He *et al. BMC Med Inform Decis Mak* 2020, **20**(Suppl 4):315
https://doi.org/10.1186/s12911-020-01292-x

BMC Medical Informatics and Decision Making

INTRODUCTION

Open Access

Selected articles from the Fourth International Workshop on Semantics-Powered Data Mining and Analytics (SEPDA 2019)

Zhe He[1]*, Cui Tao[2], Jiang Bian[3] and Rui Zhang[4]

From The 4th International Workshop on Semantics-Powered Data Analytics Auckland, New Zealand. 27 October 2019

Abstract

In this introduction, we first summarize the Fourth International Workshop on Semantics-Powered Data Mining and Analytics (SEPDA 2019) held on October 26, 2019 in conjunction with the 18th International Semantic Web Conference (ISWC 2019) in Auckland, New Zealand, and then briefly introduce seven research articles included in this supplement issue, covering the topics on Knowledge Graph, Ontology-Powered Analytics, and Deep Learning.

Background

In the era of big data, the volume, the variety, as well as the velocity of data being generated have posed major challenges for people to leverage multiple data sets for decision making [1]. Ontologies and semantic standards have been widely used to tackle some of the challenges in big data analytics such as data integration and knowledge discovery [2]. In the biomedical domain, ontologies and controlled vocabularies are a cornerstone for health information systems including clinical decision support systems and electronic health record (EHR) systems [2, 3]. Moreover, rich vocabularies and semantic information embedded in the ontologies have been leveraged to extract clinically meaningful information from heterogenous data from various sources. In particular, they are instrumental in natural language processing and text mining [4]. As a notable example, the Unified Medical Language System, developed and maintained by the U.S. National Library of Medicine, has been widely used in informatics research and applications using data in social media, scientific literature, and EHRs [5]. Applications like PubMed, which uses the UMLS indirectly, has been used by millions of users worldwide for biomedical research.

The International Workshop on Semantics-Powered Data Mining and Analytics (SEPDA) has been established as an important venue for experts to discuss semantic-based methods and applications in health data analytics [6–8]. To continue our momentum, SEPDA 2019 was held on October 26, 2019, in conjunction with the 18th International Semantic Web Conference (ISWC 2019). Submissions were solicited on the topics including Semantics-Based Data Mining and Analytics, Ontologies and Controlled Vocabularies, Data Integration, and Applications. After the peer review by the program committee members, 11 papers were accepted for presentation and publication in the SEPDA 2019 workshop proceedings [9]. After the workshop, the authors of seven selected papers were invited to extend their workshop papers to journal papers by adding additional

*Correspondence: zhe@fsu.edu
[1] School of Information, College of Communication and Information, Florida State University, 142 Collegiate Loop, Tallahassee, FL 32306-2100, USA
Full list of author information is available at the end of the article

experiments and greater details of the methods, results, and discussion. Each of the extended papers was subsequently reviewed by two experts in the field followed by multiple rounds of revisions to ensure the highest scientific rigor and clear presentation.

In this editorial, we summarize the papers included in this supplement. We categorize them into three main themes: Knowledge Graph, Ontology-Powered Analytics, and Deep Learning.

Knowledge graph

The majority of biomedical knowledge is still locked in text format such as those from textbook and scientific literature, while downstream applications such as those that provide clinical decision support still heavily rely on structured discrete data. Systems that curate knowledge graphs and knowledge bases from biomedical literature are rational intermediate steps. The paper from Rossanez et al. [10] introduced and evaluated a semi-automatic natural language processing (NLP) method that can generate knowledge graphs from biomedical texts. Their case study focused on Alzheimer's disease and their evaluation results demonstrated reasonable performance of the ontology-linked knowledge graphs.

