Faster WAH Compression Querying through the Use of Metadata

Miguel Velez and Jason Sawin
Computer & Information Sciences, University of St. Thomas

Both business and research applications have become data-intensive, which require massive storage and fast interaction. Several database indexing techniques have been developed to both compress the data and allow efficient access and processing. Bitmap index is a popular database indexing technique that leverages fast machine bitwise operations when querying large data sets. A popular method for compressing bitmaps is the Word-Aligned Hybrid (WAH) compression technique that encodes the number of runs, the consecutive number of words in a bitmap that have homogeneous bit values, in a single machine word while not compressing literals, words with heterogeneous bit values. In order to improve the querying speed of WAH compressed datasets, we developed two techniques that used two different types of metadata about the datasets. *Verbose* encoded, for every column in the bitmap, both the length of the run and the number of literals following each run. *Succinct* encoded just the number of literals after each run. Our querying algorithms use the metadata to limit unnecessary memory reads. The results of our empirical study showed an improvement in the general case of ~8% using *verbose* and ~11% using s*uccinct* over an entire data set compared to regular WAH querying. When limited to columns that contained 30% or more runs juxtaposed with 26% or more literals, there was an improvement of ~12% using *verbose* and ~14% using *succinct*.