Data Mapping Framework in a Digital Library with Computational Epidemiology Datasets

S.M.Shamimul Hasan [1,2], Sandeep Gupta [2], Edward A. Fox [1], Keith Bisset [2], Madhav V. Marathe [1,2]

[1] Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA
[2] Network Dynamics and Simulation Science Laboratory, VBI, Virginia Tech, Blacksburg, VA 24061, USA

shasan2@vt.edu, sandeep@vbi.vt.edu, fox@vt.edu, {kbisset, mmarathe} @vbi.vt.edu

ABSTRACT

Computational epidemiology employs computer models and informatics tools to reason about the spatio-temporal spread of diseases. The diversity of models, data sources, data representations, and modalities that are collected, used, and modified motivate the development of a digital library (DL) framework to support computational epidemiology. The heterogeneous content includes metadata, text, tables, spreadsheets, experimental descriptions, and large result files. There is no accepted framework that allows unified access to such content. We propose a framework for a digital library system tailored to such datasets to support computational network epidemiology.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: Systems issues.

Keywords

Digital library, Epidemiology, Simulation

1. INTRODUCTION

Computer simulations play an important role in understanding the spatio-temporal properties of diseases, especially due to the fact that, unlike in the physical sciences, real world experiments are usually not possible. Studies are conducted, in general, through the use of a simulation and require information on the population structure, agent behavior, disease transmission, and a model of the disease. Table 1 shows the type, size, and format for some of the datasets present in networked epidemiology [1]. Data access and digital library services in current setups are cumbersome due to heterogeneity and fragmentation across datasets. We propose a data mapping framework for digital library systems for computational epidemiology datasets. The proposed framework provides a unified view to access data (Figure 1).

978-1-4799-5569-5/14/$31.00

Table 1: Epidemiology Dataset

Category	Data	Size	Representation
Synthetic Population	Household, Person Activity	566 GB	Relational
Social Network and Output	Contact Network, Simulation Output	1.84 TB	File
Experiment	Experiment	240 GB	Relational

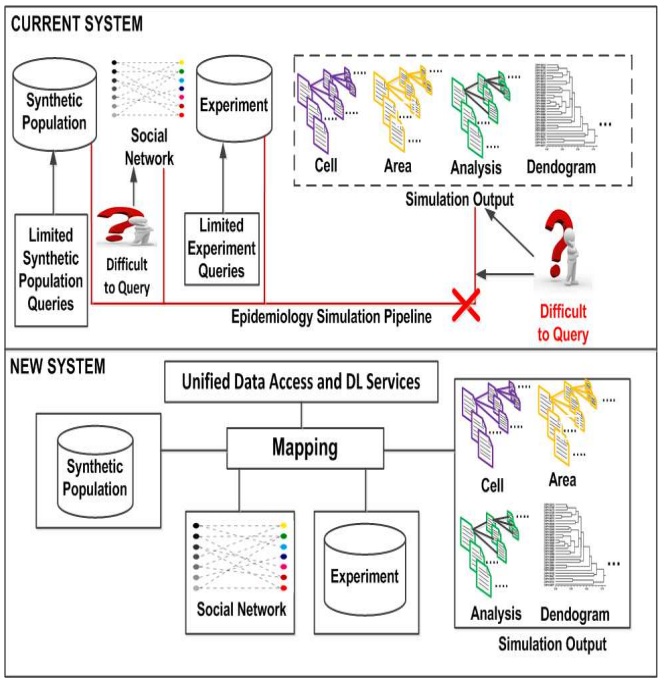

Figure 1: Unified view

2. FRAMEWORK

Data mapping provides us the flexibility to switch between various databases and execute queries on them. We investigate two different query execution paradigms: bottom-up and top-down. The bottom-up approach converts the relational datasets to an RDF graph. The top-down approach translates the queries written in SPARQL to queries over the

physical (native) representation of the datasets, written in SQL. Figure 2 illustrates both of the approaches.

