Biomedical Big Data for Clinical Research and Patient Care: Role of Semantic Computing

Satya S. Sahoo
Division of Medical Informatics
School of Medicine, Case Western Reserve University
Cleveland, OH, USA
satya.sahoo@case.edu

Abstract—**Healthcare datasets are increasingly characterized by large volume, high rate of generation and need for real time analysis (velocity), and variety. These datasets are often termed biomedical big data and include multi-modal electrophysiological signals and electronic health records. In this talk, we focus on the computational challenges associated with signal data management and the role of semantic computing in addressing these challenges. We describe a cloud computing platform called Cloudwave that has been developed to effectively manage electrophysiological big data for epilepsy clinical research and patient care.**

Keywords: Domain ontology, Semantic Web, Biomedical Big Data, Hadoop

I. Introduction

The increasing digital representation of biomedical information together with greater capabilities in diagnostics and imaging instruments has led to an exponential increase in the availability of healthcare data. Many of these datasets are called “Biomedical Big Data” and are characterized by large volume, high velocity with respect to rate of generation and time period available for interpretation, and disparate variety. Although these three V’s introduce a number of computational challenges, big data represent an important opportunity to derive significant insights to advance biomedical research and enhance patient care [1]. For example, electronic healthcare records (EHR) and multi-modal electrophysiological signal data are being analyzed to discover adverse drug events, long-term patient outcomes, and comparative effectiveness of treatments [1-3]. In this talk, we focus on the electrophysiological signal to answer three questions: (1) What are the computational challenges that need to be addressed to effectively leverage healthcare big data; (2) Why should semantic computing approaches be used to address many of these challenges; and (3) How does a biomedical big data platform called Cloudwave use open source Hadoop technology stack with domain ontology to leverage biomedical big data for healthcare research.

The Cloudwave platform is being implemented to support epilepsy clinical research. Epilepsy is the most common serious neurological disease, affecting 65 million persons worldwide with more than 200,00 new cases diagnosed every year. Electrophysiological signal data are used to record brain activity (electroencephalogram, EEG), cardiac measures (electrocardiogram, ECG), and related physiological measures (blood oxygen saturation). Signal data is generated at a high rate; a typical five day evaluation period generates 1.6GB of data and our medical collaborators at the University Hospitals of Cleveland have accumulated more than 11 Terabytes (TB) of data over past 3 years. Signal data is used as the gold standard for diagnosis, preoperative patient evaluation, medication, and related activities in patient care, hence we need to develop appropriate computing platforms to derive knowledge from these datasets.

II. Challenges in Electrophysiological Big Data Management

The computational challenges associated with electrophysiological signal data can be categorized along a number of dimensions, including:

1. A data representation format that incorporates domain semantics
2. Scalable storage for large volumes of data
3. Efficient data processing for analyzing high throughput data
4. Optimized query modules to support ad-hoc user queries
5. Real time interactive visualization with low latency network transfer

Existing signal data visualization systems and data representation formats such as the European Data Format (EDF) [4] were not designed to support the functionalities listed above. Hence, we developed the Cloudwave platform using the open source Hadoop technology stack [5] together with an epilepsy domain ontology [6] to allow healthcare users to leverage signal data for both clinical research and patient care [3]. The Cloudwave platform consists of (a) an Ontology-driven Web-based signal visualization and query module, (b) a data processing and storage module that extends the Hadoop Distributed File System (HDFS) [7], and (c) a dedicated middleware layer to efficiently retrieve and transfer data from storage to visualization module. An epilepsy domain ontology plays a central role in Cloudwave by using domain semantics to underpin multiple functionalities.

III. Role of Semantic Computing

We use the term semantic computing to describe knowledge resources, such as ontologies, and components of the Semantic Web technology stack, such as Web Ontology Language (OWL2) and reasoning [8]. Domain ontologies

978-1-4799-4003-5/14 $31.00 © 2014 IEEE
DOI 10.1109/ICSC.2014.68

use formal knowledge representation language such as OWL2 to accurately model complex domain semantics including terminological classification and domain-specific constraints. This enables ontologies to be used as a reference model for integrating heterogeneous data, to compose complex query expressions to query integrated data, and support reasoning to automatically infer implicit knowledge over large datasets [9]. In addition, ontologies such as Gene Ontology [10] have been used for data annotation to ensure consistent use of domain terminology for data sharing.

In this talk we explore new roles for domain ontologies in defining a new data representation format for electrophysiological data that facilitates efficient data retrieval and network transfer to Web-based visualization clients. We also explore the role of domain ontologies in partitioning large volumes of multi-modal data for storage on high performance distributed file systems, such as the Hadoop Distributed File Systems (HDFS).

IV. Cloudwave: Using Domain Ontologies for Managing Electrophysiological Big Data

Cloudwave is a big data platform for supporting real time access to multi-modal signal data used in multi-center collaborative epilepsy research. The Cloudwave platform supports real time access and querying of large volumes of signal data using the HDFS storage module.

