DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
UNIVERSITY OF SOUTHERN DENMARK, ODENSE

COMPUTER SCIENCE COLLOQUIUM

The Story of DeCP: a Web-Scale CBIR system

Gylfi Þór Guðmundsson

Monday, 07 March, 2016 at 10:15

U51

Abstract:

The scale of multimedia collections has grown fast over the decade. The largest is Facebook, storing over a 100 billion images and adding hundreds of millions more each day. In order to cope with such a large image collections and growth, methods for content-based image retrieval must adapt gracefully. The work to be presented is aimed at addressing this challenge. For the better part of a decade we have been working with a high-dimensional indexing technique called Distributed extended Cluster-Pruning (DeCP) that is based on a hierarchical vectorial quantizer and an approximate k-NN search algorithm. Two key observations drove the design and evolution of DeCP. The first observation is that the collections are so huge that using secondary storage is unavoidable. Taking disk related issues into account is therefore central to our work and design of DeCP. Second observation is that all CPUs are now multi-core and clusters of machines are a commonplace. Parallelism and distribution are both key factors that must be harnessed for any viable web-scale CBIR system. The design of DeCP includes the constraints associated with using secondary storage, parallelism and distribution. At its core is an hierarchical non-iterative unstructured vectorial quantization scheme. Our main contribution is to have pushed the design of DeCP from a single core C++ system, that could handle a few tens of thousands of images, to a distributed web-scale CBIR system capable of handling hundreds of millions of images, or tens to hundreds of billions of high-dimensional vectors. DeCP has been adapted to run both on Hadoop and Spark and the development has resulted 6 international publications so far, with another 2 publications in the pipes.

Host: Yongluan Zhou