Graph-based Interactive Data Federation System for Heterogeneous Data Retrieval and Analytics

Xuan-Son Vu, Addi Ait-Mlouk, Erik Elmroth, Lili Jiang
Department of Computing Science, Umeå University, Sweden
{sonvx, addia, elmroth, lili.jiang}@cs.umu.se

ABSTRACT

Given the increasing number of heterogeneous data stored in relational databases, file systems or cloud environment, it needs to be easily accessed and semantically connected for further data analytic. The potential of data federation is largely untapped, this paper presents an interactive data federation system (https://vimeo.com/319473546) by applying large-scale techniques including heterogeneous data federation, natural language processing, association rules and semantic web to perform data retrieval and analytics on social network data. The system first creates a Virtual Database (VDB) to virtually integrate data from multiple data sources. Next, a RDF generator is built to unify data, together with SPARQL queries, to support semantic data search over the processed text data by natural language processing (NLP). Association rule analysis is used to discover the patterns and recognize the most important co-occurrences of variables from multiple data sources. The system demonstrates how it facilitates interactive data analytic towards different application scenarios (e.g., sentiment analysis, privacy-concern analysis, community detection).

KEYWORDS

heterogeneous data federation, RDF, interactive data analysis

ACM Reference Format:
Xuan-Son Vu, Addi Ait-Mlouk, Erik Elmroth, Lili Jiang. 2019. Graph-based Interactive Data Federation System for Heterogeneous Data Retrieval and Analytics. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308558.3314138

1 INTRODUCTION

Motivation. The ultra-connected world has been generating massive volumes of heterogeneous data stored in different data sources. And these data sources need to be normalized and interconnected to create a federated database that can be used to analyze, extract useful knowledge, and present it as a valid element for decision making. Semantic web techniques (e.g., RDF, SPARQL) have been widely used for data federation and linkage. However, the primary issue of the semantic web is insufficient integrated solution. To tackle this issue, we applied data federation, semantic web, and data mining technologies to develop this system, which can complement with each other. The system allows users to select data sources, interact with visualized graphs, and run customized queries across federated data to meet specific needs for data analytics.

Related Work. Related with data federation, there exist some popular enterprise data virtualization tools, such as IBM InfoSphere Federation Server (https://ibm.co/2qWQbom) and Oracle Data Service Integrator (https://goo.gl/6MKXkF). As well as some open source frameworks such as Teiid (teiid.jboss.org) which we used in this paper. Similar efforts in data federation have been seen from academia such as BioMart (www.ensembl.org/biomart) and Maelstrom (www.maelstrom-research.org). In the context of RDF and Linked Open Data, various works have been proposed in the literatures [2, 3]. In addition, several tools offering RDF and linked data visualization have been developed e.g., Sgvizler [6], LODWheel [7], IsaViz [1]. RDF-Gravity[2], etc. Also, many researchers used federated SPARQL queries to analyze and visualize linked data[4]. However, considering advanced data analytics across federated data is ignored. In this demo paper, we proposed RDF-supported data visualization framework over federated databases enriched with data mining (e.g., association rules) and NLP techniques (e.g., sentiment analysis). It can efficiently federate and analyze large heterogeneous data sources for general or specific analysis needs (e.g., community detection and the like).

Dataset. We processed one of the largest social science research databases, myPersonality corpus [3], which was collected from over 6 million volunteers on Facebook (FB). The data was anonymized and sampled to share with registered scholars around the world. In this paper, we used four of its data sources including *demographic dataset, personality dataset, political views*, and *FB status updates dataset*. One more *community detection dataset* was inferred based on these four. More information about these data sources is available in the menu of Data Source of the demo system.

