October 2008

MADALGO seminar by Srinivasa Rao, Aarhus University

On Secondary Indexing in One Dimension

Let X = x_1 x_2 ... x_n be a string over a finite, ordered alphabet S. A secondary index for X answers alphabet range queries of the form: Given a range [1,r] over S, return the set I_[1,r] = {i | x_i \in [l,r]}. Secondary indexes are heavily used in relational databases and scientific data analysis. It is well-known that the obvious solution of storing a dictionary for the position set associated with each character, does not always give optimal query time. In this paper we give the first theoretically optimal data structure for the secondary indexing problem. In the I/O model, the amount of data read when answering a query is within a constant factor of the minimum space needed to represent I_[l,r], assuming that the size of internal memory is S^{Omega(1)} blocks. The space usage is O(n log |S|) bits in the worst case, and we also show how to bound the size of the data structure in terms of the 0th order entropy of X. We show how to support updates with some time-space trade-offs. We also consider an approximate version of the basic secondary indexing problem, where a query reports a superset of I_[l,r] containing each element not in I_[l,r] with probability at most eps, where eps > 0 is the false positive probability.
For this problem the amount of data that needs to be read by the query algorithm is reduced to |I_[l,r]| log (1/eps) bits.

Joint work with Rasmus Pagh, IT University of Copenhagen.