Collaboration Spotting Visual Analytics Tool Optimized For Very Large Graphs

Richard Forster

Faculty of Informatics, Eotvos Lorand University
CERN

forceuse@inf.elte.hu

As big data is getting collected in every sciences and other applications too, visual analytics is starting to gain traction to help the users understand their data. Such a tool is TIM (Technology Innovation Monitor), which is a graph visualization application for publications and patents that is being developed in a collaboration between CERN and JRC. The more generic CERN specific version of TIM is the Collaboration Spotting Visual Analytics tool. Based on the users query multiple landscapes, maps can be generated that will be used to visualize the given data. This generation is a multistep process that involves community detection and visual graph organization. As the graphs starts to become more complex and contain more nodes and edges this part of the computation becomes very costly and thus the bottleneck of the system. Community detection has become an important operation in numerous graph based applications. It is used to reveal groups that exist within real world networks without imposing prior size or cardinality constraints on the set of communities. Despite its potential, the support for parallel computers is rather limited. This is largely because the algorithm is irregular and the underlying heuristics imply a sequential nature. The Louvain method is a multi-phase, iterative heuristic for modularity optimization. It was originally developed by Blondel et al. [**?**], the method has become increasingly popular owing to its ability to detect high modularity community partitions in a fast and memory-efficient manner. To parallelize this solution multiple heuristics are used, that were first introduced in [**?**]. For graph organization the ForceAtlas algorithm is used that is part of the Gephi toolkit [**?**]. This method is responsible to assign coordinates in a 2D space for every node in such a way that they will not overlap on each other. The test data is multiple organization landscape taken from the project's dataset that involves patents and publications to visualize the connections in technologies among its collaborators. The mentioned methods are implemented in C++ and are used on a test machine running on a single processor with 8 threads, where we can achieve speedups up to 5.

References

[1] V. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. Theory Exp. (2008) P10008

[2] Hao Lu, Mahantesh Halappanavar, Ananth Kalyanaraman, Parallel heuristics for scalable community detection, Parallel Computing 47 (2015) 1937

[3] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, Mathieu Bastian, ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software, http://dx.doi.org/10.1371/journal.pone.0098679 (2014)