Big Data Seminar (BDS)

Data Discovery

Chidiebere Ogbuchi, and Tianheng Zhou

Universitat Politecnica de Catalunya, BarcelonaTech, Barcelona, Spain
chidiebere.Ogbuchi@estudiantat.upc.edu | tianheng.zhou@estudiantat.upc.edu

Abstract
Data discovery is vital for data scientists to explore and select relevant datasets from large repositories, enabling the development of accurate and generalizable mathematical models while avoiding overfitting. This review emphasizes the significance of data discovery in facilitating effective data exploration and model development in the field of data science. The methodologies and approaches employed in data discovery are discussed, with a focus on structural data joining and the role of data lakes as extensive repositories that store diverse data sources. Data lakes play an important role in data discovery by providing access to a wealth of heterogeneous datasets and enabling the development of robust models. Within the context of exploring vast repositories with diverse data sources, data discovery emerges as a critical concept thoroughly examined in this review.

Keywords: Data Discovery, Data Lake, Structured Data Discovery

1. INTRODUCTION

We live in an era of big data, where a large amount of information is recorded as data in various forms. In order to obtain valuable insights from both homogeneous and heterogeneous data sets, the rapid expansion of data volume and variety has dramatically increased the demand for data science in many organizations and domains. Data discovery is an important research field in data science. It involves an automated set of tasks to identify relevant and complementary data sets for a given analytical task from a vast repository of data [13]. It is also a process of exploring and understanding available data sources to discover new and valuable datasets that can enhance the analysis and decision-making process [17]. In data science, the goal is to leverage diverse data features (variables) to build robust mathematical models for predicting or classifying events of interest. Therefore, the ability to gather a wide range of data sets and variables becomes essential. By employing data discovery techniques, data scientists can access a rich pool of data from various sources, including proprietary and external datasets, and recognize data sets that are particular and essential for their tasks, enabling them to build more informative and powerful models that are less prone to overfitting. The comprehensive exploration and integration of relevant data through data discovery leads to better-informed decision-making processes and the development of more accurate and robust predictive models.

Furthermore, the increasing significance of data-driven decision-making has underscored the criticality of efficient data discovery processes within contemporary organizations. Data discovery is essential in enabling organizations to uncover concealed patterns and insights that can profoundly influence decision-making and drive innovation. A fundamental element of modern data discovery processes involves harnessing data lakes as repositories for organizing and accessing diverse data sources [2]. Data lakes offer a centralized and scalable infrastructure capable of accommodating structured, semi-structured, and unstructured data from various sources. They provide a flexible and cost-effective solution for data storage and management, thereby facilitating the integration and analysis of heterogeneous data sets. The concept of data lakes has gained substantial attention in both research and industry circles, with numerous studies highlighting their potential to facilitate effective data discovery processes and support data-driven decision-making.

2. LITERATURE REVIEW

This chapter highlights, from several studies, the evolution of data discovery from manual processes to catalogs and, more recently, to advanced data discovery tools that have significantly enhanced the efficiency and effectiveness of data exploration and utilization in various domains.

2.1. Traditional Data Discovery

Data discovery, in the days before the advent of data lakes, involved various methods aimed at identifying, organizing, and understanding data. These methods relied on manual cataloging, metadata extraction, search and query-based discovery, data profiling, data integration and consolidation, collaboration, and knowledge sharing.

Manual cataloging was a common approach to data discovery, involving the manual organization and categorization of data resources. This method relied on the expertise and efforts of individuals to create catalogs or inventories of available data sources [5].

Metadata extraction plays a crucial role in understanding and discovering data in structured web pages or databases. Techniques such as conditional random fields were employed to automatically extract metadata from web pages [14].

Search and query-based discovery involve using search engines or querying databases to find relevant data. Researchers and practitioners relied on keyword-based searches or structured queries to locate and retrieve specific data [18].

Data profiling was employed to gain insights into the characteristics and quality of the data. This method involved analyzing data sets to understand their structure, content, and relationships. Data profiling techniques helped identify data anomalies, inconsistencies, or missing values [28].

Data integration and consolidation were used to combine data from multiple sources to create a unified view. Techniques such as data mapping, schema matching, and data transformation were applied to integrate heterogeneous data sources into a single, coherent representation [7].

Collaboration and knowledge sharing played a vital role in data discovery. Professionals shared their domain expertise and insights, contributing to the collective understanding of available data sources. Collaboration platforms and communities of practice facilitate knowledge exchange and collaborative data discovery efforts [12].

These methods formed the foundation of data discovery before the advent of data lakes, providing insights into data sources, their characteristics, and relationships. While data lakes have since revolutionized the data discovery process, these earlier methods laid the groundwork for effective data management and understanding.

