POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Conversational Access to Structured Knowledge exploiting Large Models

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Vincenzo Manto

Advisor: Prof. Maristella Matera

Co-advisor: Emanuele Pucci

Academic year: 2022-2023

1. Introduction

In today's data-driven world, accessing and using data is becoming increasingly important for the entire society, spanning a wide range of industries. The role of structured information is still cardinal today, being at the base of the engines and systems that support entire sectors.

Nonetheless, the development of traditional methods of accessing and analyzing structured data has found it very difficult to abandon the technical approach and has crystallized on highly structured tools, such as the by-now standard relational database or SQL. Such methodologies require specialized technical knowledge and can be time-consuming and error-prone, reducing direct access to data to a very elitist practice. As a result, there is a growing need for more intuitive and easy-to-use ways to access structured data that don't require specialized skills.

The now fertile ground of artificial intelligence technologies seems to have paved the way for more natural and therefore "democratic" access to structured data. This Thesis fits into this context of continuous evolution and rapid growth, trying to define a broader possible conversational approach capable of giving business users the necessary means for efficient data-driven decision-making.

Until now, multiple experiments have resorted to machine learning methodologies aimed at transforming natural language into queries, sometimes failing to consider the set of collateral services necessary to simplify access to data.

Indeed, the potential of conversational data access will be explored and the practical capabilities of these technologies will be demonstrated, also highlighting the challenges involved in effectively implementing these technologies by reviewing the existing literature on conversational interfaces and developing a prototype system that combines and bundles different tools.

In this Thesis, the fundamental focus has been maintained on a specific class of users, namely business managers, who show a high interest in conversational access to knowledge that integrates extraction, presentation and transfer of information.

This work aims to understand better the opportunities and challenges involved in implementing a *holistic* conversational platform to access, communicate and analyze structured data and provide further insights, focusing on relational sources. Indeed, the future approach towards information would probably not only be a matter of extracting but also communicating, transferring and manipulating relevant data to transform it into knowledge and wisdom.

At the basis of this effort are the requests elicited in this work, which emerged through the careful observation of the managers' operational needs. From the elicitation process, the need for an **agnostic** tool towards data sources, **simple** in interaction, **modular**, **intuitive** and **inferential**, emerged strongly. These requirement identification has been performed by interviewing 14 business managers, trying to tackle their most relevant preferences and needs.

We have responded to these requests with an innovative approach that includes text-to-SQL translation but also its efficiency, the automation of choices regarding data visualization, data summarization and the automatic generation of data-based dashboards and slideshows entirely via conversation.

Finally, to demonstrate the effectiveness of the approach, a prototype (Queric) was developed using the most advanced technologies available, such as GPT-3.5, and subsequently we subjected it to comparative user tests against one of the most used tools in the business world i.e. Excel.

2. State of arts

The field of conversational intelligence, specifically chatbots, has undergone a significant disruption in recent years with the emergence of several conversational models. This breakthrough has drastically changed the landscape of the field, making it difficult to snapshot the current state of the art due to the rapid evolution of research and the widespread adoption of pre-trained models. The big bang disruption caused by GPT has led to the development of more sophisticated and accurate chatbots, making it difficult to keep track of the latest developments.

Significant technological advancements have been witnessed, particularly through the widespread adoption of large language models (LLMs) like GPT[6], which have showcased remarkable power even in the domain of text-to-code translation. These innovations have already made their way into certain academic endeavors, albeit to a limited extent. However, recent years have witnessed dedicated efforts to solidify supervised approaches for conversational access to data, exemplified by projects such as CODEX[3][2] from Politecnico di Milano or CAT[4].

This technological and theoretical landscape is extremely relevant and crucial in the development of a tool that can encompass from data extraction to data narration, above all considering the high competition deriving from the activities of the large software houses.

3. Approach

The proposed approach prioritizes the definition of processes and pipelines as an abstraction layer, regardless of the technologies employed. The rapid pace of innovation in this field necessitates a modular and interchangeable infrastructure not tied to a technology-specific approach. Therefore, the focus is on defining the necessary and sufficient processes, pipelines, and functionalities to fulfill the information needs of our key users - middle managers in large corporations. Each component of the proposed infrastructure is interchangeable with equivalent modules, respecting established interface limits, thereby ensuring modularity and extensibility through adapters.

The architecture is primarily composed of three fundamental modules:

1. Designer: The Designer module is at the core of the connection between the data sources and the application, and it involves defining the semantics of objects.
2. Chat: The Chat module is the central module for querying and conversational access to data, and it is closely linked to the Designer module.
3. Dashboard: The Dashboard module is an extension of the Chat module, which provides an interactive and expandable dashboard for accessing data.

