A parallel p-median clustering algorithm

Igor Vasilyev and Anton Ushakov

Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, 134 Lermontov str., 664033 Irkutsk, Russia,
{vil,aushakov}@icc.ru

In this report we address a cluster analysis problem in a sense of a famous discrete facility location problem has been in focus of many researchers for more than 50 years. Given a set $I = \{1, \ldots, m\}$ of potential sites for locating $p \leq m$ facilities, a set $J = \{1, \ldots, n\}$ of customers to be served from open facilities, and d_{ij} defining the distances (transportation costs) of serving customer $j \in J$ from the facility $i \in I$. The p-median problem consists in locating p facilities such that the overall sum of distances from each customer to its closest facility is minimized.

The problem has also become popular due to its broad applicability. Maybe one of the most important application of the p-median problem is clustering [**?**]. Though the p-median problem is often able to provide competitive and high quality solutions to the cluster analysis problem [**?**], they do not often apply it to large scale datasets due to the absence of effective methods of solving such large p-median problem instances. The most advanced state-of-the-art approaches are able to find exact or good suboptimal solutions to problems on graph with several tens of thousands nodes.

In this paper we develop an improved modification of the sequential approach proposed in [**?**] for the p-median problem and its parallel implementation. The algorithm is based on finding the sequences of lower and upper bounds for the optimal value by use of a Lagrangean relaxation method with a subgradient column generation and a core selection approach in combination with a simulated annealing. The parallel algorithm is implemented coupling the shared memory (OpenMP) with the message passing (MPI) technology. It allows us to deal with large instances on modern high performance computing clusters. The effectiveness and efficiency of parallel algorithm is tested and compared with the most effective modern methods on a set of test instances taken from the literature using the HPC-cluster "Academician V. M. Matrosov" [**?**].