2.1.1. A new Architecture: Data lake

In recent years, data lakes have emerged as a promising solution for managing and accessing diverse data sources in organizations. As the volume and variety of data available to organizations continue to grow, traditional data management approaches often struggle to handle the scale and complexity of modern data sources. Data lakes (the load-First, model-Later approach) provide a flexible and scalable solution by allowing organizations to store raw and unprocessed data in its original format. This literature review will explore the various aspects of data lakes, including their potential to facilitate effective data discovery processes and support data-driven decision-making, in order to highlight the significance of data lakes in the modern organizational landscape. We will also examine studies on approaches for data discovery in order to get a comprehensive understanding of this technique in both theories and applications.

2.2. Techniques and Approaches for Data Discovery

Several studies have highlighted different techniques useful for data discovery. In the context of the absence of any metadata, a method of inferring join plans for a set of relation instances has been proposed [3]. The method enumerates the possible join plans in order of likelihood, based on the compatibility of a pair of columns and their suitability as join attributes. It sheds light on the challenges of performing join operations in schemaless data environments and proposes efficient algorithms for discovering join plans.

Utilizing Semantic Information is an important way of conducting efficient data discovery. In [6], the concept of PEXESO, which is a framework for joinable table discovery in data lakes, was introduced. Textual values are embedded as high-dimensional vectors, and columns are joined under similarity predicates on those high-dimensional vectors to address the limitations of equi-join approaches and identify more meaningful results. By leveraging similarity measures, organizations can identify and link tables within data lakes, enabling cross-domain analysis and insights. Another study also focuses on linking datasets using word embeddings for data discovery, a process referred to as "seeping semantics" [11]. By leveraging word embeddings in the enterprise knowledge graph, the study proposes an approach to identify and establish meaningful connections between datasets within data lakes.

The study [13] reveals the problem of discovering joinable datasets at scale and approaches it from a learning perspective, relying on profiles, which are concise representations that capture the underlying characteristics of the schemata and data values of datasets and can be efficiently extracted in a distributed and parallel fashion. Profiles are then compared to predict the quality of the join operation among a pair of attributes from different datasets, and this is implemented using NextiaJD over Apache Spark.

Interactive data discovery in data lakes is another concept proposed in recent years [1], which emphasizes the need for empowering users with tools and techniques that facilitate dynamic and intuitive exploration of data lakes. By providing interactive data discovery capabilities, organizations can enable users to navigate through vast amounts of data, uncover hidden patterns, and gain valuable insights.

Besides, the formidable challenges encountered by organizations in the identification and retrieval of specific datasets within the extensive repositories of data lakes are highlighted [2]. It is revealed that the development and implementation of efficient techniques for dataset discovery are recognized as critical for seamlessly integrating and analyzing disparate datasets, thereby enabling the generation of comprehensive insights to support effective decision-making.

In addition, effective data discovery processes also play a vital role in enabling informed decision-making based on empirical evidence, with data lakes serving as repositories for organizing and accessing diverse data sources.

2.3. Data Discovery Systems, Platforms, and Architectures

Several studies have investigated novel platforms, systems, and architectures that facilitate efficient data discovery processes and enhance the overall utilization of data lakes.

One such system is Aurum, introduced in [9], which is designed to facilitate the effective exploration and retrieval of relevant data within data lakes. The system employs advanced search techniques and metadata analysis to enable users to discover and understand the content of data lakes. Another system is Auctus, presented in [4], which is a search engine specifically designed for data discovery and augmentation in data lakes. The system employs advanced indexing and ranking techniques to provide users with comprehensive search results, enabling effective data exploration and integration.

Large-scale data discovery is another important aspect investigated in the study [10]. It proposes a data discovery system to overcome the challenges of discovering data within vast and diverse data lakes. The study [28] presents a data-driven domain discovery approach for structured datasets within data lakes. It focuses on identifying domain-specific datasets and their relationships. The architecture and metadata management of data lakes are explored in the study [29], which emphasizes the importance of proper data lake architecture and metadata management for efficient data discovery and utilization.

In summary, systems such as Aurum and Auctus contribute to the development of efficient search and exploration tools. New techniques for large-scale data discovery and data-driven domain discovery further enhance the capabilities of data discovery in data lakes. Additionally, research on data lake architecture and metadata management emphasizes the importance of proper design and organization for effective data discovery processes.

In general, the referenced studies contribute to the understanding and advancement of data lakes and data discovery processes, emphasizing their significance in the modern organizational landscape.

