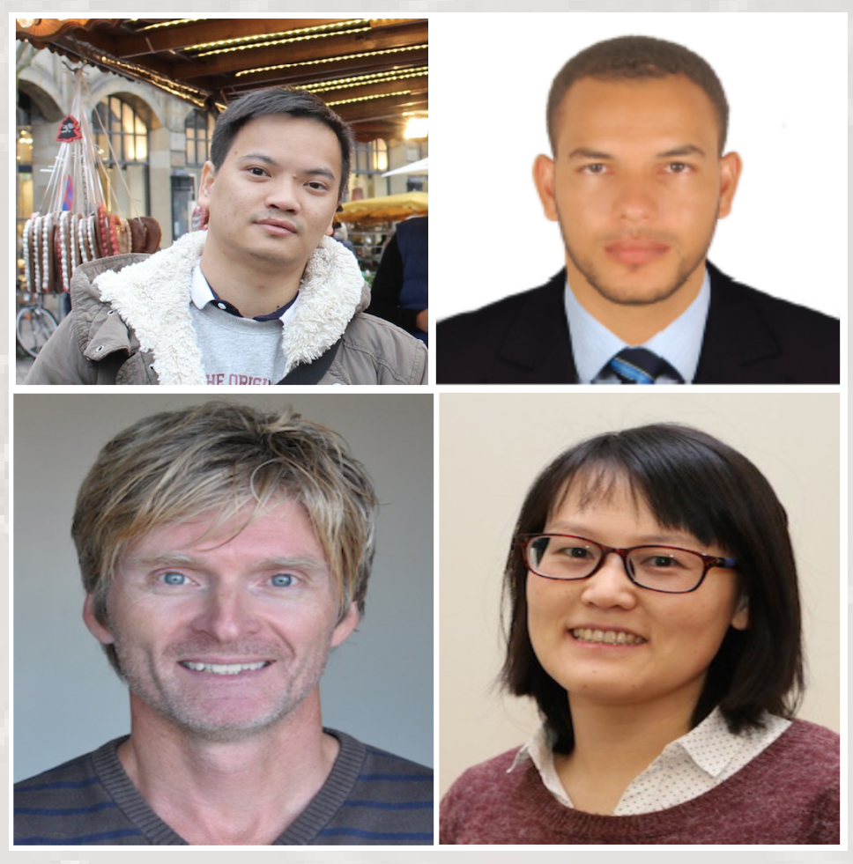


Graph-based Interactive Data Federation System for Heterogeneous Data Retrieval and Analytics

