Published online: 2019-08-16

Formal Medical Knowledge Representation Supports Deep Learning Algorithms, Bioinformatics Pipelines, Genomics Data Analysis, and Big Data Processes

Findings from the 2019 IMIA Yearbook Section on Knowledge Representation and Management

Ferdinand Dhombres[1, 2], Jean Charlet[1, 3], Section Editors for the IMIA Yearbook Section on Knowledge Representation and Management

[1] Sorbonne Université, Université Paris 13, Sorbonne Paris Cité, INSERM, UMR_S 1142, LIMICS, Paris, France

[2] Médecine Sorbonne Université, Service de Médecine Fœtale, AP-HP/HUEP, Hôpital Armand Trousseau, Paris, France

[3] AP-HP, Delegation for Clinical Research and Innovation, Paris, France

Summary

Objective: To select, present, and summarize the best papers published in 2018 in the field of Knowledge Representation and Management (KRM).

Methods: A comprehensive and standardized review of the medical informatics literature was performed to select the most interesting papers published in 2018 in KRM, based on PubMed and ISI Web Of Knowledge queries.

Results: Four best papers were selected among the 962 publications retrieved following the Yearbook review process. The research areas in 2018 were mainly related to the ontology-based data integration for phenotype-genotype association mining, the design of ontologies and their application, and the semantic annotation of clinical texts.

Conclusion: In the KRM selection for 2018, research on semantic representations demonstrated their added value for enhanced deep learning approaches in text mining and for designing novel bioinformatics pipelines based on graph databases. In addition, the ontology structure can enrich the analyses of whole genome expression data. Finally, semantic representations demonstrated promising results to process phenotypic big data.

Keywords

Big data, informatics, health information technology, genomics, ontologies

Yearb Med Inform 2019:152-7
http://dx.doi.org/10.1055/s-0039-1677933

1 Introduction

The year 2018 has produced a large amount of publications related to Knowledge Representation and Management (KRM) in Medicine. KRM focuses on the development of techniques to be used and leveraged in other medical informatics domains. In this review, we present a selection of the best papers published in 2018 in the KRM domain, based either on their impact or on the novelty of the approach they proposed in the medical knowledge representation and management field.

2 Paper Selection Method

We conducted the selection of KRM papers based on a new set of queries. In comparison with the previous editions of the IMIA Yearbook, both PubMed/MELDINE and Web of Knowledge were used to search for KRM articles published in 2018 [1]. We followed the generic method commonly used in all sections of the Yearbook to select the best papers. As for the last four years, the search was performed on MEDLINE by querying PubMed. This year, we also performed an additional query on the ISI Web of Knowledge database (WoL).

Our query includes Medical Subject Headings (MeSH) descriptors related to KRM in the context of medical informatics with a restriction to international peer-reviewed journals, including conference proceedings indexed in PubMed. Only original research articles published in 2018 (from 01/01/2018 to 12/31/2018) were considered; we excluded the following publications types: reviews, editorials, comments, and letters to the editors.

The selection of the best papers was performed among the results of the query process, in three steps. At the first step, the section editors reviewed all titles, abstracts, and publication types in order to establish a short list of 15 candidate best papers. At the second step, five expert reviewers (including the section editors) reviewed the candidate best papers using the IMIA Yearbook quality criteria scoring method. More specifically, the following aspects of the papers were evaluated: significance, quality of scientific content, originality and innovativeness, coverage of related literature, organization, and quality of the presentation. The final step of best papers' selection was achieved during a meeting gathering the whole editorial board, based on the reviews and the report of the two section editors.

3 Results

For 2018, the KRM query retrieved 928 citations from PubMed and 34 additional citations from WoL. This new optimized query set accounts for 52% decrease of retrieved papers in comparison with results of the query used in 2017, with an overall improved precision of KRM relevant papers. Section editors achieved a first selection of 100 papers based on title and abstract. After a second review of this set of papers, including full text reviews, a selection of 15 candidate best papers was established [2-16]. Five reviewers reviewed these papers and four papers were finally selected as the best papers [2-5].

In direct line with the research presented last year [1], the 2018 four best papers demonstrated even further the added-value of ontology-based integration approaches for phenotype-genotype association mining.