3. METHODS AND METHODOLOGY

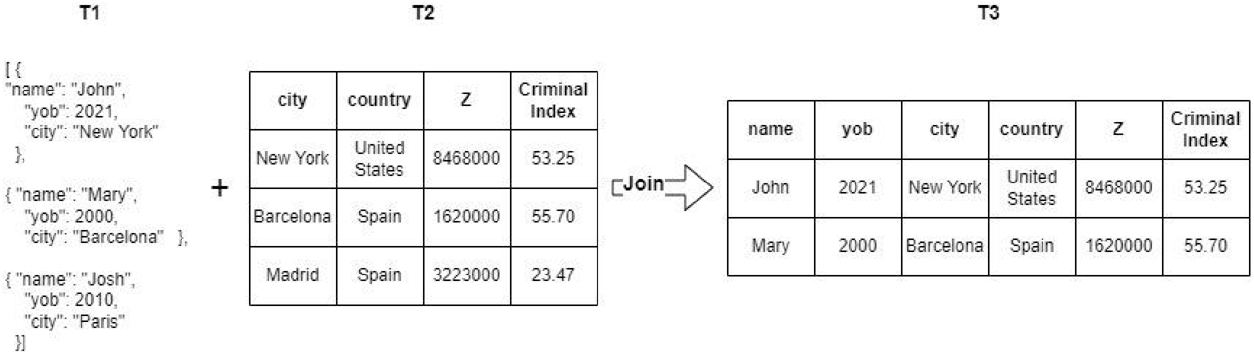
In this chapter, we will combine literature and diagrams to represent the process of data discovery with the help of examples. It will go deep into several classic works of literature and give an in-depth elaboration on its research methods.

In the field of data discovery, a major problem that needs to be solved is finding joinable attributes in different schemas for different heterogeneous data in the data lake. For example, as data analysts, we want to investigate the crime index of the city where people live. During the search process inside the data lake, we found two different data files, as shown in Figure 1, which we call T1 and T2, respectively. Among them, T1 is the information about people in JSON format, and T2 is a structured city information data table. Through observation, we found that the attribute city of T1 can be joined with the attribute city of T2 to form a merged structured data table, T3. This provides all the information we need to get the analysis job done.

Usually, tabular data in CSV format is the most common format in data lakes. As Figure 1 shows, since most other data formats in a data lake (e.g., JSON) can be converted to tabular data and then joined, a large amount of literature on data discovery in data lakes shows how to discover joinable columns between tabular data.

Therefore, the description of the methodology in this section also focuses primarily on data discovery in tabular data. Before explaining the methods proposed in the next few papers in detail, the previous analysis process in Figure 1 exposed several classic problems that may exist in structural data discovery. For example:

a. In the above discovery process, the city attributes of T1 and T2 were manually selected by us. How can we automatically identify the joinable attributes in the two structured data sets? The name of the attribute, and the characteristics of the value in the attribute (e.g., cardinality proportion, containment) may be helpful in this identification.
b. For the attribute for which we don't know the semantic characteristics (such as Z in the T2 table, which actually represents the population of each city), how do we identify the attributes that are joinable to it? How do we avoid having it join the wrong attribute? (For example, the yob attribute of T1 and the Z attribute of T2 are both integers; what if we mistakenly identify them as joinable attributes?)
c. For file T1, if we find another table T4 in the data discovery process, that also contains an attribute city, how do we evaluate whether the city attribute of table T2 is more suitable for joining, or the city attribute of table T4 is more suitable for joining?



```
[{
"name": "John",
   "yob": 2021,
   "city": "New York"
 },

{ "name": "Mary",
   "yob": 2000,
   "city": "Barcelona"  },

{ "name": "Josh",
   "yob": 2010,
   "city": "Paris"
 }]
```

city	country	Z	Criminal Index
New York	United States	8468000	53.25
Barcelona	Spain	1620000	55.70
Madrid	Spain	3223000	23.47

name	yob	city	country	Z	Criminal Index
John	2021	New York	United States	8468000	53.25
Mary	2000	Barcelona	Spain	1620000	55.70

Figure 1. Join Operation with Heterogeneous Data in the Data Lake

In order to solve these problems, we will illustrate literature on several typical methods from recent years to show the solutions.

3.1. Data discovery using Profiles

In [13], the authors propose an approach that leverages profiles as concise and informative representations of dataset schemata and values, capturing the underlying characteristics of datasets. This methodology aims to facilitate efficient and effective data discovery processes.

It involves several key steps, as summarized in Figure 2, including Profiling methods such as evaluation of cardinalities, value distribution, syntax, and relationships, as well as data discovery processes such as profile normalization, comparison, classification, and ranking. The general process outlined by the authors in [13] involves creating profiles that encompass both the structural information of the datasets (e.g., attribute names, types) and the statistical properties of the data values (e.g., distributions, cardinality). These profiles serve as compact summaries of the datasets' features and enable quick comparisons and evaluations.

