International Journal of Biomedical Science & Bioinformatics
Volume 3 : Issue 2 [ISSN 2475 - 2290]
Publication Date : 31 August, 2016

Semantic Representation of Radiotherapy data for effective data mining

Semantic Representation of Radiotherapy (subtitle)

[Geetha Mahadevaiah [a,*], Johan Van Soest [b], Dr. Andre Dekker [b], Dr. Narendranath Udupa [a], Dr. Shyam Vasudev Rao [c], Y.Kiran Kumar [a,], R.V.Prasad [a]]

Abstract— **Radiotherapy plays an important role in the treatment of cancer patients. As part of clinical workflow, patient has to undergo through diagnostic imaging procedures, which are used to identify the tumor location and size. Enormous amounts of data are generated during this procedure. The volume of medical information is so large and complex that it becomes difficult to mine for relevant information. The Digital Imaging and Communications in Medicine (DICOM) standard is widely used in medicine for storing and transmitting medical information. The DICOM-RT is the extension to DICOM standard, and dedicated to radiotherapy. In this paper, we propose a technique to store clinical relevant features from DICOM files using semantic concepts. The proposed technique defines a novel method to delayer the hierarchy of DICOM-RT for storing the clinical relevant information into triples in Resource Description Framework (RDF) repository. The methodology also proposes different combinations for storing data such as DICOM-RT with tumor information, DICOM-RT with pathology details. The proposed method uses the Semantic Web Technology to store and represent the information from DICOM-RT files along with into RDF graph and a data mining approach. Natural Language processing technique is used for the retrieval of data. We have evaluated our methodology qualitatively for 20 patients including combinations such as RTSTRUCT, tumor size data along with CT data, pathology information, by producing 25 varieties of different queries. We have analyzed quantitatively with accuracy of 90% for different hypothetical conditions using our proposed methodology.**

Keywords— DICOM-RT, Semantic Web, RDF, SPARQL, Natural Language Processing.

1.0 Introduction

Radiation therapy is one method of the cancer treatment, and plays an important role for patient during the course of the disease [1]. In this process, patients have to undergo diagnostic imaging procedures, which are performed to identify the tumor location and size. Data generated during this procedure contains large volume of information as well as complex structures, which makes it a challenging task for clinicians to query and retrieve relevant data [2]. Standards like those that Digital Imaging and Communications in Medicine (DICOM) [3] widely used in Radiology as a standard for diagnostic imaging. The DICOM standard has evolved over the years and extended to incorporate medical specialties such as radiotherapy, which led to the creation DICOM radiation therapy data (DICOM-RT) [4].

The DICOM-RT objects provides information about patient related structures identified from diagnostic data known as radiotherapy structure set (RTSTRUCT), contains radiotherapy treatment plan information (RTPLAN) and also provides total dose distributions from the planning system-dose information (RTDOSE) [5]. The DICOM-RT objects are stored in hierarchal manner; this restricts the search path while traversing the DICOM modalities and the DICOM query model itself and current DICOM tools do not support this required traversing well [6]. The literature shows planning for dose to be received by a certain region of interest in a radiotherapy treatment is to find the region of interest (ROI) name and contour points in the RTSTRUCT object. The coordinates and slices are defined in the CT objects, the treatment information stored in RTPLAN object and dose matrix in the RTDOSE objects, refer figure 1 [7]. Semantic Web technology provides access to meaningful public data and enables context based information interpretation of any data source. Graph based data representation enables dynamic modelling, and standardized ontologies allows-collaboration, sharing and reuse across applications. In the semantic world, data is modelled using the Resource Description Framework (RDF) [8], where resources and their relationships are stored in the form of a "triple" that is subject-object-predicate (SOP) [9,10].

