The VLDB Journal (2005) 14(3): 279–280
DOI 10.1007/s00778-005-0165-5

GUEST EDITORIAL

Terry Gaasterland · H. V. Jagadish · Louiqa Raschid

Special issue on data management, analysis, and mining for the life sciences

Published online: 9 September 2005

During the last decade, biologists have experienced a fundamental revolution from traditional R&D involving the study of single genes and isolated cellular mechanisms to e-biology that addresses whole cell physiology and complex biological systems. This has been accompanied by an explosion in the number and size of public data resources, and a rapid growth in the variety and volume of laboratory data from microarrays measuring transcription levels simultaneously for hundreds of genes and to protein interaction screens measuring multiple components of protein complexes.

Life science data is complex and data sets have complex inter-relationships. The data is often incomplete, uncertain, and can vary significantly across biological replicates. Data can evolve more quickly than the technologies developed to interpret the data. Thus, life sciences data is not well suited for current data models and query paradigms. The challenge for database researchers is developing appropriate technology to manage life sciences data, to make it accessible in an efficient way to scientists, and to provide the appropriate data models and data analysis tools.

The last few years has seen emerging activity towards addressing this challenge in the database, data mining, and information retrieval communities. This special issue was organized to feature papers that reflect synergies between computational advances in data management and manipulation and the fields of molecular biology, cell physiology and systems biology.

The call for Papers resulted in over 20 submissions in October 2004. Each paper was reviewed by two or three experts. As a result of the first round of review, we accepted eight papers for minor or major revision. After a subsequent review, and lengthy discussions among the reviewers, we identified four papers for inclusion in the special issue, and two papers were recommended for further review by the journal. The review process was completed in June 2005. A brief summary of these four papers follows.

Data integration from web accessible bioinformatics data sources is critical to the success of life science research. There are two main approaches to such integration. In the "warehouse model" this integration is performed ahead of time, and made available in an integrated warehouse. In the "federation model" this integration is performed on the fly when needed, by accessing the appropriate original sources of data. This issue has one paper dealing with each of these models.

In "Sync Your Data: Update Propagation for Heterogeneous Protein Databases," the authors deal with an issue that all warehousing schemes have to deal with – how to update the cached data in the warehouse when changes are made to the base data over time. This is a version of the problem of incremental maintenance of materialized views, with the additional difficulty that the base data is in remote autonomous databases that may have limited facilities for reporting updates. This paper presents an algebra for view maintenance under these challenging conditions.

The paper "Composing, Optimizing, and Executing Bioinformatics Web Services," presents an approach to automatically generate composition plans for web services, to optimize the composition plans, and to execute these plans efficiently. The authors present parameterized integration plans that can be hosted as Web services as well as two optimization techniques to improve execution time. The first technique, tuple-level filtering, analyzes the source/service descriptions in order to automatically insert filtering conditions in the composition plans; this technique usually requires incorporating sensing operations in the plan. The second technique consists of mapping the integration plans into programs that can be executed by a dataflow-style, streaming execution engine. Using real-world Web services, they show that the automatic composition approach and parameterized

T. Gaasterland (✉)
University of California, San Diego and Rockefeller University

H.V. Jagadish
University of Michigan
E-mail: jag@eecs.umich.edu

Louiqa Raschid
University of Maryland
E-mail: louiqa@umiacs.umd.edu

plans are effective in data integration from large numbers of services. Second, they show that their optimization techniques can significantly reduce the response time of the generated plans.

Computational biology often requires the performance of a number of data access, analysis, and processing tasks. Furthermore, many of these have to be repeated as the desired results are obtained through iterative analysis. In the paper, “Rule-Based Workflow Management for Bioinformatics,” the authors present a model for managing such bioinformatics workflows. Through the use of a rule-based framework, rather than requiring a completely specified process graph, this model permits the management of very complex workflows, and the possibility of iterative restarts as needed.

Sequence data sets are ubiquitous in modern life sciences applications, and the suffix tree has been used to evaluate a wide variety of queries on sequence data sets, as required for biological data analysis. However, suffix tree construction does not scale well in time or space. In “Suffix Tree Construction: Practical Methods for Small and Large Datasets,” the authors explore suffix tree construction algorithms over a wide spectrum of data sources and sizes. First, they show that on modern processors, a cache-efficient algorithm with $O(n^2)$ complexity outperforms popular linear algorithms. For larger datasets, the disk I/O requirement quickly becomes the bottleneck in each algorithm’s performance. They present a buffer management strategy for the $O(n^2)$ algorithm; the resulting new algorithm, TDD, scales to sizes much larger than have been previously described in the literature. Finally, they present a new disk-based suffix tree construction algorithm that is based on a sort-merge paradigm, and show that for constructing very large suffix trees, this algorithm is more efficient than TDD.

In short, this special issue covers a broad range of data management related issues motivated by the needs of biological science. We thank our 45 reviewers for completing their review assignments under strict deadlines and for providing us with valuable input in putting this issue together.

Finally, this special issue would not have been possible without the able assistance of Nancy Sosa, Laboratory Administrative Manager of the Gaasterland Laboratory at Rockefeller University.