Editorial Board

David Hutchison  
  Lancaster University, Lancaster, UK

Takeo Kanade  
  Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler  
  University of Surrey, Guildford, UK

Jon M. Kleinberg  
  Cornell University, Ithaca, NY, USA

Friedemann Mattern  
  ETH Zurich, Zurich, Switzerland

John C. Mitchell  
  Stanford University, Stanford, CA, USA

Momi Naor  
  Weizmann Institute of Science, Rehovot, Israel

C. Pandu Rangan  
  Indian Institute of Technology, Madras, India

Bernhard Steffen  
  TU Dortmund University, Dortmund, Germany

Demetri Terzopoulos  
  University of California, Los Angeles, CA, USA

Doug Tygar  
  University of California, Berkeley, CA, USA

Gerhard Weikum  
  Max Planck Institute for Informatics, Saarbrücken, Germany
More information about this series at http://www.springer.com/series/7409
Fusheng Wang · Lixia Yao
Gang Luo (Eds.)

Data Management and Analytics for Medicine and Healthcare

Second International Workshop, DMAH 2016
Held at VLDB 2016
New Delhi, India, September 9, 2016
Revised Selected Papers
Preface

In this volume we present the accepted contributions for the Second International Workshop on Data Management and Analytics for Medicine and Healthcare (DMAH 2016), held in New Delhi, India, in conjunction with the 42nd International Conference on Very Large Data Bases on September 5–9, 2016.

The goal of the DMAH workshop is to bring together people in the field 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 health care, public health, and everyday wellness, with clinical, physiological, imaging, behavioral, environmental, and -omic data as well as data from social media and the Web. It provides a unique opportunity for interaction between information management researchers and biomedical researchers in this interdisciplinary field.

For DMAH 2016, we received 11 papers. A rigorous, single-blind, peer-review selection procedure was adopted, resulting in seven accepted papers presented at the workshop. Each paper was reviewed by three members of the Program Committee, who were carefully selected for their knowledge and competence. As far as possible, papers were matched with the reviewer’s particular interests and special expertise. The result of this careful process can be seen here in the high quality of the contributions published in this volume.

We would like to express our sincere thanks especially to the internationally renowned speakers who gave keynote talks at the workshops plenary sessions: Prof. Yuval Shahar of Ben-Gurion University, Beer Sheva, Israel; Prof. Adi V. Gundlapalli of VA Salt Lake City Health Care System and University of Utah School of Medicine, Salt Lake City, Utah, USA; and Prof. Supten Sarbadhikari of National Health Portal, India. We would like to thank the members of the Program Committee for their attentiveness, perseverance, and willingness to provide high-quality reviews.

March 2017

Fusheng Wang
Lixia Yao
Gang Luo
Organization

DMAH 2016

Workshop Chairs
Fusheng Wang  Stony Brook University, USA
Lixia Yao  University of North Carolina at Charlotte, USA
Gang Luo  University of Utah, USA

Program Committee
Syed Sibt e Raza Abidi  Dalhousie University, Canada
Carlo Combi  University of Verona, Italy
Vassilis Cutsuridis  Foundation for Research and Technology – Hellas, Greece
Amarendra Das  Dartmouth College, USA
Anna Divoli  Pingar Research, New Zealand
Kerstin Denecke  Bern University of Applied Sciences, Switzerland
Dejing Dou  University of Oregon, USA
Peter Elkin  University of Buffalo, USA
David Greenhalgh  University of Strathclyde, UK
Jesús B. Alonso Hernández  University of Las Palmas de Gran Canaria, Spain
Guoqian Jiang  Mayo Clinic, USA
Ying Li  IBM T.J. Watson Research Center, USA
Jun Kong  Emory University, USA
Tahsin Kurc  Stony Brook University, USA
Ulf Leser  Humboldt-Universität zu Berlin, Germany
Yanhui Liang  Stony Brook University, USA
Fernando Martin-Sanchez  Weill Cornell Medicine, USA
Casey Lynnette Overby  Johns Hopkins School of Medicine, USA
Yuval Shahar  Ben-Gurion University, Israel
Xiang Li  Fudan University, China
Hua Xu  University of Texas Health Science Center, USA
Chao Yang  Amazon, USA
Lin Yang  University of Florida, USA
Zhe He  Florida State University, USA
Abstracts of Keynotes
High Yield Document Sets: An Information Extraction Strategy for Large Clinical Text Corpora