Deep learning, which can classify nodes in the knowledge graph with good predictive performance, suffers from poor interpretability. In the healthcare domain, interpretability of AI models is critical for clinical decision making. Vandewiele et al. [11] presented a new method called MINDWAL, an inherently interpretable technique for classifying nodes in a knowledge graph. This technique uses a recursive algorithm to induce multiple decision trees and then decouple the modeling with multiple using informative random walks, which will create high-dimensional binary features that can feed a classification algorithm. This model has an improved interpretability and a competitive performance in terms of accuracy compared to other baseline techniques (e.g., decision tree, random forest, transform + logistic regression, transform + random forest). This technique can be applied to knowledge graphs in the biomedical domain to classify nodes in the graph.

Ontology-powered analytics

The needs to integrate diverse data sources across different domains (e.g., genetic factors and environmental exposures) and levels (e.g., individual traits as well as their interactions with the community) are growing so that a comprehensive examination of all potential risk factors is possible. The number of these multi-level integrative data analysis (mIDA) studies is increasing; nevertheless, the data integration processes in these mIDA studies are inconsistently performed and poorly documented. Zhang et al. [12] developed the ATTEST check list for standardized reporting of the variable and data source selection and subsequently the data integration processes. The novel piece of their study is the proposal to standardize the reports using an ontology, OD-ATTEST, that paves the way to enable sharing of mIDA study reports among researchers. Only when the selection and integration choices are clearly documented, the transparency and reproducibility of the studies can be warranted.

In [13], Zhang et al. proposed a semantic relationship mining method among disorders, genes, and drugs from different biomedical datasets. First, multiple heterogeneous biomedical datasets were converted and integrated into a resource description framework (RDF) storage system. Second, nine query patterns about genes, disorders, and drugs were presented. Third, the gene-disorder-drug semantic relationship mining algorithm was designed with these query patterns. The method was verified on SemMedDB, PharmGKB, KEGG, and Uniprot for Parkinson's disease semantic relationship mining. The results demonstrated that the method has advantages in mining and integrating heterogeneous biomedical datasets.

Amith and colleagues utilized their dialogue ontology called the Patient Health Information Dialogue Ontology (PHIDO) [14] to control a software engine for dialogue management ("Conversational Ontology Operator"). Using utterance data collected from past Wizard of OZ simulations [15, 16], they described how their ontology-driven software engine could power various software agents to preform dialogue tasks from health-based counseling for the HPV vaccine [17]. Their paper also outlines a question-answering sub-system ("FOQUS") that supplements the automated counseling of HPV vaccine where patients may ask questions. FOQUS utilizes a previous developed ontology knowledge base of HPV vaccine [18] to supply answers and was tested with question utterances from the aforementioned simulation. Their prototype engine presents some early showing of an ontology-based system to manage counseling methods for machines. Their future goal is to deploy this system to a live speech-enabled system to demonstrate its functional potential.

Deep learning

Deep learning has transformed medicine in the past few years [19]. Predicting treatment effects based on patients' personalized clinical status is vital in disease management. Traditional randomized controlled trials (RCT) usually are limited to a focused population and only evaluated the treatment effects after they have occurred [20]. EHRs containing large amounts of fine-grained clinical data provide a rich source to predict treatment effects. Chu et al. [21] proposed an adversarial deep treatment effect prediction (ADTEP) model based on auto-encoder and adversarial learning (AL). They encoded physical condition and treatment information for individual patients. An AL schema was also adopted to align the generated treatment with the actual performed treatments. The ADTEP model was evaluated on two clinical datasets and the results demonstrated its superiority compared with state-of-the-art methods.

Cancer survivors often experience emotional stress, post-traumatic stress disorder (PTSD), and other mental health issues. As such, they are at a high risk of self-destruction and harming others [22]. Early detection of mental health issues and early intervention would help prevent these undesired consequences. Social web such as Twitter allows people to share their experiences and opinions while keeping anonymous. Therefore, it is a great source for identifying cancer survivors with PTSD or other mental health issues. Ismail and colleagues [23] developed and evaluated a technique based on convolutional neural networks (CNN) to automatically classify tweets related to cancer survivors living with PTSD using word embeddings for text representation. The CNN-based model with word embeddings was trained to extract text features related to PTSD using a transfer learning approach and a depression lexicon. The results showed that the proposed model outperformed baselines including NBC, SVM, MLP, and CNN with n-grams for classifying the tweets.

