Benchmarking Heterogeneous Graph Data

Amarnath Gupta

It is now widely recognized that developing a benchmark for graph-structured data is an important and challenging problem. It is realized that not all graphs are created equal, and graphs produced from protein-protein interactions, social networks, phone call and SMS messages, DBPedia, citation networks etc. are all quite different from each other in terms of their structural properties as well as the operations performed on them. [Duan et al 2011] have recently pointed out that even within a specific domain like RDF graphs, there is a wide variability amongst benchmarks themselves and between benchmarks created with synthetic data vs. real data. One approach to address the problem is to select some generic characteristics of the graph that needs to be modeled. The Graph500 benchmark (www.graph500.org/) performs breadth-first searches in weighted, undirected large graphs generated by a scalable data generator based on a Kronecker graph that implements a scale-free graph generation algorithm. However, this benchmark does not consider directed edge-labeled graphs, and other graph properties like community effects and small diameters that are found in various real data graphs [Paradies 2012]. The IBM group's benchmark [Duan et al 2011] models the "structuredness" of a real data set by a measure called *coherence*, and then generates data sets of different sizes with the same kind of coherence.

Our experience in dealing with integration of life science data reveals that a suitable benchmark for this class of data cannot be modeled by any fixed set of characteristics over the entire graph. The benchmark should instead consider ***heterogeneous graphs***, i.e., multicomponent graphs such that each component may have different characteristics and an application workload will need to access multiple components of the graph. Based on the linked data we deal with within the domain of Neuroinformatics, we make the following observations on our linked data. These observations will expectedly inform the nature of benchmark data required to adequately model the real data characteristics and access patterns for large scale linked life science data in general.

- One component of our linked graph consists of RDFS/OWL style of ontology graphs from multiple ontologies. While each ontology has at least one backbone with a labeled DAG, some ontologies have multiple such DAGs, and when multiple intermapped ontologies are used, the mapping process makes the structure of the linked graph more complex.
- A second component of the linked graph consists of instance-level connectivity information, including connectivity networks within the brain, connectivity patterns amongst cells, and so forth. While this information can be (and often is) modeled in RDF, the subgraph of the

linked graphs that just represent the connectivity pattern have a different kind of structures and community pattern than other parts of the graph.

- A third component of the linked graph related to pathway graphs, i.e., graphs of molecular interactions under different conditions. These graphs generally cannot be modeled as power-law observing structures. However, the pathway graphs often have a nested hypergraph structure, i.e., it can have nesting of subgraphs but nodes from one subgraph can belong to multiple groups.
- Like many scientific domains, citations are an important component of the linked data graph. However, an interesting characteristic of this domain is that data records directly refer to citations, and consequently citation chains often serve as the bridge to traverse from one data set to another. A realistic data set for this domain should consider the graph characteristics of the reference network alongside the data network and the citation network.

As evident from the above a heterogeneous graph data benchmark should be multi-structural in nature, so that the component graphs do not all have the same connectivity structure, and individual components can be trees, DAGs, multigraphs, and so forth. Similarly edges may have weight, labels, a full type system on the labels, leading to a wide variety of component structures. The benchmark should allow a user to specify different sizes and other parameters (such as average node fan outs, range of clustering coefficients, average height of tree components, distribution of labels, etc.) for each component. We believe that creation of benchmarking data for large-scale graph problems in life science needs to consider this heterogeneity within a single benchmark data generator.

References

[Duan et al 2011] S. Duan, A. Kementsietsidis, K. Srinivas, O. Udrea: "Apples and oranges: a comparison of RDF benchmarks and real RDF datasets". *ACM SIGMOD Conference* 2011: 145-156.

[Paradies 2012] M. Paradies: "Challenges in the Design of a Graph Database Benchmark". FOSDEM'12 – Graph Processing DevRoom, 2012. Available at: http://www.slideshare.net/graphdevroom/challenges-in-the-design-of-a-graph-database-benchmark

Author

Amarnath Gupta is a research scientist at the San Diego Supercomputer Center of UC San Diego. His Advanced Query Processing Lab specializes in semantic search engines, scientific data integration, ontology management, graph data management, and implementation techniques of the above in various architectural platforms including the cloud.