Querying the query space

Thibault Sellam
CWI
thibault.sellam@cwi.nl

Martin Kersten
CWI
martin.kersten@cwi.nl

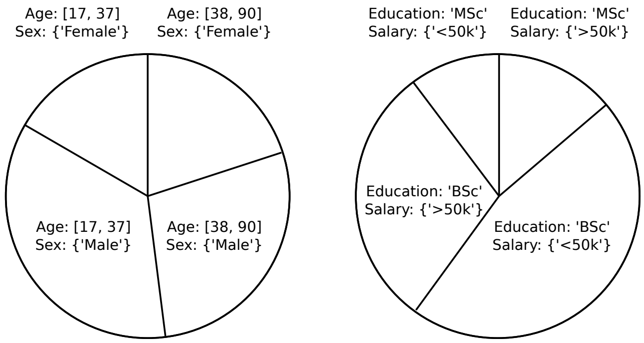

Figure 1: Two sets of queries for a census database

1. MOTIVATION AND PRINCIPLES

As data collection gets increasingly simple, exploration gets harder. In scientific data management and business analytics, the most informative queries are a holy grail. Currently, a data analyst can use two types of tools. The Spartan method is to query the database directly. This is done with e.g. SQL, Map-Reduce jobs (Java, Pig Latin[2], Hive[3]), or ad-hoc interfaces. However, querying a database requires a minimal knowledge of the data. An ill-defined query will often return an empty set, or overwhelming pages of results. On the other hand, data mining requires time (or hardware), preparation, and a precise idea of the type of knowledge to be extracted. In other terms, we are still far from a database engine to which we could simply ask: "tell me something about my data". All these methods are based on the same assumption: the user knows what he wants.

We tackle this problem with a novel approach: *we generate queries from data*. We propose several ways to query a set of tuples provided by a user. Consider for instance Figure 1. We suggest two sets of queries. Each of them focusses on one aspect of the same dataset. We call them *segmentations*. First, they give a succint summary of the data to the user. Second, the segments can themselves be submitted for analysis. This gives opportunities for deeper exploration.

2. PRODUCING GOOD SEGMENTATIONS

There are many ways to generate queries from data. The most trivial is to create a large query that covers the value domain of all the columns. Alternatively, it is possible to generate one query per tuple. None of these appoaches are satisfying. We aim at giving an overview of the data and hints for future research. Therefore, the segments should be easily *understable* by a human user. In this sense, we have other priorities than traditionnal clustering [1]. We expose five criteria:

- The items covered by a segment should be "similar"
- The segmentation should contain as many pieces as a user can understand (no more than a dozen)
- The segmentation should describe a wide range of attributes
- The segments should be balanced
- Their description should be easy to read

We introduce a heuristic to generate such queries in real time. The algorithm breaks down the domain of each attribute in two equal pieces. Then, these pieces are cut in two, on the most correlated attribute in the dataset. The process is repeated until the maximum number of pieces is reached.

3. CONCLUSION

Although this is still a work in progess, we believe our approach opens up ample opportunities for innovative theoretical, algorithmic, and systems research.

4. ACKNOWLEDGMENTS

This work was supported by the Dutch national program COMMIT.

5. REFERENCES

[1] G. Gan, C. Ma, and J. Wu. *Data clustering*. SIAM, Society for Industrial and Applied Mathematics, 2007.

[2] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins. Pig latin: a not-so-foreign language for data processing. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1099–1110. ACM, 2008.

[3] A. Thusoo, J. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy. Hive: a warehousing solution over a map-reduce framework. *Proceedings of the VLDB Endowment*, 2(2):1626–1629, 2009.