Toward Democratizing Access to Facilities Data: A Framework for Intelligent Data Discovery and Delivery

Yubo Qin, *Rutgers University, New Brunswick, NJ, 08901, USA*
Ivan Rodero and Manish Parashar, *University of Utah, Salt Lake City, UT, 84112, USA*

Data collected by large-scale instruments, observatories, and sensor networks (i.e., science facilities) are key enablers of scientific discoveries in many disciplines. However, ensuring that these data can be accessed, integrated, and analyzed in a democratized and timely manner remains a challenge. In this article, we explore how state-of-the-art techniques for data discovery and access can be adapted to facilitate data and develop a conceptual framework for intelligent data access and discovery.

Science in the 21st century is being transformed by our unprecedented ability to collect and process data from a variety of sources. At the same time, large-scale multiuser scientific observatories, instruments, and experimental platforms provide a broad community of researchers and educators with open access to shared-use infrastructure and data products generated from geo-distributed instruments and equipment. These large facilities (LF) have recently enabled significant scientific discoveries such as the detection of gravitational waves[1] and the imaging of the event horizon of a black hole.[2]

However, as the number of such LF and their scale increases along with corresponding growth in the number, distribution, and diversity of their users, ensuring that LF data can be discovered, accessed, integrated, and analyzed in a timely manner is resulting in significant demands on LF cyberinfrastructure (CI).[3] For example, the Ocean Observatory Initiative (OOI)[4] integrates over 1250 instruments, producing over 25,000 data items and over 100,000 data products. Similarly, each antenna of the Square Kilometre Array (SKA), the world's largest radio telescope project, produces raw data at the rate of approximately 0.5-1TB per second and approximately 300PB of data after preprocessing per telescope per year.[a]

Satisfying the overarching goal of LF, i.e., ensuring democratized and equitable access to their data and data products across the broadest set of users, can be challenging. Many LF (e.g., OOI) provide data-download portals and interfaces. However, discovering data/data-products using these portals can be challenging, especially for users from a different domain, due to data complexity, diversity, and volumes. Furthermore, data transfer time for the same high-resolution data can range from near-real-time streaming access to several weeks via shipped disk drives. For example, Dart *et al.*[5] demonstrated that transferring 56 TB of climate data to the NERSC computing center took up to three months due to network bandwidth and the poor performance of data transfer nodes. Additionally, access to low-latency, high-bandwidth network connections, and adequate computing and storage resources remains a significant challenge, especially for smaller and under-resourced institutions. Although national resources may be leveraged, such as those funded by the U.S. National Science Foundation (NSF), they are typically oversubscribed and are largely separate from the LF. Likewise, using them to process LF data requires users to download the data, get the data ready for their workflows, and then upload the data and the workflow to the national resource for execution. As a result, their

Digital Object Identifier 10.1109/MCSE.2022.3179408
Date of publication 1 June 2022; date of current version 31 August 2022.

[a]https://www.skatelescope.org/the-skaproject/

effective use in processing LF data is limited by the users' local resources.

Consequently, democratizing LF-enabled science requires new approaches for data discovery, access, and processing. In recent years, we have seen advances in related technologies and capabilities aimed at increasing access to commercial data and data services such as content delivery networks (i.e., a geographically distributed network of data centers and proxy servers with embedded data placement engines) and recommender systems. These technologies aid data discovery, proactively recommend data that are most relevant to the user and provide anytime/anywhere access to the data. Recent efforts also address how corresponding services can be developed for science data and to support science workflows.[6]

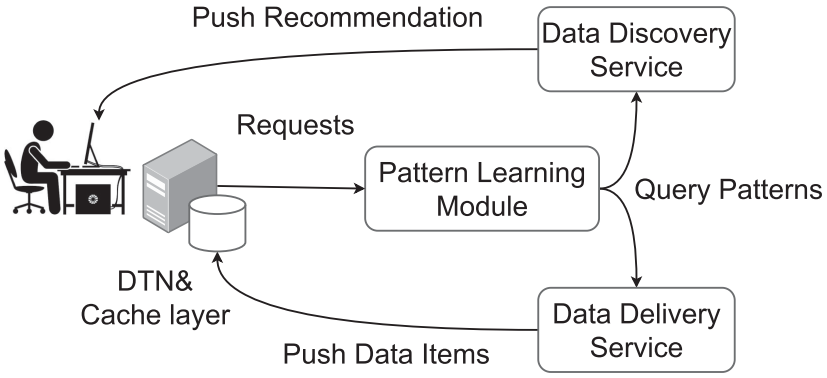

FIGURE 1. Overarching architecture of an intelligent data service framework.