Adi V. Gundlapalli

VA Salt Lake City Health Care System, University of Utah School of Medicine, Salt Lake City, UT, USA
adi.gundlapalli@hsc.utah.edu

Abstract. Extracting relevant concepts from clinical text corpora is an important strategy for quality improvement, operations, and research. With increasing use of electronic medical records and documentation by medical providers, the clinical text corpora needed to be processed is large. While it may be computationally feasible to process all available text documents, it would be beneficial to develop a strategy to identify high yield document sets for relevant concepts. Using principles of information foraging theory, we describe a method to identify high yield documents to enable processing of limited portions of large clinical text corpora for domain-specific information extraction. We present three use cases from real world examples using US Department of Veterans Affairs (VA) electronic medical record data. The VA is one of the largest health care systems in the US that treats nearly 6 million unique individuals every year; the data generated includes hundreds of millions of text notes from 83 million outpatient and 600,000 inpatient visits annually. The methods described are generalizable to electronic medical records from other health care systems.
eHealth Initiatives in Digital India

Suptendra Nath Sarbadhikari

Centre for Health Informatics, National Health Portal NIHFW, New Delhi
110067, India
supten@gmail.com

Abstract. The National eHealth Authority (NeHA) is being set up for facilitating the smooth adoption of eHealth in India. Following the putting up of the Concept Note in the public domain, and, subsequently after revising the Concept Note, based on the comments received from all the stakeholders, a National Consultation on National eHealth Authority (NeHA) was held on 4th April, 2016. Based on the deliberations of the Workshop, now a draft Bill/Act is being prepared for further processing.

An Integrated Health Information Platform (IHIP) is also being proposed so that Standards-compliant EHR systems can talk to each other and health information can be exchanged meaningfully.

All the healthcare facilities in India are now being provided a unique and permanent National Identification Number (NIN). This could be a starting point for linking data from various healthcare facilities throughout India.

As per the Digital India initiatives, the fifth pillar of eKranti includes Technology for Healthcare or eHealth having online medical consultation, online medical records, online medicine supply, and pan-India exchange for patient information.

For successful and meaningful information of health information, appropriate Standards are necessary to maintain interoperability between diverse Electronic Health Record (EHR) systems. The Ministry of Health and Family Welfare had recommended the Guidelines for EHR Standards in August 2013. Subsequently, India has become a country member of IHTSDO, which develops and maintains the SNOMED-CT. Now the Guidelines for EHR Standards are being revised and comments have been received on the proposed new edition. This will be notified soon.

Keywords: eHealth · Digital India · National eHealth Authority · Standards for Health Information Exchange
The Representation, Application, and Automated Discovery of Clinical Temporal-Abstraction Knowledge

Yuval Shahar

The Josef Erteschik Chair in Information Systems Engineering, Head, Medical Informatics Research Center, Department of Information Systems Engineering, Ben Gurion University, Beer Sheva 84105, Israel
yshahar@bgu.ac.il
http://www.ise.bgu.ac.il/faculty/shahar/

Monitoring, interpretation, and analysis of large amounts of time-stamped clinical data are subtasks that are at the core of multiple tasks important for medical care and research. Examples include the management of chronic patients using clinical guidelines, the retrospective assessment of the quality of the application of such a guideline, and the learning of new knowledge from analyzing the data regarding repeating patterns of measured data, and of meaningful abstractions derivable from these data. Such new knowledge can support additional tasks such as clustering, classification, and prediction.

My talk describes several conceptual and computational architectures developed over the past 20 years, mostly by my research teams at Stanford and Ben Gurion universities, for knowledge-based performance of these tasks, and highlights the complex and interesting relationships amongst them. Examples of such architectures include the IDAN and Momentum goal-directed and data-driven temporal-abstraction architectures, the KNAVE-II and VISITORS interactive-exploration frameworks for single and multiple longitudinal records, the KarmaLego temporal data mining methodology, and the ViTA-Lab framework, which integrates goal-directed and data-driven analysis of large numbers of multivariate time-oriented clinical data.

