Have a Chat with Clustine, a Conversational Engine for Data Exploration

Thibault Sellam
CWI, the Netherlands
thibault.sellam@cwi.nl

Martin Kersten
CWI, the Netherlands
martin.kersten@cwi.nl

ABSTRACT

1. INTRODUCTION

For many database users, writing database queries is a struggle. It is a struggle because mastering a query language requires training. It is also a struggle because it requires a precise and exhaustive knowledge of the database. Users must know exactly which table to use, which columns to inspect and which conditions to set. When users *explore* their data, that is, when they are interrogating it to discover its content, they do not have this knowledge.

To address this problem, software editors and researchers have come up *natural language interfaces* to databases. The idea is to let users interrogate their databases with plain English. It is then up to the system to interpret the query and cast it in a form that the underlying data manager can understand. This approach constitutes a huge leap forward, but it is not exempt of drawbacks. For a start, it still assumes that they user has a specific query in mind. Even if they do not know how to express this query SQL, they must have some idea of which column to use. Furthermore, it only solves half of the problem: the users express their queries in SQL, but they still obtain their results in tabular format. If the output contains a few tuples and a handful of columns, this paradigm is ideal. But what if the users are interested in broader queries?

A popular alternative is to use visual tools, such as Tableau. With these packages, users can write queries with drag and drops and obtain their results with visualizations. But this approach has its limits too. It does not solve the starting point problem, as users still need to write a query. Furthermore, visualizations are no panacea. Their main drawback come from the fact that they attempt to show everything. Consequently, they require training (though mild), attention, and they cannot scale beyond a few dozen variables.

In this paper, we introduce Clustine, the first conversational agent for data exploration. Our system remains on two pillars. First, it uses natural language during the whole exploration process. This means that it collects queries, but also *answers* in natural language. While several papers have studied the first direction, little to none of them have tackled the reverse direction. Second, our system is proactive: it makes suggestions, instead of collecting queries passively. It is then up to the user to accept or reject these suggestions.

This paper is an early-stage report: we present the main ideas behind Clustine and present preliminary experiments to show that they are feasible. Nevertheless, we will omit a few details and leave questions open for future publications. This work is partly inspired by Blaeu, a system to explore data with cluster analysis. The main difference is that Blaeu relies heavily on visualizations and expert judgment.

2. OVERVIEW

Clustine has several advantages compared to visual tools such as Tableau. It is proactive, in the sense that it suggests questions to the user. Also, its scope is ever broader that that of Tableau: it can be used by users with no literacy in data analysis, user who rely on audio (e.g., visually impaired users), and it is generally well suited to educational contexts. But the strengths of our method are also its weaknesses. As Clustine makes "editorial choices", it omits potentially interesting information about the data. Furthermore, full text is much less space efficient than charts. Therefore, Clustine is conceptually, but also materially limited.

3. ARCHITECTURE

4. EXPERIMENTS

4.1 Accuracy of the Descriptions

4.2 Speed of the System

4.3 Simulation

5. RELATED WORK

6. CONCLUSION

During the last few years, several companies have introduced virtual assistants: Cortana, Siri, Google Voice. In this paper, we investigated how to develop a similar system for data exploration.

7. REFERENCES

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235