Discussion and conclusions

In this supplement of selected articles from the Fourth International Workshop on Semantics-Powered Data Mining and Analytics (SEPDA 2019), seven papers were accepted after a rigorous peer review process. These papers demonstrated the power of the semantic methods in various applications, many of which are addressing critical challenges in healthcare such as predicting treatment effect, identifying cancer survivors living with PTSD, and mining relationships among disorders, genes, and drugs from biomedical databases. We hope these papers will have sustainable impacts not only on biomedical and health informatics but also other related fields. We also hope more researchers will be motivated by these exciting results and join our effort to improve population health and advance biomedical research with semantics-powered data analytics over disparate datasets.

Acknowledgements
The Guest Editors of this supplement would like to thank the authors and the reviewers for their scientific contribution and congratulate them on their high quality work.

About this supplement
This articles has been published as part of BMC Medical Informatics and Decision Making Volume 20 Supplement 4 2020: Selected articles from the Fourth International Workshop on Semantics-Powered Data Analytics (SEPDA 2019). The full contents of the supplement are available at https://bmcmedinformdecismak.biomedcentral.com/articles/supplements/volume-20-supplement-4.

Authors' contributions
ZH, CT, JB, and RZ. contributed to the writing of the manuscript. All authors have read and approved the final manuscript.

Funding
This manuscript was sponsored by the National Institute on Aging (NIA) of the National Institutes of Health (NIH) under Award Number R21AG061431 and in part by University of Florida-Florida State University Clinical and Translational Science Award funded by National Center for Advancing Translational Sciences under Award Number UL1TR001427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Availability of data and materials
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] School of Information, College of Communication and Information, Florida State University, 142 Collegiate Loop, Tallahassee, FL 32306-2100, USA. [2] School of Biomedical Informatics, University of Texas Health Science Center At Houston, Houston, TX, USA. [3] Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL, USA. [4] Institute for Health Informatics and College of Pharmacy, University of Minnesota, Minneapolis, MN, USA.

Published: 14 December 2020

References

1. Duan Y, Edwards JS, Dwivedi YK. Artificial intelligence for decision making in the era of Big Data-evolution, challenges and research agenda. Int J Inf Manag. 2019;48:63–71.
2. Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearb Med Inform. 2008;17(01):67–79. https://doi.org/10.1055/s-0038-1638585.
3. Amith M, He Z, Bian J, Lossio-Ventura JA, Tao C. Assessing the practice of biomedical ontology evaluation: gaps and opportunities. J Biomed Inform. 2018;80:1–13.
4. Yoo I-H, Song M. Biomedical ontologies and text mining for biomedicine and healthcare: a survey. J Comput Sci Eng. 2008;2(2):109–36.
5. Amos L, Anderson D, Brody S, Ripple A, Humphreys BL. UMLS users and uses: a current overview. J Am Med Inform Assoc. 2020;27(10):1606–11.
6. He Z, Tao C, Bian J, Dumontier M, Hogan WR. Semantics-powered healthcare engineering and data analytics. J Healthc Eng. 2017;2017:7983473. https://doi.org/10.1155/2017/7983473.
7. He Z, Tao C, Bian J, Zhang R, Huang J. Introduction: selected extended articles from the 2nd international workshop on semantics-powered data analytics (SEPDA 2017). BMC Med Inform Decis Mak. 2018;18(Suppl 2):56. https://doi.org/10.1186/s12911-018-0624-8.
8. He Z, Bian J, Tao C, Zhang R. Selected articles from the third international workshop on semantics-powered data analytics (SEPDA 2018). BMC Med Inform Decis Mak. 2019;19(Suppl 4):148. https://doi.org/10.1186/s12911-019-0855-3.
9. He Z, Bian J, Tao C, Zhang R. Proceedings of the 4th international workshop on semantics-powered data mining and analytics: CEUR workshop proceedings. https://ceur-ws.org/Vol-2427/. 21 Sept 2020
10. Rossanez A, Cesar does Reis J, de Silva Torres R, de Ribaupeirre H. KGen: a knowledge graph generator from biomedical scientific literature. BMC Med Inform Decis Mak. 2020. https://doi.org/10.1186/s12911-020-01341-5.

