Fostering Serendipity through Big Linked Data

Saleem, M., Kamdar, M.R., Iqbal, A.
Sampath, S., Deus, H.F., Ngonga Ngomo, A.C.
Semantic Web Challenge at ISWC2013

Introduction

• Significant growth of the amount of Linked Open Data (LOD) [1]

• Large LOD datasets in bio-medical domain such as Linked TCGA [5]

• Scalable data management solution is key for dealing with Big Linked Data

• PubMed publications are being published on daily basis on a variety of biomedical literature

• We present a scalable approach for the continuous integration, querying of Linked TCGA and PubMed publications

• We provide a graphical interface to support the serendipitous discovery of bio-medical hypotheses using Big Linked Data

Linked TCGA

• Cancer Genome Atlas (TCGA) [2] is a pilot project aiming to accelerate the understanding of the molecular basis of cancer through the application of genome analysis technologies

• TCGA contains:
- 9000 patients cancer data
- 30 different cancer types
- 147,645 raw data files (12.7 TB)
- This is only 46% of the expected data with new data being submitted every day [5]

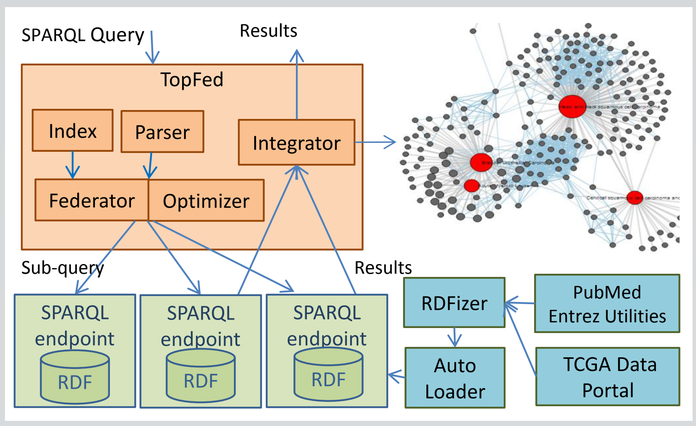
• Linked TCGA is the RDFized version of the Cancer Genome Atlas containing 7.36 Billion triples, 10 cancer types [5]

Use Cases

• Enabling Evidence-based genomic medicine
- By making use of cross-resource linking, we enable the discovery on whether a drug could be applied to more than one tumour typology, by linking the two typologies through their genomic signature

• Generation of new hypotheses
- Some types of cancer are more common than others and therefore the intricacies of their genomic signatures and genetic events are more well known
- The resource that we make available will enable researchers in the less common tumor typologies to discover association between their cancer of interest, and those that are more well studied
- Another possible arena for hypothesis generation is that of tumour cell migration

General Architecture

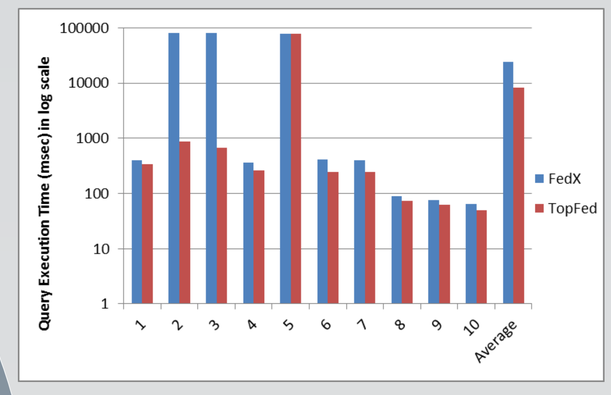


TopFed

TopFed [3] is a TCGA tailored federated SPARQL query processing engine designed for efficient integration of data from multiple TCGA SPARQL endpoints

