Approximate Query Answering over Open Data

Mengqi Zhang
University of Rochester
mzhang93@ur.rochester.edu

Pranay Mundra
University of Rochester
pmundra@ur.rochester.edu

Chukwubuikem Chikweze
University of Rochester
cchikwez@u.rochester.edu

Fatemeh Nargesian
University of Rochester
fnargesian@rochester.edu

Gerhard Weikum
Max Planck Institute
weikum@mpi-inf.mpg.de

ABSTRACT

Open knowledge, including open data and publicly available knowledge bases, offers a rich opportunity for data scientists for analysis and query answering, but comes with big obstacles due to the diverse, noisy, and incomplete nature of its data eco-system. This paper proposes a vision for enabling approximate QUery answering over Open Knowledge (Quok), with a focus on supporting analytic tasks that involve identifying relevant data and computing aggregations. We define the problem, outline a system architecture, and discuss challenges and approaches to taming the uncertainty and incompleteness of open knowledge.

ACM Reference Format:
Mengqi Zhang, Pranay Mundra, Chukwubuikem Chikweze, Fatemeh Nargesian, and Gerhard Weikum. 2023. Approximate Query Answering over Open Data. In *Workshop on Human-In-the-Loop Data Analytics (HILDA '23), June 18, 2023, Seattle, WA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3597465.3605227

1 INTRODUCTION

Data analytics has been well established in business and finance, traditionally operating over a single database or data warehouse with a small (sub-)set of well-designed and content-curated tables. With modern data science, this has been changing on both application areas and data landscapes. Today, data scientists are active on much wider use cases, from life sciences and environmental studies (energy, traffic, climate, ecosystems) all the way to humanities and social sciences. In many settings, the analyst is faced with a large data lake of highly diverse tables or even dataset-search [5, 11, 20] results from the Internet, and would like to quickly join and aggregate relevant pieces for exploration, knowledge discovery and intellectual insight. The underlying *Open Data* comprises ad-hoc datasets, reference repositories, spreadsheets and web tables from web and cloud sources, with formats from relational and JSON to HTML and CSV [19].

As an example, consider an environmental protection researcher Zoe who wants to analyze the emissions distribution of greenhouse gas per US state. The Open Data habitat offers a variety of datasets with such emission results. However, these finer sources are likely noisy and incomplete, and come with ad-hoc schemas in the form of hand-crafted column headers (some with informative names, others with generic and useless strings). Then, Zoe may also face the following tasks:

1. Identifying relevant tables in the open data lake, and selecting their relevant rows.
2. Interpreting table headers in order to identify joinable columns and relevant join paths.
3. Identifying columns for grouping and aggregation, and evaluating relevant aggregates (sum, cnt, avg, etc.).

Each of these steps would be a straightforward SQL-for-OLAP exercise if all contents were clean, complete and integrated into a single database. The challenges in our setting, though, are to cope with the large scale of ad-hoc choice of datasets with hardly interpretable metadata and a major amount of noisy or missing values, the sheer number of join paths we may find and the expensive execution of join. Fortunately, since analytic tasks often involve aggregations where trends and relative comparisons can already be insightful, there is hope for success by adopting the paradigm of *approximate query processing (AQP)* [1, 7, 14].

There is a wealth of techniques for efficient and effective AQP over single databases, such as sampling-based approaches [3, 13] and sketch-based algorithms [9, 10]. These techniques aim to strike a balance between answer accuracy and computational cost. This paper, on the other hand, explores the state-of-the-art for the nascent theme of approximate answering of aggregate queries over *Open Knowledge*, that is, large knowledge bases (*KB*) and data lakes (*DL*) with vastly diverse habitats of noisy and incomplete tables.

HILDA '23, June 18, 2023, Seattle, WA, USA

ACM ISBN 979-8-4007-0216-7/23/06…$15.00
https://doi.org/10.1145/3597465.3605227

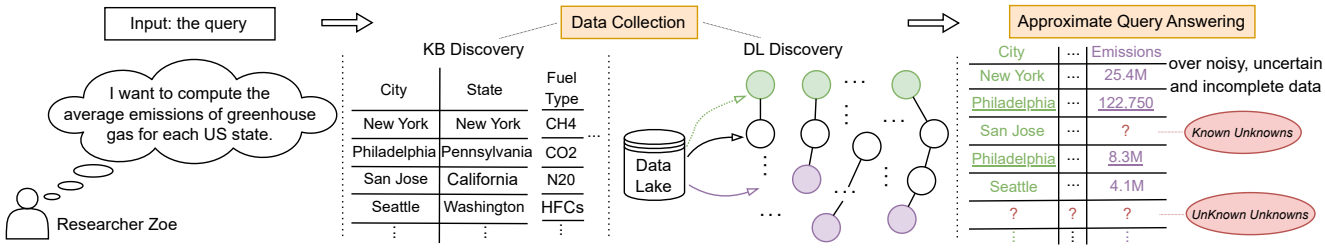

Figure 1: System Overview.