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

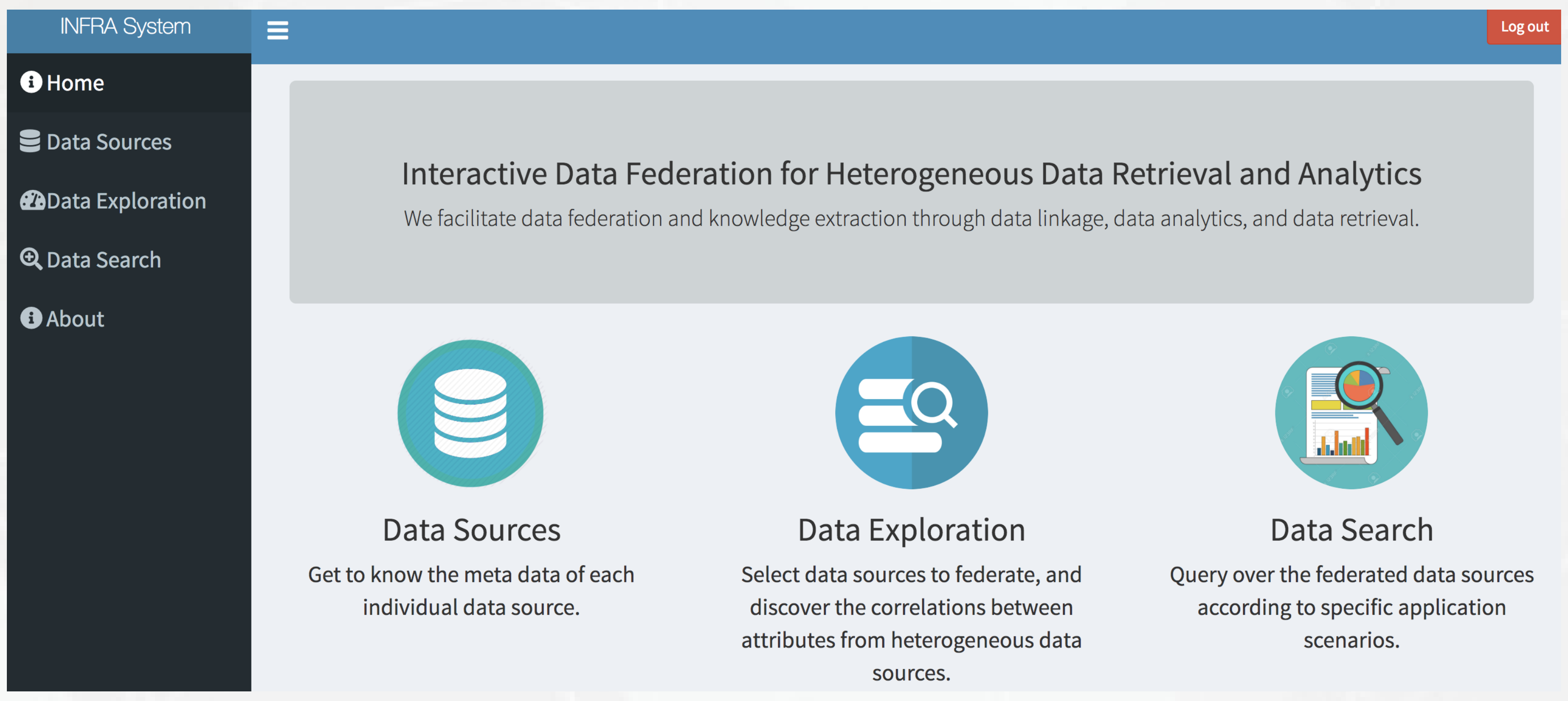
UMEÅ UNIVERSITY

DDM Database & Data Mining Research Group

INTRODUCTION

INFRA - an interactive data federation system 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.

- Demo video: `https://vimeo.com/319473546`



Interactive Data Federation for Heterogeneous Data Retrieval and Analytics