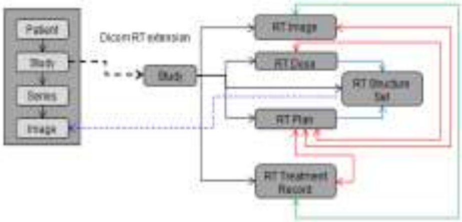

Figure 1: Illustration of DICOM-RT objects as an extension of the DICOM standard. These DICOM-RT objects correspond to Computed Tomography (CT) modality.

The study by Möller [11] describes semantic technique for annotating image with concepts using standard ontologies and stores the annotated data in RDF format such that images themselves becomes semantic rich. However their notion of context is tied with both anatomy and disease. The research work by Brunnbauer [12] focuses on representing the DICOM metadata by building ontology. They have developed a tool called as dicom2rdf for converting the DICOM metadata into RDF. This tool extracts the RDF metadata out of DICOM files and generates RDF files of large size, however data privacy and security has to be maintained during conversion. The researchers have studied [13] in storing and representing metadata of DICOM files in RDF repository. They have created an ontology called as SEDI: Semantic DICOM, to represent DICOM metadata

Authors Name/s per Affiliation:) : **Geetha Mahadevaiah**
Philips Research, Bangalore, India a,,,*Research Scholar Maastricht University, Maastricht, The Netherlands.
Johan Van Soest Dr. Andre Dekker , Maastricht University Medical Centre, Department of Radiation Oncology (MAASTRO).
Dr. Narendranath Udupa Dr. Shyam Vasudev Rao, Y.Kiran Kumar,, R.V.Prasad - Philips Research, Bangalore, India

elements and developed a proof of concept for storing the DICOM metadata in an RDF repository.

.

In this paper, the study describes a technique to leverage Semantic Web Technology to store the DICOM data for radiotherapy into RDF graphs and a method to mine the data by using the natural language based search. The retrieval of data from RDF graph can be achieved by querying using Simple Protocol and RDF Query Language (SPARQL) [14], which matches the pattern for a query in graph. In this paper, the solution to the traversing of objects is addressed by representing the objects of DICOM-RT in a generic and flexible format, in contrast to traditional relational databases, where the data is stored in rows and column. The data can also be stored as nodes that are connected to each other using 'semantic' links [15].

2.0 Material and Methods: In this section, the study propose a technique based on Semantic Web technology to model the DICOM-RT object and store it into RDF repository in the form of triples. This section also explains the process of information retrieval using SPARQL queries and translation of natural language query into SPARQL. In order to show the advantages of our proposed method, the study has hypothesized the problem into two independent cases. The proposed methodology is based on the semantic web technologies for representing the DICOM data in radiotherapy. The literature shows various DICOM-RT objects in well-defined data model. The proposed methodology focus on RT Structure Set & RT Dose objects. However, proposed system can be easily extended to other RT objects.

2.1 **Modeling of DICOM Metadata** Prerequisite for Modeling of DICOM Metadata: The DICOM-RT objects is a hierarchy data structures, which defines the directionality. The method to extract key information from each of the DICOM-RT object is crucial. In this hypothesis, the study assumes each DICOM-RT objects has "i" number of tags from which we select a set of "j" tags for our semantic modelling. The basic necessity to start semantic modelling for RT objects, is to understand the complexity of how data is stored, bring semantics into data, and make data accessible during search. Semantic modelling is similar to conceptual modeling approaches such as entity-relationship. The ontology is the core part of the Semantic system. The ontology provides the domain information in terms of concepts and the relationship between them as properties. RDF is to make statements about resources relationships. The figure 2 shows example of RDF graph creation.

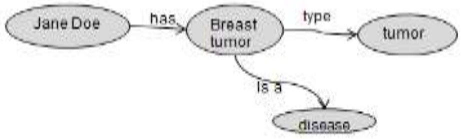

Figure 2: Shows a RDF graph for "Jane Doe has breast tumor".

Efforts of Semantic Web Community made it possible to formalize the knowledge in languages such as Web Ontology Language (OWL) [16]. In this study, "Semantic DICOM ontology" (SEDI) is used. SEDI is still under development, yet it has maintained the basic structure/concept of DICOM standard. The study have extended the current ontology for a few key RT objects.

