Similarity Estimation among OWL Descriptions of Computational Cardiology Problems in a Knowledge Base

Ch Bratsas[1], IS Hatzizisis[1], P Bamidis[1], P Quaresma[2], N Maglaveras[1]

[1]Aristotle University, Thessaloniki, Greece
[2]University of Evora, Evora, Portugal

Abstract

The purpose of this study is to evaluate ontologies in order to achieve an efficient search of Medical Computational Problems in cardiology (MCP). These ontologies provide the ability to standardize the description process of cardiology problems and algorithms by utilizing Ontology Web Language – Description Logic (OWL-DL). The final aim is to build a Knowledge Base (KB) from the OWL descriptions of the MCPs. The identification of a new cardiology problem or related to this problem included in KB is proposed to be conducted by cosine-based similarity. A novel algorithm that constructs the MCPs vectors from the ontology individuals is presented. These vectors are defined by a matrix based on problems, Concept Unique Identifier (CUI) Keywords and CUI Categories. The CUIs are automatically received by the Unified Medical Language System Knowledge Server (UMLS-KS) and stored into ontologies. Finally, an example where twelve cardiology problems vectors are constructed from the corresponding individuals of the ontology classes is presented and analyzed.

1.   Introduction

Computer cardiology research is an established knowledge engineering that has allowed researchers to identify new or significant problems, better understand the existing approaches and experimental results and obtain new, effective and efficient solutions. Some questions that researchers are familiar with include: How does one determine a new problem? Do similar or related problems exist? Is there any algorithm that solves such problem? Has the solving algorithm of the problem been implemented ever? Are there any relative publications or references?

The Answer of these questions requires literature searches in libraries and digital archives, but the scattered and the poorly organized information that exists in these sources, is considered a barrier to knowledge acquisition. It is believed that this inefficient procedure can be improved through a Knowledge Base System (KBS) based on advanced information technology.

Ontologies provide a structured framework for modeling the concepts and the relationships of some domain of expertise. Particularly, ontologies support the creation of repositories of domain-specific reference knowledge or knowledge for communication and sharing among people and applications [1]. In previous work [2] Resource Description Framework (RDF) [3] was used to accomplish a simplistic framework for the ontological description of Medical Computational Problems (MCPs) and the algorithms that solve them.

In this piece of work, this framework is further expanded with the use of Ontology Web Language (OWL) [4] in order to facilitate the description process of the users and the algorithms' implementations. Moreover, an innovative algorithm that constructs the problem vectors from the ontology individuals has been built. These vectors are used to calculate the similarity of the problems in order to achieve an efficient search in a KB and to help the knowledge problems authors to identify if a problem already exists in the KB.

2.   MCP ontology

The MCP ontology is an OWL ontology model that manages the MCPs by means of organizing and visualizing their existing knowledge. MCPs are medical problems, the solution of which deals with mathematical or statistical models, signal or image processing and estimation of corresponding parameters. The tool Protégé [5] with its OWL plug-in [6] was used to build the MCP Ontology. In order to achieve concept consensus, a standardized medical terminology, the Unified Medical Language System (UMLS) developed by the US National Library of Medicine [7] was utilized. To describe MCP references in a structured manner, the BibTex OWL-Ontology [8] was adjusted into the MCP Ontology.

The user space and three computational spaces namely the medical problem, its solution and its implementation are proposed in order to provide a detailed and accurate description of the MCP Ontology (Fig 1). The problem space consists of medical computational problems; the

solution space is composed of algorithm solutions, whereas the implementation space objects assist in carrying out the solutions. The user space objects consist of the users and their different capabilities.

