November 2009

MADALGO seminar by Elad Verbin, ITCS, Tsinghua University

The Limits of Buffering: A Lower Bound for Membership Data Structures in the External Memory Model

Abstract:

In this talk I will describe a new technique for proving data structure lower bounds based on statistical reasoning. I will present it in the context of new (as in submitted-5-days-ago new) lower bounds for dynamic membership in the external memory model; the general technique, however, might (and should!) have further applications.

The external memory model is just like the cell probe model, except that it has a free-to-access cache of size m, and the cell size is typically w=polylog(n). The cell-probe model for data structures counts only cell-accesses, so computation is free. One of the most interesting features of the external memory model is that it allows to achieve sub-constant update time by writing multiple items to the cache, and then writing them to memory at the same time using only one probe (just like what is done in practice when paging). This is called *buffering*.

There is a data structure called the buffer tree, which achieves update time roughly O(log2(n)/w) and query time O(logn); it works for multiple problems, among them membership, predecessor search, rank select, 1-d range counting, etc. . For w=log9(n), for example, the update time here is subconstant.

We prove that if one wants to keep the update time less than 0.999 (or any const<1), it is impossible to reduce the query time to less than logarithmic (namely, to o(log_{wlogn}(n/m))). Thus one has a choice between two sides of a dichotomy: either buffer very well but take at least logarithmic query time, or use the old data structures from the RAM model who do not buffer at all, and have a shot at sublogarithmic query time.

To restate, we prove that for membership data structures, in order to get update time 0.999, the query time has to be at least logarithmic.
This is a *threshold phenomenon for data structures*, since when the update time is allowed to be 1+o(1), then a bit vector or hash table give query time O(1).

The proof of our lower bound is based on statistical reasoning, and in particular on a new combinatorial lemma called the Lemma Of Surprising Intersections (LOSI) The LOSI allows us to use a proof methodology where we first analyze the intersection structure of the positive queries by using encoding arguments, and then use statistical arguments to deduce properties of the intersection structure of *all* queries, even the negative ones. We have not previously seen this way of arguing about the negative queries, and we suspect it might have further applications.

Joint work with: Qin Zhang from HKUST