Our software architecture app is built upon the classic 3-tier paradigm, comprising a modular back-end core, a front-end, a metadata DB and a set of support persistent resources. The app allows to create and configure a variety of connections in runtime, forming a network of data sources that users can query. In this paradigm, connections can be directed towards both locally and remotely hosted data sources, making it not limited to just simple use cases.

3.1. Design process

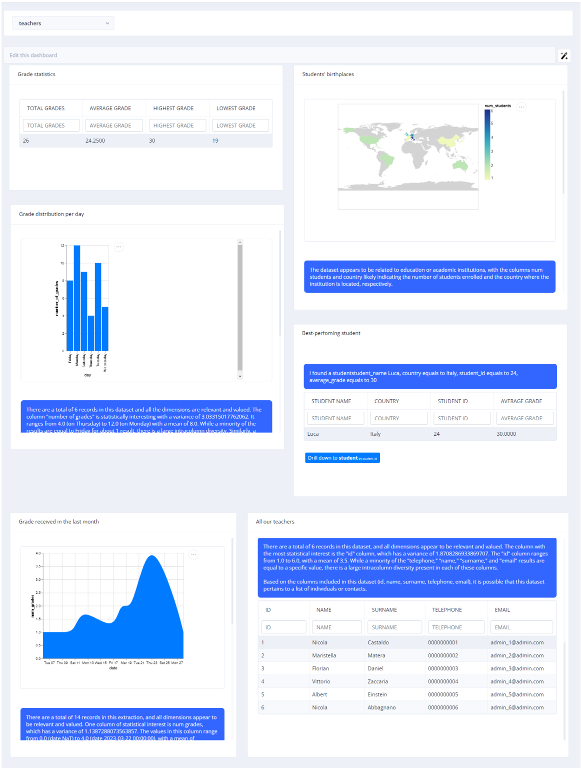
The **design process** of the meta-schema is a crucial task of our pipeline. Indeed, in order to operate, the framework requires knowledge about the semantics of the entities. Thus, the design process plays a crucial role in the platform. As mentioned earlier, considerable effort has been devoted to expediting, modifying, simplifying, and optimizing conventional ways of annotating and designing data sources. This agnosticism makes the platform more accessible to novice users who lack expertise in technical infrastructures and annotation requirements, which makes it challenging to interface with tagging methods.

3.2. Inquiry process

The inquiry or chat process represents the main interaction between users and data in the proposed solution through dialog. The inquiry is an extremely complex procedure that finds its fulcrum in the prompt-to-code model of GPT-3.5-turbo, however, requiring important computations upstream and downstream of the text-query translation. Also, when we refer to the inquiry process, we must bear in mind that not all user requests are aimed at obtaining an extraction, but some interactions may refer to more specific commands such as reassortment of display dimensions, drill down or navigation to other entities.
Indeed, in this step, the system accesses the metadata of the sources and, based on the user request, selects the most useful entities for extraction: this pruning process is essential when dealing with large sources. During this stage, the extracted data undergoes processing which includes the automatic generation of graphs and summaries. These visual supports are designed to assist the user in comprehending the information and are presented within the chat interface itself. This inferential process is crucial for creating a tool that is adaptable, independent, and valuable to the user.

3.3. Dashboard process

The significance of monitoring various aspects and dimensions of a topic is recognized, even though conversational access to data is often limited to single extractions. Interviews with middle managers reveal that they rely on interactive dashboards rather than single indicators to access information. Therefore, visual or textual extractions are preferred over tabular results, indicating a dominance of visualizations through dashboards for information aggregation and communication in many work environments. This dashboarding process combines the maximum inferential capabilities of the proposed approach, generating entire panels of indicators based solely on user requests in natural language, allowing them to modify the entire graphical interface using only words, neither buttons nor code. Finally, it is not uncommon for these dashboards to be exported as slideshows or PDF reports, therefore the proposed architecture is able to export PowerPoint entirely built autonomously starting from the single query of the single indicators, thus choosing the optimal visualization and communication methods for each type of format.



TOTAL GRADES	AVERAGE GRADE	HIGHEST GRADE	LOWEST GRADE
TOTAL GRADES	AVERAGE GRADE	HIGHEST GRADE	LOWEST GRADE
26	24.2500	30	19

The dataset appears to be related to education or academic institutions, with the columns num students and country likely indicating the number of students enrolled and the country where the institution is located, respectively.

