Distrib Parallel Databases (2014) 32:313–314
DOI 10.1007/s10619-014-7145-y

EDITORIAL

# Scalable data summarization on big data

**Feifei Li · Suman Nath**

Published online: 15 February 2014


Across different scientific domains, engineering disciplines, and application scenarios, increasingly, users have to deal with large-scale, diverse, feature-rich, and high-resolution data sets that allow for data-intensive decision-making. The so-called big data challenge is making a profound transformation in computing. Big data not only refers to data sets that are large in size, but also covers data sets that are complex in structures, high dimensional, distributed, and heterogeneous. An effective framework when working with big data is through data summaries, such as different sampling methods, histograms, sketches and synopses, low-rank subspace approximation, dimensionality reduction techniques, etc. Instead of operating on complex and large raw data directly, these tools enable the execution of various data analytics tasks through appropriate and carefully constructed summaries, which improve their efficiency and scalability. Though some of these topics have been well studied in the past, the big data phenomena opens doors for interesting new research. These challenges include, but are not limited to, how to quantify the accuracy and efficiency trade-off when summarizing big data in massively parallel and distributed environments, how to summarize features in complex heterogeneous data, how to address IO and system issues in a summarization process, how to reduce communication cost when building a summary for a large data set stored in a cluster of commodity machines (such as a key-value store), how to dynamically maintain a summary in an incremental fashion under arbitrary arrivals of new data. As a result, answering the big data challenge through scalable data summarization is becoming of paramount importance.

F. Li (✉)
School of Computing, University of Utah, Salt Lake City, UT, USA
e-mail: lifeifei@cs.utah.edu

S. Nath
Microsoft Research, Redmond, WA, USA

This special section of the Distributed and Parallel Databases (DAPD) features a strong collection of five papers, selected from 12 submissions, representing recent advances in summarizing big data. These works present novel techniques for IO-sampling, building various kinds of summaries in parallel environments, and summary-based parallel query processing and online aggregation.

The first paper presents a unified framework for building an $\ell_0$-sampler, which is is to sample near-uniformly from the support set of a dynamic multiset. This problem has a variety of applications within data analysis, computational geometry and graph algorithms. This paper has provided rigorous analyses for initiating and building $\ell_0$-samplers in the proposed framework, and empirically studied different $\ell_0$-sampling algorithms under the proposed framework.

The second paper presents the PF-OLA framework for efficient implementations of parallel online aggregations. Online aggregation provides estimates to the final result of a computation during the actual processing. The user can stop the computation as soon as the estimate is accurate enough, typically early in the execution. This allows for the interactive data exploration of the largest datasets. This paper introduces the PF-OLA framework for parallel online aggregation in which the estimation virtually does not incur any overhead on top of the actual execution.

The third paper studies how to conduct principal component analysis (PCA) in parallel. PCA is a well known technique for dimensionality reduction and feature extraction for large high dimensional data. This work extends a previous sequential method to a highly parallel algorithm that can compute PCA in one pass on a large data set based on summarization matrices. They also study how to integrate the algorithm with a DBMS; their solution is based on a combination of parallel data set summarization via user-defined aggregations and calling the MKL parallel variant of the LAPACK library to solve singular value decomposition in RAM.

The fourth paper investigates how to build entity-based summarization for web search results in efficient and scalable fashion using MapReduce. An useful technique to achieve advanced exploration to exploit the availability of structured (and semantic) data in Web search is to enrich it with entity mining over the full contents of the search results. This paper considers a general scenario of providing such services as meta-services (that is, layered over systems that support keywords search) without a-priori indexing of the underlying document collection(s). They show how to make such service feasible for large data using MapReduce.

The fifth paper presents efficient algorithms for executing set similarity join on large probabilistic data in MapReduce. Set similarity is a useful summary operator. However, set similarity join over two large data sets is expensive, especially in working with large probabilistic data. This paper examines various parallel methods in executing set similarity joins over such data in MapReduce and shows interesting results in improving the scalability and performance over the baseline method.

We would like to thank all of the authors who submitted papers to this special section for their high-quality contributions. We also thank the referees for their generous help and valuable suggestions. We are grateful to Professor Divyakant Agrawal and Amit P. Sheth, the Editor-in-Chiefs of DAPD, for their strong support for this special section.