Semantics-Driven Frequent Data Pattern Mining on Electronic Health Records for Effective Adverse Drug Event Monitoring

Jingshan Huang
School of Computing
University of South Alabama
Mobile, Alabama 36688-0002
Email: huang@southalabama.edu

Jun Huan
School of Engineering
University of Kansas
Lawrence, KS, 66047-7621
Email: jhuan@ittc.ku.edu

Alexander Tropsha
School of Pharmacy
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-7355
Email: alex_tropsha@unc.edu

Jiangbo Dang
Corporate Technology
Siemens Corporation
Princeton, New Jersey 08540-6632
Email: jiangbo.dang@siemens.com

He Zhang
School of Computing
University of South Alabama
Mobile, Alabama 36688-0002
Email: hz1101@jagmail.southalabama.edu

Min Xiong
School of Computing
University of South Alabama
Mobile, Alabama 36688-0002
Email: mx1201@jagmail.southalabama.edu

***Abstract*—Continued surveillance of post-marketing Adverse Drug Events (ADEs) is considered essential for patient safety, and Electronic Health Records (EHRs) serve as a critical source for identifying relevant information. But effective EHR knowledge discovery and data mining is not trivial because involved data usually have significantly different semantics among each other. Semantic technologies are believed to greatly assist in this regard; unfortunately, semantic technologies and conventional data mining remain largely separate disciplines, and the fusion of these two disciplines is still in its infancy. This position paper explores two semantics-driven frequent data pattern mining algorithms for EHR knowledge discovery, aiming at more effective ADE monitoring in a population. By effectively utilizing human knowledge formally encoded in EHR domain ontologies, our proposed algorithms will enhance the identification of the drug ADE causality out of large amounts of heterogeneous data sets. Through mining a large corpus of representative EHRs at semantic level, we will be able to compile a comprehensive list of ADE endpoints by obtaining critical, but originally hidden and implicit, frequent data patterns. Ultimately, our software to be developed will significantly facilitate effective ADE monitoring and prediction. Moreover, our research is expected to produce broader impacts on the pharmaceutical industry by reducing the R & D cost for new drug discovery and on transforming current pharmacovigilance methods to reduce adverse events and hence improve human health.**

Keywords*—EHR mining, ADE monitoring, semantics-driven frequent data pattern mining.

I. Introduction

Adverse Drug Events (ADEs) result in serious problems globally, causing hospitalizations and deaths, and incurring huge cost to healthcare ([1][2][3]). It is estimated that in the United States alone ADEs cause more than 770,000 injuries and deaths annually, and cost between $1.56 and $5.6 billion annually ([4][5][6][7][8]). It was reported that the incidence of serious ADEs was 6.7%, exclusive of medical errors [9]. Because many ADEs are not common (1 in 1,000) or rare (1 in 10,000), they often go undetected during clinical trials that typically include several thousand patients at most; therefore, continued surveillance (pharmacovigilance) of post-marketing ADEs is essential for patient safety [10].

Electronic Health Records (EHRs) are emerging as a critical information source for identifying drug post-market safety issues. Despite their potential significant relevance, limited work has been done partially due to the difficulties in effectively mining EHR data. One challenges is the wide range of data sources in EHRs including large amounts of narrative data, non-uniform report across ADE endpoints, patients with multiple medications, and many confounding factors (such as patient demographic information, genetic background, occupation, and disease history) in shadowing the real causality of observed toxicological endpoints. These involved data are not only distributed among various sources, but more importantly, they have significantly different semantics (intended meanings) among each other. In a nutshell, pharmacy scientists are facing significant barriers in effective knowledge discovery out of large amounts of semantically heterogeneous EHR data.