Contributions. We build up a federation system, mapping the multiple heterogeneous, distributed, and autonomous data sources into a unified federated database interface, where user can choose data sources in their area of interest; we provide data analytics including data exploration and search, which empowers users with ability to explore the data via data mining algorithms (i.e., association rules), and search by queries to lead advanced analytics. we implement interactive visualization, which allows users to plot the result in different formats like tables, graphs, scatter plots, and download the results. The system is scalable by adding other data sources, applying other data mining algorithms, and aiming at other data analytic scenarios.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3314138

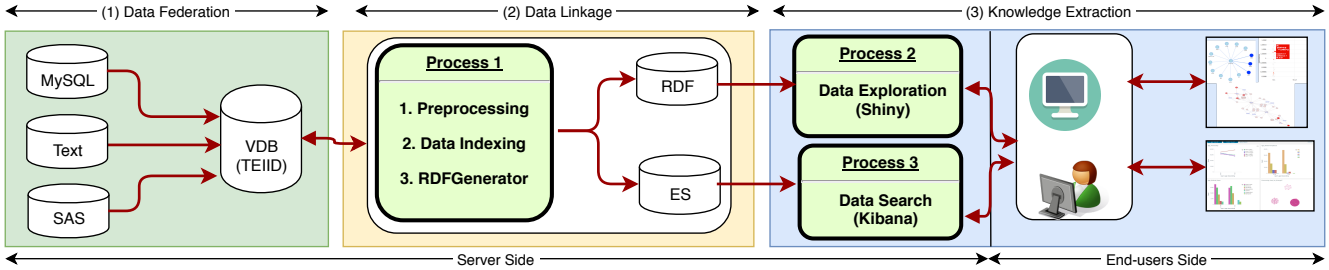

Figure 1: Overview of system architecture: (1) Data Federation, (2) Data Linkage, and (3) Data Analytics.

2 SYSTEM DESCRIPTION

The system architecture is shown in Figure 1 with 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. 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 simple/advanced queries of their (research) interests, to get data analytic results. Figure 2 presents the user interface.

Ahead of the above procedure, we apply NLP techniques [10] for raw data preprocessing (e.g., data normalization, standardization, non-utf8 characters removal etc.) and infer additional variables such as sentiment analysis and privacy concern based on the given variables. More information about these variables can be found in the menu of *Data Source* in the system.

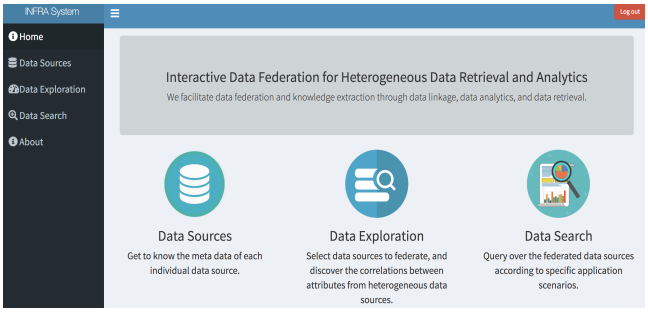

Figure 2: User interface of the system

[1]www.w3.org/2001/11/IsaViz
[2]https://www.salzburgresearch.at/publikation/rdf-gravity-3/
[3]https://www.psychometrics.cam.ac.uk/productsservices/mypersonality

2.1 Data Federation and Data Linkage

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 real-time across distributed data sources without copying or otherwise moving data from its system of record.

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.

2.2 Data Analytics

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. Association rules technique was initiated by Agrawal [1] to analyze transactional databases. It usually defined as an implication of the form $A \rightarrow B$ such as $A, B \subset I$ and $A \cap B = \emptyset$. Every rule is composed of two different sets of items A and B, where A is called antecedent or left-hand-side (LHS) and B called consequent or right-hand-side (RHS). In order to select interesting rules from the set of all possible rules, constraints on various measures of significance and interest are used. The best-known constraints are minimum thresholds on support and confidence. The support is defined as the proportion of transactions in the database which contain the items A, and the confidence defines how frequently items in B appear in a transaction that contains A. 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.

Data Search. After exploring the federated data sources, users

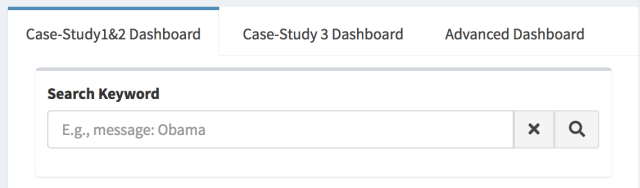

Figure 3: *Data Search* interface of the system

come to the page of *Data Search* and issue queries over the indexed data on ElasticSearch. As shown in Figure 3, for demonstration purpose, we created two dashboards of *Case-Study 1&2* and *Case-Study 3* for users. Each dashboard provides a search box as well as filter configurations. 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. The third dashboard supports general advanced search (e.g., Boolean search).

