Enriching the Knowledge of a Domain Expert in a Recommendation System Based on Knowledge-Graph via Integrating a Domain-Specific Ontology

Sivan Albagli-Kim, Dizza Beimel
Department of Computer and Information Sciences
Ruppin Academic Center
Emek Hefer 4025000, Israel
email: sivana@ruppin.ac.il, dizzab2ruppin.ac.il

Sivan Albagli-Kim, Dizza Beimel
Dror (Imri) Aloni Center for Health Informatics
Ruppin Academic Center
Emek Hefer 4025000, Israel
email: sivana@ruppin.ac.il, dizzab2ruppin.ac.il

***Abstract*—This work in progress expands a previous study, where we proposed a decision-making framework designed to address the following need: an end-user contacts a domain expert to help him solve a problem. The end-user and the domain expert establish an interaction between them, consisting of questions and answers. This interaction is required to be effective, and to this end, the number of questions must be limited. The purpose of the framework is to suggest the next question to the domain expert, while he interacts with the end user. The framework consists of inference algorithms, making use of the domain expert's knowledge, which is structured into a knowledge graph. During the interaction, the end-user provides data that is fed into the graph as evidence and serves the inference algorithms to refine the next recommended question to the domain expert. The proposed extension refers to the addition of an existing ontology, describing the relevant domain, to the framework's base of knowledge. In particular, we want to take advantage of the knowledge, existing in the ontology (i.e., concepts and their relations), to enrich the framework's ability to offer a greater and more accurate range of questions. In the paper, we describe the proposed extension, followed by a case study.**

Keywords-knowledge graph; semantic reasoning; medical diagnostic; decision support systems; ontologies.

I. INTRODUCTION

The world of "big data" produces many challenges [1], One of them refers to the integration of big data in the technological realm dealing with decision-making processes to leverage these processes. Considering different needs, there are several types of decision-making processes, each requiring a suitable setup [2].

Our ongoing research [3] focuses on decision-making processes with the following setup: the process involves two entities - an *end user* (which is also the process initiator) and a *domain expert* (which assists the end-user to solve a problem); the entities establish an interaction, consisting of questions and answers, and is required to be as limited as possible (in time, the number of questions, money, etc.).

Given the above setup, we propose a semantic technology-based framework, which assists the domain expert in solving the end-user's problem, by suggesting a set of questions (inferred from the integrated big data) for the end user, such that the cycles of questions and answers will be reduced.

Our framework includes three components: (a) a formal representation of the relevant domain expert's knowledge using semantic technology, specifically a *knowledge graph,* which has emerged as a natural way of representing connected data [4], (b) an interactive set of algorithms, using the knowledge graph, and initial knowledge provided by the end user. The framework suggests relevant questions to the end user, while his/her answers advance the domain expert in the decision-making process and become input for the next iteration. The iterations will stop once the domain expert is satisfied, and a decision is made; (c) a domain-specific *ontology*, which is integrated into the knowledge graph. The ontology enriches the knowledge graph, thereby expanding the set of questions the domain expert can ask the end-user. The larger the question space, the more accurate decisions the domain expert can make.

The framework can support several domains that comply with the required setup; to demonstrate this, we chose to focus on the medical domain. To that end, we built a knowledge graph, which consists of two types of nodes representing *diseases* and *symptoms*. The directional edges, going from a symptom node to a disease node, represent a symptom that characterizes the disease. It is possible that a specific symptom can characterize several diseases. The goal of the decision-making process is to assist the domain expert to decide on a *diagnosis* (i.e., provide an explanation for a given set of symptoms based on analyzing available data).

The terms: disease, symptom, and diagnosis can be generalized, thus being used to represent other domains. For instance, in the domain of appliance repairs: the symptom represents a problem, the disease represents a malfunction, and the diagnosis is a fault identification.

The rest of the paper begin with reviewing knowledge representations (Section 2). We then briefly introduce the proposed framework (Section 3) and its new extension. Then, we provide further details on the KG enrichment by using the Symptoms Ontology (Section 4). In Section 5, we compare the previous version of our work with the current one. lastly, in section 6, we discuss our contribution and future work.