2.1.2 Methodology:

The steps for delayer the hierarchal structure of DICOM-RT objects while storing the information as RDF triples is as follows:

The RTSTRUCT contains the information about the region of significance in radiation therapy such as contours, ROI's, tumor volumes (GTV, CTV, and PTV). Each structure set is linked with a frame of reference to the scanned images (usually CT). Similarly the RTDOSE contains the information related to the radiation dose specifically a 3D dose matrix again defined in the frame of reference of the scanned images and sometimes DVH (Dose Volume Histogram) sequences that include parameters such as the mean, maximum and minimum dose the ROI. Each DVH sequence structure is linked to the reference ROI. DVH is user defined module and provides the differential or cumulative dose volume histograms (DVHs).This proves various associated modules of the RT objects follows a hierarchy within the RT object sand with other RT objects. The complex DICOM-RT objects structure is converted into RDF triples for our analysis.

2.1.2.1 Hypothetical analysis - Graphical representation of DICOM-RT: The study proposes two hypothesis as problem statement into two independent cases. The detail of hypothesis are as follows:

2.1.2.2 Hypothesis I: Store and Represent the DICOM RT metadata into RDF by delayering the hierarchy in RTSTRUCT and RTDOSE files and retrieve the information via SPARQL query. Two patients (including RTSTRUCT and RTDOSE data along with CT data) is used for this hypothetical study. The hierarchal structure of RTSTRUCT and RTDOSE files is analyzed, and stored the attribute values in the form of triples. By creating a common node between RTSTRUCT and RTDOSE to other attributes such as ROI Name, DVH Minimum Dose, the hierarchy is reduced to only one level. The mapping of dose sequence and structure sequence to a reference node makes it easily accessible as this node will behave as a SPARQL endpoint. The study proposes a methodology to develop a prototype, where the semantic parser converts the DICOM data into triples and store it into RDF repository. These triples are exposed and easily accessible via SPARQL endpoint. Endpoint enables the users to query a graph by writing a SPARQL query, as shown in figure 3. The open Sesame framework [17] is used for querying and analyzing RDF data. This study has been evaluated qualitatively by producing 25 varieties of different queries and quantitatively as well. Evaluation results are shown in result section and discussed in detail.

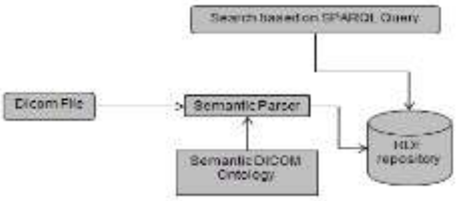

Figure 3: Setup for proposed method

2.1.2.3 Hypothesis II: Store and Represent the DICOM RT metadata into RDF from RTSTRUCT files and add the tumor size information and retrieve the information via SPARQL query. This hypothesis is analyzed using 20 patients (RTSTRUCT and tumor size data along with CT data is available). To test our hypothesis, this study has analyzed that RDF triples stores the statements about resources, which means, tumor size information is represented into triples as well. The literature shows tumor size which is one of the important factors for making any decision during treatment and usually stored in pathologically reports, can also be represented in triples. In this study, the actual tumor size information are read from a pathology database (excel file). This study links tumor size information as observation procedure finding. The procedure for set up of proof of concept remain the same for hypothesis I, but with a difference that, it includes the ROO ontology as well. This study has been evaluated by producing 10 varieties of different queries. Evaluation results are shown in result section and discussed in detail.

2.1.2.4 Search Linked RDF Data using SPARQL Query: Triples enable the users to query a RDF graph by writing a SPARQL query. It's an RDF query language, able to retrieve and manipulate data stored in RDF format. It allows the user to write the queries against data, by defining "key terms". In our semantic representation, data is linked between RTSTRUCT with RTDOSE objects to identify the dose values for specific structure. In this section, the method to form queries to retrieve the data from RDF files is described. The study is to determine the key terms and relations. The patient ID, Structure name, and Dose values were key terms which user wants to query over RDF data. The next step would be the term mapping; this can be done by finding all the resource nodes which are linked via specified predicate. The final SPARQL queries are the actual queries to be executed on the target endpoints. The query will match the specified pattern on graph and retrieve the data.