We facilitate data federation and knowledge extraction through data linkage, data analytics, and data retrieval.

Data Sources

Get to know the meta data of each individual data source.

Data Exploration

Select data sources to federate, and discover the correlations between attributes from heterogeneous data sources.

Data Search

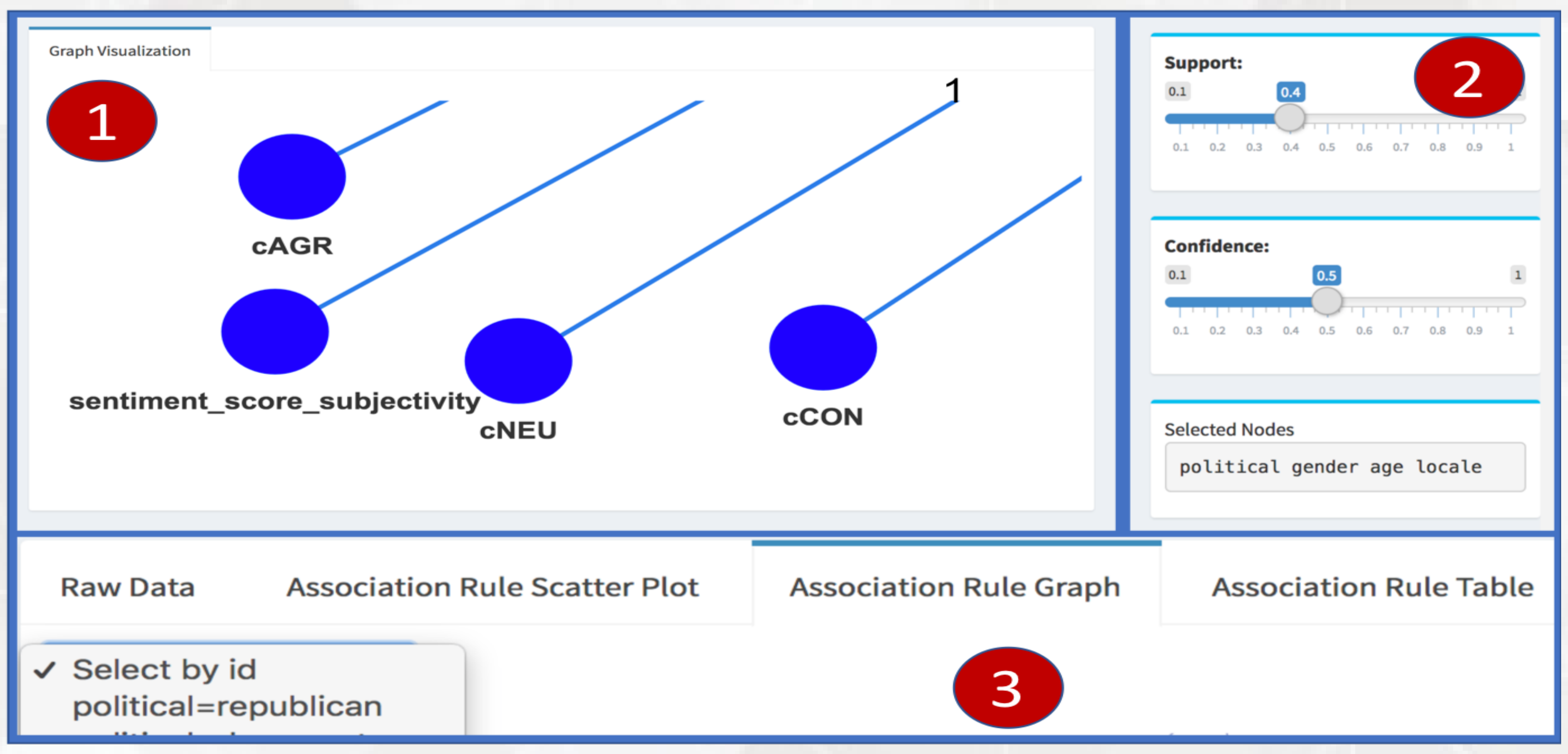
Query over the federated data sources according to specific application scenarios.