4 Discussion and Outlook

4.1 Best Papers Selection for 2018

The paper authored by Arguello Casteleiro *et al*., and selected as a best paper, aims at automatically identify term variants or acceptable alternative free-text terms for gene and protein names in PubMed biomedical publications [2]. The use of a domain knowledge ontology, the Cardio Vascular Disease Ontology (CVDO), was associated with the best results. This study led to performance improvements for both Continuous Bag of Words (CBOW) and Skip-gram on a gene/protein synonym detection task by adding knowledge formalized in the CVDO and without modifying the word embeddings created. Hence, the CVDO supplies context that is effective in inducing term variability for both CBOW and Skip-gram while reducing ambiguity. In an other best paper, Le *et al*., presents Spfy, a platform that exploits semantic technologies with a graph database that allows rapid phenotype identification through a novel bioinformatics pipeline, as well as efficient storage and downstream comparative analysis of thousands of genome sequences [3]. In their paper, Osumi-Sutherland *et al*., use OWL-based (Ontology Web Language) reasoning on Gene Ontology (GO) to generate novel, biologically relevant groupings of GO terms to support mapping with a controlled vocabulary [4]. The GO term groupings generated by this approach can be used in over-representation analysis to detect cell and tissue type signatures in whole genome expression datasets. Also one of the best papers for 2018, the work of Yu *et al*., [5] presents a phenotyping algorithm (PheNorm) that does not require expert-labeled samples at the training step. This completely annotation-free 2-step classification method for phenotyping involves an initial normalization step of highly predictive features of the target phenotype, followed by a denoising step to leverage additional information contained in the remaining candidate features. This work introduces a method especially relevant for big data processing, which is the case for EHR-driven (Electronic Health Record) phenotyping. The four best papers are listed in table 1 and detailed in the appendix.

4.2 Main Trends in KRM in 2018

Among the 11 other candidate best papers from the short list for 2018, we observed four directions in research, (i) the research on ontology-based data integration for phenotype-genotype association mining; (ii) the design of ontologies and their application, the common direction of our field; (iii) the works regarding the semantic annotation of texts; and (iv) an experience about the deployment of OMOP-CDM (Observational Medical Outcomes Partnership Common Data Model) in Germany.

4.2.1 Semantics for Genomic Data Management

In a long article, Al Kawam *et al*., [15] tackled fundamental bioinformatics challenges involving semantic representations: genomic data generation, storage, representation, and utilization in conjunction with clinical data. For each aspect, they provided a detailed discussion on the current research directions, outstanding challenges, and possible resolutions. This paper seeks to help narrow the gap between genomic applications, which are being predominantly utilized in research settings, and the clinical adoption of these applications.

In differential diagnoses and disease gene prioritization, the Human Phenotype Ontology (HPO) is often used to compare a phenotype profile against gold-standard phenotype profiles of diseases or genes. In his article [7], Köhler investigated how this comparison can be improved by exploiting structure and information existing in annotation datasets or full text disease descriptions. He tested a study-wise annotation model for diseases annotated with HPO classes and for genes annotated with GO classes. This paper adds weight to the need for enhancing simple flat list representations of disease or gene annotations.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2019 in the section 'Knowledge Representation and Management'. The articles are listed in alphabetical order of the first author's surname.

Section Knowledge Representation and Management
▪ Arguello Casteleiro M, Demetriou G, Read W, Fernandez Prieto MJ, Maroto N, Maseda Fernandez D, Nenadic G, Klein J, Keane J, Stevens R. Deep learning meets ontologies: experiments to anchor the cardiovascular disease ontology in the biomedical literature. J Biomed Semantics 2018;9(1):13.
▪ Le KK, Whiteside MD, Hopkins JE, Gannon VPJ, Laing CR. Spfy: an integrated graph database for real-time prediction of bacterial phenotypes and downstream comparative analyses. Database (Oxford) 2018;2018:1-10.
▪ Osumi-Sutherland DJ, Ponta E, Courtot M, Parkinson H, Badi L. Using OWL reasoning to support the generation of novel gene sets for enrichment analysis. J Biomed Semantics 2018;9(1):10.
▪ Yu S, Ma Y, Gronsbell J, Cai T, Ananthakrishnan AN, Gainer VS, Churchill SE, Szolovits P, Murphy SN, Kohane IS, Liao KP, Cai T. Enabling phenotypic big data with PheNorm. J Am Med Inform Assoc 2018;25(1):54-60.

Cheng *et al*., [8] addressed the question of finding similarities of terms between different ontologies (e.g. HPO, Disease Ontology (DO), …etc.). They took advantage of the gene functional interaction network (GFIN) to explore such inter-ontology similarities of terms. They proposed InfAcrOnt to infer similarities between terms across ontologies and acquired similarities between terms across ontologies through modeling the information flow within the network by random walk. Comparisons of InfAcrOnt results and prior knowledge on pair-wise DO-HPO terms and pair-wise DO-GO terms showed high correlations.