In order to correlate the above computational spaces between each other the inverse property is used. More specifically, the hasTargetProblem property is stated to be the inverse of the hasExistingSolution property, and the hasTargetSolution property is stated to be the inverse of the hasImplementations property. The first relation is used to associate the problem space with the solution space, whereas the second relation is used to associate the solution space with the implementation space. Thus, if X is related to Y by the hasTargetProblem property or hasTargetSolution property, then Y is related to X by the hasExistingSolutions property or hasImplementations property respectively. These relations are used to answer the following questions: Is there any algorithm that solves this problem? Is the algorithm solving the problem implemented?
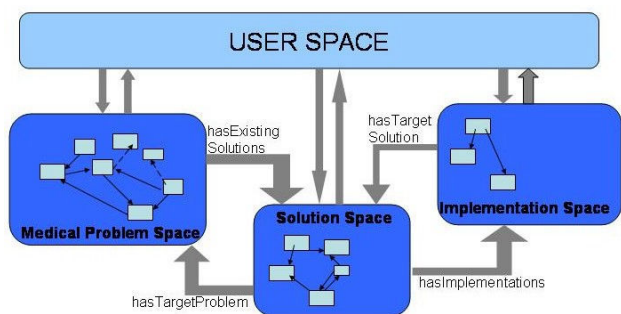

Figure 1: MCP Ontology.

The estimation of similarity among MCPs is based on the instances of the MCP ontology problem space. The main class of the problem space is the MCP class. Figure 2 shows in a table the properties that associate this class with other classes. In addition, the significance of each association is given in the same table.

Knowledge Objects as domain Classes	Properties that links the individuals between two classes	Linking Classes	Meaning
MCP classs - main class of MCP space	hasDescribesby	Knowledge Author	By whom the MCP is described
	hasExistingSolutions	Algorithm	Which is MCP's solution
	hasMCPProfile	Profile	Name, Language, Date, Description, Modification, References. Reference is based on BibTex Ontology.
	hasMedicalConcepts	Medical Keyword	Every Medical Concept is a MCP Keyword which has UMLS Description
	hasCategory	Category	Every MCP belongs to one category, which has UMLS description

Figure 2: The problem space of the MCP ontology.

The Object Properties hasMedicalConcepts and hasCategory are the inverse properties of the hasMCPProblem and hasReferredMCP properties correspondingly. Thus, the instance of the MedicalKeyword class shows the keywords that belong in any MCP, and the instance of the Category class illustrates the category/ies in which a MCP belongs.

3. Vector constructing algorithm

To construct the MCPs vectors sets P, K, C are defined. The MCPs set $P = \{p_1, p_2, ..., p_n\}$, where n is the number of MCPs in the KB, contains all the instances (<*MCP rdf:ID="ProblemID "*>) of the MCP class. The set $K = \{k_1, k_2, ..., k_m\}$, where m is the number of keywords, includes the Concept Unique Identifiers (CUIs) of the keywords, and the set $C = \{c_1, c_2, ..., c_z\}$, where z is the number of categories, includes the CUIs of the categories. The CUIs are received by the UMLS_Root class.

The function $f(i,j,z)$ is defined to be equal to one when Problem i of set P belongs to Category z of set C and Keyword j of set K belongs to Problem i of set P, and equal to zero otherwise (equation 1). The above function is constructed by the instances of the MedicalKeyword class and the Category class.

$$f(i,j,z) = \begin{cases} 1, \text{if } (P_i \in C_z) \wedge (K_j \in P_i). \\ 0, \text{if } (P_i \notin C_z) \vee (K_j \notin Pr_i) \end{cases} \quad (1)$$

The matrix P_{ijz}, of n×m×z dimension, includes the MCPs vectors existing in KB. For instance, the i row corresponds to the MCP_i vector $\vec{P_i}$ that belongs to Category z. The elements of the matrix P_{ijz} are computed by the equations 2,3,4.

$$P_{ijz} = \begin{cases} ws_j, \text{if } f(i,j,z)=1. \\ 0, \text{if } f(i,j,z)=0 \end{cases} \quad (2)$$

$$ws_j = \frac{n_c - (Freq(K_j) - 1)}{n_c} \quad (3)$$

$$Freq(K_j) = \sum_{i=1}^{n} f(i, j, z=c) \quad (4)$$

The MCPs description is rather short, since every MCP is clearly defined by Expert Users within a paragraph of 3 or 4 lines. So, the keywords are not repeated frequently and as a result every Medical Concept of a MCP description is of equal importance as a keyword. The ws_j, corresponds to the weight of the keyword in a particular problem in a specific Category. The ws provides the keyword with its sparsity and is calculated from equation (2). In this equation n_c represents the number of MCPs

that belong in category c, whereas Freq(K_j) represents the MCPs containing the Keyword K_j.