There are a total of 6 records in this dataset and all the dimensions are relevant and valued. The column "number of grades" is statistically interesting with a variance of 3.03315017762062. It ranges from 4.0 (on Thursday) to 12.0 (on Monday) with a mean of 8.0. While a minority of the results are equal to Friday for about 1 result, there is a large intracolumn diversity. Similarly, a

I found a studentstudent_name Luca, country equals to Italy, student_id equals to 24, average_grade equals to 30

STUDENT NAME	COUNTRY	STUDENT ID	AVERAGE GRADE
STUDENT NAME	COUNTRY	STUDENT ID	AVERAGE GRADE
Luca	Italy	24	30.0000

There are a total of 14 records in this extraction, and all dimensions appear to be relevant and valued. One column of statistical interest is num grades, which has a variance of 1.1387288073563857. The values in this column range from 0.0 (date NaT) to 4.0 (date 2023-03-22 00:00:00), with a mean of

There are a total of 6 records in this dataset, and all dimensions appear to be relevant and valued. The column with the most statistical interest is the "id" column, which has a variance of 1.8708286933869707. The "id" column ranges from 1.0 to 6.0, with a mean of 3.5. While a minority of the "telephone," "name," "surname," and "email" results are equal to a specific value, there is a large intracolumn diversity present in each of these columns.

Based on the columns included in this dataset (id, name, surname, telephone, email), it is possible that this dataset pertains to a list of individuals or contacts.

ID	NAME	SURNAME	TELEPHONE	EMAIL
ID	NAME	SURNAME	TELEPHONE	EMAIL
1	Nicola	Castaldo	0000000001	admin_1@admin.com
2	Maristella	Matera	0000000002	admin_2@admin.com
3	Florian	Daniel	0000000003	admin_3@admin.com
4	Vittorio	Zaccaria	0000000004	admin_4@admin.com
5	Albert	Einstein	0000000005	admin_5@admin.com
6	Nicola	Abbagnano	0000000006	admin_6@admin.com

Figure 1: Autogenerated panel from "Create a dashboard about students situation" request on a school database

3.4. Architecture

The proposed platform is built upon the interaction and integration of multiple components, working together to not only extract data conversationally but also enhance the presentation and navigation of information. Within this architecture, large models (LM) play a significant role, as they translate verbal commands into executable code. These models have achieved remarkable performance levels[6], making them ideal for conversational tasks. For the prototype implementation, the GPT-3.5 model was selected.

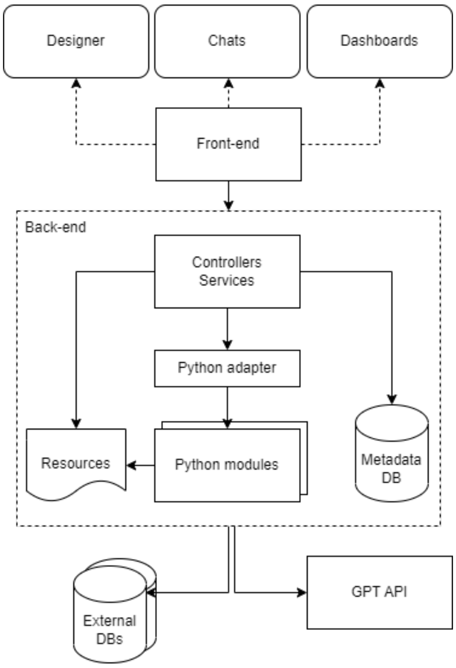


Figure 2: Simplified platform architecture

While LMs serve as the core of natural language processing, the processes preceding and following their usage are managed by a range of interchangeable modules. Each module specializes in tasks such as schema pruning, data visualization, data summarization, dashboard extraction, and export. These modules are coordinated through facade design patterns and an adapter, which ensure the system's extensibility and modularity are maintained and consolidated. The overall platform architecture is presented in Figure 2.

4. Evaluation

To evaluate the effectiveness of the prototype created on the basis of the proposed approach, technical tests have measured the time efficiency and theoretical complexity of the proposed algorithms. User-based studies have been performed to compare the prototype (Queric) and Excel. The tests involved business managers who were asked to extract and visualize requests consistent with analysis tasks they perform on a daily basis.

