of a datamation, with the animated presentation of each operation and its corresponding data changes. The datamation is displayed in a *Natural Language Interface* (Fig. 1(c)) that users are allowed to edit (i.e., add, delete, modify) the operations in S_{op} or input a new question. Once operations are edited, the datamation is updated accordingly.

3 Natural Language Interface and Interactions

The natural language interface of Urania, as shown in Fig. 2, consists of three major views: the *data view* (Fig. 2-1), the *key-frame view* (Fig. 2-2), and *datamation view* (Fig. 2-3). The *data view* is designed to illustrate the uploaded data, enabling users to easily access the data during the analysis process. Users can preview the data and enter a question of interest into the input box (Fig. 2-(a)).

The *key-frame view* (Fig. 2-2) illustrates the analysis operations in the generated pipeline and allows users to edit them. In addition, the drop-down menus under key-frames, as shown in Fig. 2-2(b), offer parameter access for operation refinement and extended exploration.

The *datamation view* (Fig. 2-3) visualizes the data analysis pipeline as a datamation and enables users to add new analysis operations to the pipeline. Moreover, users can use shortcut buttons (Fig. 2-3(c)) to add new operators to the existing data analysis pipeline. Notably, users can build a fresh data analysis pipeline using these buttons, bypassing the need for a data question. Alterations or additions to operations trigger updates in the key-frames and the datamation depicted in the interface.

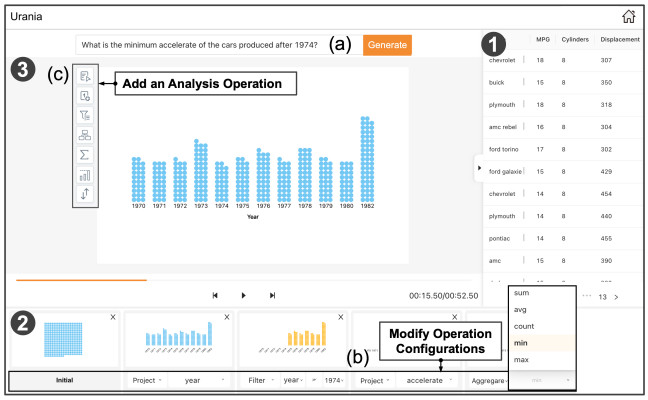

Figure 2: The natural language interface consists of three major views: (1) the data view; (2) the key-frame view; (3) the datamation view.

4 Expert Interviews

In order to evaluate the usability of Urania, we conducted a semi-structured interview with three domain experts denoted as **E1**, **E2**, and **E3**. **E1** is an industry analyst with three years of consulting experience, **E2** is a software developer specializing in data analysis and exploration systems for four years, and **E3** serves as the CEO of a company involved in the development and retailing of visual analytics dashboards.

Procedure and Task. The interviews were conducted offline, engaging in one-on-one meetings with each expert. Two systems, Urania and Tableau Ask Data [1], were made available to the experts for hands-on experience. Each expert was encouraged to explore a provided dataset using natural language queries with both systems. After the data exploration phase, the experts were invited to an interview, during which their feedback on the two systems was collected, focusing on three aspects: (1) the quality of the answers in terms of interpretability and reliability, (2) the interactions, and (3) the usability of the systems.

Example Cases. We select two questions that were correctly resolved by both systems and illustrate their answers in Fig. 3, Fig. 4. In these figures, the answers from Urania are marked in black, whereas the answer provided by Ask Data are marked in red.

Fig.3 presents the findings of a query posed by entity **E1**, specifically regarding the maximum passenger count for flights arriving from the United States. Fig.3(1-5) illustrates the keyframes derived from Urania's data analysis process. Urania initially grouped the records by countries (*Frame* 2), then filtered for flights originating from the USA (*Frame* 3). Subsequently, it determined the passenger count for each flight (*Frame* 4) and computed the maximum value (*Frame* 5). In contrast, Ask Data's response, depicted in Fig. 3(A), directly provides the calculated result of 229 for the input question.