4.2.2 Ontology Design and Documentation

Five of the candidate best papers present research in the ontology design domain [6, 10-12, 16]. This theme is common in the KRM section. While it was declining in recent years, it has become more present in 2018 than in previous years. In each case, articles described the motivation for building the ontology and developed the application that uses it for validation.

Traverso *et al*., [12] developed a Radiation Oncology Ontology (ROO). This ontology takes into account a few standard ontologies as the Fundational Model of Anatomy (FMA), the National Cancer Institute thesaurus (NCIt), and others terminologies. Authors demonstrated the possible conversion of clinical data following the FAIR principles (Findability, Accessibility, Interoperability, and Reusability), by using a combination of ontologies and Semantic Web (SW) technologies. This work proposes, using SW technologies based on existing ontologies, to efficiently and easily query data from different sources (relational databases) without knowing a priori their structures.

Bibault *et al*., [10] developed a Radiation Oncology Structures (ROS) Ontology. This ontology also relies on several standard ontologies as FMA, Radlex, and others. Authors provided annotations of EHR radiation oncology data with ROS concepts and integrated them into their clinical data warehouse. Finally, they showed the utility of the ontology in order to integrate dosimetric data.

Jing *et al*., [6] described the motivation and the building of OntoKBCF, an ontology for the cystic fibrosis domain. They illustrated the lack of sufficient clinical actionable knowledge that is related to molecular genetic information. The Cystic fibrosis ontology (OntoKBCF) is just a use case example, but given its structure, it should be relatively straightforward to extend the prototype to cover different genetic conditions. The principles underpinning its development could efficiently serve the design of knowledge bases for alternative human monogenetic diseases.

Facing the significant time cost to build ontologies, Zhao *et al*., proposed a data-driven sublanguage pattern mining method that can be used to create a knowledge model [16]. They combined standard Natural Language Processing (NLP) and semantic network analysis in their model generation pipeline. The results suggest that their pipeline is able to produce a comprehensive content-based knowledge model to represent context from various sources in the same domain.

In line with the publication of ontologies, Matentzoglu *et al*., [11] proposed the Minimum Information for Reporting an Ontology (MIRO) guidelines as a means to facilitate a higher degree of completeness and consistency between ontology documentation, including published papers, and ultimately a higher standard of report quality. These guidelines result from a survey among the KRM community that is detailed in the article. An illustrative review of 15 recently published ontology description reports from three important journals in the Semantic Web and Biomedical domain analyzed them for compliance with the MIRO guidelines. Only 41.38% of MIRO items were covered by these reports.

4.2.3 Semantics and Clinical Notes

Two candidate best papers presented a research involving semantic formalization associated with NLP approaches to process clinical texts. The first paper by Catling *et al*., [13] explored methods for representing clinical text using hierarchical clinical coding ontologies. This study demonstrates that hierarchically-structured medical knowledge can be incorporated into statistical models to produce improved performance for automated clinical coding. However, the authors reported that the data processing was difficult: they used a supervised learning approach with manually-as-signed clinical codes for the training dataset. Consequently, learning good representations of rare diseases in clinical coding ontologies from data alone remains challenging.

Viani *et al*., [9] proposed an ontology-driven approach to identify events (and their attributes) from episodes of care in medical reports written in Italian. Authors developed an ontology that can be easily enriched and translated. For this language, shared resources for clinical information extraction are not easily accessible. The proposed approach performed well on the considered Italian medical corpus, with a percentage of correct annotations above 90% for most considered clinical events.

4.2.4 Interoperability and Data Integration

The paper from Maier *et al*., [14] reports the experiment of implementing an OMOP/OHDSI-based pilot within a consortium of eight German University hospitals. Authors evaluated the applicability to support data harmonization and sharing among University hospitals, and they identified potential enhancement requirements. In order to facilitate the work of hospital centers, they provided a virtual machine preconfigured with the OMOP database and the OHDSI tools as well as the jobs to import the data and conduct the analysis. This work is encouraging, even if taking into account important vocabularies for Germany remains to be done. Such a paper shows the difficulties of moving from a model to a real implementation.

5 Conclusions

In the KRM selection for 2018, research on semantic representations demonstrated their added-value for enhanced deep learning approaches in text mining and for designing novel bioinformatics pipelines based on

graph database. In addition, the ontology structure can enrich the analyses of whole genome expression data. Finally, semantic representations demonstrated promising results to process phenotypic big data.

Acknowledgements
We would like to thank Martina Hutter and Adrien Ugon for their support and the reviewers for their participation in the selection process of best papers for the Knowledge Representation and Management section of the IMIA Yearbook.

