Data Discovery

Mohammad Bayat

ABSTRACT

This report dives into the challenges inherent in data discovery, a crucial initial step in the Big Data analytics process, by synthesizing insights from three pivotal papers from the Research Session F2 of the 2023 Very Large Databases conference. One of the papers utilized within this report discusses a discovery system called CMDL which enables a systematic approach to finding relationships between unstructured and structured within giant data lakes. Cross Model Data Discovery or CMDL enables searching for data similarities within textual and tabular data. Evaluation of CMDL on three-world data lakes shows that the system is significantly more eective than a searched-based technique that already exists. CMDL is more accurate and robust than the existing search systems.

The subsequent papers introduce innovative solutions targeted towards optimizing the quality and consumption of the data. The second paper discusses an annotation framework that helps with column semantics, a solution that would enhance the overall quality of the data. RECA searches for inter-table context information and similar schema to enhanced column annotation. The paper evaluates RECA by experimenting on two web datasets which relieve an outstanding performance of F-score 0.85 and 0.937 which outperform the existing annotation frameworks. The last paper discussed in the report is on the Python framework, JoinBoost, which provides an interesting way of data consumption for data modeling. The performance of JoinBoost was evaluated on DuckDB and the result shows that not only JoinBoost is 3X faster than random forests, but also it is over of magnitude faster than the existing In-DB ML system.

KEYWORDS

In-Database Machine Learning, Unstructured and Structured Data, Data Lakes, Column Labeling

ACM Reference Format:

Mohammad Bayat. 2024. Data Discovery. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 OVERVIEW

Big Data is an enormous field that could lead to outstanding outcomes. As impressive as its outcome gets, there are so many challenges within the field that require the use of advanced technology to overcome the technical hassles, particularly when confronted with the difficult task of extracting meaningful information from a sea of data originating from multiple sources and formats, such as textual, tabular, and beyond. It would take numerous pre-processing steps for data to be consumable, beginning with operations within data lakes, where the pivotal task lies in discerning relationships between structured and unstructured data. This crucial step lays down the foundational groundwork upon which subsequent data cleaning and normalization phases build, leading to multiple tables representing the data and their inherent relationships [**?**].

Despite establishing relationships potentially being perceived as the most crucial pre-processing step, the next step in this complex process mandates a rigorous focus on the quality of normalized data. Here, data quality can be enhanced by employing strategies such as adopting a systematic approach to column naming between related tables [9], ensuring that the pre-processed data is optimal for consumption by Machine Learning (ML) algorithms. Despite the current drawbacks in the way ML libraries use normalized data exported from the DBMS, the utilization of machine learning has significantly improved the usability of the data, resulting in remarkable performance across various contexts. When ML algorithms want to use data from the DBMS, the data need to be materialized [7]. In other words, all the tables within the DBMS need to be concatenated into one giant table which is a complex operation. The complexity of the operation would make it prone to error. Furthermore, the materialization of the database is very space expensive. For example, the space cost for the join materialization of data in a simple database of 1.2GB is over 1TB [7]. The third and most important drawback that requires attention is the privacy concern because when the data is materialized it needs to be exported into the ML library which raises privacy concerns. One of the most efficient solutions to address this concern is In-DB ML computation which is the motivation of frameworks such as JoinBoost that is introduced in the third conference paper [7] discussed in this report. The practicality and efficiency of In-DB ML would depend on the optimization of the ML algorithm or the availability of the GPU accelerator.

The subsequent sections of this report will compare and contrast the practical and innovative solutions presented in the three papers from the 2023 Very Large Databases Conference, presenting pivotal insights and methodologies that have the potential to significantly mitigate the challenges encountered throughout the data discovery journey in Big Data analytics.

2 COMMON THEMES

There were three common themes that I observed while reviewing the three conference papers. These common themes are directly related to one another. At first, each author starts to discuss the importance of the topic of choice and how a potential solution can help in the advancement of the data discovery field. After briefly discussing the importance of the topic, they tend to introduce the research question that will be discussed in the paper.

The next common is how each author makes the argument toward a solution that they are proposing. Each author tries to talk

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, July 2017, Washington, DC, USA

© 2024 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

about the solution in a step-wise manner. In other words, they first talk about the current challenges they could potentially face if they want to move toward a potential solution. Then, they will talk about whether the potential solution has potential drawbacks. If there are any drawbacks, the same approach will be taken which is identifying the potential challenges and trying to optimize their solution in terms of efficiency.

The third common theme that I observed was how the authors try to look at the existing solutions in the industry and use the possible gap that could be used to propose a better solution. This is not only used when they are trying to propose their solution but also when they are evaluating the solution.

3 DISCORDANT THEMES

The discordant themes that I observed were the different approaches that each author was taking to enhance the existing solution in the field of data discovery. The conference papers take contrasting approaches to improve the sophisticated task of discovery in different stages of the data cycle.