Fig. 4 displays the responses to the query posed by **E2**. Urania grouped the records by education (*Frame* 2) and career (*Frame* 3), retrieved individual graduate salaries (*Frame* 4), and calculated the average values for each group (*Frame* 5). Ask Data also provided a list of average salaries for different groups. The data can be visualized through a bar chart, heatmap, or table view. For a concise presentation within the figure's constraints, we chose the table view.

Interview Feedback. In the interview, we received a number of valuable comments from experts that were summarized below:

The quality of answers. In terms of interpretability and reliability, experts broadly agreed that the Urania offers a more intuitive and

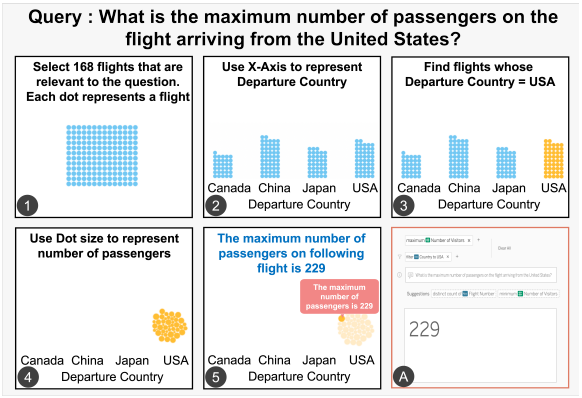

Figure 3: Urania (1-5) and Ask Data (A) responded to E1's question.

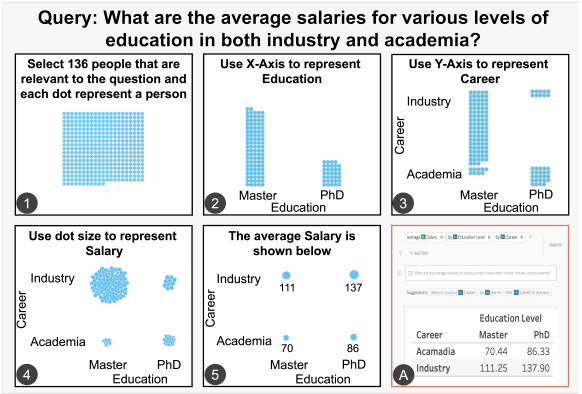

Figure 4: Urania (1-5) and Ask Data (A) responded to E2's question.

trustworthy way to perform EDA. They appreciated its ability to animate the analysis process, which not only allows for easy verification of the results but also safeguards users against potential false discoveries. The system's step-by-step approach also aids in understanding the derivation of the answer and gleaning useful insights.

The interactions in systems. All the experts were satisfied with exploring the data via natural language questions, which facilitate easy exploration. They found the interactive features of the results particularly beneficial. This interactivity in Urania negates the need for repetitive query input, thus streamlining the exploration process. Moreover, Urania's personalized auto-generated sessions made the experts feel as if they maintained control over the exploration, likening their role to being in the "driver's seat".

System usability. All experts thought that Urania is more useful and helpful in EDA compared with Ask Data. The experts highlighted the importance of the exploration process over just obtaining an answer, especially for new datasets. They also acknowledged the balance Urania strikes between manual and automatic exploration, noting its value in personalizing and rendering auto-generated exploratory sessions interactive.

References

[1] Tableau. `https://www.tableau.com/`. [Online; accessed 11-December-2022].

[2] C. Liu, Y. Han, R. Jiang, and X. Yuan. Advisor: Automatic visualization answer for natural-language question on tabular data. In *PacificVis*, pp. 11–20. IEEE, 2021.

[3] T. Milo and A. Somech. Automating exploratory data analysis via machine learning: An overview. In *SIGMOD*, pp. 2617–2622, 2020.

[4] X. Pu, S. Kross, J. M. Hofman, and D. G. Goldstein. Datamations: Animated explanations of data analysis pipelines. In *CHI*, pp. 1–14, 2021.

[5] T. Wolfson, M. Geva, A. Gupta, M. Gardner, Y. Goldberg, D. Deutch, and J. Berant. Break it down: A question understanding benchmark. *TACL*, 8:183–198, 2020.