To measure the similarity between datasets, the authors propose various similarity measures, considering attributes' names, values, formats, statistical properties, and semantic characteristics. These measures enable the identification of joinable attributes and help avoid erroneous joins.

The integration of profile-based techniques into their NextiaJD data discovery systems enables scalable data discovery. The profiles facilitate the matching and ranking of datasets based on their similarity to a given target dataset, aiding data scientists in identifying relevant datasets for their analysis tasks. The approach leverages parallel processing frameworks, such as Apache Spark, to ensure efficient profiling and discovery processes, even with large-scale datasets.

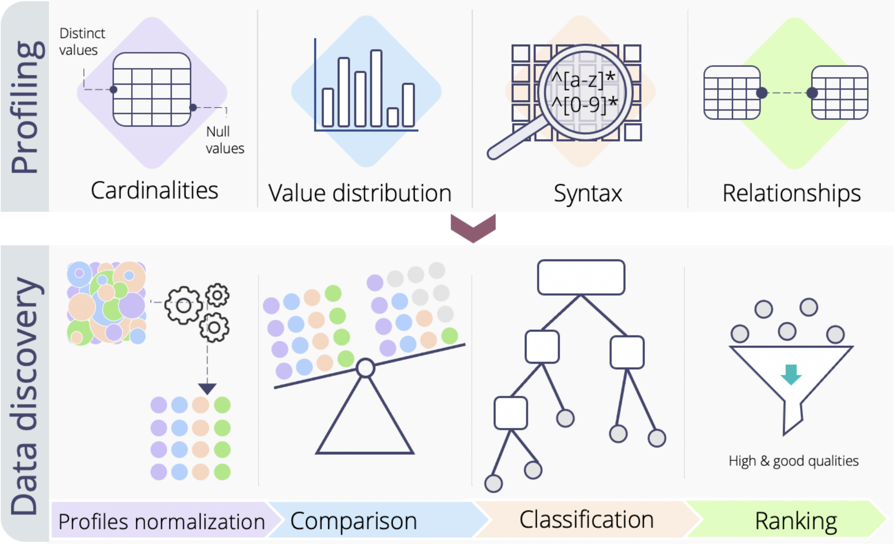


Figure 2: Profiling and Data Discovery Stages Implemented by NextiaJD [13]

For example, to conceptually address the join table challenge posed in this chapter, the process will involve:

a. Automatic identification of joinable attributes using a novel similarity measure that considers:

- Profiling on datasets to gather information such as cardinalities, value distributions, syntax, and relationships among the attributes.
- Normalizing the attribute profiles by standardizing the collected information for comparison purposes. Compare the normalized attribute profiles to identify attributes with similar characteristics, such as the name of the attribute and the characteristics of the values within the attribute.
- Applying similarity measures, such as considering cardinality proportion and containment, to determine the joinable attributes.

In our case, applying this approach to T1 and T2, we can easily identify the joinable attribute as “city”.