References

1. Dhombres F, Charlet J. As Ontologies Reach Maturity, Artificial Intelligence Starts Being Fully Efficient: Findings from the Section on Knowledge Representation and Management for the Yearbook 2018. Yearb Med Inform 2018;27(1):140-5.
2. Arguello Casteleiro M, Demetriou G, Read W, Fernandez Prieto MJ, Maroto N, Maseda Fernandez D, et al. Deep learning meets ontologies: experiments to anchor the cardiovascular disease ontology in the biomedical literature. J Biomed Semantics 2018;9(1):13.
3. Le KK, Whiteside MD, Hopkins JE, Gannon VPJ, Laing CR. Spfy: an integrated graph database for real-time prediction of bacterial phenotypes and downstream comparative analyses. Database (Oxford) 2018;2018:1-10.
4. Osumi-Sutherland DJ, Ponta E, Courtot M, Parkinson H, Badi L. Using OWL reasoning to support the generation of novel gene sets for enrichment analysis. J Biomed Semantics 2018;9(1):10.
5. Yu S, Ma Y, Gronsbell J, Cai T, Ananthakrishnan AN, Gainer VS, et al. Enabling phenotypic big data with PheNorm. J Am Med Inform Assoc 2018;25(1):54-60.
6. Jing X, Hardiker NR, Kay S, Gao Y. Identifying Principles for the Construction of an Ontology-Based Knowledge Base: A Case Study Approach. JMIR Med Inform. 2018;6(4):e52.
7. Kohler, S. Improved ontology-based similarity calculations using a study-wise annotation model. Database (Oxford) 2018;2018.
8. Cheng L, Jiang Y, Ju H, Sun J, Peng J, Zhou M, et al. InfAcrOnt: calculating cross-ontology term similarities using information flow by a random walk. BMC Genomics 2018;19(Suppl 1):919.
9. Viani N, Larizza C, Tibollo V, Napolitano C, Priori SG, Bellazzi R, et al. Information extraction from Italian medical reports: An ontology-driven approach. Int J Med Inform 2018;111:140-8.
10. Bibault JE, Zapletal E, Rance B, Giraud P, Burgun A. Labeling for Big Data in radiation oncology: The Radiation Oncology Structures ontology. PLoS One 2018;13(1):e0191263.
11. Matentzoglu N, Malone J, Mungall C, Stevens R. MIRO: guidelines for minimum information for the reporting of an ontology. J Biomed Semantics 2018;9(1):6.
12. Traverso A, van Soest J, Wee L, Dekker A. The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques. Med Phys 2018;45(10):e854-e62.
13. Catling F, Spithourakis GP, Riedel S. Towards automated clinical coding. Int J Med Inform 2018;120:50-61.
14. Maier C, Lang L, Storf H, Vormstein P, Bieber R, Bernarding J, et al. Towards Implementation of OMOP in a German University Hospital Consortium. Appl Clin Inform 2018;9(1):54-61.
15. Al Kawam A, Sen A, Datta A, Dickey N. Understanding the Bioinformatics Challenges of Integrating Genomics into Healthcare. IEEE J Biomed Health Inform 2018;22(5):1672-83.
16. Zhao Y, Fesharak, NJ, Liu H, Luo J. Using data-driven sublanguage pattern mining to induce knowledge models: application in medical image reports knowledge representation. BMC Med Inform Decis Mak 2018;18(1):61.

Correspondence to:
Dr. Ferdinand Dhombres
Médecine Sorbonne Université, INSERM and APHP
Hôpital Universitaire Armand Trousseau
service de médecine foetale
26 rue du Dr Arnold Netter
75012 Paris, France
E-mail: ferdinand.dhombres@inserm.fr

Appendix: Content Summaries of Best Papers for the 'Knowledge Representation and Management' section of the 2019 IMIA Yearbook

Arguello Casteleiro M, Demetriou G, Read W, Fernandez Prieto MJ, Maroto N, Maseda Fernandez D, Nenadic G, Klein J, Keane J, Stevens R

Deep learning meets ontologies: experiments to anchor the cardiovascular disease ontology in the biomedical literature

J Biomed Semantics 2018;9(1):13

This work combines semantic modelling (Cardiovascular Disease Ontology, CVDO) and learning algorithms (word embeddings). The authors aim at automatically identifying term variants or acceptable alternative free-text terms for gene and protein names from PubMed biomedical publications. Ontologies, such as CVDO, capture domain knowledge in a computational form and can provide context for gene/protein names as written in the literature. This study investigates: i) if word embeddings from Deep Learning algorithms can provide a list of term variants for a given gene/protein of interest; and ii) if biological knowledge from the CVDO can improve such a list without modifying the word embeddings created. The results are of significant performance improvements for deep learning algorithms on a gene/protein synonym detection task, by adding knowledge formalized in the CVDO (leveraging the formal relations between genes and proteins). Hence, the CVDO supplies the context that is effective to induce term variability for algorithms while reducing ambiguity. As a result, CVDO can be enriched with new discovered synonyms (*skos:altLabel*). This work relies on a generic approach to be reused with other medical ontologies.