From a computational viewpoint, increasingly available data offer unprecedented opportunities for computer scientists to investigate effective approaches to enhance pharmacy scientists' knowledge discovery skills in a much broader context. In particular, emerging semantic technologies (based on domain ontologies) are believed to greatly facilitate conventional knowledge discovery. Semantic technologies place an emphasis on data semantics (intended meanings) instead of data syntax (forms in which data are represented), helping transform original data into semantics-enhanced data that are machine-readable and machine-processible. Semantics-enhanced data in turn will make possible the identification of more insightful connections among original data, and knowledge discovery can then be improved. In fact, there exist numerous EHR domain ontologies and respective on-going research ([11], [12], [13], and [14] for example). But there are significant knowledge

978-1-4799-1310-7/13/$31.00 ©2013 IEEE

gaps. In particular, semantic technologies and data mining remain largely separate disciplines. Although it is widely stated that exploring the fusion of these two disciplines is worthwhile, many researchers ([15], [16], and [17] for example) have pointed out that this line of work is still in its infancy.

One important focus of knowledge discovery and data mining is to extract frequent data patterns (a.k.a. frequent itemsets, sometimes shorten as k-itemsets) [18]. Frequent data patterns play critical roles in knowledge discovery as a basis for many mining tasks such as association rule mining and link discovery. Therefore, we propose in this position paper two innovative semantics-driven algorithms to discover frequent data patterns from semantics-enhanced EHR data.

The rest of this paper is organized as follows. Section II summarizes state-of-the-art research in semantics-driven data mining; Section III describes our proposed algorithms; and finally, Section IV concludes with future research directions.

II. Related Work in Semantics-Driven Data Mining

To systematically integrate domain knowledge into conventional data mining is a challenging problem. Ontologies and semantic technologies are believed to offer great assistance in this regard, and many research activities have been conducted. Existing semantics-driven data mining algorithms have used ontological concept generalization (based on the isa relationship) and concept inference. The former can reduce the data dimensionality and the latter aims to discover patterns through deductive reasoning rather than inductive learning. Hotho et al. [19] described an ontology-based text clustering approach, COSA, which was one of the earliest to utilize concept generalization through mapping terms in the text to ontological entities. COSA was later extended in [20] to assign higher weights to class-specific core words, and [21] further introduced a method to automatically augment ontologies to fit the specific purpose during the concept generalization. A P2P architecture was developed in [22] where queries were specifically forwarded to semantically related overlays. Authors in [23] presented an association rule mining algorithm to filter undesired results based on templates defined using general ontological concepts. SPARQL-ML [24] enabled ontology-based inference to outperform statistical relational learning models without inferencing mechanisms. Taghva et al. [25] reported the design of an ontology to encode expert rules to predict proper email categories. Despite all these research outcomes, as mentioned in Section I, many researchers have pointed out that the effective fusion of semantic technologies with data mining remains largely unsolved.

III. Proposed Semantics-Driven Data Mining Algorithms for EHR Knowledge Discovery

A. EHR domain ontologies and semantics-enhanced EHR data

As discussed in Section I, currently there exist numerous EHR domain ontologies. In particular, openEHR specifications, maintained by the openEHR Foundation [14], were specifically designed for health informatics to describe the management and storage, retrieval, and exchange of health data in EHRs. The openEHR specifications include information and service models for EHRs, demographics, clinical workflows, and archetypes. The success of openEHR was partially due to the formal acceptance of CEN 13606 as a European and ISO standard, which is based on many aspects of the openEHR design approach. The standard is also a snapshot of the openEHR archetype specifications, and the openEHR Foundation is working closely with CEN, ISO, HL7, OMG, and other standards organisations on EHR-related and clinical modelling standards. Additionally, openEHR provides convenient download mechanisms in the project website. In brief, openEHR, among other existing ontological models (e.g., [11], [12], and [13]), provide EHR domain ontologies as a solid foundation of semantic technologies to be applied.