II. BACKGROUND: KNOWLEDGE REPRESENTATION

According to Davis [5], a Knowledge Representation (KR) serves five roles: as 1) a surrogate to enable an entity to determine the consequences of a plan or idea; 2) a set of ontological commitments about how and what to see in the world; 3) a fragmentary theory of intelligent reasoning; 4) a medium for efficient computation; and 5) a medium for human expression.

 ISBN: 978-1-68558-056-8

In this section, we review methods of KR: knowledge graphs, ontologies, and semantic technology.

A. Knowledge Graph

Knowledge Graphs (KG) represent information by converting data into a coded form, in particular by formulating relationships between entities into graph structures. KGs, also known as semantic graphs, generate interest among academic and industrial researchers, who deal with a wide variety of topics that all have the need to represent knowledge in common.

KGs have the property of providing semantically structured information. This property enables KGs to provide creative solutions for important tasks, such as answering questions [6], recommendation systems [7] and information retrieval [8]. Knowledge graphs are also considered to hold great promise for building smarter machines. KGs are also considered to offer great promise for building more intelligent machines.

B. Ontology

An ontology [9] is an explicit, machine-interpretable specification of a conceptualization—that is, the entities, or concepts, that are presumed to exist in some area of interest, their attributes, and the relationships amongst them. Ontology defines a common vocabulary for humans and machines that need to share information in a domain. The key reasons to develop ontologies includes [10]: 1) to enable the sharing of common understanding about the structure of information, among people or software agents; 2) to allow reusing of domain knowledge; and 3) to analyze domain knowledge.

C. Semantic Technology

Semantic technology represents a family of technologies that seek to derive meaning from information. That is, manage knowledge and join different data streams to perform inference. Representing knowledge is naturally done using the domain ontologies, and since the ontologies are based on a graph model, it is common to use a graph model to represent and store the data. By using graph representation for both the data and the domain knowledge, graph algorithms are used in order to infer new insights.

III. THE FRAMEWORK

In this section, we briefly introduce the proposed framework in [3], which includes a collection of algorithms and the flow between them. Then, we describe our extension to the framework, which is our current work.

We aim for interaction-based decision-making processes. The interaction is between a domain expert and an end-user, and results in a limited number of iterations consisting of questions that the framework suggests the domain expert ask the end-user. The decision-making process will progress according to the end-user's answers.

When we analyzed these types of processes, we concluded that they can be generically modeled as a collection of symptoms and diseases. Eventually, the process goal is to assist the domain expert to decide on a diagnosis (i.e., provide an explanation for a given set of symptoms based on analyzing available data). Questions that may arise during the diagnosis process are of the type: Does the end-user have a particular symptom?

The above terms (i.e., symptoms, diseases, questions, and diagnoses) produce a jargon that can naturally be used in the medical diagnostic domain, yet it is also suitable for other domains, such as appliance repairs: the symptom represents a problem, the disease represents a malfunction, the diagnosis is a fault identification, and a typical question can be: Does the end-user have a particular problem with his appliance?

In the rest of this section, we describe the framework presented in [3] along with its algorithms, and the extension of this work.

We start with building a knowledge graph from raw data, which will assist in exploring the relationships between diseases and symptoms. Following this, we use the Louvain hierarchical clustering [11] on the KG (Algorithm 1) to find communities (i.e., clusters of diseases that have similar symptoms). Then, given the symptoms reported by the end-user (called evidence symptoms), we find the possible diseases that are compatible with the evidence symptoms using inference on the KG (Algorithm 2). At this point, we infer the most probable community to include the end-user disease and suggest to the domain expert a question (symptom) that indicates this community (Algorithm 3). Lastly, we find the best diseases and symptoms that the end-user might have, to suggest to the domain expert (Algorithm 4), to address the improvement of the diagnostic process.

