Scale Independence: Using Small Data to Answer Queries on Big Data

Floris Geerts

Department of Mathematics & Computer Science, University of Antwerp, Belgium
floris.geerts@uantwerpen.be

Abstract

Large datasets introduce challenges to the scalability of query answering. Given a query $Q$ and a dataset $D$, it is often prohibitively costly to compute the query answers $Q(D)$ when $D$ is big. To this end, one may want to use heuristics, “quick and dirty” algorithms which return approximate answers. However, in many applications it is a must to find exact query answers. So, how can we efficiently compute $Q(D)$ when $D$ is big or when we only have limited resources?

One idea is to find a small subset $D_Q$ of $D$ such that $Q(D_Q) = Q(D)$ where the size of $D_Q$ is independent of the size of the underlying dataset $D$. Intuitively, when such a $D_Q$ can be found for a query $Q$, the query is said to be scale independent [1, 2, 9]. Indeed, for answering such queries the size of the underlying database does not matter, i.e., query processing is independent of the scale of the database.

In this talk, I will survey various formalisms that enable large classes of queries to be scale independent. These formalisms primarily rely on the availability of access constraints, a combination of indexes and cardinality constraints, on the data [8, 9]. We will take a closer look at how, in the presence of such constraints, queries can often be compiled into efficient query plans that access a bounded amount data [6, 8], and how these techniques relate to query processing in the presence of access patterns [3, 4, 7]. Finally, we illustrate that scale independent queries are quite common in practice and that they indeed can be efficiently answered on big datasets when access constraints are present [5, 6].

1998 ACM Subject Classification
H.2.4 [Database Management] Systems – Query Processing,
H.2.3 [Database Management] Languages – Query Languages, H.2.2 [Database Management] Physical Design – Access methods

Keywords and phrases Scale independence, Access constraints, Query processing

Digital Object Identifier 10.4230/LIPIcs.ICDT.2016.2

Category Invited Talk

References