Le KK, Whiteside MD, Hopkins JE, Gannon VPJ, Laing CR

Spfy: an integrated graph database for real-time prediction of bacterial phenotypes and downstream comparative analyses

Database (Oxford) 2018;2018:1-10

Spfy is a platform that rapidly performs the common reference laboratory tests owing to its database of diverse pre-computed results, and the ability to incorporate user data. This platform handles all analysis tasks by dividing them into subtasks, which are subsequently distributed across a built-in task management process. All results are converted into individual graphs and stored within a large graph database according to previously created ontologies: the Genomic Epidemiology Ontology (GenEpiO), the Feature Annotation Location Description Ontology (FALDO), and the Microbial Typing Ontology (TypOn). These ontologies provide the relevant metadata for genotypes, location, biomarkers, host, and source. In its presented version, Spfy contains 10,243 Escherichia Coli genomes, for which in-silico serotypes and Shiga-toxin subtypes, as well as the presence of known virulence factors and antimicrobial resistance determinants have been computed. Spfy includes analyses modules that are also self-contained and can be used in existing platforms. This work demonstrates that Spfy, by leveraging semantic technologies with a graph database, facilitates rapid phenotype identification, as well as the efficient storage and downstream comparative analysis of thousands of genome sequences.

Osumi-Sutherland DJ, Ponta E, Courtot M, Parkinson H, Badi L

Using OWL reasoning to support the generation of novel gene sets for enrichment analysis

J Biomed Semantics 2018;9(1):10

The Gene Ontology (GO) consists of over 40,000 terms for biological processes, cell components, and gene product activities linked into a graph structure by over 90,000 relationships. It has been used to annotate the functions and the cellular locations of gene products. The graph structure is used by a variety of tools to group annotated genes into sets whose products share function or location. These gene sets are widely used to interpret the results of genomics experiments by assessing which sets are significantly over- or under-represented in results lists. F Hoffmann-La Roche Ltd. has developed a manually maintained controlled vocabulary (RCV) for use in over-representation analysis. The formal structure of GO and logical queries in OWL allow to map RCV terms to sets of GO terms. Finally, gene sets derived from the resulting GO terms sets can be used to detect the signatures of cell and tissue types in whole genome expression data. This article is very interesting and demonstrates all the added-value of ontological representation to three axes: (i) it shows a practical use case of ontology-based reasoning and how the authors can solve problems with widely available standards and tools (OWL2 EL, ELK); (ii) in mapping from the RCV to the GO, the authors found and resolved over 200 omissions in the axiomatization; and (iii) the approach to automate mapping between RCV and GO, replacing the unsustainable manual mapping process.

Yu S, Ma Y, Gronsbell J, Cai T, Ananthakrishnan AN, Gainer VS, Churchill SE, Szolovits P, Murphy SN, Kohane IS, Liao KP, Cai T

Enabling phenotypic big data with PheNorm

J Am Med Inform Assoc 2018;25(1):54-60

This paper addresses the difficulty to obtain a gold standard to train machine learning processes. The authors introduced a silver standard approach without human solicitation. They present PheNorm, a phenotyping algorithm that does not require expert-labeled samples at the training step. The input for the PheNorm algorithm consists of unlabeled data on a set of potentially informative features, either automatically curated or designed by experts. Online articles about the target phenotype from publicly available knowledge sources, such as Wikipedia and Medscape, are scanned with Natural Language Processing (NLP) software to extract medical concepts recorded in the Unified Medical Language System. These concepts are potentially related to the target phenotype. Then, narrative notes from the

Electronic Heath Record (EHR) database are processed with NLP software, which identifies mentions of the above medical concepts. With such a material, the most predictive features, such as the number of International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes or mentions of the target phenotype are normalized to resemble a normal mixture distribution with high area under the receiver operating curve (AUC) for prediction. The transformed features are then denoised and combined into a score for accurate disease classification. The authors validated the accuracy of PheNorm with four phenotypes: coronary artery disease, rheumatoid arthritis, Crohn's disease, and ulcerative colitis. The results suggest that PheNorm can potentially reduce the machine learning algorithm development process and demonstrate the capacity for EHR-driven annotations to scale to the next level – phenotypic big data.