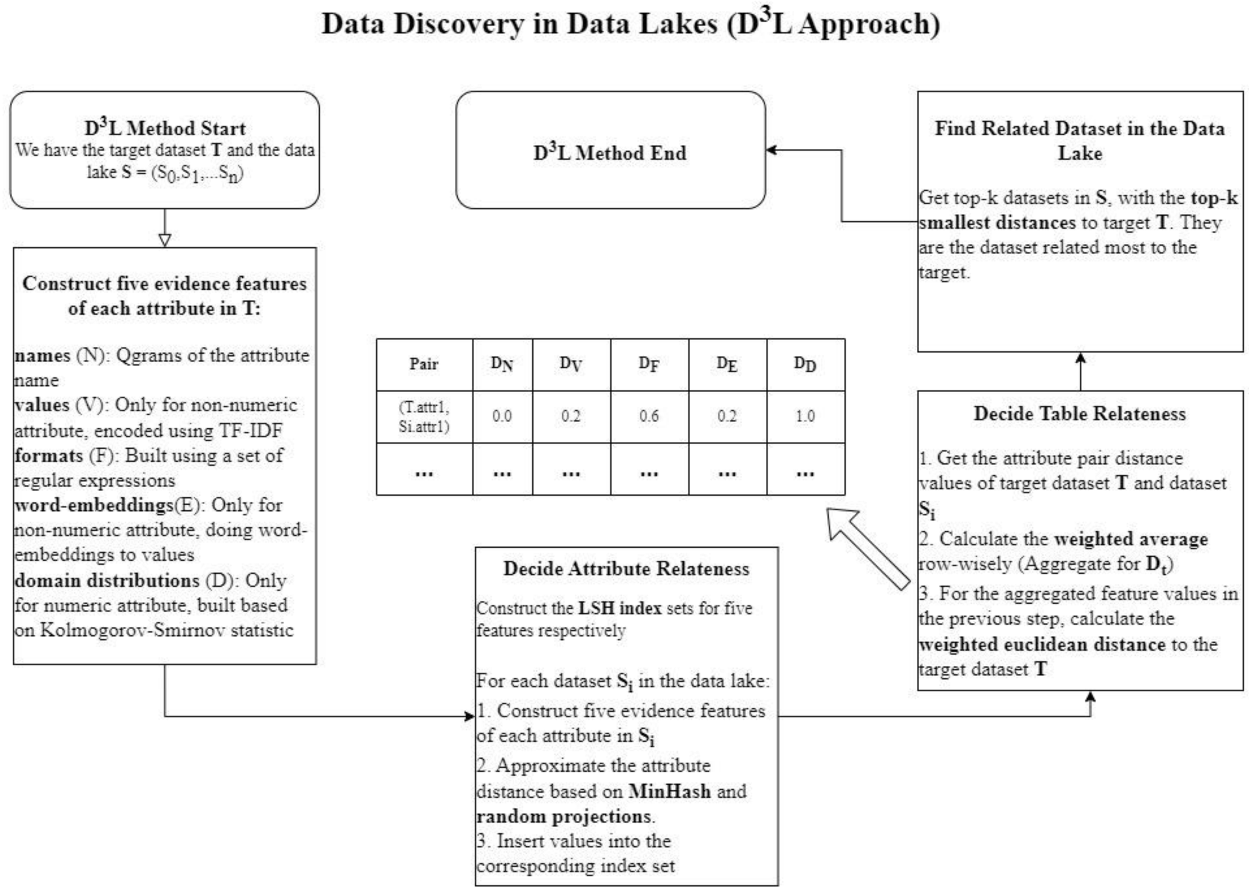
b. Identifying accurate joinable attributes with unknown semantic characteristics (e.g., attribute Z representing the population in T2), will involve repeating the underlying identification of joinable attributes as earlier. However, in addition, we compare the normalized attribute profile of the unknown attribute with other attributes to evaluate their compatibility based on data types, value distributions, and syntax. Afterward, we utilize the collected information to avoid mistakenly joining the unknown attribute with an incompatible attribute.

c. If we encounter another table, we need another evaluation of the suitability of joinable attributes (e.g., table T4). We repeat the same attribute profiling as earlier for a fair comparison. However, we must consider factors such as data quality, data completeness, and relevance to the specific data discovery task to determine which city attribute (e.g., from T2 or T4) is more suitable for joining.

Following these conceptual processes can address the far more challenging data discovery problem not stated in the methodology.

3.2. Dataset Discovery using Local Sensitive Hashing

In [2], the authors show a data discovery approach using local sensitive hashing called Dataset Discovery in Data Lakes (**D3L**). Figure 3 shows the primary process of this approach.



Data Discovery in Data Lakes (D^3L Approach)

D^3L Method Start
We have the target dataset **T** and the data lake S = (S_0,S_1,...S_n)

Construct five evidence features of each attribute in T:

names (N): Qgrams of the attribute name
values (V): Only for non-numeric attribute, encoded using TF-IDF
formats (F): Built using a set of regular expressions
word-embeddings(E): Only for non-numeric attribute, doing word-embeddings to values
domain distributions (D): Only for numeric attribute, built based on Kolmogorov-Smirnov statistic

Decide Attribute Relateness

Construct the **LSH index** sets for five features respectively

For each dataset S_i in the data lake:
1. Construct five evidence features of each attribute in S_i
2. Approximate the attribute distance based on **MinHash** and **random projections**.
3. Insert values into the corresponding index set

Pair	D_N	D_V	D_F	D_E	D_D
(T.attr1, Si.attr1)	0.0	0.2	0.6	0.2	1.0
...	...	...	...	...	...

Decide Table Relateness

1. Get the attribute pair distance values of target dataset **T** and dataset S_i
2. Calculate the **weighted average** row-wisely (Aggregate for D_t)
3. For the aggregated feature values in the previous step, calculate the **weighted euclidean distance** to the target dataset **T**

Find Related Dataset in the Data Lake

Get top-k datasets in **S**, with the **top-k smallest distances** to target **T**. They are the dataset related most to the target.

D^3L Method End

Figure 3: D3L Approach Process Diagram

In the D3L method, the authors take structural data as an example and transform the data discovery problem in the data lake into matching the top-k data frames closest to the target data frame. Moreover, this approach answers all the questions that we proposed at the beginning of Chapter 3.