Figure 1 – User interface of the system with 3 main modules: Data Sources, Data Exploration, and Data Search.

DEMONSTRATION

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.

Case-study 1: association rules based sentiment analysis



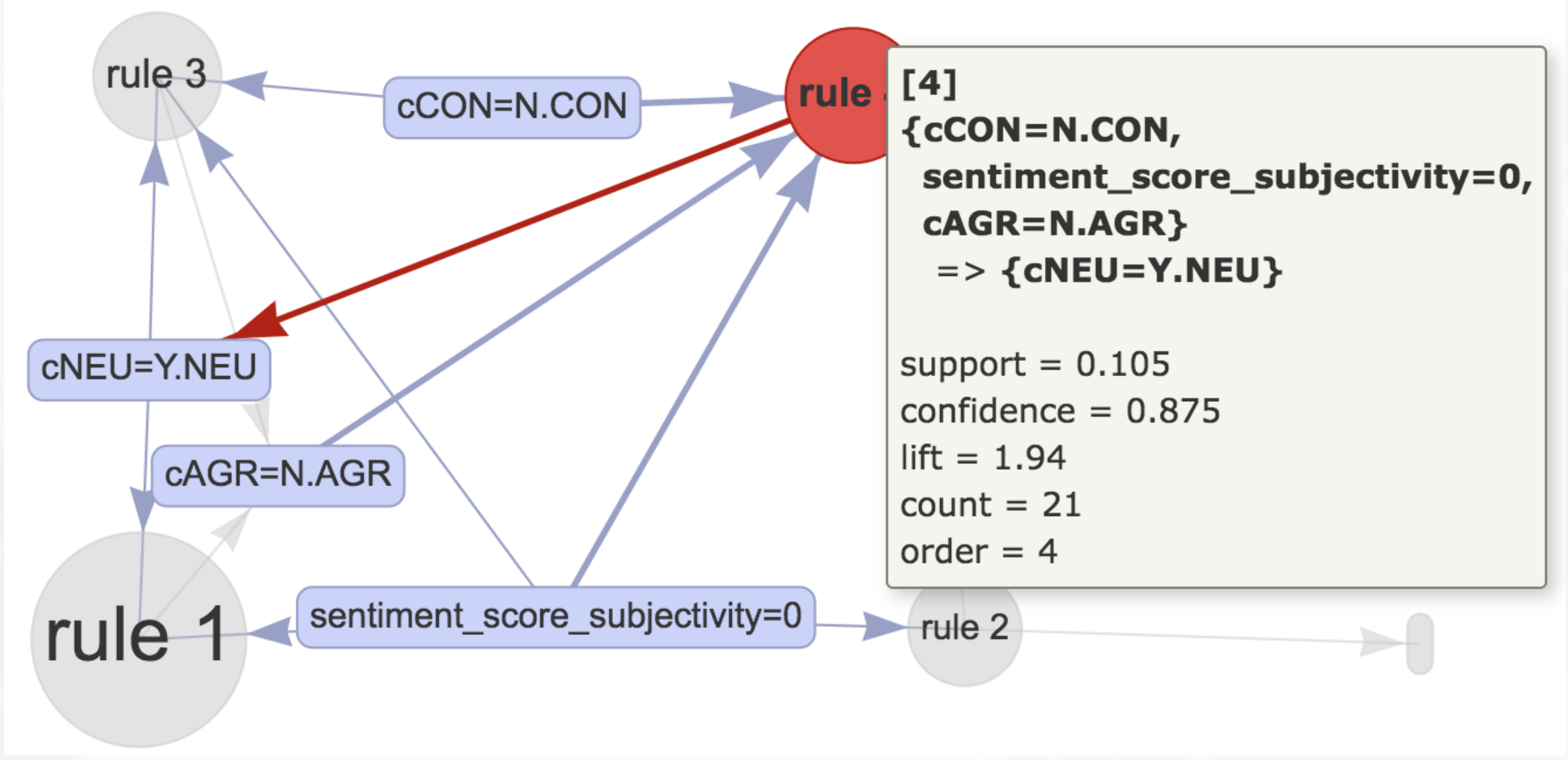
Selected Nodes: political gender age locale

