Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

"One differentiator in task 16 is that we are leveraging the richness provided by semantic graph models in order to obtain deeper insights from scientific datasets. Edge and node types are exploited to guide search and quantify results and this is very suitable for the massively multithreaded architecture for which we design our algorithms."

- PNNL Task Lead
Sinan al-Saffar

U.S. DEPARTMENT OF ENERGY

Multi-Relational Knowledge Discovery

Developing multithreaded graph-based algorithms to query, analyze, and interpret semantic models of massive biomedical data

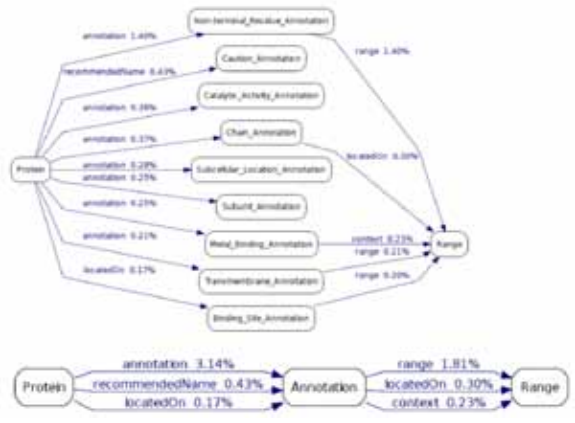
At a glance

At Pacific Northwest National Laboratory (PNNL), CASS-MT researchers are developing advanced high performance approaches and computational methods to discover knowledge from massive datasets. Discovery is achieved through semantic querying, link prediction, clustering, and other mining and relational learning algorithms. We seek improvements in both speed and depth of results provided to the domain experts. We closely collaborate with our partners in the biomedical community and at the Mayo Clinic.

What we do

Biological or medical data comes from various structured and unstructured sources. Our first step is to model the problem as a semantic graph. A semantic graph is simply a labeled and directed graph. Labels on edges and nodes in the graph represent the kinds of entities and kinds of relationships in the data. This is in contrast to simple untyped social network graphs where the presence of connections is modeled but typing is not. This is significant, as typing not only provides a richer model, but also acts as a gateway into logical inference in the presence of ontologies. Ontologies can be understood as mathematical structures to expand the semantic graph model based on the present typing. In summary, we are using richer models, in conjunction with methods that take this richness into account, to give better answers. This results in new computational demands, which we address by building graph-based algorithms to run on our Cray-XMT supercomputer, which is designed to perform well on graph-like algorithms. Semantic graphs are a good natural fit for both our problems and execution platform.

How we do it



Scientific datasets are abundant but their structure and richness varies greatly. Plain text from web crawls and scientific publications, spreadsheets, or databases are all sources of data. Many times the data producers convert these data into semantic graphs in the hope that the graph constitutes a mathematical model supporting quantification and prediction. Ideally, when there is a rich structure in these models, the graph can be quantified and exploited to reason about the originating problem. However many times this is not the case, such as semantic graphs dominated by a single link with a bookkeeping label, e.g. "seeAlso," instead of a knowledge-rich domain label. These degenerate to social networks and need not have labels in the first place (if all the links mean the same thing, why distinguish them?). Thus the first step in mining a semantic graph model is to discover if it is a rich model that sufficiently represents a problem. One approach to ensure quality of the model is to generate it in collaboration between semantic technology experts and domain experts. This is the approach we found

most useful in this effort. For example, given a biomedical data warehouse, we generate the equivalent semantic graph by using the field names as graph edge labels, and using the primary/foreign key relations to provide the connectedness structure in the graph nodes identifier. For example, given a biomedical data warehouse, we generate the equivalent semantic graph by mapping the field names from the database tables to edge labels in the graph, while the primary/foreign key relations provides further graph structure through connecting the corresponding graph nodes as identified by their equivalent URIs. This graph structure allows us to investigate whether execution of mining and querying on such graphs on graph-friendly architectures such as the Cray-XMT is more efficient on average compared to relational queries on the original database. Beyond efficiency, the semantic graph counterpart of the original database facilitates knowledge fusion by enabling navigation into other linked data entities on the semantic web. In addition to investigating ad-hoc semantic querying and mining on the XMT supercomputer, we are investigating the efficiency of such graph friendly platforms to aid in the next-generation genome sequencing computations. The idea is to align a target DNA sequence against a reference genome sequence. The reference sequence is the human genome already sequenced and the target sequence represents your personalized genome to be discovered. There exists shared sequences of various lengths between the two and unique sequences representing individual variations. We are modeling this problem as a graph by having nodes represent either a repeat or a unique sequence and edges the connections between those nodes in order to form the final sequence. By deciding what defines the optimal partitioning between repeat and unique nodes, a graph based algorithm can discover the most likely sequence of the target genome.

Applications

- Cohort identification and classification
- Personalized genome sequencing
- Drug discovery and understanding of biochemical processes

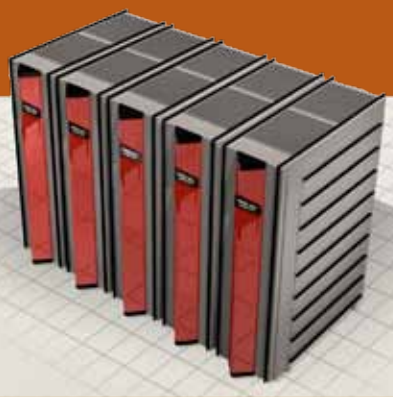
CASS-MT is dedicated to research on systems software, programming environments, and applications in a High-Performance Computing (HPC) multithreaded architecture environment.

We offer the only Open Science Cray XMT system, a one-of-a-kind supercomputer consisting of 128 multithreaded processors, 1 TB RAM, and a 7.7 TB Lustre parallel filesystem.

The Cray XMT supercomputer has the potential to substantially accelerate data analysis and predictive analytics beyond the limitations of traditional computing. Multithreaded processors allow multiple, simultaneous processing, helping researchers find solutions to the world's most complex challenges faster. The XMT can process irregular, data-intensive applications that have random memory access patterns. Unlike many applications where data delivery is dependent on memory speed, the Cray XMT's multi-threaded architecture tolerates memory access latencies by switching context between multiple threads that work continuously, overlapping the memory latency and preventing the processor from being held up while it waits for data to arrive.

The multithreaded technology powering our Cray XMT is ideally suited to perform pattern matching, scenario development, behavioral prediction, anomaly identification, and graph analysis.

Try it for yourself. We seek to create collaborations and provide expertise for porting and optimizing applications. The opportunity to use our Cray XMT system is available to internal and external research partners.



John Feo,
CASS-MT Director
(509) 375-3768
John.feo@pnnl.gov
cass-mt.pnnl.gov/

Sinan al-Saffar
Task Lead
Pacific Northwest National Laboratory
206-528-3356
sinan@pnnl.gov

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*