For the joinability of a pair of attributes in the target data frame and the data frame in the data lake, the authors propose five features that can be used to measure it, which are names, values, formats, word-embeddings, and domain distributions.

- **names (N):** The name of the attribute. The same or similar attribute names are more likely to imply joinability. The authors represent it as a collection of q-grams, allowing for a better similarity measure.
- **values (V):** The value of the textual attribute. If the pair of attributes contains more equal or similar values, then they are likely to be joinable. The authors express it as the token collection after the TF-IDF operation.
- **formats (F):** The format of the attribute. e.g., email format follows a fixed regular expression; if the pair of attributes both follow the email's regular expression, then there is a good chance they are joinable. So the authors define a set of regular expressions to represent it.
- **word-embeddings (E):** The author converts the attribute values into a word embedding vector to measure the similarity of the content.
- **domain distributions (D):** It is also used to indicate value similarity, but only when the pair of attributes are both numeric. The authors measure it using the K-S (Kolmogorov-Smirnov) statistic.

For defining attribute relatedness, the Jaccard distance is used in the N, V, and F sets; the cosine distance is used in the E set, and the K-S statistic is used in the D set. The authors cleverly use the LSH index so that the less efficient pairwise comparison is avoided. Also, MinHash and random projections can approximate the original Jaccard distance and cosine distance with only a slight accuracy loss, but provide more efficiency in computations.

After calculating the attribute distance of two data frames, we can get the distance table showing the five feature distances of different attribute pairs. They are all in the range of 0 to 1. In order to aggregate from attribute-level relatedness to table-level relatedness, first aggregating through each pair and then aggregating through each feature is understandable. For the aggregation through rows, the authors collect all related attributes' distances in the data lake for each attribute in the target table, and use the cumulative probability distributions of these distances to select the weight, in order to balance the weakly related attributes. For the final Euclidean distance calculation, the authors define the target table T's coordinates as (0,0,0,0,0), and also use the same weight selection technique. Finally, the top-k-related data frames in the data lake can be obtained according to all table-level distances.

The authors answer our question proposed at the beginning of Chapter 3:

a. Those five features of the attribute pair help in the context of joining. Both the semantics and value characteristics of the column are included in the features. With the help of a clearly defined distance measure, we can get the relatedness between two attributes automatically.
b. Even if the semantics of those two attributes (e.g., attribute name) don't match, some other features containing value characteristics will help compensate for it. An attribute format is also just one feature in this calculation framework. This D3L method's calculation framework does not only contain these two characteristics that human eyes can easily recognize.
c. According to this distance-based attribute relatedness measure framework's calculation, we will be able to measure which attribute is more suitable for joining based on the calculated distance.

3.3. Method Comparison

In the previous two subsections, we focused on analyzing two approaches - [13] and [2] - for structured data discovery in data lakes. Here is our comparison of the similarities and differences between the two approaches:

- **Similarity:** Both extract the structural characteristics (such as attribute name and attribute format) and statistical characteristics of the data to form profiles or features. They both propose measurements to calculate the relatedness. They both incorporate a ranking mechanism to obtain relevance or similarity for matching as well as metrics to determine the relatedness of datasets.

- **Difference:** Those extracted characteristics are constructed differently. The relatedness measurements are different; [13] is similarity-based, whilst [2] is distance-based. The profile-based method has an additional step of normalization. The profile-based method is more scalable, and uses Apache Spark to leverage efficiency. In contrast, the D3L approach uses LSH-index to avoid inefficient pair-wise comparison, and uses MinHash and Random Projections to approximate the distance calculation to increase the computational efficiency.

4. DISCUSSION

4.1. Challenges and Opportunities in Data Discovery

The process of data discovery is not without its challenges. Researchers have investigated various aspects of data discovery and identified several challenges and opportunities in this domain. The study [8] addresses the challenge of conformance constraint discovery and proposes a methodology for measuring trust in data-driven systems. Conformance constraints ensure that data conforms to specified rules and standards, and their discovery is crucial for ensuring data quality and reliability. The study highlights the importance of trust in data-driven systems and provides insights into measuring trust through the discovery of conformance constraints.

The studies [22, 27] focus on relational data enrichment through discovery and transformation. It addresses the challenge of enriching existing relational datasets with additional relevant information. By leveraging discovery and transformation techniques, the research enables the integration of external data sources into relational databases, enhancing the richness and completeness of the data.

Data lake organization and management are key challenges in data discovery, and Nargesian et al. have made significant contributions in this area. The studies [23, 24, 25, 26] emphasize the need for effective organization and management strategies for data lakes. They propose approaches and techniques for organizing data within data lakes, ensuring efficient data discovery and utilization. These studies contribute to addressing the challenges of data lake organization and provide insights into effective data management practices.