The objective of this article is to explore how these approaches, coupled with an understanding of the data and their usage, can be effectively used to democratize access to LF data/products and accelerate the science enabled by LF. In this article, we build on concepts and technology presented in our previous work[8,9] to construct an intelligent data discovery and delivery framework composed of 1) user query analysis techniques that model access patterns and associated localities and affinities; 2) optimized data caching, data prefetching, and data steaming mechanisms to support optimized push-based data delivery; and 3) a data recommendation framework based on the collaborative knowledge-aware graph-attention network (CKAT) recommendation model to facilitate data discovery. We also present an evaluation[b] of the effectiveness and performance of these components using access traces from two NSF-funded large-scale observatories, the OOI, and the Geodetic Facility for the Advancement of Geoscience (GAGE). The results show how the data discovery and delivery framework and the mechanisms it provides can broaden access to LF. The key contribution of this article is a novel approach and framework for the discovery, access, and delivery of science data that leverage advances in the enterprise data technologies and complements existing services (e.g., search, metadata catalogs, knowledge bases, etc.). The article provides experiences and preliminary results to show that the framework is viable and useful and can be the basis for additional research and development efforts toward a production solution.

[b]The sources are available at https://gitlab.com/sci-data.

AN INTELLIGENT DATA DISCOVERY AND DELIVERY FRAMEWORK FOR LF

Several research advances in CI technologies and services can be leveraged to address LF data discovery and delivery challenges. For example, the demilitarized zone (DMZ) network model[7] creates a dedicated network to enable high-throughput data transfer for scientific data flows. A data transfer node (DTN)[8] provides an access point for users connecting to a DMZ network and is responsible for managing and optimizing data transfers. DTNs can also be used to analyze users' requests in the network; identify the patterns, localities, and affinities described above; and host services that use this information to improve data access performance. For example, frequently accessed data can be cached at DTNs. Furthermore, the analysis can be used to develop strategies for predicting future queries and for prefetching data to DTN storage closer to the user. Finally, the analysis of user data query patterns can be used for recommending other relevant data to users and to host such recommendation services at the DTNs. Finally, the DMZ and DTNs have been used to develop a federated data collaboration architecture, where DTNs are used to support access to data within a collaboratory. Specifically, the Virtual Data Collaboratory (VDC) project[3] implements a data-DMZ that supports data sharing and data-driven collaborations through DTNs. It also integrates data from other sources such as LF.

Building on these CI elements, we propose an intelligent data discovery and delivery framework, as illustrated in Figure 1. The *Analysis Module* analyzes user accesses to identify patterns as well as access localities and affinities. This knowledge is then used by the *Intelligent Data Delivery Service* to cache frequently used data, predict a user's subsequent queries, and prefetch the corresponding data. The *Data Discovery Service* uses the analysis along with domain-specific data models and user associations to generate a knowledge graph to implement a data recommendation service. Together, these services can help move us toward democratizing access to LF data following the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles. These components are discussed in the following sections.

