Privacy-aware Data Federation Infrastructure*

Lili Jiang[1], Xuan-Son Vu[1], Addi Ait-Mlouk[1]
Anders Brändström[2], and Erik Elmroth[1]

[1]Department of Computing Science, Umeå University, Sweden
[2] Demographic and Aging Research Centre, Umeå University, Sweden
[1]{lili.jiang, sonvx, addi.ait-mlouk, elmroth}@cs.umu.se
[2] anders.brandstrom@umu.se

Abstract. Motivated by the need for cross-database analysis on heterogeneous data and facilitating distributed data usage, we are developing a data federation infrastructure for practical usage and research testbed. The major themes of this project are conducting academic research on data federation techniques and privacy preservation on data sharing. In this article, we present the project background, technology challenges, overall system framework, and academic solutions.

Keywords: Data federation · Privacy preservation· Heterogeneous data analysis · Knowledge graph

1 Background and Significance

A vastly increasing volume of digital data has been generated and collected by people across organizations (e.g., governments, academic institutions, business corporations, web users) for different purposes, in different schema, and using different methodologies. Facing the situation of information glut and a knowledge shortage, data integration and sharing becomes essential and valuable in handling large scale of heterogeneous data. It would be extraordinary beneficial if researchers could globally access the data and conduct research for real applications, such as chronic disease prevention, society security, demographic change, mobility and economic status surveillance, and elderly care. These demands pose real challenges on federating large scale of heterogeneous data sources. Especially data federation is threatened by user concerns from a privacy perspective, with sharing an increasing amount of information regarding their profile information, health, service usage and activities. Netflix released anonymized data for recommendation contest in 2007, but after two researchers identified several Netflix users in the dataset matching with IMDB, the Netflix contest had to be cancelled. And a deep neural network was trained to discern sexual orientations with accuracy higher than 80%, based on an internet dating website.

The goal of our project is to construct a privacy-aware data federation infrastructure based on federating distributed heterogeneous data sources, and offering

* supported by the Federated Database project funded by Umeå University, Sweden

data access/analysis to users with privacy concern [1]. Regarding test dataset, we used myPersonality project data corpus[2], collected from over 6 million volunteers on Facebook (FB) to form one of the largest social science research databases in history. The data was anonymized and sampled to share with collaborators registered scholars around the world.

2 Academic and Technical Challenges

2.1 Data Heterogeneity

- Model conflict: individual data sources use different data models, e.g., relational data model (SQL), unstructured (plain text), and/or semi-structured data (XML, SAS, SPSS).
- Schema conflict: different data sources use different names to represent the same attribute/concept, and vice versa (e.g., price & "cost", "date of birth" & "age").
- Instance conflict: conflicts on data value interpretation (e.g., "kg" and "gram", "Erik Johannsson" in different data sources), missing attributes, duplicates, contradicting records.

2.2 Privacy Concern

Federating and sharing data could improve scientific research, but the cost of obtaining consent to use individually identifiable information can be prohibitive. Sharing health-care and consumer data enables early detection of disease outbreak, but without provable privacy protection it is difficult to extend these surveillance measures nationally or internationally .

3 Framework Overview

Figure 1 presents our proposed privacy-aware data federation framework. There are three goals including data federation, data linkage, and data analytics. On the server side, data from multiple data sources were preprocessed and connected through a VDB in the local deployed Teiid server (teiid.jboss.org). Different techniques are firstly applied to process raw data including generating RDF from raw data and indexing text-based data (Process 1). Association rule analysis is applied to support data exploration, such as exploring hidden patterns and co-occurrences of variables from multiple data sources in visualized ways (Process 2). After data exploration, users can go to data search and issue queries across RDF endpoints for general/specific data analytics, which are powered by semantic web techniques (Process 3). Based on the three processes mentioned above, on the end-user side, users are enabled to view metadata of data sources, explore the individual and federated data, and further construct queries of their (research) interests, to get data analytic results with privacy concern.