The whole framework is divided into two main parts: the first part, the pre-processing part, is carried out once the framework is launched; while the second part, the processing part, is carried out each time a new request arrives in the framework. This current work expands on our previous work. As mentioned, it semantically enriched the knowledge maintained within the framework. In particular, the new addition expanded the pre-processing part, in step 2 (See Figure 1 for the new architecture of the pre-processing part).

A. Pre-Processing Part

Input: A list of diseases and their symptoms.

Step 1: Construct a knowledge graph (KG) of diseases and symptoms. The left hand side of Figure 4 exhibits an example of a such KG.

Step 2: Enrich the KG with symptoms Ontology [12]. (See Section IV for more details).

Step 3: Cluster the diseases into groups (called communities), according to their symptoms: diseases with similar symptoms will be in the same community (Algorithm 1, from [3]).

B. Processing part

This part is presented in detail in [3].

Input: k evidence symptoms

Step 1: Find the most probable diseases: the possible diseases that are compatible with evidence symptoms (Algorithm 2).

Step 2: Infer and suggest to the domain expert (repeatedly as required) a question (symptom) that indicates the most probable community to include the end-user disease (Algorithm 3).

Step 3: Infer and suggest to the domain expert a list of diseases the end-user might have and their related questions (symptoms), sorted by relevance (Algorithm 4).

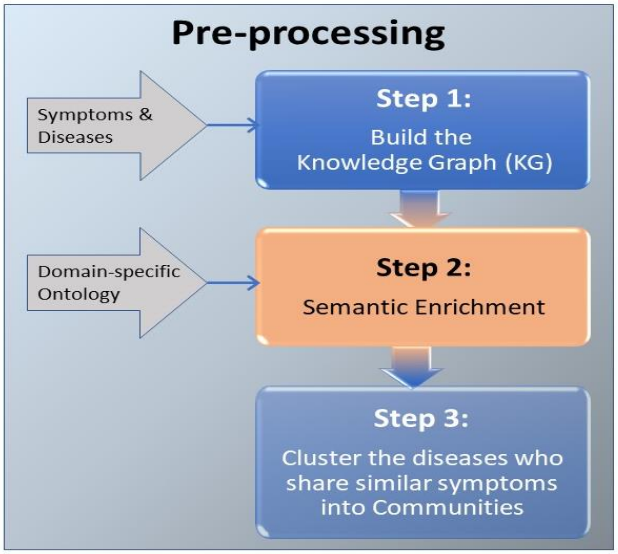

Figure 1. The architecture of the new pre-processing part

C. Contribution of the Semantic Technology Extention

Adding the ontology to the KG, as part of the *Pre-Processing* part (see Figure 2), has a main role in enriching the semantic knowledge of the domain expert. The KG is data driven knowledge, based on the historical examination of domain experts [13], and does not consist of the structure of the symptoms themselves (hierarchy). Adding this knowledge to the KG assists the recommendations process by inferring new relations, and thus inferring new relevant diseases to the domain expert.

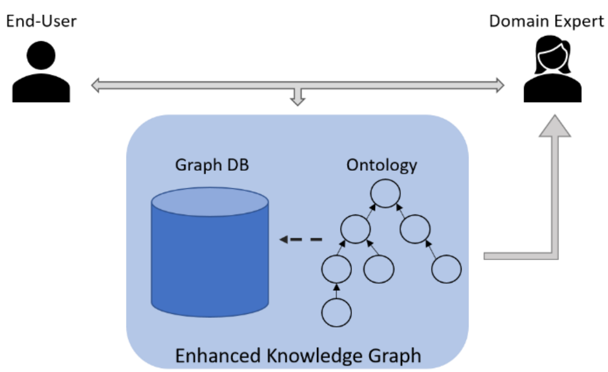

Figure 2. The Semantic Technology Architecture

IV. Graph Enrichment Algorithm

In this section, we present Step 2 of the *Pre-Processing part*: the enrichment of the KG using the Symptoms Ontology [12]. The Symptoms Ontology (SYMP) consists of nodes representing the symptoms, and edges representing an isA relation between symptoms. Thus, the ontology represents the hierarchy of the symptoms. The right-hand side of Figure 4 exhibits an example of such ontology. After constructing the KG (step 1 of the *Pre-Processing* part), and storing it using Neo4j Graph Database, the ontology SYMP is added to the database, and then we perform the following procedures:

