Distrib Parallel Databases (2018) 36:5–8
https://doi.org/10.1007/s10619-017-7217-x

Guest editorial: large-scale data curation and metadata management

Mohamed Eltabakh[1] · **Boris Glavic**[2]

Published online: 12 January 2018

Introduction

We are delighted to present this special issue of the distributed and parallel databases journal (DPDB) on large-scale data curation and metadata management. Data curation and annotation are becoming essential mechanisms for capturing a wide variety of metadata related to data. This metadata may carry different semantics ranging from tracking the data's lineage and provenance, quality information, exchanging knowledge and discussion messages among scientists, attaching related articles or documents, linking to relevant statistics about the data, and highlighting erroneous or conflicting values. This metadata may be represented in many different formats including free-text values, articles or binary files, images, structured information such as provenance, or semi-structured content such as email messages.

The creation and maintenance of annotated databases and metadata repositories require a great deal of effort (and cost) from many scientists and domain experts. Yet, the gain from the maintained annotations is still very limited, because of the lack of comprehensive solutions that automate large-scale metadata management tasks such as storage of annotations, extracting meaningful information from large sets of annotations, and their propagation through operations. Thus, the virtue of the hidden knowledge in this metadata is still uncharted. The growing volume, profound complex-

✉ Mohamed Eltabakh
meltabakh@cs.wpi.edu

Boris Glavic
bglavic@iit.edu

[1] Computer Science Department, Worcester Polytechnic Institute (WPI), Worcester, MA 01609, USA

[2] Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616, USA

ity, increasing heterogeneity, and hidden semantics of emerging metadata repositories create many unprecedented challenges. Unfortunately, existing techniques are not well prepared to deal with the aforementioned characteristics, and they fall short in providing advanced processing, semantics extraction, and deep analytics over the wealth of such metadata.

The aim of this special issue is to give an overview of current research in distributed data curation and metadata management that addresses the challenges outlined above. The special issue calls for innovative solutions and insightful contributions to address the aforementioned challenges. The scope of the issue covered a wide range of subtopics including:

- Metadata summarization, mining, and deep analytics
- Distributed data curation and metadata management
- Big data curation and annotation in emerging infrastructures, e.g., MapReduce, Spark, and NoSQL engines
- Metadata validation and verification
- New annotation management and metadata models
- Proactive and automated data curation and annotation
- Metadata cleaning and quality control strategies
- Metadata-driven query processing and optimizations
- Storage and indexing techniques for big metadata
- Crowd-Sourcing models for data curation and annotation
- Scalable metadata and provenance visualization
- Interoperability of metadata formats and engines

Following an open call for papers and a rigorous peer-review process, seven papers were selected to be part of the special issue covering a wide selection of approaches from crowd-based curation of metadata, distributed metadata repositories, handling updates in collaborative databases, and web-scale provenance reconstruction.

The accepted papers in the issue are the following:

- Crowd enabled curation and querying of large and noisy text mined protein interaction data
- Web-scale provenance reconstruction of implicit information diffusion on social media
- AUDIT: approving and tracking updates with dependencies in collaborative databases
- MetaStore: An adaptive metadata management framework for heterogeneous metadata models
- A statistically-based ontology matching tool
- COACT: a query interface language for collaborative databases
- P-PIF: a ProvONE provenance interoperability framework for analyzing heterogeneous workflow specifications and provenance traces

Summary of contributions

The accepted papers introduced the following contributions:

Crowd enabled curation and querying of large and noisy text mined protein interaction data

The authors proposed a model for crowd-enabled curation of text minded biological databases on demand using a declarative language. The paper introduces a trust model for managing and querying the tentative and uncertain annotations. The proposed system *CrowdCure* and its query language *CureQL* provide a foundation for possible extensions to domains other than biological databases.

Web-scale provenance reconstruction of implicit information diffusion on social media

The authors present an efficient method for reconstructing the provenance of information diffusion in social media. This work addresses a challenging problem, namely, how to reconstruct provenance when not all interactions of users with the data are available to the provenance management system. While this is not the first approach for reconstructing provenance based on limited knowledge about interactions, it stands out by providing an efficient solution that scales to large datasets.

AUDIT: approving and tracking updates with dependencies in collaborative databases

This paper presents a system for tracking and managing updates to curated databases based on the Update Pending Approval (UPA) model presented by the authors. In this approach, all updates have to approved or rejected by privileged users. The system is implemented on-top of HBase. The solution presented by the authors is unique in that it combines management of non-linear version histories over large datasets with ideas from curated databases.

COACT: a query interface language for collaborative databases

This paper studies querying the version histories of data that is updated based on the UPA model mentioned above.

MetaStore: an adaptive metadata management framework for heterogeneous metadata models

The authors present the MetaStore system, which is a metadata management framework for scientific data repositories. The framework provides a foundation for supporting a wide range of heterogeneous metadata types, e.g., administrative, descriptive, structural, and provenance-related metadata. MetaStore is based on NoSQL database and RDF models and it supports both static and dynamic metadata.

A statistically-based ontology matching tool

The authors propose a mechanism for pair-wise ontology matching. Ontologies have become a popular means of knowledge sharing and reuse within a single domain or across different domains. The proposed work explores the use of predictive statistical model to establish proper alignment between two ontologies. The proposed matching mechanism produces one-to-many cardinality mapping by leveraging partitioning and parallelism into the matching process.

P-PIF: a ProvONE provenance interoperability framework for analyzing heterogeneous workflow specifications and provenance traces
The authors present the Prov2ONE algorithm for translating BPEL workflows into ProvONE, a standard for representing provenance of workflows (both retrospective and prospective). The presented P-PIF system implementing this algorithm improves provenance interoperability by capturing provenance over multiple workflow management systems and translating it into the ProvONE standard.

Conclusion

We would to thank the authors for their valuable contribution to this special issue, as well as the reviewers for their thoughtful reviews and discussions. The articles of this special issue provide a representative selection of research in this field. We hope that the reader will find them inspiring and that the work presented here will stimulate future research in Data Curation and Metadata Management at scale.