The similarity among the MCPs in the KB can be estimated by calculating the cosine of the MCPs Vectors (equation 5, where $\bullet$ is the dot product).

$$Sim(P_i, P_j) = Cos(\vec{P_i}, \vec{P_j}) = \frac{\vec{P_i} \bullet \vec{P_j}}{|\vec{P_i}| \cdot |\vec{P_j}|}, \forall i, j = \{1,2,...,n\} \quad (5)$$

The set $QK = \{qk_1, qk_2, ..., qk_s\}$, includes the CUIs of the quested MCP keywords (s is the keywords' number). The vector of the quested MCP is defined by the matrixes SP (equation 6) and QK_2 $\vec{QP} = [QP \quad QK_2]$.

$$QP_j = \begin{cases} 1, \text{if } k_j = QK_i \\ 0, \text{if } k_j \neq QK_i \end{cases}, \forall i = \{1,2,...,s\} \quad (6)$$

The QK_2 is a set of aces such in number as the CUIs Keywords of the quested MCP that are not included in the K set. When the P_i vectors are completed with as much zeros as the number of the QK_2 aces, the similarity among the quested MCP and the MCPs included in the KB can be estimated by equation 5.

4.   MCP insertion scenario

The above algorithm is implemented in order to build a Knowledge Base System from MCP ontology. To build this KBS Java and Jena API [9] were employed. The search process is used to identify if an inserted problem is new one or is already exist in the KB (figure 3).

Figure 3: The search process for an insertion MCP.

The procedure in details is the following:
1. The Knowledge Problem Author defines the quested MCP by giving its category and a short description.
2. The UMLS Connection Application automatically acquires, from the quested MCP short description, the concepts which included in UMLS-KS (all the stop words of the quested MCP short description such as "a" or "and" are removed). To achieve the connection with UMLS-KS, the UMLS – API [7] is utilized.
3. The Knowledge Problem Author selects which of the concepts above are MCP's keywords. Moreover, he/she has the ability to define additional keywords described in UMLS (The P_q Vector can be defined).
4. The MCP ontology provides all the instances of the classes necessary for the MCPs' vectors creation.
5. The similarity among the quested MCP and the MCPs available in the KB is calculated.
6. The Knowledge Problem Author decides whether one of the highest ranked MCPs is exactly the same with the quested problem or not.

The search procedure of MCPs can also be applied in finding algorithms that solve a particular MCP.

5.   Algorithm application

An example of estimating the similarity of a quested problem among twelve cardiologic MCPs is given in this section (Figure 4a).