A. Add Symptom Nodes to the KG

- For all edges $e = (s_i, s_j)$ in SYMP, such that $s_j \in KG$ and $s_i \notin KG$:
  - Add s_i as a symptom node to KG.
- For all edges $e = (s_i, s_j)$ in SYMP, such that $s_i \in KG$ and $s_j \notin KG$:
  - Add s_j as a symptom node to KG

B. Add isA Relations between Symptoms in the KG, according to the Ontology

- For all edges $e = (s_i, s_j)$ in SYMP, such that $s_j \in KG$ and $s_i \in KG$:
  - Add the edge (si, sj) to KG, labeled *isA*.

Figure 3 presents the legend we use in Figure 4 and Figure 5. Figure 4 and Figure 5 present the construction and the integration of the ontology into the KG.

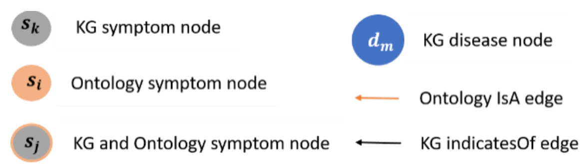

Figure 3. Legend

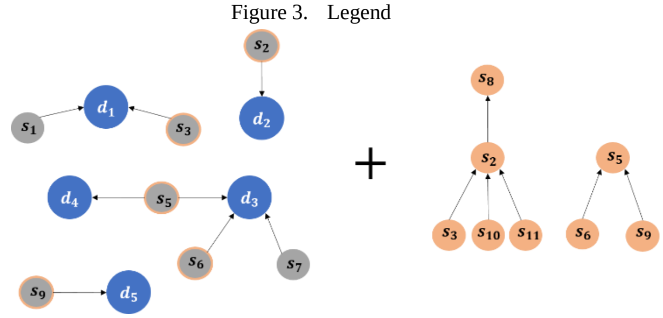

Figure 4. On the left side the KG, on the right side the Ontology

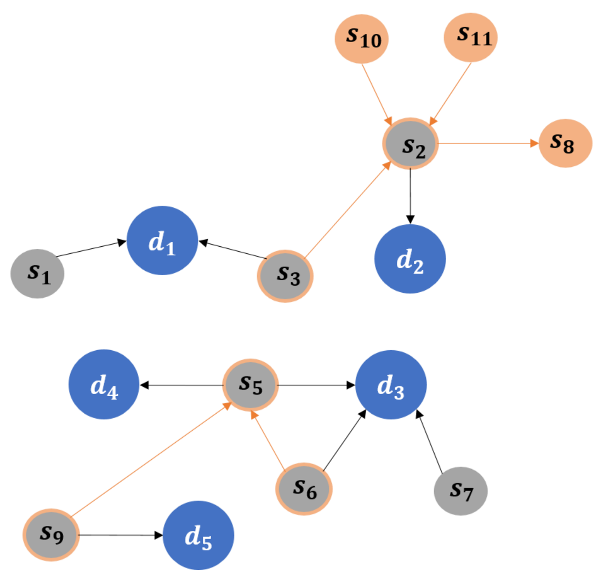

Figure 5. The Knowledge Graph Enrichment

V. COMPARING THE PREVIOUS WORK WITH THE CURRENT

The new edges that were added to the knowledge graph (the ontological edges - painted in orange), semantically enriched the space of questions, and are used by the domain expert. In the previous work, the selected symptoms were those that reinforce the most probable disease (independent of the other symptoms the patient has). In the current work, the symptoms that will be examined are those that will strengthen the most probable disease, and are also semantically related to the other symptoms of the patient.

To illustrate the influence of this semantic enrichment, let's consider the following scenario: A patient arrives with the following two symptoms: s_1, s_2 and s_5. These are our evidence symptoms. Therefore, the patient's probable diseases are d_1, d_2, d_3 and d_4. Since s_2 is an evidence, and s_3 *isA* s_2, the symptom s_3 is more likely to be considered to the domain expert in the hypothesis that d_1 is the patient's disease. In addition, since s_5 is an evidence, and s_6 *isA* s_5, the symptom s_6 is more likely to be considered to the domain expert in the hypothesis that d_3 is the patient's disease.