My talk also points out the progression, from individual-patient monitoring, diagnosis, and therapy, to multiple-patient aggregate analysis and research, and finally to the learning of new knowledge. This progression can be viewed as a positive-feedback loop, in which newly discovered knowledge can be exploited, on one hand, to improve the process of individual-patient management, and on the other hand, to learn additional meaningful (temporal) knowledge.
Abstracts of Invited Papers
Abstract. In clinical decision making, information on the treatment of patients that show similar medical conditions and symptoms to the current case, is one of the most relevant information sources to create a good, evidence-based treatment plan. However, the retrieval of similar cases is still challenging and automatic support is missing. The reasons are two-fold: First, the query formulation is difficult since multiple criteria need to be selected and specified in short query phrases. Second, the discrete storage of multimedia patient records makes the retrieval and summary of a patient history extremely difficult. In this paper, we present a retrieval system for electronic health records (EHR). More specifically, a retrieval platform for EHRs for supporting clinical decision making in treatment of cervical spine defects with the information extracted from textual data of patient records is implemented as prototype. The patient cases are classified according to cervical spine defect classes, while the classification relies upon rules obtained from the corresponding defect classification schema and guidelines. In a retrospective study, the classifier is applied to clinical documents and the classification results are evaluated.
Building an i2b2-Based Integrated Data Repository for Cancer Research: A Case Study of Ovarian Cancer Registry

Na Hong¹,², Zheng Li¹,³, Richard C. Kiefer¹, Melissa S. Robertson¹, Ellen L. Goode¹, Chen Wang¹, and Guoqian Jiang¹

¹ Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
{hong.na, li.zheng, kiefer.richard, robertson.melissa1, egoode, wang.chen, jiang.guoqian}@mayo.edu
² Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, China
³ Department of Gynecologic Oncology, The Third Affiliated Hospital of Kunming Medical University, Kunming, Yunnan, China

Abstract. In this study, we describe our preliminary efforts in building an i2b2-based integrated data repository that supports centralized data management for ovarian cancer clinical research, and discuss important lessons learnt that would inspire the evaluation and enhancement for future generic cancer-specific data repository. We collected multiple types of heterogeneous clinical data, including demographic, outcome, chemo-treatment and lab-test information for ovarian cancer. To better integrate different data types, we conducted data normalization procedures through reusing standard codes and creating mappings between local codes and standard vocabularies. We also developed the extract, transform and load (ETL) scripts to load the data into an i2b2 instance. Through further analytic practices, we evaluated major expectations of the systems according to common clinical research needs, including cohort query and identification, clinical data-based hypothesis-testing, and exploratory data-mining. We also identified and discussed outstanding issues we will address through additional enhancement of existing i2b2 system.
Contents

Knowledge Discovery of Biomedical Data

Exploiting HPO to Predict a Ranked List of Phenotype Categories for LiverTox Case Reports ................................................................. 3
   Casey Lynnette Overby, Louiza Raschid, and Hongfang Liu

Patient Records Retrieval System for Integrated Care in Treatment of Cervical Spine Defect ................................................................. 10
   Yihan Deng and Kerstin Denecke

Managing, Querying and Processing of Medical Image Data

IEVQ: An Iterative Example-Based Visual Query for Pathology Database . . . 29
   Cong Xie, Wen Zhong, Jun Kong, Wei Xu, Klaus Mueller, and Fusheng Wang

Storing and Querying DICOM Data with HYTORMO ..................................... 43
   Danh Nguyen-Cong, Laurent d’Orazio, Nga Tran, and Mohand-Said Hadid

Cloud-Based Whole Slide Image Analysis Using MapReduce ......................... 62
   Hoang Vo, Jun Kong, Dejun Teng, Yanhui Liang, Ablimit Aji, George Teodoro, and Fusheng Wang

Information Extraction and Data Integration for Biomedical Data

Drug Dosage Balancing Using Large Scale Multi-omics Datasets .................. 81
   Alokkumar Jha, Muntazir Mehdi, Yasar Khan, Qaiser Mehmood, Dietrich Rebholz-Schuhmann, and Ratnesh Sahay

A Dynamic Data Warehousing Platform for Creating and Accessing Biomedical Data Lakes ................................................................. 101
   Pradeeban Kathiravel and Ashish Sharma

Building an i2b2-Based Integrated Data Repository for Cancer Research: A Case Study of Ovarian Cancer Registry .......................... 121
   Na Hong, Zheng Li, Richard C. Kiefer, Melissa S. Robertson, Ellen L. Goode, Chen Wang, and Guoqian Jiang
Health Information Systems

AsthmaCheck: Multi-Level Modeling Based Health Information System . . . 139
Tanveen Singh Bharaj, Shelly Sachdeva, and Subhash Bhalla

Author Index .......................................................... 155