Investigating a Plausible Reasoning Approach for Performing Semantics-based Health Data Analytics

Hossein Mohammadhassanzadeh[1]

[1] NICHE Research Group, Faculty of Computer Science, Dalhousie University, Canada
hassanzadeh@dal.ca

Abstract. In the development of data-driven models for clinical decision support, incompleteness of the data is always a consideration. Plausible reasoning is the manifestation of the "plasticity" element of humans' reasoning capability to reason over incomplete data and discover new relationships. Hence, plausible reasoning can provide a practical approach to extend the coverage of knowledge-base of a clinical decision support system by abstracting plausible associations from heath data. However, an effective implementation of plausible reasoning relies on fine-grained conceptual relationships expressing how different concepts are semantically related. The Semantic Web offers effective formalisms to represent semantically annotated knowledge at various levels of expressivity, and to automatically reason over the knowledge and perform semantic analytics based on data. In our research, we investigate the potential of implementing plausible reasoning within the Semantic Web framework to handle the missing knowledge, especially when working with the open-world assumption. We propose a semantics-based data analytics framework, an innovative semantic reasoning method, to investigate certain Description Logic-based reasoning capabilities in the Semantic Web to discover hidden associations. We will evaluate the efficacy of the proposed framework in healthcare to perform effective semantic analytics using partial health data to make better decisions in disease diagnosis and long-term care. We demonstrate the efficiency of SeDan by answering intelligent medical questions posed by BioASQ challenges using Disease ontology, DrugBank and Semantic MEDLINE databases.

Keywords: Semantic Analytics, Plausible Reasoning, Semantic Web, OWL.

1 Motivation and Problem Statement

The massive volume of diverse data from clinical practice, healthcare and biomedical research is an opportunity for medical big-data analytics. However, due to the intrinsic nature of data that may be incomplete and inaccurate, the interpretation of data and its associations might be a serious challenge [1]. In this regard, innovative methods, algorithms and tools are needed to facilitate knowledge representation, exchange and reasoning, which is understandable for both human and machine [2].

There are a number of ways that new facts can be inferred when we have complete knowledge, however within an open-world assumption we need to account for incomplete knowledge that may lead to non-deductive reasoning, and plausible reasoning is one such reasoning approach. Plausible Reasoning (PR) is a form of non-demonstrative reasoning that it provides a mechanism to infer new knowledge, albeit a weaker inference, given partial knowledge. PR follows the physicians' thinking process to generate new hypothesis; if the existing knowledge is not sufficient, then the physicians leverage their own tacit knowledge to discover the correlations within existing medical data, draw new relationships and infer the missing knowledge [1].

The Semantic Web (SW) provides formalisms to semantically represent knowledge with various levels of granularity and reason over it to infer novel solutions. SW supports logic-based constructs to represent semantically annotated knowledge and offers built-in support for deductive reasoning conforming the open-world assumption [3].

In this research, we propose the concept of semantics analytics as the analysis of semantically annotated data (i.e., data represented in RDF) to infer new knowledge, whilst adhering to the SW's open-world assumption about knowledge incompleteness, by using expressive semantics and semantics relevant reasoning methods. We believe that RDF Schema and the Web Ontology Language allow for expressing additional semantics on top of the RDF knowledge base to achieve semantic analytics.

This research aims to investigate the potential of implementing plausible reasoning within the SW, targeting a semantic analytics framework for health data analytics, especially when working with large health datasets. In line with this objective, we aim to: (i) introduce additional markups (plausible extension to OWL) that extend OWL semantics to better capture and represent plausible semantics, (ii) develop a semantic analytics framework using query-rewriting algorithm to discover new associations between underlying domain-specific data, (iii) evaluate framework using health data.

1.1 Rationale

The proposed framework has the potential to be used for decision-making and problem solving in situations when we have incomplete knowledge. For the sake of domain consistency, we focus on healthcare applications for the following reasons:

- Semantic analytics is very relevant to healthcare, as it is mainly a knowledge-intensive domain. The opportunity to capture and leverage semantics via query processing is crucial for supporting disease diagnosis and long term care [4].
- A vast amount of health data is available from many diverse information systems. Effective semantic analytics of data enables the discovery of potential relationships to provide insights that can assist healthcare providers to make better decisions.

2 Plausible Reasoning

Plausible reasoning is non-demonstrative, ampliative and non-monotonic, that identifies the links between the question and the stored knowledge, and draw the line of inference based on conceptual semantics. PR performs inferencing by using a set of frequently recurring patterns that do not occur in formal logic [5]. [1] classifies the plausible patterns into 3 groups: hierarchy-based patterns, order-based and hybrid.