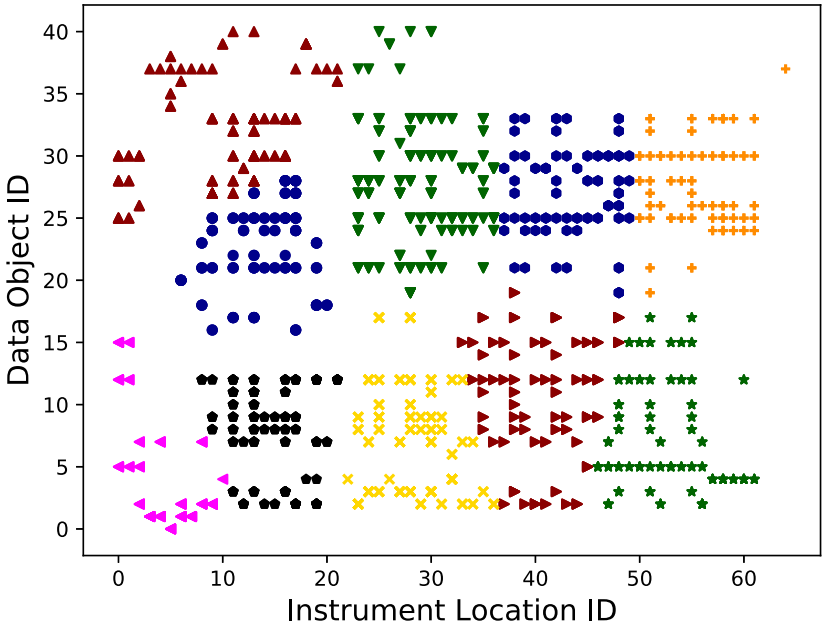

FIGURE 2. Representation of users' requests from a fragment of the OOI trace. Each cluster represents a distinct user (please see in color).

ANALYZING LF DATA AND DATA ACCESS

Analyzing LF data and user data usage and query patterns is essential to understand correlations and predicting user behaviors. Such analysis also enables the identification of ineffective practices or bottlenecks and thus can support the improvement of system performance. Our goal in this analysis is to classify users based on their queries or requests and model these queries to identify affinities that can anticipate future requests. To achieve this goal, we analyzed one year of requests from the OOI and GAGE access traces. OOI and GAGE support data discovery and access interactively through their web data portal and programmatically using an application programming interface (API). Our analysis showed that whereas most of the accesses in the traces were via the data portal, 90.1% of the data downloads used APIs and were triggered by workflows or scripts. As user access data using APIs (termed as *program users*) are the major data consumers, we focus on improving the access performance for these users. Program user requests are generated programmatically, these request patterns tend to remain consistent over time and can be used to develop models to predict future requests.

To identify consistent request patterns, we analyzed program user requests in the traces using different parameters, such as the time intervals between queries, the set of queried data, and the queried time range. We identified three access patterns: regular, overlapping, and real-time requests. Regular requests represent the most common request type and query new data since the last request without any overlap. Real-time requests are regular high-frequency requests typically used to monitor the occurrence of specific events. Overlapping requests are similar to regular requests but have overlaps in the queried time range across consecutive requests.

We also found significant overlap in the data queried in both traces. On the one hand, this overlap in queried data comes from overlapping time ranges across a user's consecutive requests. On the other hand, it results from similar data requests generated by different users (i.e., groups of users request similar data items). The overlap allows us to leverage data-caching mechanisms immediately to improve data access performance and reduce redundant queries and data transfers.

We also analyzed correlations across data queries and identified three key types of affinities:

1) *Facility instrument locality:* Data from instruments that are located close together tend to be queried together. LF typically deploy multiple

instruments in an area with high research value. As a result, users studying that area will naturally download data from some or all the instruments within the area. Consequently, instrument locality defines spatial affinities among data and data products and results in corresponding correlations across data queries. Our analysis of the OOI and GAGE access traces shows that, on average, users make 43.1% and 36.3% of their queries for data objects from instruments located in one region, and 51.6% and 68.8% of their queries are to the same data type, respectively. We also observed a temporal correlation across user queries in our analysis. Figure 2 shows the requested data objects (using instrument name and instrument location) from a fragment of the OOI trace clustered by users. The observed patterns suggest the existence of a spatial correlation across the requests as users request multiple data objects from one region and the same type of data object from nearby regions. We also observed a temporal correlation in our analysis.

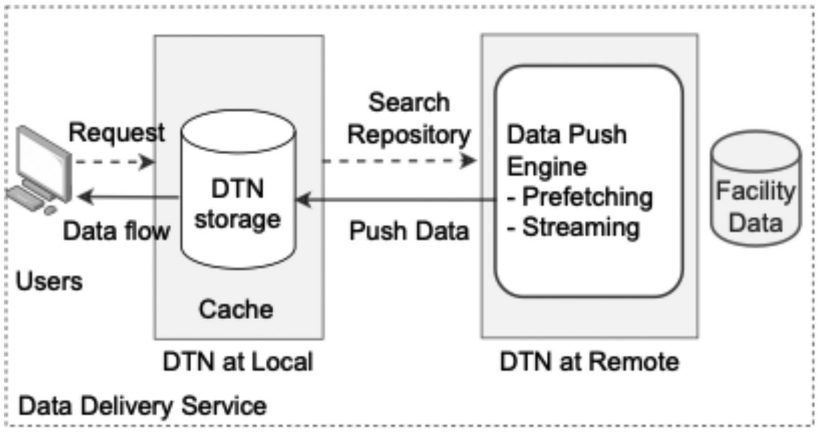

FIGURE 3. Architecture of the intelligent data delivery service.