Besides domain ontologies, there is another important component in semantic technologies: semantic data annotation (a.k.a. tagging), which is the process of tagging source files with predefined metadata such as names, entities, attributes, definitions, descriptions, and so forth. The annotation provides existing pieces of data with additional information from metadata. Such metadata are usually from a set of predefined ontological concepts or instances of ontological concepts. Semantic data annotation has attracted a large amount of research and resulted in many software products, for example, Annotea [26], SemTag [27], Ont-O-Mat [28], MnM [29], and KIM [30]. It is reasonable to assume that metadata from various EHR data sources (discussed in Section I) can somehow be annotated with existing EHR domain ontologies. The two proposed semantics-driven data mining algorithms in this paper will make use of existing EHR domain ontologies and their respective annotation results on conventional EHR data (referred to as semantics-enhanced EHR data in this paper).

B. Ontology-based k-itemset enrichment (OKE)

Beyond k-itemsets that can be generated from conventional mining methods, OKE will uncover additional k-itemsets from original data by effectively exploring and integrating various ontological relationships. In particular, not only the isa relationship is taken into account (as most of the state-of-the-art semantics-driven data mining research has done [19][20][21][22][23][24][25]), but also many other relationships, especially those ADE domain-dependent ones (e.g., *evidenceOfADE*, *mechanismOfMolecularFunction*, and *causeOfADE*) will be considered as well. As a result, extra, meaningful k-itemsets will be generated.

Step 1. Original k-itemsets. Use the widely accepted, conventional frequent itemset mining algorithm, Apriori [18], to generate a set of k-itemsets from conventional EHR data, of which each item corresponds to a column in original data. According to the abovementioned assumption of semantics-enhanced EHR data, each column would have been connected with some entity in domain ontologies.

Step 2. Semantic enrichment by domain ontologies. Find all "one-step" semantic neighbors for each entity in domain ontologies. Each relationship, either domain-dependent or domain-independent, between two entities is a directed edge in the ontology graph. Starting from an entity *E*, find all entities that are one edge away from *E*, denoted as *OneStepNeighbor(E)*, considering both incoming and outgoing edges. Then, find all columns connected with entities in *OneStepNeighbor(E)* and add these columns back to corresponding k-

itemsets. This new collection of k-itemsets will then be incorporated with additional domain knowledge formally encoded in domain ontologies.

Step 3. Expanding semantic connections. Step 2 can be repeated to find "two-step" semantic neighbors, "three-step" semantic neighbors, and so forth. There are two termination criteria: the user can predefine an integer, n, which is greater than or equal to two, to find "n-step" semantic neighbors; or when the user considers the new collection of k-itemsets as adequate.

Conventional multilevel association mining algorithm makes use of concept hierarchy (corresponding to isa) to find associations discovered at higher levels of abstraction, which is useful when it is difficult to find strong associations at lower levels. OKE is similar in that both methods utilize concept hierarchy. However, OKE goes far beyond a simple concept hierarchy. *Many more relationships considered by OKE are encoded in ontologies and will not be available to and utilized by conventional mining methods.*

C. Semantic hypergraph-based k-itemset generation (SHKG)

In conventional knowledge discovery and data mining, there exist widely accepted algorithms (Apriori for example) to discover frequent itemsets. But a frequent itemset cannot be identified in case the support of this itemset is low. It is not uncommon that situations exist where an itemset with low support indeed reflects some novel discovery. To remedy this issue, we propose a semantic hypergraph-based k-itemset generation algorithm: *without high support in original data (i.e., direct support), an itemset can still be possibly considered as a frequent itemset should adequate indirect support be identified from domain knowledge encoded in ontologies*.

Step 1. Data represented in a hypergraph. According to the analysis in [31], a hypergraph, *H*, can be effectively utilized to represent original data tuples, where each hyperedge, *e*, corresponds to one tuple in original data.

Step 2. Semantically enriched hypergraph. The hypergraph will be semantically enriched: (i) Each hyperedge contains a set of vertices, each of which corresponds to a column in original data; each column in turn would have been connected with some ontological entity. (ii) Using the same procedure described in the above OKE algorithm, find *OneStepNeighbor(E)* for every entity in domain ontologies, and then find columns connected with *OneStepNeighbor(E)* and add them back to original hyperedges. A new, semantically enriched hypergraph, *H'*, will be obtained. Each hyperedge in *H'* corresponds to a tuple in original data and is further combined with domain knowledge formally encoded in ontologies. Each such semantically enriched hyperedge represents a candidate k-itemset.