A. Cloudwave Storage and Data Processing Workflow

The Cloudwave storage module extends HDFS with dedicated support to efficiently store signal data that can be retrieved based on user queries. The storage module implements a MapReduce-based data processing workflow to process EDF signal data generated from epilepsy centers and extracts data segments corresponding to specific signal channels. For example, Cloudwave extracts and stores EEG, ECG, and blood oxygen signal data in separate files in HDFS [11]. Cloudwave has developed a JavaScript Object Notation (JSON)-based expressive representation format called Cloudwave Signal Format (CSF) to store channel-specific data together with metadata information in a single file. The metadata information includes clinical events, which are mapped to classes modeled in the epilepsy domain ontology. This facilitates effective querying of signal data using the visual interface module.

B. Cloudwave Query and Visualization Module

Clinical researchers analyze electrophysiological signal data using a visualization interface that supports querying for specific clinical events, such as the start of an epileptic seizure. The visual interface also supports combining signal data from multiple channels according to “signal montages”. To support these functionalities, the Cloudwave query and visualization interface extends an open source visualization software called Highcharts. We describe the architecture of the visualization and query interface that allows users to search for clinical events using ontology classes and subsumption reasoning based on OWL2 semantics.

In addition, the visualization interface implements signal filtering features that remove noise and optimize signal features to support appropriate rendering of signal data for clinical researchers. The visualization interface is integrated with the HDFS-based storage module to support efficient retrieval of data segments for multi-channel signal visualization.

C. Cloudwave Middleware Layer

Effective data partitioning is essential for developing big data applications that allow the use of high performance “shared-nothing” distributed computing resources. The widely used EDF files for electrophysiological signals store data as consecutive signal segments corresponding to small time periods [4]. In Cloudwave, we have implemented two partitioning schemes that correspond to ontology classes modeled in the epilepsy domain ontology to support querying and retrieval of signal data segments for use in the visualization module. This ontology-based data partitioning approach using the concepts of signal montages and “epochs” is a promising approach for using domain semantics to effectively manage electrophysiological big data.

V. Conclusions

In this talk, we discuss the increasing importance of biomedical big data in both clinical research and patient care. Using the Cloudwave platform as an exemplar for electrophysiological signal data, we discuss the computational challenges that need to be addressed for effective management of big data. We describe the role of semantic computing, represented by domain ontologies and Semantic Web technologies, in the Cloudwave project for data representation, storage, query, and visualization.

Acknowledgment

This work was supported by NIH/NINDS grant number 1-P20-NS076965-01 and NIH/NCATS grant number UL1TR000439. We acknowledge Professor Samden Lhatoo, Professor Guo-Qiang Zhang, and the Cloudwave team members.

References

[1] P. E. Bourne, "What Big Data means to me," *Journal of American Medical Informatics Association,* vol. 21, p. 193, 2014.

[2] R. W. White, Tatonetti, N.P., Shah, N.H., Altman, R.B., Horvitz, E., "Web-scale pharmacovigilance: listening to signals from the crowd " *Journal of American Medical Informatics Association,* vol. 20, pp. 404-408, 2013.

[3] S. S. Sahoo, Jayapandian, C., Garg, G., Kaffashi, F., Chung, S., Bozorgi, A., Chen, C., Loparo, K., Lhatoo, S.D., Zhang, GQ, "Heartbeats in the Cloud: Distributed Analysis of Electrophysiological “Big Data” using Cloud Computing for Epilepsy Clinical Research," *Journal of American Medical Informatics Association (Special Issue on Big Data),* vol. 21, pp. 263-71, 2014.

[4] B. Kemp, Olivan, J., "European data format 'plus' (EDF+), an EDF alike standard format for the exchange of physiological data.," *Clinical Neurophysiology,* vol. 114, pp. 1755-61, 2003.

[5] *Apache Hadoop*. Available: http://hadoop.apache.org/, accessed on April 20, 2014.

[6] S. S. Sahoo, Lhatoo, S.D., Gupta, D.K., Cui, L., Zhao, M., Jayapandian, C., Bozorgi, A., Zhang, GQ., "Epilepsy and

seizure ontology: towards an epilepsy informatics infrastructure for clinical research and patient care.," *Journal of American Medical Association,* vol. 21, pp. 82-9, 2014.

[7] K. Shvachko, Kuang, H., Radia, S., Chansler, R., "The Hadoop Distributed File System," presented at the IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), NV, 2010.

[8] P. Hitzler, Krötzsch, M., Parsia, B., Patel-Schneider, P.F., Rudolph, S., "OWL 2 Web Ontology Language Primer," World Wide Web Consortium W3C2009.

[9] O. Bodenreider, "Quality assurance in biomedical terminologies and ontologies.," Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda2010.

[10] M. Ashburner, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.," *Nat Genet.,* vol. 25, pp. 25-9, May 2000.

[11] C. P. Jayapandian, Chen, C.H., Bozorgi, A., Lhatoo, S.D., Zhang, G.Q., Sahoo, S.S., "Cloudwave: Distributed Processing of “Big Data” from Electrophysiological Recordings for Epilepsy Clinical Research Using Hadoop.," in *American Medical Informatics Association (AMIA) Annual Symposium*, Washington DC, 2013, pp. 691-700.