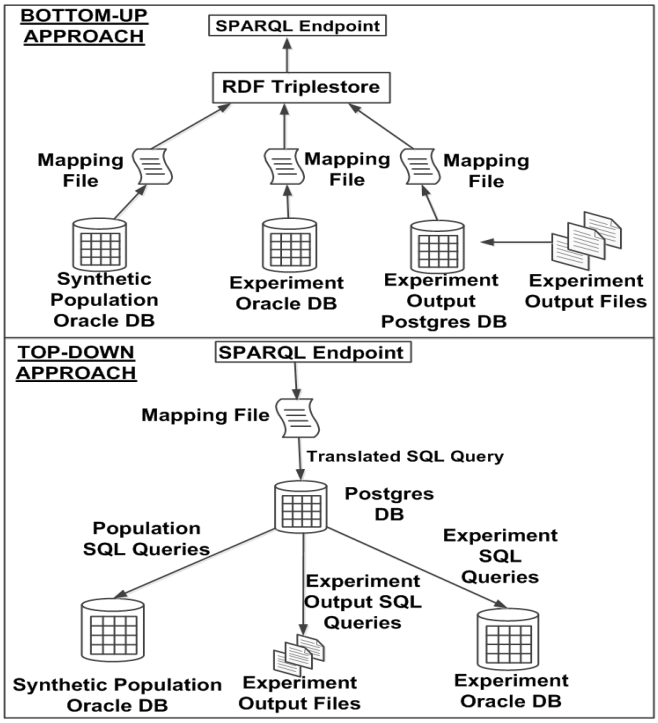

Figure 2: Framework

3. EXPERIMENTAL STUDY

Datasets: To evaluate our framework, we considered a real-time epidemiology simulation study conducted in the Seattle area. The study assumed that influenza transmits in various regional populations through person-person contact. Simulation is conducted by the EpiFast engine [2] that can perform large scale realistic epidemic simulations on distributed memory systems. We collect 2.33 GB of synthetic population data (relational), 4.31 MB of experiment data (relational), and 44 MB of simulation output data (file). We use the D2RQ Mapping Language to convert relational and file data to RDF graphs [3], Virtuoso Open-Source Edition 6.1.6 as RDF data engine [4], and the SPARQL query language.

Table 2: RDF Graph Information

Databases	RDF Graph Size (GB)	Number of Triples	RDF Graph Generation Time (Minutes)
Seattle Synthetic Population	177	661,848,662	317
Output	3.10	12,979,996	6
Experiment	0.01	66,654	0.37

Experiment Result: The D2RQ tool produces the mapping files for each data source. The tool then, using the mapping files, queries the data sources to generate the RDF graphs. Table 2 shows that for large synthetic populations relational data creates a large number of RDF triples, taking considerable time to generate such an RDF graph. To measure the strength of the framework we execute various queries collected from epidemiology scientists. All of our queries are written in the SPARQL language. We implement the bottom-up approach through the Virtuoso tool. It enables us to execute SPARQL queries over the complete epidemiology workflow datasets. The top-down approach is implemented with the D2RQ query facility. That allows us to execute SPARQL queries against a relational database by using D2RQ mapping files. We find that in the bottom-up approach we have large storage cost and it is problematic to execute queries on large graphs. On the other hand in the top-down approach we don't have any RDF graph storage cost and it works properly even for large graphs. However, execution for top-down is not as efficient as bottom up. Table 3 illustrates a comparison of the two approaches.

Table 3: Query Runtime

Queries	Bottom-up Approach (SPARQL Query Runtime in Seconds)	Top-down Approach (SPARQL Query Runtime in Seconds)
How many people of a particular demographic are sick?	0.04	7.18
Find who infected whom of a particular demographic	0.38	9.18
How many people get infected on a particular simulation day?	0.03	5.76

4. CONCLUSION

In this paper we describe a data mapping framework that is developed as a part of epidemiology digital library. We show that it facilitates large scale simulation data unification. Our experiment results show the strengths and weaknesses of the framework.

5. REFERENCES

[1] "Interface to synthetic information systems (ISIS), 2014." [Online]. Available: http://ndssl.vbi.vt.edu/apps/isis/

[2] K. R. Bisset, J. Chen, X. Feng, V. A. Kumar, and M. V. Marathe, "Epifast: A fast algorithm for large scale realistic epidemic simulations on distributed memory systems," in *Proceedings of the 23rd International Conference on Supercomputing*, ser. ICS '09. New York, NY, USA: ACM, 2009, pp. 430–439. [Online]. Available: http://doi.acm.org/10.1145/1542275.1542336

[3] C. Bizer and R. Cyganiak, "D2R Server - publishing relational databases on the semantic web," 2006. [Online]. Available: http://www4.wiwiss.fu-berlin.de/bizer/pub/Bizer-Cyganiak-D2R-Server-ISWC2006.pdf

[4] "Virtuoso open-source edition, 2014." [Online]. Available: http://www.openlinksw.com/dataspace/doc/dav/wiki/Main/