a MCP	Keyword (CUI)	MCP	Keyword (CUI)
MCP1: ECG diagnosis of tricyclic antidepressant cardiotoxicity	Antidepressive Agents, Tricuclic(C003290) ECG(C0013738) Risk (C0035547) Cardiotoxicity(C0876994	MCP7: Predicting probability of acute myocardial infarction when left bunch branch block is present	MI(C0027051) Probability(C0033204) ECG(C0013738) AMI(C0155626) LBBB(C0023211)
MCP2: Identification left atrial enlargement	Left Artial Hypertrophy(C0238705) ECG(C0013738) Hypertrophy(C0020564) Enlargement procedure	MCP8: Predict the risk mortality after an acute myocardial infraction with thrombolytic therapy	Risk(C0035647) ECG(C0013738) Mortality(C0025565) Acute Myocardial Infraction(C0155626) MI(C0027051) Thrombolytic Agent
MCP3: Determination of QRS electrical axis, QRS complex in the frontal (coronal plane)	QRS axis(C0429012) ECG(C0013738) Coronal Plane(C0459350) Set of frontal planes(C1183148)	MCP9: Predict the risk mortality after an acute myocardial infraction	Risk(C0035647) ECG(C0013738) Mortality(C0025565) Acute Myocardial Infraction(C0155626) MI(C0027051)
MCP4: Identification of left ventricular hypertrophy	ECG(C0013738) Hypertrophy(C0020564) LVH(C0149721) Heart Ventricle(C0018827	MCP10: Identification of left ventricular hypertrophy when left bunch branch block is present	ECG(C0013738) Hypertrophy(C0020564) LVH(C0149721) Heart Ventricle(C0018827 LBBB(C0023211)
MCP5: Determination of QRS electrical axis, QRS complex in the horizontal plane	QRS axis(C0429012) ECG(C0013738) Horizontal(C0205162) Set of horizontal planes(C11831449)	MCP11: Identification of right ventricular hypertrophy	ECG(C0013738) Hypertrophy(C0020564) RVH(C0162770) Heart Ventricle(C0018827
MCP6: Predicting probability of acute myocardial infarction	MI(C0027051) Probability(C0033204) ECG(C0013738) AMI(C0155626)	MCP12: Identification ventricular tachycardia	Heart Ventricle(C0018827) Tachycardia(C0039231) ECG(C0013738) Tachycardia, Ventricular(C0042514)

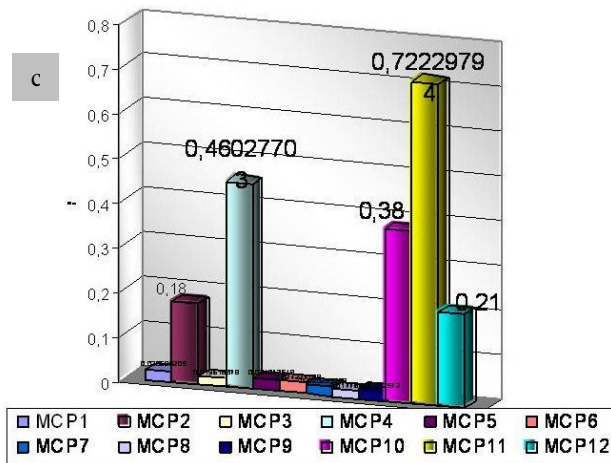| b	WS																							
MCP1	1	0,08	0,83	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MCP2	0	0,08	0	0	1	1	1	0,75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MCP3	0	0,08	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
MCP4	0	0,08	0	0	0	0	0,75	0	0	0	0,92	0,75	0	0	0	0	0	0	0	0	0	0	0	0
MCP5	0	0,08	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
MCP6	0	0,08	0	0	0	0	0	0	0	0	0	0,8	0,92	0,8	0	0	0	0	0	0	0	0	0	0
MCP7	0	0,08	0	0	0	0	0	0	0	0	0	0,8	0,92	0,8	0,92	0	0	0	0	0	0	0	0	0
MCP8	0	0,08	0,83	0	0	0	0	0	0	0,92	0,8	0	0,8	0	0,92	1	0	0	0	0	0	0	0	0
MCP9	0	0,08	0,83	0	0	0	0	0	0	0,92	0,8	0	0,8	0	0,92	0	0	0	0	0	0	0	0	0
MCP10	0	0,08	0	0	0	0	0,75	0	0	0,92	0,75	0	0	0	0,92	0	0	0	0	0	0	0	0	0
MCP11	0	0,08	0	0	0	0	0,75	0	0	0,75	0	0	0	1	0	0	1	0	0	0	0	0	0	0
MCP12	0	0,08	0	0	0	0	0	0	0	0,75	0	0	0	1	0	0	0	1	1	0	0	0	0	0
QuestedMCP	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	1

Figure 4:a) Matrix of twelve MCPs and their Keywords (CUIs) b) MCPs and Quested MCP Vectors c) Similarity among Quested MCP and MCPs.

