## **Relational Search and its Application to Investigative Intelligence Scenarios**<sup>\*</sup>

**Discussion** Paper

Stéphane Campinas, Matteo Catena and Renaud Delbru

Siren (https://siren.io), Galway, Ireland

Data-driven investigation intelligence systems are critical to the Intelligence and Defense domains to detect and prevent existing, evolving, and emerging criminal activities such as cyber-threats, human trafficking, money laundering, frauds, and crime rings. Several technical and usability challenges must be tackled to implement an effective investigative intelligence system. For instance, the ever-increasing volume and complexity of information form a fertile ground for malicious actors to blend in and operate, and investigations often involve connecting the dots on both large quantities of structured (well-defined records), semi-structured (logs), and unstructured data (textual and other media). In a Law Enforcement scenario, this could mean connecting tables of vehicles, cases, and traffic camera license plate readings, while it is common to use concepts such as IPs, MD5 hash values, or user IDs to tie together security logs in a Cybersecurity scenario. Therefore, investigative systems must provide intuitive and scalable ways to search, explore, and analyze large relational datasets. With information that is often interconnected in essence, such systems must also enable a mixed workload of search, data and graph analytics to support users in examining records and their relationships from different perspectives. Finally, analysts often interact with the system by following an explorative and iterative process that represents their train of thoughts. Consequently, investigative systems must have fast response times to avoid impeding the mental process of the analysts.

With roots in Information Retrieval research applied to the faceted exploration of Linked Data [1, 2], Siren provides an investigative intelligence platform called Investigate [3, 4] which is based on Elasticsearch and on the Siren's Elasticsearch plugin called Federate [3, 5]. The platform gives analysts several data interaction paradigms, such as search, analytic dashboard, set-to-set navigation, and graph visualization, and combines them into a unified and coherent interaction model. For example, the set-to-set navigation helps analysts to express their information needs visually, by connecting together different areas of interest. Applying filters on a set has an impact on all (in)directly connected sets, which allows an analyst to view only the relevant information for a given need. For instance, in a Financial Investigation scenario, we could have data on companies that have received investments from investors but are also mentioned by articles and have headquarters in cities. Siren Investigate can move from a set of companies to the set of records connected to it – for example, the investments received by Irish companies

IIR2023: 13th Italian Information Retrieval Workshop, 8th - 9th June 2023, Pisa, Italy

<sup>&</sup>lt;sup>\*</sup>This work was presented at the Industry Day of the  $45^{\text{th}}$  European Conference on Information Retrieval (ECIR'23).  $\bigcirc$  name.surname@siren.io

<sup>© 023</sup> Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

that are mentioned in articles regarding Artificial Intelligence. The graph visualization instead helps an analyst to have an overview of the computed sets, to see if any pattern or clusters emerge. In our example, graph visualization can help to picture the answers to questions like *"Which investors invested in which companies? Are they investing in pairs or groups? Are there groups that appear to be investing in competing companies?"*.

Efficient relation search capabilities are essential for implementing both set-to-set and graph visualization paradigms. Regrettably, document-oriented databases like Elasticsearch present some limitations in joining data. For instance, joining must be planned in advance at document indexing-time, and the documents to be joined must reside in the same index and on the same shard for scalability reasons [6]. This is adequate for hierarchical relationships but does not easily support more complex relationship models (e.g., networks) without incurring data duplication. The Federate plugin addresses these constraints by incorporating a query-time distributed semi-join operation between different inverted indices, and an effective caching mechanism for join results [7]. Consequently, Federate significantly enhances the flexibility and performance of the platform, facilitating complex tasks that are crucial for investigative intelligence such as extracting the shortest paths between entities in a graph.

In this talk, we will explore how the Siren Federate plugin integrates relational search capabilities into Elasticsearch and how it empowers the Siren Investigate platform to tackle some of the challenges in Investigative Intelligence. Specifically, we will discuss:

- Enabling analysts to devise custom data models that describe relationships between records,
- Supporting set-to-set navigation for interactive exploration and defining information needs,
- Utilizing graph visualization to uncover patterns within the data.

## References

- E. Oren, R. Delbru, S. Decker, Extending faceted navigation for rdf data, in: Proceedings of the 5th International Conference on The Semantic Web, ISWC'06, Springer-Verlag, Berlin, Heidelberg, 2006, p. 559–572.
- [2] R. Delbru, S. Campinas, G. Tummarello, Searching web data: An entity retrieval and high-performance indexing model, Web Semant. 10 (2012) 33–58.
- [3] Siren Platform Version 13 Product Overview, 2022. URL: https://info.siren.io/content/ siren-13-product-overview-whitepaper, last accessed: May 30, 2023.
- [4] Introduction to Siren Platform, 2020. URL: https://docs.siren.io/siren-platform-user-guide/ 13.2/index.html, last accessed: May 3, 2023.
- [5] Introduction to Siren Federate, 2020. URL: https://docs.siren.io/siren-federate-user-guide/ 31/siren-federate/introduction.html, last accessed: May 3, 2023.
- [6] Elasticsearch Guide: Joining queries, 2023. URL: https://www.elastic.co/guide/en/elasticsearch/reference/current/joining-queries.html, last accessed: May 3, 2023.
- [7] Joining 15 Billion Records in Seconds: the New Siren Benchmark, 2022. URL: https://info. siren.io/content/siren-benchmark-whitepaper, last accessed: May 30, 2023.