In addition, the challenges and opportunities in data discovery are multifaceted. Researchers have addressed various aspects, including conformance constraint discovery, relational data enrichment, and data lake organization and management. By developing methodologies, techniques, and strategies, these studies contribute to improving data discovery processes and enhancing the value of data-driven systems.

4.2. Future Directions and Research Trends

The field of data discovery is constantly evolving, with new research trends and future directions emerging. We investigate this through several studies and highlight aspects of potential improvements in data discovery.

The study [15] considers the specific characteristics of big data, such as volume, velocity, and variety. It opines on the need to develop tailored approaches for data discovery in large-scale and complex data environments. Also, the study [20] emphasizes the importance of data reuse and exploration techniques for discovering and utilizing existing data services. By promoting the reuse of data services, researchers aim to enhance the efficiency and effectiveness of data discovery processes, reduce redundancy, and accelerate data-driven decision-making.

In [21], it proposes a data discovery platform empowered by knowledge graph technologies that can improve data discovery by capturing and representing the relationships and semantics of data elements. The use of knowledge graph technologies enables more intelligent and context-aware data discovery, facilitating more accurate and relevant results. In [24], it focuses on the organizational aspects of data discovery and proposes strategies for enhancing organizational capabilities for effectively leveraging data lakes. By considering factors such as roles, responsibilities, and processes within organizations, data discovery practices will be optimized, and the value extracted from data lakes will be maximized. Moreover, by exploring the challenge of demand-driven data provisioning in data lakes based on specific user demands, according to [19], we can optimize approaches and techniques, as well as streamline data discovery processes, to improve data access and availability in data lakes.

The study [16] opines that we can enable seamless data exploration and analysis for efficient discovery and integration of tables within data lakes, by developing advanced search and join algorithms. This will contribute to improving the discoverability and usability of data in data lake environments.

5. CONCLUSION

In this review, we discussed multiple aspects of data discovery, including its importance and benefit in different contexts and new techniques in its multiple sub-domains, especially in data lakes. For a detailed illustration of those new techniques, we mainly focused on data discovery with structural data in data lakes. We proposed three critical questions with an example for researching the methodology and comprehensively discussed a profile-based technique and a distance-based technique. Finally, the challenges and future directions of the data discovery domain are included.

The comprehensive exploration and integration of relevant data through data discovery leads to better-informed decision-making processes and the development of more accurate and robust predictive models. By leveraging multiple data discovery techniques, we can capture and represent the relationships and semantics of data elements, enabling more intelligent and context-aware discovery. Organizational capabilities and the efficiency of data discovery architectures can be enhanced to allow us to effectively leverage data lakes so that we can optimize data discovery practices and maximize the value extracted from them. Additionally, the development of advanced search and join algorithms can improve the discoverability and usability of data in data lake environments, which is beneficial for data exploration and analysis.

In conclusion, data discovery is crucial to facilitating efficient and effective data exploration, integration, and analysis. Leveraging technologies and strategies such as efficient discovery architectures and advanced matching algorithms can further enhance the data discovery process and increase the value derived from data lakes.

References

1. A. Ionescu, Katsifodimos, A., & Houben, G.-J. (2021). Interactive Data Discovery in Data Lakes. Very Large Data Bases.
2. Bogatu, A., A. A. Fernandes, A., W. Paton, N., & Konstantinou, N. (2020). Dataset Discovery in Data Lakes. arXiv: Databases. https://doi.org/https://doi.org/10.1109/icde48307.2020.00067
3. C. Acar, A., & Motro, A. (2009). Efficient discovery of join plans in schemaless data. International Database Engineering and Applications Symposium. https://cs.gmu.edu/~ami/research/publications/pdf/ideas09.pdf. https://doi.org/10.1145/1620432.1620434
4. Castelo, S., Rémi Rampin, Aécio Santos, Bessa, A., Chirigati, F., & Freire, J. (2021). Auctus: A search engine for data discovery and augmentation. Very Large Data Bases.
5. Smith, J., & Johnson, A. (2005). Data Management: A Practical Guide. New York, NY: Wiley.
6. Dong, Y., Takeoka, K., Xiao, C., & Oyamada, M. (2021). Efficient Joinable Table Discovery in Data Lakes: A High-Dimensional Similarity-Based Approach. International Conference on Data Engineering (CDE 2021). https://arxiv.org/abs/2010.13273. https://doi.org/10.1109/icde51399.2021.00046
7. Lenzerini, M. (2002). Data integration: A theoretical perspective. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 233-246). 41
8. Fariha, A., Tiwari, A., Radhakrishna, A., Gulwani, S., & Meliou, A. (2020). Conformance Constraint Discovery: Measuring Trust in Data-Driven Systems. Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM). arXiv: Databases.
9. Fernandez, R., Abidjan, Z., Koko, F., Yuan, G., Madden, S., & Stonebraker, M. (2018). Aurum: A Data Discovery System. International Conference on Data Engineering. https://doi.org/https://doi.org/10.1109/icde.2018.00094
10. Fernandez, R., Abidjan, Z., Madden, S., & Stonebraker, M. (2016). Towards large-scale data discovery. International Conference on Management of Data. https://doi.org/https://doi.org/10.1145/2948674.2948675
11. Fernandez, R., Mansour, E., Qahtan, A., K. Elmagarmid, A., F. Ilyas, I., Madden, S., Ouzzani, M., Stonebraker, M., & Tang, N. (2018). Seeping Semantics: Linking Datasets Using Word Embeddings for Data Discovery. International Conference on Data Engineering. https://doi.org/https://doi.org/10.1109/icde.2018.00093
12. Lave, J., & Wenger, E. (1991). Situated Learning: Legitimate Peripheral Participation. Cambridge University Press.
13. Gil Flores, J., Nadal, S., & E Romero, O. (2020). Scalable Data Discovery Using Profiles. https://arxiv.org/pdf/2012.00890.pdf.
14. Kim, W., & Lee, Y. (2008). Metadata extraction for structured web pages using conditional random fields. Journal of the American Society for Information Science and Technology, 59(8), 1213-1227.