The twelve above MCPs as well as their algorithms were originated by the Institute of Medical Algorithms. For more than eight years, the institute has collected and processed over 4,000 algorithms, including about 200 cardiology-specific algorithms that are available on the following URL www.medal.org [10].

The quested MCP is: "Identification of right ventricular hypertrophy when mitral stenosis is present", with Keywords: ECG, Hypertrophy, RVH, Heart Ventricle, mirtal stenosis, stenosis. The number of the Keywords is *m=24*, whereas the Keywords stenosis and mitral stenosis are not included in the set of Keywords K. Thus, in the 12 vectors of the cardiologic MCPs 2 extra zero elements are added, while the rest of the elements are calculated using the equations (2),(3),(4) (figure 3b). For example, the ws_8 (Keyword "Hypertrophy") is 0.75

$$\text{Freq}(K_8) = \sum_{i=1}^{12} f(i,8,1) = 4 \text{ and } ws_8 = \frac{12-(4-1)}{12} = \frac{9}{12} = 0.75$$

The most similar MCP is MCP11: "Identification of right ventricular hypertrophy", which has similarity 0.722 (figure 3c). The Knowledge Problem Author decides that the quested MCP is not included in the KB. So, it is inserted in the KB.

6. Conclusions

The research in question deals with the construction of ontologies that describe MCP using OWL. The aim of this work is to provide better MCP search capabilities, by introducing and exploiting cosine based similarity. This approach inherits the benefits of UMLS, which standardizes the medical terminology. The UMLS was employed as well as the MCP ontology for the search processing, during which the similarity among the MCPs and the quested MCP vectors was presented by an example. Moreover, the algorithm that constructed the above vectors was presented.

The final aim of this project is to build a Medical Computational Problem Solving Knowledge Portal. This Portal will constitute a Problem Solving Environment (PSE) which will provide all the computational facilities necessary to solve a target class of Medical Problems. Thus, Medical Scientists and Algorithm Scientists will be able to organize and perform MCP Solutions. This portal could be used in clinical research or more broadly in Medical research, by providing existing solutions of MCPs. Moreover, it could be utilized as a Problem Based Learning tool or as a collaborative learning tool, by providing the knowledge of the MCP.

References

[1] Davies J, Fensel D, Harmelen F, Towards the semantic web, Ontology-driven Knowledge Management , Wiley & Sons Ltd, (2003)

[2] Ch Bratsas, P Quaresma, G Pangalos, N Maglaveras, Using Ontologies to Build a Knowledge Base of Cardiology Problems and Algorithms, In Proc. of Computers inCardiology, IEEE Comp. Soc. Press, (2004), 31, 609-612

[3] Ora Lassila and Ralph R.Swick, Resource Description Framework (RDF) Model and Syntax Specification Feb 1999.

[4] World Wide Web Consortium. OWL Web Ontology Language Semantics and Abstract Syn-tax.W3C Recommendation 10 Feb, 2004.

[5] Protégé an ontology tool, http://protege.stanford.edu

[6] H. Knublauch, M. A. Musen, and A. L. Rector. Editing description logics ontologies with the Protégé OWL plugin. In International Workshop on Description Logics, Whistler, BC, Canada, 2004.

[7] Unified Medical Language System UMLS http://www.nlm.nih.gov/research/umls/

[8] BibTex Definition in Web Ontology Language (OWL) Version 0.1 (2004). hhtp://visus.mit.edu/bibtex/0.1/.

[9] Jena A Semantic Web Framework for Java. http://jena.sourceforge.net/

[10] Svirbely J, Sriram MG. MEDAL, a compendium of medical algorithms for access over the Internet. Proc AMIA Symp. 1999;p1172..

Address for correspondence.

Charalampos Bratsas
Aristotle University of Thessaloniki.
Lab of Medical Informatics, POB 323
Thessaloniki, 54124 Greece.
E-mail: mpampis@med.auth.gr