2) *Domain data model:* Data produced by LF instruments and observations are typically used to derive data products, and the "recipes" used in this derivation are defined in the facilities' data models. For example, studies in oceanology use conductivity, temperature, and depth data to calculate water salinity and density. These domain-specific relationships define data-model affinities between data items and result in corresponding correlations across data queries. For example, conductivity, temperature, and depth data are correlated and are likely to be requested together to calculate water salinity and density.

User association: A classic association used in collaborative filtering recommendation models is that users with similar interests download similar data items. This association indicates that if two users have similar characteristics, such as research interests, they will likely request the same data items. Identifying such associations is difficult since facilities typically do not keep track of users or ask them to create profiles. However, our analysis shows that determining user similarity according to their geographical proximity is possible because LF users are typically part of larger research groups and/or projects and potentially part of the organization or institute. Such users would, with high probability, request similar data items. Consequently, we can leverage user locality as an indicator of data affinity and the resulting correlation across queries. Our analysis of the OOI and GAGE access traces validates user association and shows that users within the same research group (or same organization) tend to have similar data-query patterns.

INTELLIGENT DATA DELIVERY SERVICE

The data delivery service aims to improve the user data access performance by prefetching data to DTNs close to users and enabling them to access data primarily from the cache rather than retrieving them from the remote data source. The prefetching mechanism is based on user request history. As discussed above, over 90.1% of the volume of data downloaded is in response to queries from program users, i.e., programmatic queries generated by automated programs or scripts. These queries, by their nature, are predictable. As a result, by prefetching the relevant data items and caching requested data that can potentially serve future requests, the local cache at the DTN storage can serve a large fraction of user requests.

The data delivery service is designed based on these insights, as illustrated in Figure 3. The architecture consists of two primary functional components: the cache layer and the data push mechanism. The cache layer spans DTNs at the data sources (i.e., the LF) and at user locations, and forms a distributed interconnected cache network using the storage available at the DTNs. The goal of the data placement

strategy is to place the data at local DTNs that are close to potential users, and to keep data with a high probability of being accessed in the future in the cache network. The overall data placement strategy is composed of local caching based on the least recently used (LRU) policy and the creation of virtual groups. Virtual groups are groups of users who have common data interests and are geographically close to each other. We can place data objects of interest to a virtual group at a DTN that has the best connectivity to the corresponding set of users. We use *k*-means clustering to identify virtual groups. The data push mechanism is responsible for prefetching and streaming data based on user access patterns' analysis.

Clients run on DTNs at the user side and preprocess user requests by searching for the requested data in the cache layer. If the requested data are not present in the cache, the request is forwarded to a server, which runs at DTNs at the data source, e.g., the LF. The server also runs the data prefetching mechanisms and manages the placement of cached data. Although we aim to improve data delivery using a federation of DTNs, our approach is complementary to (and can be integrated with) existing data services such as those provided by open science grid (OSG), open storage network (OSN), and others.

The evaluation of the data delivery service is based on a simulated VDC with seven geographically distributed DTNs interconnected via a wide-area network. It uses the OOI and GAGE request traces to evaluate the effectiveness and performance of the data delivery service under various operational conditions and using different scenarios. The results show that the delivery service improves data delivery performance as well as the quality of service along different dimensions as compared to current practices. Key results of our evaluation are summarized below:

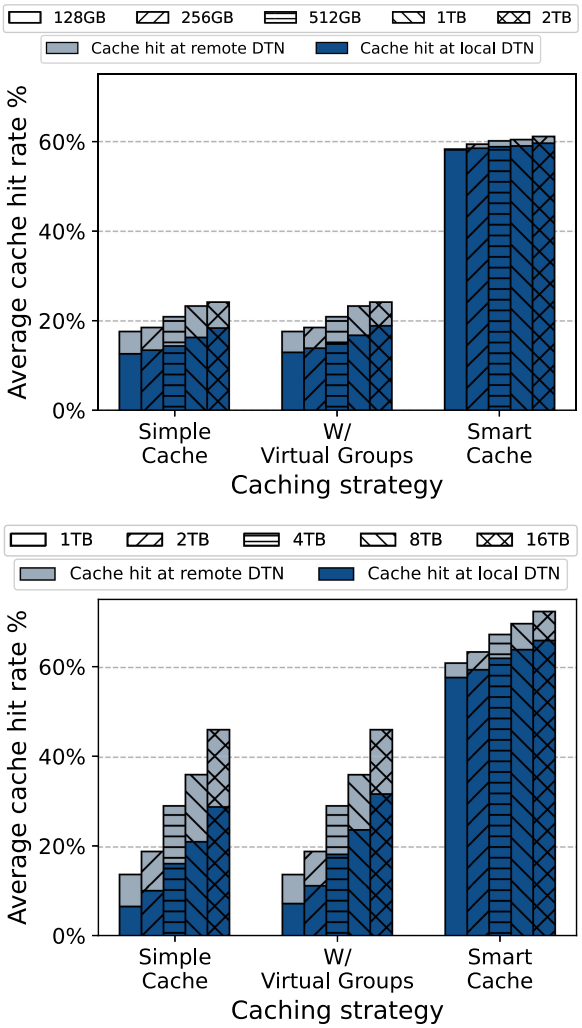

FIGURE 4. Percentage of data movement from the local cache for OOI (top) and GAGE (bottom).

1) *Higher throughput and lower latencies achieved:* The data delivery service improves data delivery throughput by over 2600 times, and reduces the latency from request submission to data access by 38%. These improvements are primarily the result of requests being served using cached data.
2) *Data delivery performance is more robust to network variation:* The data delivery service is more robust to the network variations. Data access performance does not dramatically change with changing network conditions. The overall network bandwidth requirements are also reduced, especially over the wide area, as redundant data requests are eliminated.
3) *Reduced load and network traffic at the LF:* The data delivery service reduces the total requests and corresponding network traffic at the LF, as many requests are satisfied using cached data at the DTN on the client side.

We also studied how prefetching improves local data reuse. Figure 4 shows the average percentage of requests that are served from a local cache, using virtual groups and caching with prefetching (referred to as "Smart Cache" in Figure 4). The results indicate that prefetching enables users to obtain a larger fraction of data from their local cache. Instead of passively searching cached data, the prefetching mechanism proactively pushes data to the user. Thus, it ensures that users can access more of their data locally regardless of whether they are reused from the previous requests. Furthermore, the prefetching mechanism can achieve near-

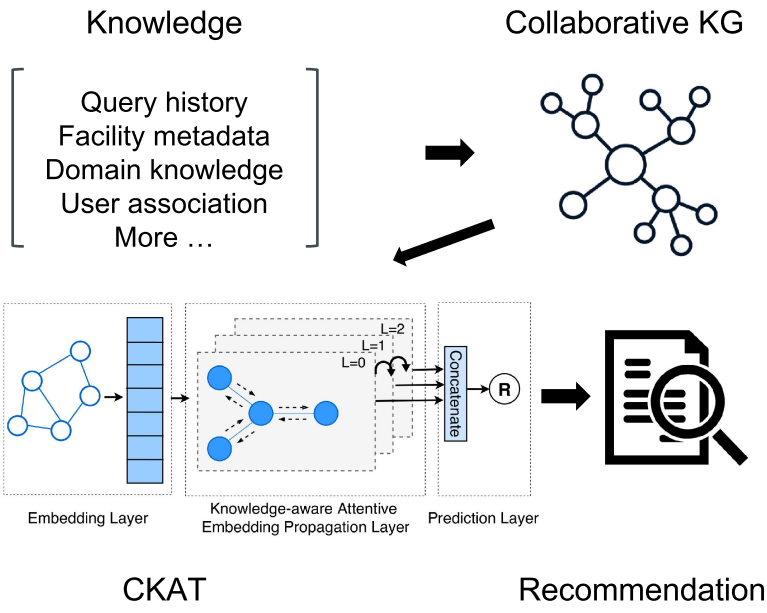

FIGURE 5. Overview of the recommendation process based on based on the CKAT model.

optimal performance with small cache size. Please refer to previous work by Qin *et al.*[9] for more details.

Overall, we observe that the proposed data delivery service significantly improves performance and enhances service robustness in response to complex real-world operational variations.