3. Results

The de-layering of the hierarchal structure of DICOM-RT objects and converting it into RDF triples has been successfully implemented. The resulting triples are stored in a RDF repository. The concepts and their relationships were maintained in the RDF graph as they were defined in the SEDI ontology. The proposed method helps to link data of RTSTRUCT and RTDOSE via a reference node, so while executing a query for patients with dose and their corresponding organs, patient lists with values related to those tags were displayed. When performing a query for patient with DVHs and their corresponding organ, all relevant doses and structures value were returned. These results were manually verified for- the correct answers to the questions posed. We have shown the results separately for both the hypothesis. For the first hypothesis we have evaluated 25 different queries where the information about doses and structures were retrieved. And for the second hypothesis we have evaluated 10 different queries where the information about structures and patient related tumor size were retrieved. In figure 4 and 5 we have shown the graph for number of query vs query results.

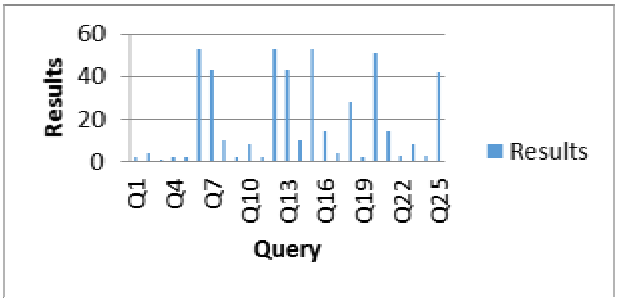

Figure 4: Number of query vs query results for hypothesis 1.

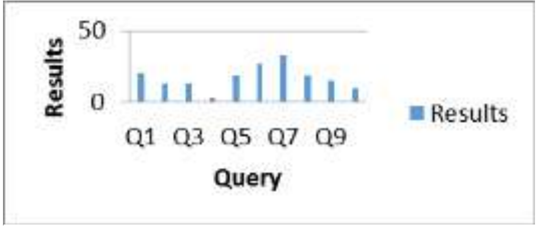

Figure 5: Number of query vs query results for hypothesis 2.

3.1 **Statistical Analysis:**

The algorithm is evaluated statistically using SPSS tool. The accuracy is calculated for different hypothetical conditions. The statistical evaluation is based on the number of queries, and query output which results in accuracy of 90% for given limited set of queries and covered a limited range of linguistic variability in natural language questions.

4. Discussion

We have successfully implemented a semantic parser to model the DICOM-RT object tags into RDF triples by reducing the hierarchy of DICOM structure and data mine by executing the SPARQL Query, thus confirming our hypothesis. The use of semantics in medical domain is not often explored by the researchers and the proposed method is unique and novel for sematic search of medical image and information. This study focused on a few key attributes for radiotherapy present in DICOM files and pathology reports which are relevant for clinicians. The advantage of proposed methodology for given study is ease of information retrieval compared to other techniques. The semantic representation of data in an RDF repository is more flexible than relational databases, which is helpful for clinicians. The proposed model is more generic model, where we integrate different ontologies. This research work can be extended to focus on modelling of DICOM-RT objects by integrating different ontologies so that data can be linked enabling seamless information retrieval by computers and the retrieval of data from RDF repository by writing natural language queries to automate the generation of SPARQL query for the desired search. Currently our study works for a very limited set of queries and have covered a limited range of linguistic variability in natural language questions. There is possibility to improve the engine to allow handling more complex query during search.