Step 3. Similarity measure design. Three similarity measures are defined to represent the strength of bond between a pair of two potentially associated items: (i) ontology-based similarity, S_{OnDis}, calculated by two item vectors' Euclidean distance; (ii) average commute time similarity, S_{CT}, calculated by the commute-time distance between two vertices in *H'*; and (iii) pseudoinverse-based inner-product similarity, S_{L+}, calculated by the Moore-Penrose pseudoinverse of hypergraph Laplacian. While S_{CT} and S_{L+} were first utilized in [31], the similarity measure, S_{OnDis}, is a new one proposed in this paper. S_{OnDis} is designed with the following idea: each item corresponds to an ontological entity; therefore, items can be naturally represented in vectors according to the formal semantics defined in ontologies. In other words, S_{OnDis} is the similarity distance between two ontological entities.

Step 4. From 2-itemsets to k-itemsets. A 2-itemset is an itemset that contains two semantically associated items. Utilizing each of the three similarity measures designed in Step 3, a collection of 2-itemsets can be calculated. Then, an overall similarity between every two candidate items can be further calculated as the average value out of three similarity values, i.e., $S_{Overall} = \frac{1}{3}(S_{OnDis} + S_{CT} + S_{L+})$. This overall similarity is the measure to define *G*, a pruned induced graph from *H'*, where an edge *(u,v)* is preserved only if the overall similarity between items u and v is greater than a threshold. Finally, given the collection of 2-itemsets, a collection of k-itemsets (k is greater than or equal to three) can be generated by calculating cliques or connected components of *G*. *Note that this collection of k-itemsets contains those k-itemsets that would not have been discovered otherwise when their direct support is low.*

Step 5. Expanding semantic connections. Steps 2 to 4 can be repeated to find "two-step" semantic neighbors, "three-step" semantic neighbors, and so forth. There are two possible termination criteria: the user can predefine an integer, n, which is greater than or equal to two, to find "n-step" semantic neighbors; or when the user considers the collection of k-itemsets as adequate.

D. Summarization

The novelty of two proposed algorithms is to improve the quality of discovered frequent data patterns by effectively utilizing prior human knowledge (explicitly and formally encoded in EHR domain ontologies). *Patterns, which would not have been realized otherwise, will be discovered to reinforce novel data connections.* Frequent data patterns play critical roles in knowledge discovery as a basis for many mining tasks such as association mining and link discovery. Thus, the implementation of these two proposed algorithms will represent an indispensable step forward in addressing the big challenge discussed earlier in this paper: semantic technologies and data mining remain largely separate. In addition, the proposed similarity measure, S_{OnDis}, will serve as an important supplement to the widely accepted (direct) support measure in conventional data mining.

IV. Conclusion

Continued surveillance of drug post-marketing ADEs is considered essential for patient safety, and EHRs serve as a critical source for identifying relevant information. Unfortunately, effective EHR knowledge discovery and data mining is highly challenging because involved data have significantly heterogeneous semantics among each other. Semantic technologies are promising in offering indispensible assistance in this regard; however, semantic technologies and conventional data mining remain largely separate disciplines. Despite that the fusion of these two disciplines is in great need, this line

of work is still in its infancy. Therefore, this position paper explores two semantics-driven frequent data pattern mining algorithms for EHR knowledge discovery, leading to more effective ADE monitoring in a population. By effectively utilizing human knowledge formally encoded in existing EHR domain ontologies, our proposed algorithms will enhance the identification of the drug ADE causality, which is usually embedded in noisy data, out of large amounts of heterogeneous data sets. Through mining a large corpus of representative EHRs at semantic level, we will be able to compile a comprehensive list of ADE endpoints by obtaining critical, but originally hidden and implicit, frequent data patterns. Ultimately, our software to be developed will significantly facilitate effective ADE monitoring and prediction. Additionally, our research is expected to produce broader impacts on the pharmaceutical industry by reducing the R & D cost for new drug discovery and on transforming current pharmacovigilance methods to reduce adverse events and hence improve human health.