A little after the collection stage of the data life-cycle, there comes the step of processing. The paper on finding relatedness between the data [**?**] is fixated on the third stage of the data life cycle, which is the processing phase. As data moves from one stage to another stage in the data life cycle, the data repository changes as well which defines the environment and the appropriate approach for enhancing data discovery. Mohamed Y. E's [**?**]paper looks at the processing phase within the data lakes. The paper's focus was to introduce a combination approach that consists algorithm and architecture approach that could be a more efficient way for data discovery in massive data lakes. The data discovery concept is also applicable in the analysis phase. In the paper on in-database machine learning computation, Huang [7] concentrated on the analysis stage of the data life-cycle. The authors in this paper focus only on a small set of machine learning algorithms within the database reflecting the complexity of the analysis stage. A solution tailored to this particular aspect of machine learning could potentially influence the implementation of other complex data models.

Each author examines a specific stage of the data life-cycle, contributing to various challenges in developing that stage. Zezhou H. [7] notes that algorithmic and systematic barriers contribute to the limitations in achieving sophisticated modeling computations akin to those occurring outside the database system. Although the authors of CMDL [**?**] concentrate on the processing stage, the challenges they encounter originate from the collection phase. Here, a vast amount of data is collected from diverse sources, which is one of the difficulties cited by the authors. Given that data is gathered from various resources, this leads to another challenge discussed in the paper: the absence of descriptors in the raw data stored in data lakes. Although the collected data gets processed before the storage stage, the focus of the processing step within the massive data lakes is to find accurate relatedness in sea of data. Yushi S. [7] looks at the data management within the database. Data management is a crucial step for discovering meaningful insights. However, one of the challenges that Yushi S. [7] mentions the he complexities of managing broad tables and intricate inter-table relationships in the RECA framework.

4 NON-OVERLAPPING THEMES

The non-overlapping themes that I observed are factorized gradient boosting, discover elements(DEs), Jaccard similarity, named entity schema, space complexity, and time complexity. The JoinBoost paper covers several topics that are important to the efficiency of data modeling in database systems including factorized gradient boosting, and residual updates. These topics are covered in detail in the paper and are not found in the other papers in the search results. The distinctness of the problems addressed in the JoinBoost paper is likely the reason for the amount of disjointedness among the papers. These ideas are covered in only one or two of the selected conference proceedings, though they are important factors in their respective works and frequently occurring themes in computer science. The distinctness of the problems addressed in the three papers is likely the reason for the amount of disjointedness among them.

The CMDL conference paper [**?**] tends to have a broader scope of work in different a environment than the JoinBoost [7] and RECA [9] conference paper. The intuition behind CMDL is based on the notion of discoverable elements(DEs) which is an abstract unit of data discovery. The main focus of the paper is to use the DE units to find different three key relationship such as Doc to Table, Joinable table, Unionable tables.

5 RELATED WORK

With the advancement of the Internet of Things more data is being collected with high velocity and data lakes have been implemented to cope with such high velocity data. The presence of data lakes has opened up a lot of opportunities in the field of data management and data discovery. Some of the recent works that have contributed to more advancement of data discovery and management include the 2021 research on Metadata-Management within the data lakes[5], Starmie [8] which look at the data from a similar angle. The metadata management paper [5] tries to manage the data and identify abstract knowledge from data in the dataset for different types of use cases but does not provide a way to utilize the metadata to find a relationship between the current table and other tables in the data lake. Starmie [8] is a paper that looks also at data within the data lakes to find similarity scores using an encoder a column encoder which uses a pre-training strategy to capture rich metadata information about the table by looking at multiple columns. Then, it follows a contrasting approach to compare the gained metadata with other tables and assign reunion-ability scores between tables.

As previously mentioned, data discovery can take place in different stages of the data life cycle. Many interesting researches have been conducted on data discovery within the database similar to the JoinBoost conference 2023 paper [7]. One of the recent conference papers from 2022 that I reviewed was the paper on User-Defined Operators (UDO) [8] which resolves several issues in modern data analytics, including the inefficiency of data exportation to CSV files, the limited ease of use of SQL for analytics, and the compromise of ACID properties when data is removed from the database. It enables

the user to write the queries using high-level programming languages like Python, and rewrite them into pure SQL by utilizing the UDO. A slightly older conference paper from 2012 [4] addressed the inefficiency of training data models away from the database. Since the paper is older, the approach tried to bring the machine learning view referred to as "tensoral view (TViews)". By having such a view, data could be treated as tensors, which is a multi-dimensional array enabling a way to perform linear algebra operation which is the backbone of every machine learning algorithm. MLog [4] as the author mentions is an extension of SQL. Unlike the most recent paper which not only proposes a way of doing machine learning in the database but JoinBoost [7] and UDO [8] that consider solution usability, the authors of MLOG [4] approach is short of that offering.

6 LEGAL CONSIDERATIONS

In this discussion, I would like to discuss about few important recent laws that help with data privacy. Data could be used to discover some confidential information about each individual. The information gained from the discovery could be then used to promote or advertise products or it could be used in some AI applications to automate the task in different industries such as healthcare, insurance, and employment which could have a serious impact on people's lives.