5. Conclusion

Efforts in informatics are focused on structural representation of medical data particularly in oncology environment, where radiotherapy plays a major role. We performed preliminary tests to evaluate the effectiveness of Semantic Web Technology to store the data for radiotherapy from DICOM files into RDF graph and retrieval of data back by using the natural language during search, which runs the SPARQL query to match the pattern on graph. To show the effectiveness, we have also implemented a system for semantic modelling of DICOM-RT objects and stored the resulting triples in RDF repository based on our hypothesis. Data is easily accessible via a SPARQL query, enabling applications in e.g. the tumor board process, information mining and research based on longitudinal studies of patient data. The storage of relevant patient data in RDF format enables longitudinal clinical studies. The tumor volume information like Gross tumor volume (GTV), Clinical target volume (CTV) from DICOM-RT files were successfully extracted, accessed, and analyzed using Sematic Web Technology. Such information is required to find the correct treatment plan for a patient. The future step would be to develop a system to seamlessly extract and store patient medical record, histopathology data, and radiotherapy data from the existing data repositories in hospitals and enable easy end-user data mining leveraging Semantic Web Technologies.

References

[1] Michael J et.al. The role of radiation therapy in the management of lung, prostate and colorectal cancer in South Dakota. South Dakota journal of medicine 2010; 60-66.

[2]Nuyts, Sandra. "Use of Imaging Data in Radiotherapy Planning of Head and Neck Cancer: Improved Tumour Characterization, Delineation and Treatment Verification." Head and Neck Cancer Imaging. Springer Berlin Heidelberg, 2006. 345-359.

[3] The National Electrical Manufacturers Association, Digital Imaging and Communications in Medicine (DICOM), NEMA Publications, PS3.1-PS3.12, 2011.

[4] Law, Maria YY, and Brent Liu. "DICOM-RT and Its Utilization in Radiation Therapy 1." RadioGraphics 29.3 (2009): 655-667.

[5] Maria et.al. DICOM-RT and Its Utilization in Radiation Therapy. 2009; 29(3):655-667.

[6] Maria et.al. DICOM-RT–based Electronic Patient Record Information System for Radiation Therapy 2009; 29(4):961-972.

[7] Data Mining DICOM RT objects for quality control in radiation oncology. Proc. SPIE 8319, Medical Imaging 2012: Advanced PACS-based Imaging Informatics and Therapeutic Applications, 83190Q, February 23, 2012.

[8] Brickley D, Guha R.V, McBride B, (2014, Feb. 25), W3C RDF Schema 1.1 [Online]. Available: http://www.w3.org/TR/2014/REC-rdf-schema-20140225/

[9] Patel-Schneider, P. and Fensel, D. layering the semantic web: Problems and directions. In First International Semantic Web Conference (ISWC2002), Sardinia, Italy, 2002; 16–29.

[10] Van Soest, Johan, Tim Lustberg, Detlef Grittner, M. Scott Marshall, Lucas Persoon, Bas Nijsten, Peter Feltens, and Andre Dekker. "Towards a semantic PACS: Using Semantic Web technology to represent imaging data." Studies in health technology and informatics 205, pp. 166-170, 2013.

[11] Möller, Manuel, and Saikat Mukherjee. "Context-Driven Ontological Annotations in DICOM Images-Towards Semantic Pacs." In HEALTHINF, pp. 294-299.

[12] Brunnbauer, Michael. "DICOM metadata as RDF." In GI-Jahrestagung, pp. 1796-1804. 2013.

[13] Sauermann, L., Cyganiak, R., & Völkel. Cool URIs for the semantic web. 2011.

[14] http://www.w3.org/TR/sparql11-overview/

[15] Semantic flooding: Search over semantic links. Dept. of Inf. Eng. & Comput. Sci., Univ. of Trento, Trento, Italy, IEEE 26th International Conference on Conference: Data Engineering Workshops (ICDEW), 2010.

[16] McGuinness, D. L. and van Harmelen, F. (2004). OWL Web Ontology Language overview. W3C recommendation,WorldWideWeb Consortium.

[17] Sesame, Available:http://rdf4j.org/