An immediate future research work is to continue the ongoing implementation of the proposed algorithms and evaluate them on real-world EHR data. An even further research direction is to construct a comprehensive pharmacovigilance knowledgebase, which includes a list of drugs, their cellular and molecular activity profiles, their associated ADEs, and the related analytic software for efficient data mining of EHRs.

REFERENCES

[1] T. Mjorndal, M. D. Boman, S. Hagg, M. Backstrom, B. E. Wiholm, and A. Wahlin, "Adverse drug reactions as a cause for admissions to a department of internal medicine," *Pharmacoepidemiol Drug Saf*, vol. 11, no. 1, pp. 65–72, 2002.

[2] S. Schneeweiss, J. Hasford, M. Gottler, A. Hoffmann, A. K. Riethling, and J. Avorn, "Admissions caused by adverse drug events to internal medicine and emergency departments in hospitals: a longitudinal population-based study," *Eur J Clin Pharmacol*, vol. 58, no. 4, pp. 285–291, 2002.

[3] P. Impicciatore, I. Choonara, A. Clarkson, D. Provasi, C. Pandolfini, and M. Bonati, "Incidence of adverse drug reactions in paediatric in/outpatients: a systematic review and meta-analysis of prospective studies," *Br J Clin Pharmacol*, vol. 52, no. 1, pp. 77–83, 2001.

[4] D. C. Classen, S. L. Pestotnik, R. S. Evans, J. F. Lloyd, and J. P. Burke, "Adverse drug events in hospitalized patients. Excess length of stay, extra costs, and attributable mortality," *JAMA*, vol. 277, no. 4, pp. 301–306, 1997.

[5] D. J. Cullen, B. J. Sweitzer, D. W. Bates, E. Burdick, A. Edmondson, and L. L. Leape, "Preventable adverse drug events in hospitalized patients: a comparative study of intensive care and general care units," *Crit Care Med*, vol. 25, no. 8, pp. 1289–1297, 1997.

[6] D. J. Cullen, D. W. Bates, S. D. Small, J. B. Cooper, A. R. Nemeskal, and L. L. Leape, "The incident reporting system does not detect adverse drug events: a problem for quality improvement," *Jt Comm J Qual Improv*, vol. 21, no. 10, pp. 541–548, 1995.

[7] D. W. Bates, D. J. Cullen, N. Laird, L. A. Petersen, S. D. Small, and D. Servi, "Incidence of adverse drug events and potential adverse drug events. Implications for prevention. ade prevention study group." *JAMA*, vol. 274, no. 1, pp. 29–34, 1995.

[8] D. W. Bates, N. Spell, D. J. Cullen, E. Burdick, N. Laird, and L. A. Petersen, "The costs of adverse drug events in hospitalized patients. adverse drug events prevention study group." *JAMA*, vol. 277, no. 4, pp. 307–311, 1997.

[9] J. Lazarou, B. H. Pomeranz, and P. N. Corey, "Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies," *JAMA*, vol. 279, no. 15, pp. 1200–1205, 1998.

[10] W. K. Amery, "Why there is a need for pharmacovigilance," *Pharmacoepidemiol Drug Saf*, vol. 8, no. 1, pp. 61–64, 1999.

[11] Trajano EHR Domain Ontology. [Online]. Available: http://trajano.us.es/ĩsabel/EHR/

[12] C, González, B. G. Blobel, and D. M. López, "Ontology-based framework for electronic health records interoperability," *Stud Health Technol Inform*, vol. 169, pp. 694–698, 2011.

[13] J. Grabenweger and G. Duftschmid, "Ontologies and their application in electronic health records," in *eHealth: Connecting health tools & services and users, eHealth 08*, Portoroz, Slovenia, May 2008.