✓ Select by id
political=republican

Figure 2 – Select 4 variables on Data Exploration to find hidden patterns between them.

Case-study 2: personality and sentiment analysis

- In Figure 3, Alice is suggested that Neurotic people might be the ONEs who are **NOT cCON + NOT cAGR + NO sentiment subjectivity** in their FB Posts. Data Search can be used to confirm this hidden pattern.

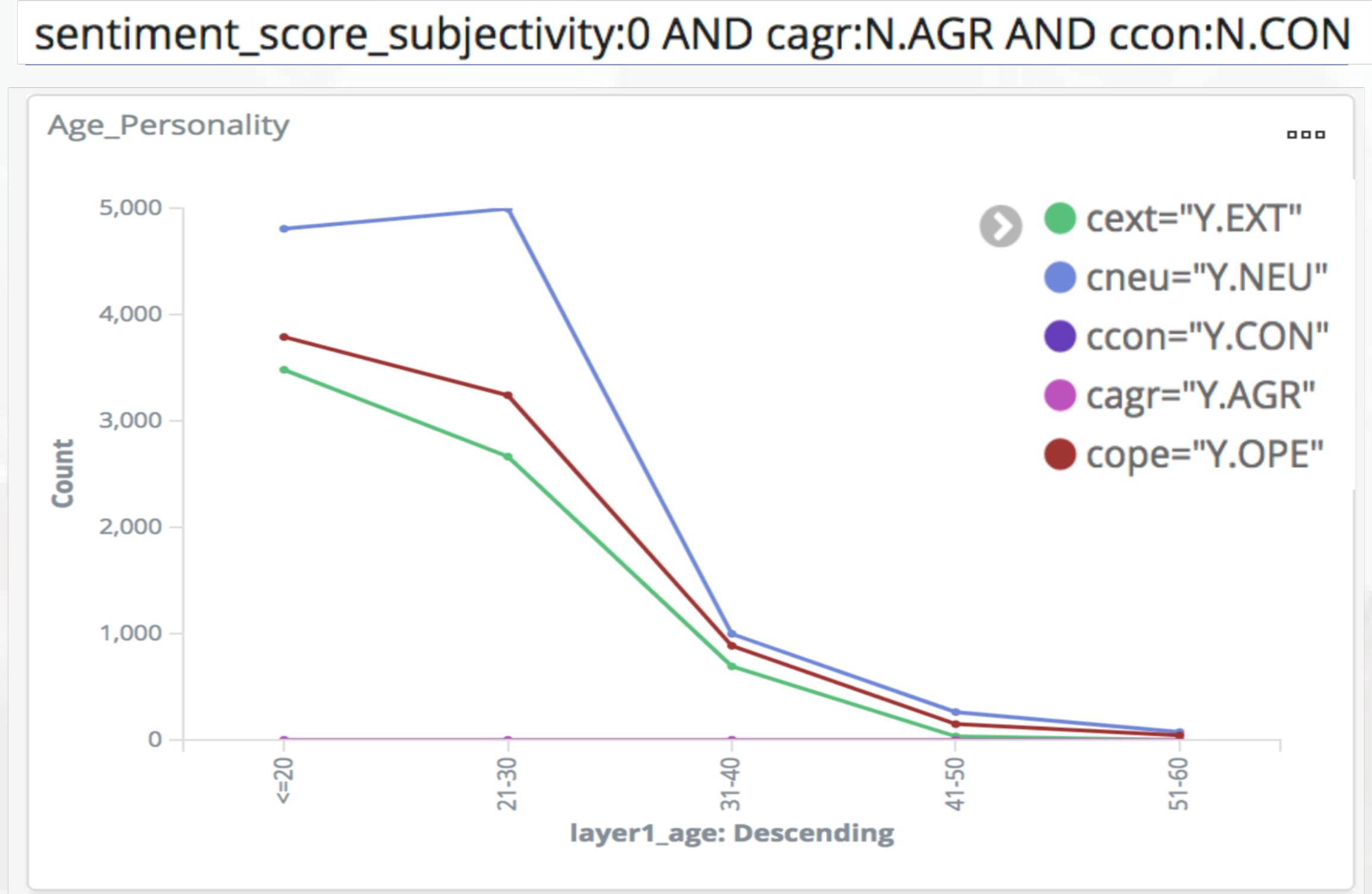


[4]
{cCON=N.CON,
 sentiment_score_subjectivity=0,
 cAGR=N.AGR}
 => {cNEU=Y.NEU}

support = 0.105
confidence = 0.875
lift = 1.94
count = 21
order = 4

Figure 3 – Case study of association rule based sentiment analysis on selected variables.

- In Figure 4, in Data Search, using 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 shows 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: e.g., mainly neurotic people have political views of "doesn't care" and "democratic".



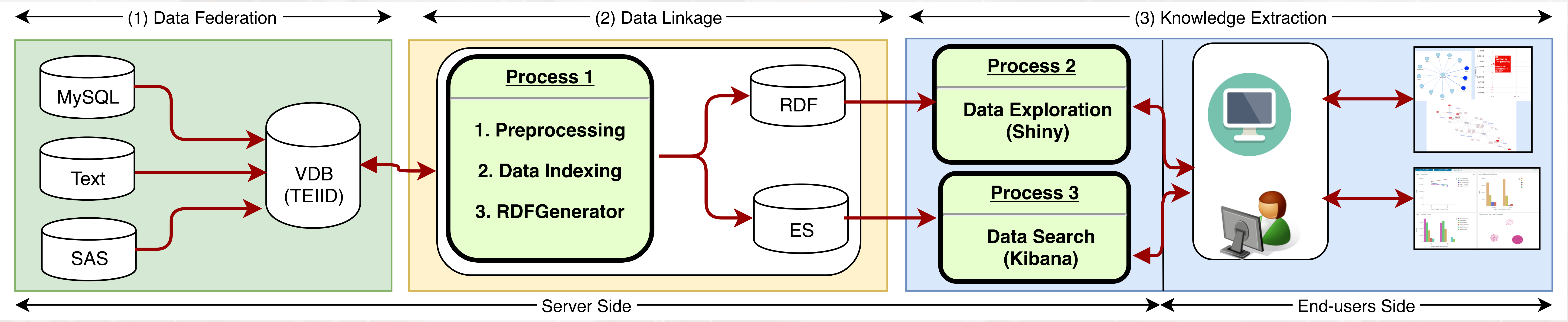
sentiment_score_subjectivity:0 AND cagr:N.AGR AND ccon:N.CON

Figure 4 – Case study of using Data Search to reconfirm the results from Data Exploration

Case-study 3: privacy-concern analysis

- This case-study shows how privacy-concern analysis is done through community detection and data analysis based on variables (i.e., age, gender, personality, FB status).
- It is found that 56.9% of users who mentioned the word "Obama" are females, in which 19.95% of them have negative sentiment (i.e., because of the high unemployment rate during his presidency).

SYSTEM TECHNICAL ARCHITECTURE



CONCLUSION

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).
- It allows users to easily construct and explore data over heterogeneous data sources.
- It is scalable by easily adding new heterogeneous data sources, and customizing data representation and analytics by users.

REFERENCES

- Xuan-Son Vu, Lili Jiang, and Anders Brändström, and Erik Elmroth: *Personality-Based Knowledge Extraction for Privacy-preserving Data Analysis*, In: Proceedings of K-CAP 2017, Austin, Texas, United States.
- Xuan-Son Vu, Lili Jiang: *Self-adaptive Privacy Concern Detection for User-generated Content*, Proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing, Hanoi, Vietnam, **best student paper award**.

This research is funded by Umeå University in Sweden on feder- ated database research. The system was deployed on resources provided by SNIC Science Cloud at HPC2N center. The authors also thank Michal Kosinski, David Stillwell and the myPersonality project for data sharing.