11. Vandewiele G, Steenwinckel B, De Turck F, Ongenae F. MINDWALC: mining interpretable, discriminative walks for classification of nodes in a knowledge graph. BMC Med Inform Decis Mak. 2020. https://doi.org/10.1186/s12911-020-01134-w.
12. Zhang H, Guo Y, Prosperi M, Bian J. An ontology-based documentation of data discovery and integration process in cancer outcomes research. BMC Med Inform Decis Mak. 2020. https://doi.org/10.1186/s12911-020-01270-3.
13. Zhang L, Hu J, Xu Q, Li F, Rao G, Tao C. A semantic relationship mining method among disorders, genes, and drugs from different biomedical datasets. BMC Med Inform Decis Mak. 2020. https://doi.org/10.1186/s12911-020-01274-z.
14. Amith M, Roberts K, Tao C. Conceiving an application ontology to model patient human papillomavirus vaccine counseling for dialogue management. BMC Bioinform. 2019;20(21):1–16.
15. Amith M, Anna Z, Cunningham R, Rebecca L, Savas L, Laura S, Yong C, Yang G, Julie B, Roberts K. Early usability assessment of a conversational agent for HPV vaccination. Stud Health Technol Inform. 2019;257:17.
16. Amith M, Lin R, Cunningham R, Wu QL, Savas LS, Gong Y, Boom JA, Tang L, Tao C. Examining potential usability and health beliefs among young adults using a conversational agent for HPV vaccine counseling. AMIA Summits Transl Sci Proc. 2020;2020:43.
17. Amith M, Lin R, Cui L, Wang D, Zhu A, Xiong G, Xu H, Roberts K, Tao C. Conversational ontology operator: patient-centric vaccine dialogue management engine for spoken conversational agents. BMC Med Inform Decis Mak. 2020. https://doi.org/10.1186/s12911-020-01267-y.
18. Dennis W, Cunningham R, Julie B, Amith M, Cui T. Towards a HPV vaccine knowledgebase for patient education content. Stud Health Technol Inform. 2016;225:432.
19. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19(6):1236–46.
20. He Z, Tang X, Yang X, Guo Y, George TJ, Charness N, Quan Hem KB, Hogan W, Bian J. Clinical trial generalizability assessment in the Big Data era: a review. Clin Transl Sci. 2020;13(4):675–84.
21. Chu J, Dong W, Wang J, He K, Huang Z. Treatment effect prediction with adversarial deep learning using electronic health records. BMC Med Inform Decis Mak. 2020. https://doi.org/10.1186/s12911-020-01151-9.
22. Gene-Cos N. Post-traumatic stress disorder: the management of PTSD in adults and children in primary and secondary care. National Collaborating Centre for Mental Health. London & Leicester: Gaskell & The British Psychological Society, 2005,£ 50.00, pp 168. ISBN: 190467125. Psychiatr Bull. 2006;30(9):357.
23. Ismail NH, Liu N, Du M, He Z, Hu X. A deep learning approach for identifying cancer survivors living with post-traumatic stress disorder on Twitter. BMC Med Inform Decis Mak. 2020. https://doi.org/10.1186/s12911-020-01272-1.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.