[14] openEHR Foundation. [Online]. Available: http://www.openehr.org/

[15] H. Daniels, A. Feelders, and M. Velikova, "Integrating economic knowledge in data mining algorithms," in *Proc. 8th International Conference on Society for Computational Economics: Computing in Economics and Finance, CEF 02*, Aix-en-Provence, France, 2002.

[16] L. Cao, "Actionable knowledge discovery and delivery," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 149–163, March 2012.

[17] S. Strohmeier and F. Piazza, "Domain driven data mining in human resource management: a review of current research," *Expert Systems with Applications*, vol. 40, no. 7, pp. 2410–2420, June 2013.

[18] J. Han, M. Kambe, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann Publishers, 2011.

[19] A. Hotho, S. Staab, and A. Maedche, "Ontology-based text document clustering," *Knstliche Intelligenz*, vol. 4, pp. 48–54, 2002.

[20] J. Wen, Z. Li, and X. Hu, "Ontology based clustering for improving genomic IR," in *Proc. 20th IEEE Symposium on Computer-Based Medical Systems, CBMS 07*, Maribor, Slovenia, June 2007, pp. 225–230.

[21] J. Fang, L. Guo, X. Wang, and N. Yang, "Ontology-based automatic classification and ranking for web documents," in *Proc. 4th International Conference on Fuzzy Systems and Knowledge Discovery, ICFSKD 07*, 2007, pp. 627–631.

[22] J. Li and S. Vuong, "Ontology-based clustering and routing in peer-to-peer networks," in *Proc. International Conference on Parallel and Distributed Computing Applications and Technologies, ICPDCAT 05*, 2005, pp. 791–795.

[23] B. Shen, M. Yao, Z. Wu, Y. Zhang, and W. Yi, "Ontology-based association rules retrieval using protègè tools," in *Proc. 6th IEEE International Conference on Data Mining, ICDMW 06*, Washington, D.C., USA, 2006, pp. 765–769.

[24] C. Kiefer, A. Bernstein, and A. Locher, "Adding data mining support to SPARQL via statistical relational learning methods," in *Proc. 5th Annual European Semantic Web Conference, ESWC 08*, Tenerife, Spain, 2008, pp. 478–492.

[25] K. Taghva, J. Borsack, J. Coombs, A. Condit, S. Lumos, and T. Nartker, "Ontology-based classification of emails," in *Proc. International Conference on Information Technology: Coding and Computing, ITCC 03*, Las Vegas, NV, USA, April 2003, p. 194.

[26] J. Kahan, M.-R. Koivunen, E. Prud'Hommeaux, and R. R. Swickd, "Annotea: An Open RDF Infrastructure for Shared Web Annotations," in *Proc. The Tenth International World Wide Web Conference, WWW 01*, Hong Kong, China, May 2001, pp. 623–632.

[27] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, K. S. McCurley, S. Rajagopalan, A. Tomkins, J. A. Tomlin, and J. Y. Zien, "A Case for Automated Large Scale Semantic Annotations," *Journal of Web Semantics*, vol. 1, no. 1, pp. 115–132, December 2003.

[28] S. Handschuh, S. Staab, and F. Ciravegna, "S-CREAM – Semi-automatic CREAtion of Metadata," in *Proc. The European Conference on Knowledge Acquisition and Management, EKAW 02*, Madrid, Spain, Octorber 2002.

[29] M. Vargas-Vera, E. Motta, J. Domingue, M. Lanzoni, A. Stutt, and F. Ciravegna, "MnM: Ontology Driven Tool for Semantic Markup." in *Proc. The Workshop of Semantic Authoring, Annotation & Knowledge Markup, SAAKM 02*, Lyon, France, July 2002.

[30] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff, "Semantic Annotation, Indexing, and Retrieval," *Journal of Web Semantics*, vol. 2, no. 1, pp. 49–79, December 2004.

[31] H. Liu, P. LePendu, R. Jin, and D. Dou, "A hypergraph-based method for discovering semantically associated itemsets," in *Proc. 11th IEEE International Conference on Data Mining, ICDM 11*, Vancouver, British Columbia, Canada, December 2011, pp. 398–406.