Hierarchy-based patterns, generalization and specialization, move between the nodes in hierarchical structure, from parent to child or vice versa, to perform a hierarchical plausible inference. Order-based patterns, interpolation and a fortiori, leverage measurable properties (partial order) to compare concepts regarding their size, order, location, etc. and infer new pieces of knowledge. However, hybrid patterns, (dis)similarity, will be performed using both hierarchical and partial order relations to infer a plausible answer; they probe hierarchy or consider the concepts that are analogues regarding some measurable properties.

3 SeDan: Semantics based Data Analytics Framework

To achieve the semantic analytics, we propose a framework (Fig. 1) that implements a plausible reasoner to infer new knowledge from RDF knowledge bases. This reasoner develops plausible reasoning patterns by manipulating the underlying graph directly with SPARQL query rewriting using OWL DL constructs.

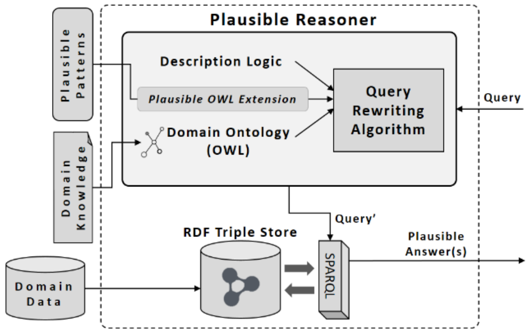

Fig. 1. Proposed semantics based data analytics framework

The proposed framework mainly includes three modules: knowledge sources, plausible reasoner and user interface. *Knowledge sources* provide terminological constructs to be consumed during the reasoning process, and assertional knowledge to be used to evaluate the extended query. The *plausible reasoner* (discussed more in the following section) delivers semantics analytics by running a query rewriting algorithm to perform plausible patterns and infer a set of so-called certain solutions. The system accepts the query with a list of desired plausible patterns via the *user interface*, and in return, delivers the plausible answer(s) and their justifications.

3.1 Plausible Extension to OWL

Standard reasoning capabilities within OWL 2 QL profile support various types of ontology-based inference – *rdfs:subClassOf* represents hierarchical relations and *owl:sameAs* conducts similarity. However, QL does not support all the semantics required in plausible patterns, like *partial order* or *context*. Therefore, it is needed to support OWL axioms in the cases that it has not enough expressivity. In this regard, plausible extension to OWL includes defining new classes, followed by defining new properties that use new (and existing) classes to express new relations. Table 1 demonstrates a subset of the proposed extension to OWL.

Table 1. Subset of the proposed plausible extension to OWL (PL-OWL)

Class Name	Supper Class		On Property	
OrderedProperty	ObjectProperty		-	
Context	Class		hasContext	
Property Name	**Type**	**Domain**	**Range**	**Inverse Property**
standsBefore	Ordered Property	Entity	Entity	standsAfter
standsAfter	Ordered Property	Entity	Entity	standsBefore
hasContext	Object Property	Entity	Context	-

In Table 1, *standsAfter* and *standsBefore* are instances of *OrderedProperty* demonstrating how entities are comparable w.r.t a measurable property (i.e., bigger, older). Also, *hasContext* indicates specific context in which ordered property is meaningful.

3.2 Query Rewriting Algorithm

Query Rewriting (QR) algorithms use ontological constructs to transform a given query to an expanded version that extracts both explicit (what a KB knows) and implicit (what it assumes) knowledge from the data [6]. Within the SW framework, OWL 2 QL profile, underpinned by DL-Lite family of description logics, provides a query rewriting mechanism to query data through an ontology. The OWA made in DLs makes QL suitable to work with partial knowledge in the SW scenarios [7], [8].

Inspired by GCLRR algorithm [9], SeDan's *plausible reasoner* transforms a query into a Union of Conjunctive Queries (UCQs) by applying the TBox ($\mathcal{T}$) axioms to the body atoms of the query. Starting with the initial query, the algorithm replaces the body atom of the query D, with new atom D'. The atom D' should be (i) semantically related to D ($\exists\alpha \in \mathcal{T}\ \alpha(D, D')$), and (ii) applicable to the preferred plausible patters. (i.e., *plowl:standsAfter* is applicable to interpolation). The new conjunctive query resulting from replacing an atom will be added to R, the set of conjunctive queries. This algorithm keeps formulating new queries until there is no UCQ to be added.

