Distributed and Parallel Databases (2019) 37:233–234
https://doi.org/10.1007/s10619-019-07269-8

Guest editorial: special issue on data management and analytics for healthcare

Fusheng Wang[1] · **Gang Luo**[2]

Published online: 9 May 2019

Healthcare enterprises are producing large amounts of data through electronic medical records, medical imaging, health insurance claims, surveillance, among others. Such data have high potential to transform current healthcare to improve healthcare quality and prevent diseases, and advance biomedical research. Medical Informatics is an interdisciplinary field that studies and pursues the effective use of medical data, information, and knowledge for scientific inquiry, problem solving and decision making, driven by efforts to improve human health and well-being.

This special issue of the Journal of Distributed and Parallel Databases sought original and high-quality research articles cross-cutting information management and medical informatics to discuss innovative data management and analytics technologies highlighting end-to-end applications, systems, and methods to address problems in healthcare and biomedical research.

In addition to open calls, this special issue invited best papers published in the International Workshop on Data Management and Analytics for Medicine and Healthcare (DMAH 2015-2017), in conjunction with the International Conference on Very Large Data Bases.

The paper "scalable and flexible management of medical image big data" by Dejun Teng et al. provides two alternative solutions for managing and querying large scale medical images in DICOM, the default image data standard for medical images. The work tries to tackle two major challenges on medical image data management: the exploding data volumes and the lack of flexibility on metadata search. The first approach uses parallel and hybrid relational and XML data management, which achieves high flexibility using XML and high scalability through data partitioning.

✉ Fusheng Wang
fusheng.wang@stonybrook.edu

Gang Luo
luogang@uw.edu

[1] Department of Biomedical Informatics and Computer Science, Stony Brook University, Stony Brook, USA

[2] Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, USA

The second approach is built on a document store using JSON based data model for flexibility and sharding based data partitioning for scalability. While the first approach achieves the best query performance, the second approach has the best loading performance.

The paper "MaReIA: a cloud MapReduce based high performance whole slide image analysis framework" by Hoang Vo et al. focuses on a highly scalable and cost-effective approach for analyzing whole slide images, an emerging medical imaging data type that is revolutionizing pathology. Whole slide images are extremely large, and a single image may contain millions of objects to be extracted with image segmentation algorithms. The work takes an overlapping partitioning based approach combined with spatial indexing based method for parallelizing image segmentation with gracefully handling of boundary objects. The experiments also demonstrate that cloud computing is highly attractive with its scalability and cost-effectiveness.

Authors Pradeeban Kathiravelu et al. propose a hybrid ETL approach in their paper, "On-demand big data integration—A hybrid ETL approach for reproducible scientific research". The work targets scientific research, which requires access, analysis, and sharing of large scale data that are often heterogeneous and geographically distributed. Instead of using an eager ETL approach to build an integrated data repository, which has inefficient bootstrapping, the authors propose a hybrid lazy and eager approach, which supports incremental integration and loading of metadata and data from the data sources. By incorporating a human-in-the-loop approach for selective data integration, the system outperforms both the eager ETL and lazy ETL approaches.

The topic of in-database patient similarity analysis is studied by Robert Fitch et al. in their paper "Concept acquisition and improved in-database similarity analysis for medical data", with a goal to achieve efficient identification of cohorts of similar patients for personalized medicine. The work focuses on increasing the performance of calculating the pairwise similarity values through SQL, which is highly expressive and natural for users. The experiments with two database systems demonstrate that the column store based implementation was competitive with traditional data mining tools.

Collectively, these four papers illustrate the diverse range of challenges on biomedical data management, integration and analytics, and demonstrate how database oriented approaches can address these challenges.

We would like to thank everybody who contributed to the special issue, the authors for their valuable contributions, the Springer team for making it happen, and Editor-in-Chief Mohamed Mokbel for pushing the idea of this special issue.