VI. CONCLUSION AND DISCUSSION

This section summarizes our results including our contribution, and present our future work.

A. Summary

In most areas of life, one can find decision-making processes, which makes this topic interesting for relevant research. At the same time, since these processes are found in many worlds of content, there is a wide and rich variety of decision-making processes, characterized by different needs. Therefore, in any attempt to support this topic, we must focus on a specific subtopic, characterized by specific requirements.

In the current (and ongoing) work, we focus on decision-making processes with the following configuration: an end-user and a domain expert are involved in the process, which establishes an interaction between them, consisting of questions and answers, to address a problem of the end-user. The domain expert uses the suggested framework to make the interaction as limited as possible (in time, the number of questions, money, etc.).

B. Contribution

In our previous work [3], we introduced for the first time the framework we built, including a detailed description of the algorithms that we developed as part of the framework, which enable inference of big data. The innovation of [3] stems from the use of semantic technologies, including a graphical data model, combined with unique algorithms.

In the current work, we introduce an extension to our framework, such that a domain-specific ontology is integrated into the knowledge graph, and hence expands the space of questions the domain expert can ask, resulting in a more accurate inference algorithm.

C. Future work

We want to develop the current research, in particular, to explore the contribution of the ontology to the decision-making process, and to run a case study on the knowledge graph we created in the previous study, after incorporating the ontology into that graph.

In addition, we wish to explore the possibility of using weighted edges in the knowledge graph for representing the cost of each question.

REFERENCES

[1] I. A. T. Hashem, et al., "The rise of "big data" on cloud computing: Review and open research issues," Inf. Syst. 47, pp. 98–115, 2015.

[2] D. J Power, "Decision Support Systems: Concepts and Resources for Managers," Greenwood Publishing Group: Westport, CT, USA, 2002.

[3] S. Albagli-Kim and D. Beimel, "Knowledge Graph-based Framework for Decision-making Process with Limited Interaction," Mathematics | Special Issue: From Edge Devices to Cloud Computing and Datacenters: Emerging Machine Learning Applications, Algorithms, and Optimizations. 10 (21),pp. 3981,2022.

[4] I. Robinson, J. Webber, and E. Eifrem, "Graph Databases: New Opportunities for Connected Data," O'Reilly Media, Inc.: Middlesex County, MA, USA, 2015.

[5] R. Davis, H. Shrobe, and P. Szolovits, "What is a Knowledge Representation?", AI Magazine, vol. 14, no. 1, pp. 17-33, 1993.

[6] A. Gashkov, A. Perevalov, M. Eltsova, and A. Both, "Improving Question Answering Quality through Language Feature-Based SPARQL Query Candidate Validation," The Semantic Web. ESWC 2022. Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; Volume 13261. https://doi.org/10.1007/978-3-031-06981-9_13.

[7] Q. Guo, et al., "A survey on knowledge graph-based recommender systems," IEEE Trans. Knowl. Data Eng. 34, pp. 3549–3568, 2020.

[8] L. Dietz, A. Kotov, and E. Meij, "Utilizing knowledge graphs for text-centric information retrieval," In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, pp. 1387–1390, July 2018.

[9] T. R. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing," Intl J Human-Computer Studies, vol. 43, no. 5-6, 1995.

[10] N. F. Noy and D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology," Stanford Medical Informatics Technical Report SMI-2001-0880 2001.

[11] H. Lu, M. Halappanavar, and A. Kalyanaraman, "Parallel heuristics for scalable community detection," Parallel Computing, 47. pp. 19-37, 2015

[12] Symptom Ontology, Ontology Lookup Service (OLS), https://www.ebi.ac.uk/ols/ontologies/symp, retrieved: 02,2023

[13] Kaggle, https://www.kaggle.com, retrieved: 02,2023