3 DEMONSTRATION OF THE SYSTEM

For demonstration, let's say Alice, a psychologist researcher, who wants to research the correlation of personality and stresses in people's life based on social network behaviours. Alice is enabled to apply given data mining algorithms (i.e., association rules) to explore the hidden patterns across selected variables of interests. Next, Alice will further extrapolate the returned results in *Data Search* to have more detailed information of those people. These two steps will be described in case-study 1 and case-study 2 accordingly.

3.1 Case-study 1: association rules based sentiment analysis

In this case study, Alice will select her variables of interests including sentiment_score_subjectivity (sentiS), cNEU (neuroticism), cCON (conscientiousness), cAGR (agreeableness) [8] in the graph visualization panel (1) to be inferred through association rules. We applied the Apriori algorithm [1] to extract frequent variables, which satisfies the minimum support requirement specified by the user, and then generate association rules based on the user-specified confidence threshold. As shown in Figure 4, in *Data Exploration*, the four variables were chosen to run the inference. Parameter configuration panel (2) was displayed on right side including selecting data source, setting support and confidence thresholds for association rules. Above the configuration panel, a component of "User Guide–Help" gives user guidance of exploring this page. Afterwards, Alice moves to the bottom panel (3) and clicks "association rule graph" tab to see the results like in Figure 5. In this figure, the rectangles represent variables and the circles represent association rules. Larger size of circle imply more data records matching the rule, while the darker circle represent more importance of the rule.

Regarding the research question, Alice found a hidden pattern between three variables regarding neurotic people (cNEU), which is a personality trait that reflects one's ability to deal with emotion states, such as stress and anxiety. The pattern suggests that people who are not agreeableness (N.AGR) and do not have strong purposes in the way they say on FB status (sentiS=0) then most likely they are neurotic people. This hidden pattern however might not be sufficiently significant to draw such a conclusion to define neurotic people. Therefore, Alice will verify this discovered pattern in *Data Search* to confirm this pattern across data sources. The case-study 2 will explain in more detail how Alice confirms the explored pattern.

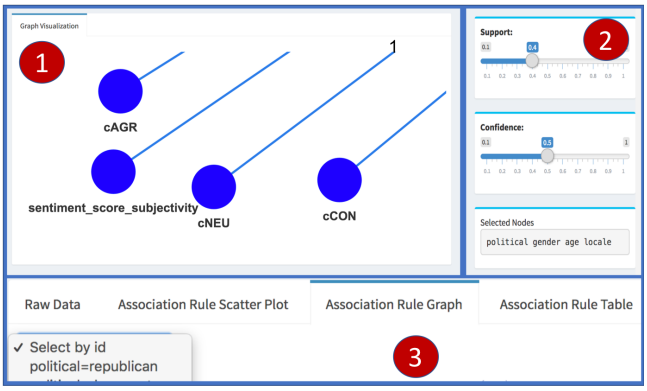

Figure 4: Selected variables on *Data Exploration*

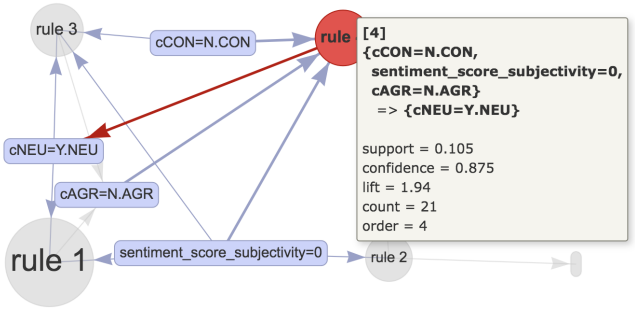

Figure 5: Case study of association rule based sentiment analysis on selected variables.