INTELLIGENT DATA DISCOVERY

As the number of data and data-products available at LF grows, discovering data/data-products of interest can be extremely challenging. The data discovery service aims to recommend data and data products to users that are most relevant to their research interests. However, most popular e-commerce recommendation models are based on linked data and rich metadata about a user's personal history and preferences. Such data may not be available and relevant when recommending data and data products from LF, and the existing models do not directly translate for such recommendations. In the case of LF users, data requests are based on research needs. Furthermore, facilities typically do not keep track of user histories or require users to create personal profiles listing their preferences. As a result, the data discovery service uses knowledge about user-query patterns and correlations across user queries along with domain-specific data models.

As noted earlier, our analysis of user requests to production LF identified three key affinities that characterize query behaviors: instrument locality, domain data model, and user association. Harvesting these affinities is critical to automating the data discovery process, and they can be obtained from a combination of information sources, including the facility instrument metadata, user query traces, and external sources such as Wikipedia. This information is then captured in a knowledge graph, which is an effective method for representing such information, capturing the facts as nodes and relationships among facts as paths in the graph. Several recently proposed recommendation models leverage knowledge graphs to carry auxiliary information to help address the cold-start and data-sparsity challenges.[10] In our case, the knowledge graph contains information about the three types of affinities described above.

To design a data discovery service capable of recommending relevant data to LF users, we developed a Collaborative Knowledge-aware ATtention network (CKAT) recommendation model. The overall recommendation generation process is summarized in Figure 5 and described below.

We first represent each knowledge source with an individual knowledge graph. To correlate the knowledge sources, we can consolidate these individual knowledge graphs into a Collaborative Knowledge Graph (CKG) using entity alignment.[10–12] The CKG enables diverse information to be connected in the graph to form a collaborative signal. The CKG construction process also allows us to examine different knowledge combinations, which is key to achieving sound recommendations. Paths in the CKG represent the connection of two data

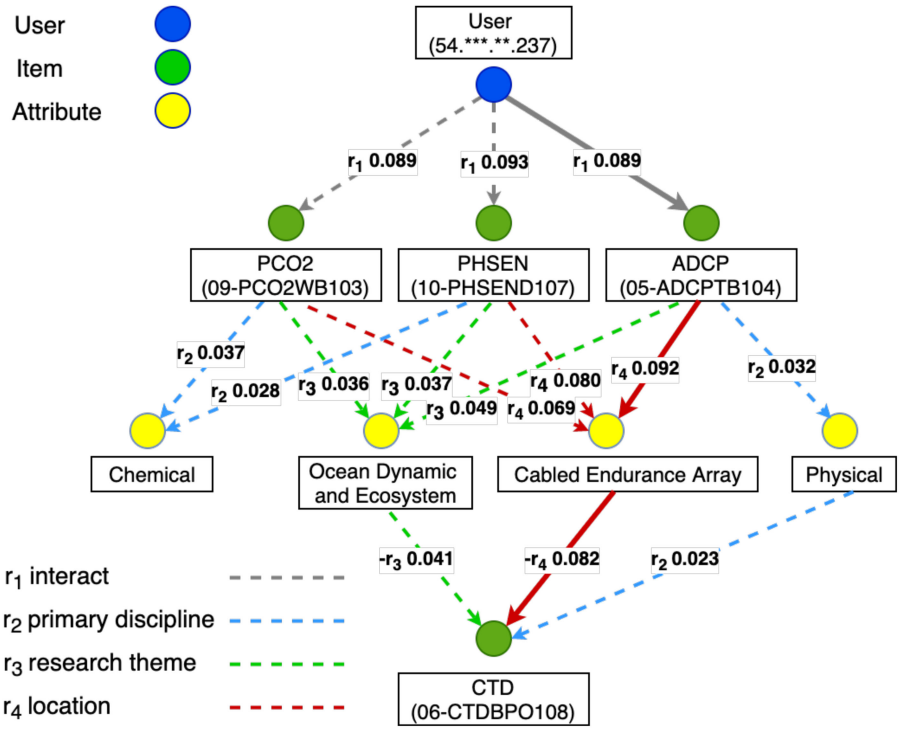

FIGURE 6. Sample recommendation outcome using OOI data. The arrows show the attention scores.