4 Evaluation Plan and Preliminary Results

In the real medical world, doctors are required to answer intelligent medical questions over large health data, which is usually incomplete and uncertain. Hence, to evaluate the efficacy of SeDan, we try to simulate a real medical setting. In this order, we pose questions from BioASQ [10] challenges over Semantic MEDLINE database [11]. DrugBank [12] and Disease Ontology [13] underpin the query rewriting algorithm.

In the preliminary experiments, we focused only on the questions that ask about *treatment* or *diagnosis*. As a result, 44 questions were retrieved; including 18 questions asking about *causes* of diseases and 26 questions asking about *treatments*. Among these questions, only 13 questions (%30) were answered using only existing triples stored in the knowledge base – deductive inference. Twelve out of 18 (%67) *causes* questions and 19 out of 26 (%73) *treatments* questions were not resolvable. Using SeDan framework, the plausible reasoner expanded the query answering coverage of SemMedDB by resolving %50, %11 and %26 of initially unanswered questions asking about *causes*, *treatments* and all the unanswered question, respectively.

5 Conclusions and Reflections

Our real word experiment proved that even a (very) large knowledge base (in our case SemMedDB with over 85 million records) suffers from incompleteness and may not be able to answer all the intelligent questions. This drawback is due to the lack of support for handling uncertainty resulting from missing associations between data attributes. To extend the knowledge coverage of medical knowledge-bases and enhance patient health outcomes, we introduced the SeDan framework that supports automated clinical decision support via semantics-based data analytics.

From the theory development view, SeDan implements plausible patterns using OWL constructs and SPARQL to provide principled means to represent and reason with incompleteness. Proposed plausible extension to OWL provides full-fledge support to implement plausible patterns within the SW. From an applied perspective, due to the flexible graph-based data format capable of incorporating new relations, support for rich semantics and automatic DL-based reasoning, the SW technologies provide excellent support for PR to draw semantic inferences from large data.

Acknowledgement. *I would like to express my gratitude to my supervisor Prof. Raza Abidi, and also Dr. Samina Abidi their guidance, encouragement and advice.*

References

[1] H. Mohammadhassanzadeh, W. Van Woensel, S. R. Abidi, and S. S. R. Abidi, “Semantics-based plausible reasoning to extend the knowledge coverage of medical knowledge bases for improved clinical decision support,” *BioData Min.*, vol. 10, no. 1, p. 7, 2017.

[2] A. Holzinger and I. Jurisica, “Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions,” *Interact. Knowl. Discov. data Min. Biomed. informatics*, pp. 1–18, 2014.

[3] N. Al Haider, S. Abidi, W. Van Woensel, and S. S. Abidi, “Integrating existing large scale medical laboratory data into the semantic web framework,” in *Big Data (Big Data), 2014 IEEE International Conference on*, 2014.

[4] O. Mohammed, “Semantic web system for differential diagnosis recommendations,” Lakehead University, 2012.

[5] M. Virvou and K. Kabassi, “Adapting the human plausible reasoning theory to a graphical user interface,” *IEEE Trans. Syst. Man, Cybern. A Syst. Humans*, vol. 34, no. 4, pp. 546–563, 2004.

[6] H. Pérez-Urbina and E. Rodrıguez-Dıaz, “Evaluation of query rewriting approaches for OWL 2,” *Proc. of SSWS+ HPCSW*, 2012.

[7] S. Grimm and B. Motik, “Closed World Reasoning in the Semantic Web through Epistemic Operators,” *OWLED*, 2005.

[8] M. Bienvenu, “Ontology-Mediated Query Answering: Harnessing Knowledge to Get More From Data,” in *nternational Joint Conference on Artificial Intelligence*, 2016.

[9] H. Pérez-Urbina, B. Motik, and I. Horrocks, “A Comparison of Query Rewriting Techniques for DL-lite,” *Descr. Logics*, 2009.

[10] G. Tsatsaronis *et al.*, “An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition,” *BMC Bioinformatics*, vol. 16, no. 1, p. 138, Dec. 2015.

[11] T. Rindflesch, H. Kilicoglu, and M. Fiszman, “Semantic MEDLINE: An advanced information management application for biomedicine,” *Inf. Serv. Use*, vol. 31, no. 1–2, pp. 15–21, 2011.

[12] V. Law, C. Knox, Y. Djoumbou, and T. Jewison, “DrugBank 4.0: shedding new light on drug metabolism,” *Nucleic Acids Res.*, vol. 42, no. D1, pp. D1091–D1097, 2013.

[13] W. Kibbe, C. Arze, V. Felix, and E. Mitraka, “Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D1071–D1078, 2014.