3.2 Case-study 2: personality and sentiment analysis

In this case study, we show how *Data Search* can be used to confirm the hidden pattern that Alice found out using the association rules. As shown in Figure 6, in *Data Search*, after issuing the query "sentiment_score_subjectivity:0 AND cagr:N.AGR AND ccon:N.CON" (Q1), Alice got a line chart showing the fraction between different personality traits and age group. It clearly shows that the neurotic group of people is the major group among five personality traits. Furthermore, Alice can find more information related to this group of neurotic people. For instance, a graph called "Age_Political_Views" shows that, mainly *neurotic people* have political views of "doesn't care" and "democratic". This extrapolation is a

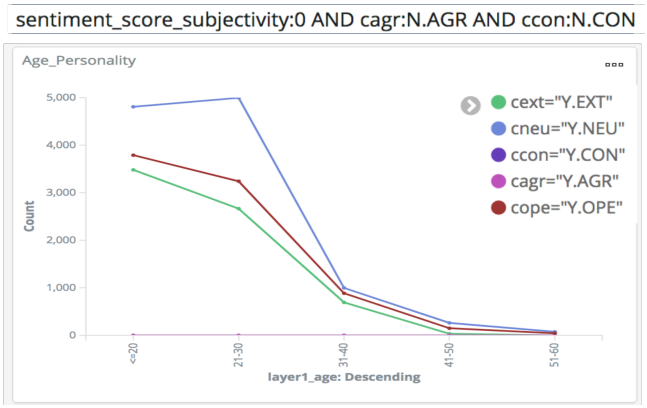

Figure 6: Case study of using *Data Search* to reconfirm the results from *Data Exploration*

powerful feature that *Data Search* can provide for data analysts. Due to the limited space, we cannot show all analytic graphs that Alice can see based on her query. Thus, we specially create a dashboard called *Case-Study 1&2 Dashboard* for this demonstration. Anyhow, Alice can also issue the query Q1 to "Advanced Dashboard" and find more related information to the neurotic group of people to confirm the hidden pattern.

Remarks: In order to have a finally verified answer on the research question, Alice must also run a number of statistical analysis to reconfirm the statistical signification of the pattern. Nevertheless, Alice's job now already becomes much easier. In future, we plan to add statistical analysis to the system as well.

3.3 Case-study 3: privacy-concern analysis

This case study shows how privacy-concern analysis is done through community detection and data analysis based on different variables (i.e., age, gender, personality, FB status). It is clearly that privacy-concerns on social network is an important factor to protect people's privacy on personal data [9]. Since there was not any user defined privacy-concerns information in the given dataset, we apply an approach from [10] to infer people's privacy based on their FB status and personality. As shown in Figure 7, in *Data Search*, the *Case-Study 3 Dashboard* displays four different charts to characterize privacy-concerns of people across data sources and their demographic information. The chart (1) shows the personality difference based on different age groups, chart (2) shows the number of communities based on different privacy-concerns, chart (3) shows the correlation between privacy-concerns and age, and chart (4) shows the association between gender, privacy, and sentiment embodied in their FB status. Especially, users can simply search based on keywords. For instance, if a user issues a query "Obama" and sees people's sentiment polarity (e.g., positive, neutral, negative) when they talked about Obama. It is found that 56.9% of users who mentioned the word "Obama" are females, in which 19.95% of them have negative sentiment when they mentioned about Obama. After digging deeper into the raw data, we realized that FB users complained about Obama mainly because of high unemployment rate in USA during his presidency, such as one post saying "because of obama there are no more jobs left in america".

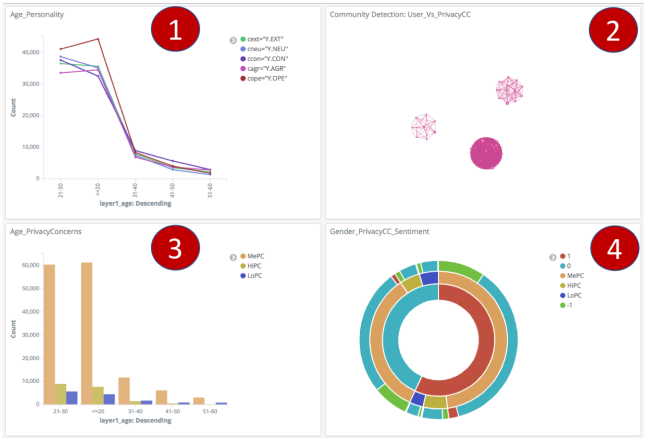

Figure 7: Case study of privacy-concern analysis

3.4 Implementation and demo environment

The system was implemented in Java and R. Data federation mechanism was built on Teiid. We developed the interactive and user-friendly interfaces using R shiny (shiny.rstudio.com/) and ArulesViz [5]. We built our own RDFGenerator to generate RDF files from VDB (Virtual Data Bases), and stored them in Apache Jenna Fuseki (jena.apache.org). To support *Data Search*, we employed ElasticSearch and Kibana for high performance data retrieval. Thanks to the plugin based architecture, Kibana can be easily extended to suit particular needs. For instance, according to the different data analytic requirements from different end-users, we support extended customized dashboard for each user need.