items, whereas first-order connectivity occurs when data items are directly connected, high-order connectivity occurs when there is a path between two data items across multiple nodes in the graph. The advantage of combining various knowledge sources using the CKG is the ability to identify connections between two indirectly related data items, which requires capturing long-distance paths (i.e., the high-order connectivity) in the graph. This can be achieved using a graph neural network (GNN). However, before sending the CKG to the GNN, we embed the graph representation into a vector representation. This embedding layer of the CKAT model initializes and parameterizes each node of the CKG using a vector representation. Specifically, we use the TransR embedding model by Lin *et al.*,[11] which considers relations in two distinct spaces, i.e., entity space and relationships spaces, and performs the translation from the entity space to the relationships space, thus reflecting the importance of two entities in different relationships. In the case of LF data usage, two data items can be used together for different research purposes and for which the correlation or importance of the two data items is different. For example, the correlation of physical environmental variables can be relevant for climate change research but can also help in the understanding of shorter-term effects such as the impact of invasive species on migratory species through predation. Therefore, being able to distinguish between these differences is important.

The knowledge-aware attentive embedding propagation layer of the CKAT model refines each node's representation by aggregating messages from its neighborhoods in the CKG and applies a knowledge-aware attention mechanism to learn the varying importance of each neighbor during a propagation. Recent work by Wang *et al.*[12] has demonstrated that GNNs can capture high-order connectivity through high-order information propagation. However, key issues may impact learning performance, such as irrelevant paths (i.e., noise) that can impact the ability to find actual correlations. Since nodes can be connected via different paths in the graph, not all of them have the same importance in a certain relation. Thus, we apply the attention mechanism[10] to enable the GNN training process to focus on the important relations.

Finally, the prediction layer of the CKAT model outputs the user–item pair prediction score by estimating the likelihood of an interaction based on the final representation.

An experimental evaluation of the CKAT-based data discovery service shows that it can effectively

recommend data to users. Its accuracy is over 6.12% higher than the state-of-the-art models such as those based on collaborative filtering, factorization, regularization, and propagation methods. Additional details about the evaluations and the baseline models are presented in Qin *et al.*[9]

The evaluation also shows that knowledge combination plays a key role in the results, indicating that knowledge sources must be carefully selected to obtain sound results. The results also differed between facilities, indicating that a preselection process is needed for each facility to achieve optimal results. A large number of knowledge sources do not necessarily provide better recommendation results. Only related knowledge sources are needed. Our experiments illustrate that when we purposely insert "noise" (i.e., irrelevant knowledge) to the best knowledge combination, the recommendation worsens, which emphasizes the importance of the knowledge selection process. Furthermore, when we disable the attention mechanism, the recommendation result is impacted by every knowledge source input, which illustrates that the attention mechanism does help eliminate the impacts of noisy knowledge sources and helps improve training accuracy. Figure 6 illustrates the high-order connectivity in inferring user preferences, i.e., using the attention score to represent the affinity in OOI data. The figure shows how CTD instrument data (used to measure the electrical conductivity, temperature, and pressure of seawater) are recommended when the user previously queried acoustic doppler current profiler (ADCP) instrument data, which are obtained at the same location (cabled endurance array). We observed that instrument locality is more influential than other general attributes.

As stated earlier, models using a GNN have been widely adopted by e-commerce and social media. The CKAT model demonstrates a new methodology and direction to assist users in discovering facility data through exploiting diverse knowledge sources. However, several challenges exist when targeting scientific data, as noted in Qin *et al.*[9]

CONCLUSION

Scientific exploration in the 21st century is increasingly leveraging data acquired from multiple distributed and diverse data sources and data-driven workflows that discover, access, and integrate this data. As a result, LF and CI have become an essential part of this exploration and will play an important role in future scientific discoveries. In this article, we explore solutions to these challenges associated with data discovery and access with the overarching goal of democratizing this access to the data and the data sources. Specifically, we explore how knowledge of user access behaviors coupled with advances in CI can be leveraged to achieve democratized access to LF data, and discuss a conceptual framework for intelligent data access and delivery.

However, several open challenges remain before access to LF data is truly democratized. For example, these include:

- The harvesting of metadata and the creation of the CKAT need to be automated. In the approach described in this article, we used a manual process to harvest this metadata from the LF.
- The applicability and effectiveness of the presented concepts and approach across different LF (beyond those presented in this article) and domains need to be verified. While we believe the approach will translate to other LF beyond those studied in this article, we have not done this yet.
- The proposed framework builds on a data fabric with the required capabilities and services. While VDC, which was leveraged in this work, has conceptualized and prototyped these components, a production deployment will be needed.

