Distributed Big Data & ML architecture for ConnectionLens

Internship supervisors :

Angelos Anadiotis, Assistant Professor at Ecole Polytechnique, Institut Polytechnique de Paris (https://www.linkedin.com/in/angelos-christos-anadiotis)

Ioana Manolescu, Senior researcher at Inria and Professor at Ecole Polytechnique, Institut Polytechnique de Paris (https://pages.saclay.inria.fr/ioana.manolescu/)

Project summary

Context

ConnectionLens [1] is a system which finds connections between user-specified search terms across heterogeneous data sources. ConnectionLens treats a set of heterogeneous, independently authored data sources as a single virtual graph, where nodes represent fine-grained data items (relational tuples, attributes, key-value pairs, RDF, JSON or XML nodes…) and edges correspond either to structural connections (e.g., a tuple *is in* a database, an attribute *is in* a tuple, a JSON node *has a parent*) or to similarity (*sameAs*) links. To further enrich the content, ConnectionLens also applies entity extraction which enables the detection of people, organizations, etc. mentioned in the text, whether full-text or text snippets typically found in RDF or XML documents.

ConnectionLens stores the graph in a relational database (PostgreSQL) and searches the graph for connections by using the Grow-and-Aggressive-Merge (GAM) algorithm [1][2][3].

Objectives

Graph construction involves several steps. First, different data sources are collected by crawling online sources. Second, the text downloaded is parsed based on the format of the data source. Third, named entities are extracted from all text nodes, regardless their data source, using trained language models. The named entities are shared across the whole graph. Fourth, entities with the same label, but different meaning are separated by leveraging a disambiguation module. Finally, nodes, whose labels have similarity over a given threshold, are connected with a *sameAs* edge.

Currently, the pipeline is centralized, that is, it is deployed on a single, scale-up server. The goal of the thesis is to scale the end-to-end graph construction pipeline, to leverage a distributed computational infrastructure (i.e., cluster). Achieving this goal requires two main tasks:

- Design and implement a distributed version of the graph construction algorithm, including database and Machine Learning (Information Extraction) computations
- Store the graph in a distributed store, such as Impala, Cassandra, HBase, while also considering pure graph stores such as MongoDB and Neo4j.

References

[1] Angelos Christos Anadiotis, Oana Balalau, Catarina Conceição, Helena Galhardas, Mhd Yamen Haddad, Ioana Manolescu, Tayeb Merabti, Jingmao You: Graph integration of structured, semistructured and unstructured data for data journalism. Information Systems, 2021. https://doi.org/10.1016/j.is.2021.101846

[2] Angelos Christos Anadiotis, Oana Balalau, Theo Bouganim, Francesco Chimienti, Helena Galhardas, Mhd Yamen Haddad, Stephane Horel, Ioana Manolescu, Youssr Youssef: Empowering Investigative Journalism with Graph-based Heterogeneous Data Management. IEEE Data Engineering Bulletin, to appear. ArXiv: https://arxiv.org/abs/2102.04141

[3] Angelos Christos Anadiotis, Oana Balalau, Théo Bouganim, Francesco Chimienti, Helena Galhardas, Mhd Yamen Haddad, Stéphane Horel, Ioana Manolescu, Youssr Youssef Discovering Conflicts of Interest across Heterogeneous Data Sources with ConnectionLens *ACM International Conference on Information and Knowledge Management (CIKM 2021)*, Nov 2021, Online, Australia. ⟨10.1145/3459637.3481982⟩, demo video at : https://youtu.be/5B0KRow0dv8

Skills

The internship requires skills in as many as possible of the following areas, plus ability to learn (under our guidance, and based on a selected reading list) things as needed:

- Relational database management systems
- Graph data management
- Distributed systems
- Machine learning techniques for natural language processing

The project can be solved by one or two students

The project is proposed as an M2 internship, and it can be tackled by *one or two students*. In the presence of two students, each student will focus on a different aspect of the system, namely extraction and storage.

Practical details

We hope to work with the intern(s) in person, but we can also adapt to remote collaboration, should that become necessary because of sanitary circumstances.

Our lab is situated at 1, rue Estienne Honoré d'Orves, 91120 Palaiseau (https://team.inria.fr/cedar/contact/)