2 SYSTEM OVERVIEW

Figure 1 illustrates our proposed architecture. We are given a data lake $\mathcal{L}$, a knowledge base $\mathcal{K}$, and a query $Q = \{H, G, T, agg\}$, where H is a categorical attribute, called subject attribute, G is a group-by attribute, T is a numerical attribute, called aggregate attribute, and agg is an aggregate function applicable on T. The goal is to compute the aggregate of T over attribute G, namely $(\gamma_{G,agg(T)}\sigma_H(\mathcal{K}, \mathcal{L}), CI)$, where CI refers to the confidence interval. For example, in the query of Figure 1, `City` and `Fuel Type` are the subject attributes, `State` is the group-by attribute, `Emissions` is the aggregate attribute, and `Average` is the aggregate function.

The framework consists of two main steps: 1) fusing the data of KB and DL to obtain relevant and necessary data for query answering in an efficient way and 2) computing approximate aggregates in the presence of noisy, incomplete, and uncertain data.

3 RESEARCH CHALLENGES

Data Collection: The first step is to integrate heterogeneous and noisy data of the lake with the knowledge base and construct a dataset containing subject, aggregate, and group-by attributes. Due to the sheer size of data and space of integration, we develop a framework for obtaining a random sample from this space where we can analyze confidence intervals on calculated aggregates. This involves first performing *KB discovery*, that is extracting the entities related to the subject and group-by attributes. Next, during *DL discovery*, the system fuses the data obtained from KB with datasets in DL. The goal is to obtain numerical values of the aggregate attributes for all or at least a subset of entities found for subject and group-by attributes in KB. This step involves relevant dataset discovery [20, 25] and navigating the join graph of DL [26] to find ways of integrating datasets. The first challenge is the pruning of the space of all possible join queries (join paths). To deal with the sheer number of join queries and expensive joins, we propose two optimizations. Since syntactic [25] and semantic join [18] discovery techniques may result in a large number of false positives, we will leverage table representation learning techniques [15] to identify meaningful joins and prune the space. Second, we develop a practical technique with provable guarantees for sampling over the union of joins for the problem of union sampling [12]. The goal is to ensure the uniformity of samples without generating and executing the complete join graph.

Approximate Query Answering over Uncertain and Incomplete Data: The data collection step may generate incomplete or uncertain data. For example, the data lake may return multiple `emission` values for a specific `city`. The main challenge is to approximate aggregates in the presence of the incompleteness of results as well as inaccuracies and uncertainties in the data collection phase, rising from the lack of standard formatting and semantics. To handle uncertainty, one simple way is to assign evidence scores to extracted tuples and incorporate the score in aggregation. In addition, in some scenarios, the extracted data is incomplete. For example, neither KB nor DL may contain the `emissions` value for some `cities`. We call such cases ***known unknowns*** and rely on the rich body of research on missing value imputation [2, 16] to predict one value [21] or a range of values [4, 23] for missing value in the aggregate attributes. Alternatively, we may consider multiple underlying data generators for the values of an attribute mitigating the potential bias regarding all values being from the same distribution.

The incompleteness may impact the subject attribute. For example, some `cities` that exist in the real world may not exist in the KB or DL. Therefore, the data collection fails to return their `emissions`. Such cases are called ***unknown unknowns***. Chung et al. proposed a technique for estimating the number and values of the missing data items by considering the overlap between different data sources [8]. The idea is based on the Species Accumulation Curve in Ecology [24], where the intuition is that the rate of discovery of new species decreases as the cumulative effort of the search increases. Meanwhile, Chao et al. proposed approximating the number of unknown unknowns based on the sample coverage [6, 8], such that when the number of duplicates is greater than the number of singletons, the sample coverage is higher, indicating that the sample is more complete. Then, the Good Turing estimator [17, 22] can be applied to estimate the sample coverage. To handle unknown unknowns, we consider random samples obtained from join queries generated during data collection as input data sources and adopt the Good Turing estimator [17, 22] and sample coverage techniques to compute aggregates and their confidence intervals.

ACKNOWLEDGE

This research is supported in part by NSF 2107050.

REFERENCES

[1] Swarup Acharya, Phillip B Gibbons, Viswanath Poosala, and Sridhar Ramaswamy. 1999. Join synopses for approximate query answering. In *SIGMOD*. 275–286.

