Overview

The ability to find and access relevant information is crucial. Traditional approaches to storing, querying, and inferencing fall short when faced with web-scale data. We present a system that combines the computational power of large clusters with an efficient data structure for storing and querying. We use this system to reduce the BTC2009 dataset to information about people with both full names and email addresses, our domain of interest.

Inferencing

- Added BTC-related ontologies and simple upper ontology about people to BTC2009 dataset.
- Used approach from [1] to materialize RDFS inferences.
- Inferencing took 349 seconds (≈5.8 minutes) using 3,712 processors on CCNI’s Opteron blade cluster.
- Resulting dataset of 1,620,437,279 triples containing 41,848,813 people.

Reduction via Query

- Used approach in [2] to extract information about people with full names and email addresses from materialized dataset using parallel CONSTRUCT query, reducing the dataset to our domain of interest.
- Querying took 940 seconds (≈15.7 minutes) using 8,192 processors on CCNI’s Blue Gene/L.
- Resulting dataset of 784,783 triples containing 260,127 people with both full names and email addresses.

Analysis with BitMat

- Converted reduced dataset to a BitMat in 25 seconds.
- BitMat is a compressed RDF structure from [3].
- Dataset compressed to 33 MB from 113 MB.
- Ran queries on BitMat in subsecond time.

Conclusion

- Clusters and supercomputers can be used to quickly reduce datasets.
- BitMat allows for efficient storage and querying of RDF.
- The entire reduction took ≈22 minutes.

References