Parallel post processing of Correlation Clustering

Géraud Le Falher

August 7, 2015

There are (at least) two methods described in the literature

1. LocalSearch (introduced by Gionis *et al.* (2007, page 13), who also describe an efficient implementation), also called Best One Element Move, which "consist of removing one vertex from a cluster and either moving it to another cluster or to a new singleton cluster" (Elsner *et al.* 2009, page 3).

2. One can also merge clusters until no further gain can be achieved as Mathieu *et al.* (2010) do to solve Correlation Clustering in an online setting.

Say we have n nodes and m clusters. For the first method, we can build a $n \times (m+1)$ matrix A where the $A_{i,j}$ $j \leq m$ is the gain of putting node i in cluster j and $A_{i,m+1}$ the gain of creating a singleton with i. Likewise in the second case, we can build $B \in \mathbb{R}^{m \times m}$ such that $B_{i,j}$ is the gain of merging clusters i and j.

Each row of those matrices can be computed independently and therefore in parallel[1]. Then the main thread would the k best moves. If $k > 1$, this is only an approximation but we expect that there is not much side effect as long as k is not to big.

The bottleneck of this procedure is that computing A or B is somewhat expensive, albeit straightforward. A possible improvement would be to take advantage of the fact that the clustering didn't change that much to only recompute the relevant parts of these matrices[2].

References

[1] M. Elsner and W. Schudy, "Bounding and comparing methods for correlation clustering beyond ilp", pp. 19–27, 2009 (cit. on p. 1).

[2] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation", *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, 4–es, 2007 (cit. on p. 1).

[1] Maybe using the Java 8 `stream` https://docs.oracle.com/javase/tutorial/collections/streams/parallelism.html

[2] Although I'm not it's as easy as it sounds.

[3] C. Mathieu, O. Sankur, and W. Schudy, "Online correlation clustering", in *27th International Symposium on Theoretical Aspects of Computer Science - STACS 2010*, Inria Nancy Grand Est & Loria, 2010, pp. 573–584 (cit. on p. 1).