Brief information of the web-based prototype:

- System video URL: https://vimeo.com/319473546
- Multiple Operating System(s): Linux, macOS, Windows
- State-of-the-art browsers e.g., Chrome (recommended)
- Memory: from 1GB

4 CONCLUSION

In this paper, we have proposed a graph-based data federation system for heterogeneous data analytics. It is an open-source SPARQL query builder and result-set visualizer for heterogeneous data sources (i.e., myPersonality corpus) which allows users to easily construct and explore data over heterogeneous data sources. The system is scalable by easily adding new heterogeneous data sources, and customizing data representation and analytics by users, especially researchers based on their own interests. The future work will be adding more text-based data sources, extending more data mining algorithms options for data exploration to accompany data search.

ACKNOWLEDGEMENT

This research is funded by Umeå University in Sweden on federated database research. The authors also thank Michal Kosinski, David Stillwell and the myPersonality project for data sharing.

REFERENCES

[1] Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 487–499.

[2] Miika Alonen, Tomi Kauppinen, Osma Suominen, and Eero Hyvönen. 2013. Exploring the Linked University Data with Visualization Tools. In *The Semantic Web: ESWC 2013 Satellite Events*, Philipp Cimiano, Miriam Fernández, Vanessa Lopez, Stefan Schlobach, and Johanna Völker (Eds.). Springer Berlin Heidelberg, 204–208.

[3] Josep Maria Brunetti, Sören Auer, and Roberto García. 2012. The Linked Data Visualization Model. In *Proceedings of the 2012th International Conference on Posters & Demonstrations Track - Volume 914 (ISWC-PD'12)*. Germany, 5–8.

[4] Marija Djokic-Petrovic, Vladimir Cvjetkovic, Jeremy Yang, Marko Zivanovic, and David J. Wild. 2017. PIBAS FedSPARQL: a web-based platform for integration and exploration of bioinformatics datasets. *Journal of Biomedical Semantics* 8, 1 (20 Sep 2017), 42.

[5] Michael Hahsler and Radoslaw Karpienko. 2017. Visualizing association rules in hierarchical groups. *Journal of Business Economics* 87, 3 (01 Apr 2017), 317–335.

[6] Martin G. Skjæveland. 2015. Sgvizler: A JavaScript Wrapper for Easy Visualization of SPARQL Result Sets. In *The Semantic Web: ESWC 2012 Satellite Events*, Elena Simperl, Barry Norton, Dunja Mladenic, Emanuele Della Valle, Irini Fundulaki, Alexandre Passant, and Raphaël Troncy (Eds.). Springer Berlin, 361–365.

[7] Magnus Stuhr, Dumitru Roman, and David Norheim. 2010. LODWheel - JavaScript-based Visualization of RDF Data -. In *Proceedings of the Second International Conference on Consuming Linked Data - Volume 782 (COLD'11)*. Germany, 73–84.

[8] Xuan-Son Vu, Lucie Flekova, Lili Jiang, and Iryna Gurevych. 2018. Lexical-semantic resources: yet powerful resources for automatic personality classification. In *Proceedings of the 9th Global WordNet Conference*. 173–182.

[9] Xuan-Son Vu and Lili Jiang. 2018. Self-adaptive Privacy Concern Detection for User-generated Content. In *Proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing, Long papers, p., March 2018*.

[10] Xuan-Son Vu, Lili Jiang, Anders Brändström, and Erik Elmroth. 2017. Personality-based Knowledge Extraction for Privacy-preserving Data Analysis. In *Proceedings of the Knowledge Capture Conference (K-CAP 2017)*. ACM, USA, Article 45, 4 pages.