Abstract

We provide simple and fast polynomial-time approximation schemes (PTASs) for several variants of the max-sum diversification problem which, in its most basic form, is as follows: given n points in a Euclidean space and an integer k, select k points such that the average distance among the chosen points is maximized. This problem is commonly applied in web search and information retrieval in order to select a diverse set of representative points from the input. In this context, it has recently received a lot of attention. We present new techniques to analyze natural local search algorithms. This leads to a $(1 - O(1/k))$-approximation for distances of negative type, even subject to a general matroid constraint of rank k, in time $O(nk^2 \log k)$, when assuming that distance evaluations and calls to the independence oracle are constant time. Negative-type distances include as special cases Euclidean and Manhattan distances, among other natural distances. Our result easily transforms into a PTAS. It improves on the only previously known PTAS for this setting, which relies on convex optimization techniques in an n-dimensional space and is impractical for large data sets. In contrast, our procedure has an (optimal) linear dependence on n. Using generalized exchange properties of matroid intersection, we show that a PTAS can be obtained for matroid intersection constraints as well. Moreover, our techniques, being based on local search, are conceptually simple and allow for various extensions. In particular, we get asymptotically optimal $O(1)$-approximations when combining the classic dispersion function with a monotone submodular objective, which is a very common class of functions to measure diversity and relevance. This result leverages recent advances on local-search techniques based on proxy functions to obtain optimal approximations for monotone submodular function maximization subject to a matroid constraint.