Results

TopFed [3] vs FedX [6] based on 10 Linked TCGA queries

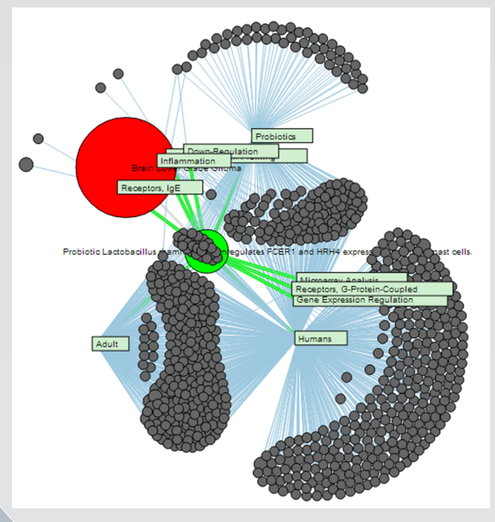


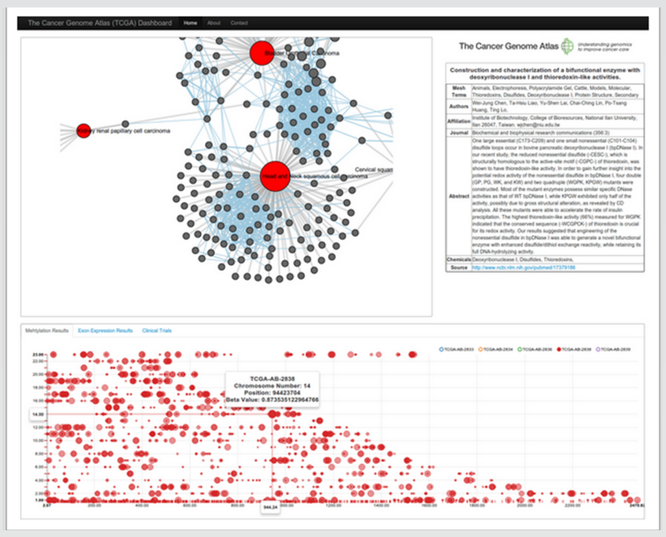
• TopFed outperform FedX signicantly on 90% of the queries

• On average, the query run time of TopFed is about one third to that of FedX

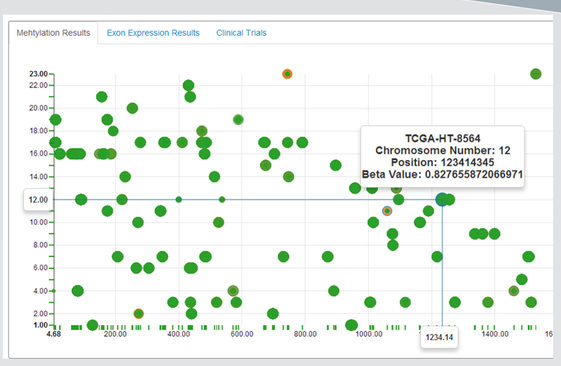
• TopFed best run-time (query 2, query 3) is more than 75 times smaller than that of FedX

Visualizations





http://srvgal78.deri.ie/tcga-pubmed/



Big Data Track Requirements

• Data Volume: We combine 7.36 billion triples from Linked TCGA with extraction results from more than 23 million publications from PubMed

• Data Variety: The Linked TCGA data was extracted from raw text files of different structures. We processed the metadata associated with PubMed publications and transform them into RDF. Unstructured data (publication abstracts) is processed to extract mentions of gene names and cancers

• Data Velocity: TCGA data doubles in size every 2 months. Moreover, the rate of new paper publication is in the order of 10k/month [4]

Future Work

• Integrate extension of Linked TCGA (over 30 billion triples [5]) with other digital libraries

• Explore new ways to further improve our data visualization platform

• Integrate other NLP approaches

References

[1] http://stats.lod2.eu/
[2] http://cancergenome.nih.gov/
[3] https://code.google.com/p/topfed/
[4] http://www.nihms.nih.gov/stats/
[5] Saleem, M., Shanmukha, S., Ngonga Ngomo, A.C., Almeida, J.S., Decker, S., Deus, H.F.: Linked cancer genome atlas database. In: I-Semantics 2013 (2013)
[6] Schwarte, A., Haase, P., Hose, K., Schenkel, R., Schmidt, M.: Fedx: Optimization techniques for federated query processing on linked data. In: ISWC 2011

Additional Desireable Features

• Usability: Our visualizations allows:
- To gain an overview of TCGA related publications on topics of interest
- To formulate hypotheses w.r.t. the interaction between cancers types, genes, drugs, etc.

• Value: Allow bio-medical experts to explore billions of triples and get a concise overview of the relations between bio-medical resources, publications and MESH terms

• Functionality: Integrated view and exploration of Linked TCGA, related publications and metadata with ensured data freshness through automatic updates. Scalable management of distributed data through TopFed

This work is supported by BIG project: http://aksw.org/Projects/BIG.html