[2] Paul D Allison. 2000. Multiple imputation for missing data: A cautionary tale. *Sociological methods & research* 28, 3 (2000), 301–309.

[3] Brian Babcock, Surajit Chaudhuri, and Gautam Das. 2003. Dynamic Sample Selection for Approximate Query Processing. In *SIGMOD*. 539–550.

[4] Parikshit Bansal, Prathamesh Deshpande, and Sunita Sarawagi. 2021. Missing value imputation on multidimensional time series. *arXiv preprint arXiv:2103.01600* (2021).

[5] Dan Brickley, Matthew Burgess, and Natasha F. Noy. 2019. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *WWW*. 1365–1375.

[6] Anne Chao and shen-Ming Lee. 1992. Estimating the Number of Classes Via Sample Coverage. *J. Amer. Statist. Assoc.* 87 (03 1992), 210–217. https://doi.org/10.1080/01621459.1992.10475194

[7] Surajit Chaudhuri, Bolin Ding, and Srikanth Kandula. 2017. Approximate query processing: No silver bullet. In *Proceedings of the 2017 ACM International Conference on Management of Data*. 511–519.

[8] Yeounoh Chung, Michael Lind Mortensen, Carsten Binnig, and Tim Kraska. 2018. Estimating the impact of unknown unknowns on aggregate query results. *TODS* 43, 1 (2018), 1–37.

[9] Graham Cormode. 2010. Sketch Techniques for Approximate Query Processing. In *Foundations and Trends in Databases*.

[10] Graham Cormode, Minos Garofalakis, Peter J. Haas, and Chris Jermaine. 2011. Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches. In *Foundations and Trends in Databases*.

[11] Dong Deng, Raul Castro Fernandez, Ziawasch Abedjan, Sibo Wang, Michael Stonebraker, Ahmed K. Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, and Nan Tang. 2017. The Data Civilizer System. In *CIDR*.

[12] Shiyuan Deng, Shangqi Lu, and Yufei Tao. 2023. On Join Sampling and the Hardness of Combinatorial Output-Sensitive Join Algorithms. In *PODS*, Floris Geerts, Hung Q. Ngo, and Stavros Sintos (Eds.). ACM, 99–111.

[13] Ihab F. Ilyas, George Beskales, and Mohamed A. Soliman. 2008. A Survey of Top-k Query Processing Techniques in Relational Database Systems. *ACM Comput. Surv.* (2008).

[14] Kaiyu Li and Guoliang Li. 2018. Approximate query processing: What is new and where to go? A survey on approximate query processing. *Data Science and Engineering* 3 (2018), 379–397.

[15] Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. 2020. Deep Entity Matching with Pre-Trained Language Models. *PVLDB* (2020), 50–60.

[16] R Malarvizhi and Antony Selvadoss Thanamani. 2012. K-nearest neighbor in missing data imputation. *Int. J. Eng. Res. Dev* 5, 1 (2012), 5–7.

[17] David Mcallester and Robert Schapire. 2000. On the Convergence Rate of Good-Turing Estimators. (06 2000).

[18] Pranay Mundra, Jianhao Zhang, Fatemeh Nargesian, and Nikolaus Augsten. 2023. KOIOS: Top-k Semantic Overlap Set Search. In *ICDE*. To appear.

[19] Fatemeh Nargesian, Erkang Zhu, Renée J. Miller, Ken Q. Pu, and Patricia C. Arocena. 2019. Data Lake Management: Challenges and Opportunities. *PVLDB* 12, 12 (2019), 1986–1989.

[20] Fatemeh Nargesian, Erkang Zhu, Ken Q. Pu, and Renée J. Miller. 2018. Table Union Search on Open Data. *PVLDB* 11, 7 (2018), 813–825.

[21] Jason W Osborne. 2013. *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Sage.

[22] Amichai Painsky. 2022. Convergence Guarantees for the Good-Turing Estimator. *Journal of Machine Learning Research* 23, 279 (2022), 1–37.

[23] Andreas Pfaffel, Marlene Kollmayer, Barbara Schober, and Christiane Spiel. 2016. A missing data approach to correct for direct and indirect range restrictions with a dichotomous criterion: A simulation study. *PloS one* 11, 3 (2016), e0152330.

[24] Karl I Ugland, John S Gray, and Kari E Ellingsen. 2003. The species–accumulation curve and estimation of species richness. *Journal of animal ecology* 72, 5 (2003), 888–897.

[25] 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.

[26] Erkang Zhu, Ken Q. Pu, Fatemeh Nargesian, and Renée J. Miller. 2017. Interactive Navigation of Open Data Linkages. *PVLDB* 10, 12 (2017), 1837–1840.