[1] https://vimeo.com/319473546
[2] https://www.psychometrics.cam.ac.uk/productsservices/mypersonality

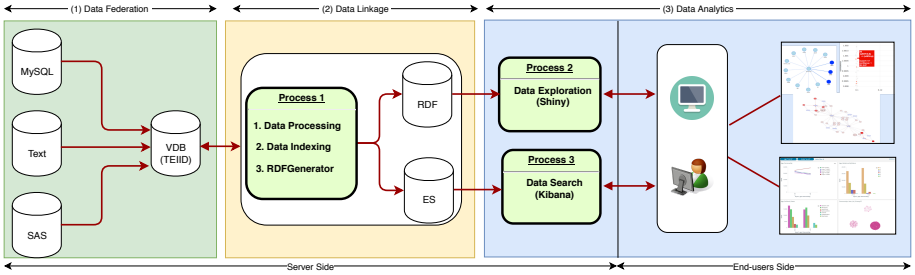

Fig. 1. The Overview of System Architecture

– **Data Federation and Data Linkage**: 1) Data Federation: our data federation component was built based on an open source framework Teiid, which is a data virtualization system that allows applications to use data from multiple, heterogeneous data stores. We created virtual database (VDB) for data federation, where data is accessed and virtually integrated in realtime across distributed data sources without copying or otherwise moving data from its system of record. 2) Data Linkage: after data federation, we applied semantic web techniques (e.g., RDF, SPARQL) for data linkage, which is a method for publishing structured data using vocabularies like schema.org that can be connected together and interpreted by machines. We created our own RDF generator, which standardizes the raw data to a unified RDF format and stored in RDF database as shown in Figure 1, so they can be read automatically by computers and enable data from different sources to be connected and queried. Afterwards, we apply the inverted indexing schema from ElasticSearch (www.elastic.co) (ES) for data indexing, which is an open source,distributable, and highly scalable search engine. The indexed results are stored to ES database. These two steps are critical to facilitate efficient data analytics in the following.
– **Data Analytics and Privacy Preservation**: 1) Data Exploration: after getting to understand the data sources, users are guided to apply data mining techniques to explore patterns and correlations over data variables. We take association rules technique as an example to show how data mining techniques discover the relationship between variables in federated data. In this system, we apply Apriori algorithm [1] to extract association rules among variables over the federated data. For example, given the FB users with specific variables like age (e.g., 31-40), gender (e.g, female), and relation status (e.g., married), the system may present association rule graph with quantified scores of the fourth variable "personality", which indicates the prone of specific personality (e.g, high-score neuroticism, low-score agreeableness etc.) of these types of users. 2) Data Search: after exploring the federated data sources, users come to the page of Data Search and issue queries over the indexed data on ElasticSearch. We have the pure long-text based variable (i.e., FB status update) to support text-based query, and additionally users

will add filters to setup constraints for other variables to customize their own query. More details and study cases can be found in our recently published work [2]. 3) Privacy Preservation: to address the privacy issue in our infrastructure, we have been applying differential privacy. The key challenge for differential privacy is determining a reasonable amount of privacy budget to balance privacy preserving and data utility. Most of the previous work applies unified privacy budget to all individual data, which leads to insufficient privacy protection for some individuals while over-protecting others. Firstly, we developed an improved differential privacy algorithm through personality-based knowledge extraction on personal sensitive data [3], where our proposed personality-based privacy preserving approach automatically calculates privacy budget for each individual. Further, we proposed a simple yet efficient generalized approach to apply differential privacy to text representation (i.e., word embedding) on user generated contents (UGC) [4]. Our experimental results show that the trained embedding models are applicable for the classic text analysis tasks (e.g., regression). The proposed approaches of learning differentially private embedding models are both framework- and data- independent, which facilitates the deployment and sharing. We have source code available at https://github.com/sonvx/dpText.

4 Conclusion

In this project, we have proposed a graph-based data federation system for heterogeneous data retrieval and analytics. After creating a Virtual Database (VDB) based on multiple data sources, we built our own RDF generator to unify data, together with SPARQL queries, to support semantic search by processing text data using natural language processing (NLP). Association rules analysis is used to recognize the most important relationships between properties from multiple data sources. We are in the progress of improving the infrastructure for more analysis assistance to researchers and more solid privacy preservation function. In near future, we plan to open the infrastructure for researcher to test.

References

1. Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB 1994). 487499.
2. Vu, X.S., Ait-Mlouk, A., Elmroth, E., Jiang, L.: Graph-based interactive data federation system for heterogeneous data retrieval and analytics. In: Demo Track, In: Proceedings of the The Web Conference 2019.
3. Vu, X.S., Jiang, L., Brandstrom, A., Elmroth, E.: Personality-based knowledge extraction for privacy-preserving data analysis. In: Proceedings of the Knowledge Capture Conference. pp. 45:1-45:4. K-CAP 2017.
4. Vu, X.S., Tran, S.N., Jiang, L.: dpUGC: Learn differentially private representation for user generated contents. In: Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing, April, 2019.