In 2023, ten states introduced new regulations in response to the surge of AI applications. Virginia and California were the first two states in the United States to introduce new data privacy regulations. This new regulation came into effect from January 1, 2023. Virginia Consumer Data Protection ACT (VCDPA) and California Privacy Right Acts (CPRA) are the two important regulations because the proposed regulations set two different paths that other states could choose to follow.[6]

The California Consumer Privacy ACT (CCPA) was the data privacy act signed in 2018 and went into effect two years later on January 1,2020. Later on, in November 2020, California legislators approved CPRA which went into effect on January 1, 2023.[1] According to CCPA regulations, the consumers have the right know-how, and why their data is being collected. They also have the right to request their from be deleted and refuse the sell of their data. CCPR add more right to CCPA which include stronger regulations for businesses regarding sensitive personal data and data retention.[1]

The VCDPA is the new regulation signed by Virginia lawmakers which came into effect on January 1, 2023. It give the consumers who used certain application that collect their data the right to opt out of having their data used for profiling which could be used for some automated decisions. Also, it provides data protection assessment for 5 different types of data processing which include the sale of personal data, and processing the personal data for purposes of targeted advertising or profiling. data protection assessment is important because it helps companies keep consumer data safe and respect their right [2].

There are a few key differences between CPRA and VPCA. The first key difference is related to the collection of sensitive data. Under VCPA it is required for businesses to disclose information about the collection, use, and processing of the data This is not a requirement passed by VCPDA which is very scary from an individual's aspect. A business can collect your data and share it with other companies who consume these data to make decisions that would impact your life [6]. I believe that VCPDA has to add this to its regulations if it wants to protect and respect consumer privacy rights.

7 ETHICAL CONSIDERATIONS

Ethical considerations should be central to any work that is happening in the field of computing. More specifically, when it comes to the topic of data discovery the role of ethical consideration becomes even more important as any discovery from the data could have a direct impact on an individual life. By providing more sophisticated tools for data discovery, computing professionals are giving more tools for businesses to have more accurate insight about people. Businesses that use this sophisticated should respect the privacy of individual tools which is mentioned in ACM [3] ethical principle 1.6. Companies that use machine learning systems such as the tools that let tree-based modeling in-database should take extraordinary care as mentioned in Acm [3] principle 2.5. Machine learning applications could be used as profiling tools and this could have real-life consequences.

The authors of JoinBoost paper conference paper were taking ethical principles in mind while proposing an in-database modeling solution. The intuition of their proposal is based on the acm principle 1.2 which mentions that computing professionals must report any risk that might result in harm. The authors mention that the drawback of current practices of data modeling which materialize and export the data away from the base would put the data at risk. Introducing in-database data modeling will avoid the privacy risks that could arise via the current data modeling approaches.

8 FINAL REMARKS

After reviewing all three conference papers, it is interesting how wide the applicability of data discovery is starting from data lakes to databases. Although when reading the JoinBoost conference article [7], it is not as clear as the other two papers how it related to data discovery, with a more intuitive view it becomes clear why this paper is set under the topic of data discovery as data discovery in this paper refers to the information discovered from the analysis of the data using sophisticated data modeling within the database. The other two papers tend to work on different aspects of data discovery than the JoinBoost [7] article. Even though CMDL [**?**] focuses on finding the relationship between the unstructured and structured data in data lakes which could then be put in the data warehouses their solution tends to not be optimized. One possible optimization is to find a systematic approach that could also name the columns with meaningful labels which could make the task of data querying more efficient.

REFERENCES

[1] https://epic.org/the-state-of-state-ai-laws-2023/: :text=the

[2] https://wideangle.co/blog/what-is-purpose-privacy-impact-assessment.

[3] https://www.acm.org/code-of-ethicsh-2.-professional-responsibilities.

[4] Mlog: Towards declarative indatabase machine learning. *Proc. VLDB Endow. 10*, 12 (2017).

[5] Eichler, R., Giebler, C., Gröger, C., Schwarz, H., and Mitschang, B. Modeling metadata in data lakes—a generic model. *Data Knowledge Engineering 136* (2021), 101931.

[6] Grieco, J. https://www.mineos.ai/articles/vcdpa-vs-ccpa-comparing-virginia-and-california-privacy-laws: :text=the

[7] Huang, Z., Sen, R., Liu, J., and Wu, E. Joinboost: Grow trees over normalized data using only sql. *Proceedings of the VLDB Endowment 16*, 11 (2023), 3071–3084.

[8] Sichert, M., and Neumann, T. User-defined operators. *Proceedings of the VLDB Endowment 15*, 5 (2022), 1119–1131.

[9] Sun, Y., Xin, H., and Chen, L. Reca: Related tables enhanced column semantic type annotation framework. *Proceedings of the VLDB Endowment 16*, 6 (2023), 1319–1331.