The effectiveness of the proposed approach can be increased by distilling knowledge about the connection between LF data, data products, and associated research. Furthermore, the proposed methods can deliver personalized recommendations by creating researcher profiles from publications and modern techniques such as natural language processing and knowledge representation learning.

We envision that the proposed framework and services will become a pervasive data CI available to all researchers as part of a national data fabric.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Grant OAC 1835692 and Grant OAC 1640834.

REFERENCES

1. B. P. Abbott *et al.*, "Observation of gravitational waves from a binary black hole merger," *Phys. Rev. Lett.*, vol. 116, no. 6, 2016, Art. no. 061102, doi: 10.1103/PhysRevLett.116.061102.

2. K. Akiyama *et al.*, "First m87 event horizon telescope results. IV. Imaging the central supermassive black hole," *Astrophys. J. Lett.*, vol. 875, no. 1, 2019, Art. no. L4, doi: 10.3847/2041-8213/ab0e85.
3. M. Parashar *et al.*, "The virtual data collaboratory: A regional cyberinfrastructure for collaborative data-driven research," *Comput. Sci. Eng.*, vol. 22, no. 3, pp. 79–92, May/Jun. 2020, doi: 10.1109/MCSE.2019.2908850.
4. I. Rodero and M. Parashar, "Data cyberinfrastructure for end-to-end science," *Comput. Sci. Eng.*, vol. 22, no. 5, pp. 60–71, Sep./Oct. 2020, doi: 10.1109/MCSE.2019.2892769.
5. E. Dart, M. F. Wehner, and Prabhat, "An assessment of data transfer performance for large-scale climate data analysis and recommendations for the data infrastructure for CMIP6," 2017, *arXiv:1709.09575*.
6. K. Fauvel *et al.*, "A distributed multi-sensor machine learning approach to earthquake early warning," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 403–411, doi: 10.1609/aaai.v34i01.5376.
7. L. Smarr *et al.*, "The pacific research platform: Making highspeed networking a reality for the scientist," *Pract. Experience Adv. Res. Comput.*, vol. 29, no. 8, pp. 1–29, 2018, doi: 10.1145/3219104.3219108.
8. Y. Qin *et al.*, "Leveraging user access patterns and advanced cyberinfrastructure to accelerate data delivery from shared-use scientific observatories," *Future Gener. Comput. Syst.*, vol. 122, pp. 14–27, 2021, doi: 10.1016/j.future.2021.03.004.
9. Y. Qin, I. Rodero, and M. Parashar, "Facilitating data discovery for large-scale science facilities using knowledge networks," in *Proc. IEEE Int. Parallel Distrib. Process. Symp.*, 2021, pp. 651–660, 2021, doi: 10.1109/IPDPS49936.2021.00073.
10. G. Wang *et al.*, "CKAN: Collaborative knowledge-aware attentive network for recommender systems," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 219–228, doi: 10.1145/3397271.3401141.
11. Y. Lin *et al.*, "Learning entity and relation embeddings for knowledge graph completion," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015 pp. 2181–2187, doi: 10.5555/2886521.2886624.
12. X. Wang *et al.*, "KGAT: Knowledge graph attention network for recommendation," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 950–958, doi: 10.1145/3292500.3330989.

YUBO QIN received the Ph.D. degree from Rutgers University, New Brunswick, NJ, 08901, USA. His research focused on addressing data discovery and geo-distributed data sharing challenges. Contact him at yubo.qin@rutgers.edu.

IVAN RODERO is research computer scientist at the Scientific Computing and Imaging (SCI) Institute , University of Utah, Salt Lake City, UT, 84112, USA. His research focuses on data-driven science and engineering, high performance parallel and distributed computing, and advanced cyberinfrastructure. He is senior member of IEEE and ACM. Contact him at ivan.rodero@utah.edu.

MANISH PARASHAR is the director and the chair in computational science and engineering, Scientific Computing and Imaging (SCI) Institute, and a professor with the School of Computing, University of Utah, Salt Lake City, UT, 84112, USA. His research interests are in the broad areas of parallel and distributed computing, and computational and data-enabled science and engineering. He is fellow of the AAAS, ACM, and IEEE. Contact him at manish.parashar@utah.edu.