15. K. Leung, C. (2018). Data Science for Big Data Applications and Services: Data Lake Management, Data Analytics and Visualization. Advances in Intelligent Systems and Computing. https://doi.org/https://doi.org/10.1007/978-981-15-8731-3_3
16. Zhu, X. (2019). Search and Join Algorithms for Tables in Data Lakes. Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM).
17. Kimball, R., Ross, M., Thornthwaite, W., Mundy, J., Becker, B., & Webb, J. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (3rd Edition). John Wiley & Sons.
18. Batini, C., & Scannapieco, M. (2016). Data Quality: Concepts, Methodologies, and Techniques. Springer.
19. Stach, C., Bräcker, J., Eichler, R., Giebler, C., & Mitschang, B. (2021). Demand-Driven Data Provisioning in Data Lakes. The 23rd International Conference on Information Integration and Web Intelligence. https://doi.org/https://doi.org/10.1145/3487664.3487784
20. Liu, Y.-H., Hsin-Liang (Oliver) Chen, Kato, M., Wu, M., & Gregory, K. (2021). Data Discovery and Reuse in Data Service Practices: A Global Perspective. Proceedings of the Association for Information Science and Technology. https://doi.org/https://doi.org/10.1002/pra2.510
21. Mansour, E. (2021). A Data Discovery Platform Empowered by Knowledge GraphTechnologies: Challenges and Opportunities. Very Large Data Bases.
22. Nargesian, F. (2019). Relational Data Enrichment by Discovery and Transformation. In Proceedings of the 35th IEEE International Conference on Data Engineering (ICDE).
23. Nargesian, F., Q. Pu, K., Ghadiri Bashardoost, B., Zhu, E., & Renée J. Miller. (2018). Data Lake Organization. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD). arXiv: Databases.
24. Nargesian, F., Q. Pu, K., Zhu, E., Ghadiri Bashardoost, B., & Renée J. Miller. (2018). Optimizing Organizations for Navigating Data Lakes.
25. Nargesian, F., Q. Pu, K., Zhu, E., Ghadiri Bashardoost, B., & Renée J. Miller. (2020). Organizing Data Lakes for Navigation. International Conference on Management of Data. https://doi.org/https://doi.org/10.1145/3318464.3380605
26. Nargesian, F., Qian Pu, K., Ghadiri Bashardoost, B., Zhu, E., & J. Miller, R. (2022). Data Lake Organization. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/https://doi.org/10.1109/tkde.2021.3091101
27. Nargesian, F., Zhu, E., Renée J. Miller, Q. Pu, K., & C. Arocena, P. (2019). Data lake management. Proceedings of the VLDB Endowment. https://doi.org/https://doi.org/10.14778/3352063.3352116
28. Ota, M., Müller, H., Freire, J., & Srivastava, D. (2020). Data-driven domain discovery for structured datasets. Proceedings of the VLDB Endowment. https://doi.org/https://doi.org/10.14778/3384345.3384346
29. Pegdwendé N. Sawadogo, & Jérôme Darmont. (2021). On data lake architectures and metadata management. Journal of Intelligent Information Systems. https://doi.org/https://doi.org/10.1007/s10844-020-00608-7