The test consists of 4 tasks to be performed on both systems, evaluating timing and success rate. The overall results are encouraging (Queric $\overline{x} = 124.58s, \sigma = 14.71$, Excel $\overline{x} = 146.06s, \sigma = 19.67$) highlighting how, in the same amount of time, participants were able to generate correct results and refine them more naturally. Moreover, these data show how the more complex the requests are, the better the conversational prototype performs. These results record a success rate of 73.95% for Queric against 64.58% on Excel. Subsequently, users were submitted to a questionnaire combining qualitative analysis tools such as NASA-TLX[5] and SUS[1]. These questionnaires highlight how the system is above the minimum level of satisfaction (69.5), recording a 70.19, but how Excel is more appreciated (74.89).

The users' feedback was also carefully analyzed using Latent Semantic Analysis (LSA) to identify the most critical issues. One prominent concern that emerged was related to data visualization and its potential effects on users' operational autonomy and developers' employment, as well as the transparency and predictability of these systems. Despite these concerns, the majority of feedback received was positive, acknowledging the effectiveness of the results and the methodologies employed.

5. Conclusion

Conversational intelligence has gained significant traction in various domains representing a crucial step forward in human-computer interaction and, when applied to data analysis, allows users to extract insights from data by simply conversing with the system, simplifying the analysis process and enhancing the accuracy and speed of data analysis.

In this Thesis, we propose an integrated and novel architecture and a paradigm capable of exploiting conversational medium to access and elaborate structured data, extract insights, and generate comprehensive data-driven interfaces.

This comprehensive approach aims to create innovative conversational access, enabling the extraction, utilization, and processing of business and operational data in a more inclusive and simplified manner.

By exploring the forefront of advancements in conversational AI, the presented model and prototype show the feasibility and desirability of this objective. In the foreseeable future, accessing transactional data could be facilitated through purely conversational tools that simultaneously extract, process, and comprehend the information, generating knowledge and valuable insights.

This work exemplifies how data extraction can evolve, emphasizing the importance of storytelling and conveying information effectively. The findings from tests reinforce the growing optimism and willingness to embrace this new approach to interacting with data.

5.1. Limitations

During the development of our framework, we encountered several structural and design choices that influenced the limitations of the system. Ensuring data privacy and access privileges posed a significant challenge, emphasizing the need for robust security protocols to protect sensitive information.

Another limitation is the handling of multimedia files: the system currently has the capability to process images as part of requests, but some limitations may affect the effectiveness of some requests. The nature of conversational inquiries also introduces a trial-and-error approach that can be time-consuming and frustrating, as interpreting user intent can be challenging.

The prompt-query-result system we implemented may experience higher latencies with larger schemas and data volumes, leading to delays in receiving responses. Moreover, our schema pruning approach based on entity semantics may reduce scalability, particularly for large databases with high cardinality.

Lastly, presenting complex data sets effectively and clearly proves to be a difficult challenge, impacting application performance and serendipity.

5.2. Future developments

Conversational agents have advanced significantly in recent years, utilizing machine learning and natural language processing to access and analyze data. Applications like GPT plugins have enabled querying data frames using conversational interfaces, but there is still room for improvement.

Key areas for enhancement include resource selection, schema annotation, integration of data mining modules, integration of open and associative data sources, voice and virtual assistant features, and prompt-based language models.

There is also an increasing interest in integrating voice and virtual assistant features into a text-based conversational platform for user-friendly data analysis, highlighting a large margin for improvement and innovation. By addressing these areas, conversational agents can become more powerful tools for interacting with complex data sets, transforming not only how we engage with information but completely revolutionizing exploration and discovery through conversation.

6. Bibliography

References

[1] John Brooke. Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, 189:11, 1995.

[2] Nicola Castaldo. A conceptual modeling approach for the rapid development of chatbots for conversational data exploration. Master's thesis, Politecnico di Milano - Scuola di Ingegneria Industriale e dell'Informazione, 4 2019.

[3] Nicola Castaldo, Florian Daniel, Maristella Matera, and Vittorio Zaccaria. Conversational data exploration. In Maxim Bakaev, Flavius Frasincar, and In-Young Ko, editors, *Web Engineering - 19th International Conference, ICWE 2019, Daejeon, South Korea, June 11-14, 2019, Proceedings*, volume 11496 of *Lecture Notes in Computer Science*, pages 490–497. Springer, 2019.

[4] Marius Gassen, Benjamin Hättasch, Benjamin Hilprecht, Nadja Geisler, Alexander Fraser, and Carsten Binnig. Demonstrating cat: Synthesizing data-aware conversational agents for transactional databases. *Proc. VLDB Endow.*, 15(12):3586–3589, 8 2022.

[5] Sandra G Hart and Lowell E Staveland. Nasa-task load index (nasa-tlx); 20 years later. *Human Factors*, 1988.

[6] OpenAI. Gpt-4. 2023.