iSMART : intelligent Semantic MedicAl Record reTrival

Yuan Ni Guotong Xie Shengping Liu Hanyu Li Jing Mei Gang Hu Haifeng Liu Xueqiao Hou
IBM China Research Lab, China
{niyuan, xieguot, lihanyu, liusp, meijing, hugang, liuhf, houxueq}@cn.ibm.com

1. ABSTRACT

We present iSMART, a system for *i*ntelligent *S*emantic *M*edic*A*l *R*ecord re*T*rival. Health Level 7 Clinical Document Architecture (CDA) [4], a standard based on XML, is well recognized for the representation and exchange of medical records. In CDAs, medical ontologies/terminologies, e.g. SNOMED CT [2], are used to specify the semantic meaning of clinical statements. To better use the structure and semantic information in CDAs for a more effective search, we propose and implement the iSMART system. Firstly, we design and implement an XML-to-RDF convertor to extract RDF statements from medical records using declarative mapping. Then, we design a reasoner to infer additional information by integrating the knowledge from the domain ontologies based on the extracted RDF statements. Finally, we index the inferred set of RDF statements and provide the semantic search on them. A demonstration video is available online [1].

2. INTRODUCTION

Currently, it is widely believed that the broad adoption of the electronic health record (EHR) will improve the healthcare quality and reduce the healthcare cost. Many countries advocate the widespread of EHRs in healthcare industry. EHR contains a large amount of information that is interested in by various kinds of people : (1) patients would like to know their health status which could be found in EHR; (2) doctors would like to retrieve some medical literature and current best practices; (3) researchers would like to access the clinical data for research that can accelerate the level of knowledge of effective medical practices. As a large number of medical records are available electronically, the search engine is indispensable to help users find their required information effectively and efficiently.

Health Level 7 Clinical Document Architecture (CDA) is the standard for information exchange in healthcare. Traditional keyword search engine could address only the text content of CDAs, while the structure information and ontology reference in CDAs are not leveraged. To better use the structure and semantic information in CDAs for a more effective search, we propose and implement the iSMART system. Firstly, we design a convertor to extract RDF statements from CDAs which are in XML format. The convertor is based on the declarative mapping which enables easy customization for RDF extraction. Secondly, as fragments of CDA documents are associated with the ontological concepts that are defined in some ontology such as SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) [2], which is considered to be the most comprehensive, multilingual clinical healthcare terminology, we design an EL$^+$ reasoner to infer additional information by combining the background knowledge from the ontology. Finally, we leverage the Semplore [6] to provide an efficient hybrid querying on the inferred set of RDF statements. The hybrid query combines both the keyword search and structure navigation.

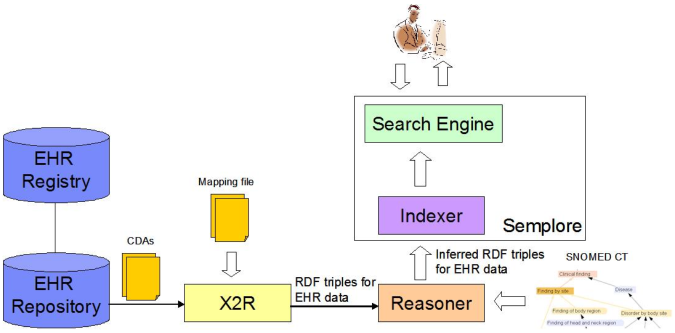

Figure 1: iSMART System Architecture

3. SYSTEM ARCHITECTURE AND IMPLEMENTATION

As illustrated in Fig 1., the iSMART consists of three components, i.e. X2R, reasoner and Semplore, which correspond to three steps to process CDA documents to enable semantic search.

X2R takes charge of extracting the RDF triples from the CDAs that are stored in some EHR repository in XML format. To the best of our knowledge, GRDDL (Gleaning Resource Descriptions from Dialects of Languages) [3] is the only existing solution, which basically constructs the RDF triples by concatenating texts in XSLT scripts. However, it ignores the semantics provided in the ontologies for the RDF triples. This leads to the procedure of writing mapping scripts between XML and RDF labor-intensive, error-prone and maintenance-difficult.

In X2R, the transformation is determined by a declarative mapping which could be easily customized by users to satisfy their extraction requirements. Given the structure of CDAs and the ontology of the output RDF data, the mapping file defines the three kinds of transformation rules : (1) the mapping from some XML entity to the instance of an RDF class; (2) the property to connect the instances or the instances and their values; (3) the URI pattern of instances. Given a valid mapping and XML data, X2R engine will search the XML data to generate RDF resources based on the class locations in the mapping, and assign them URIs according to the given naming method. After that, each property in the mapping is processed to either attach the corresponding values to the existing resources, or connect the existing resources together based on the values of object properties.

Reasoner leverages the domain ontology, i.e. SNOMED CT, to enrich the set of extracted RDF triples, and generates a set of inferred RDF triples. As the SNOMED CT ontology has a formal expressiveness of the Description Logic language EL$^+$, we employ an EL$^+$ reasoner to generate inferred RDF documents. Considering that SNOMED CT includes more than 300, 000 concepts and millions of triples could be generated from CDA documents, we build our EL$^+$ Reasoner on top of a relational database system for scalability issues. Firstly, we store the SNOMED CT ontology and the extracted RDF triples into a relational database. The SNOMED CT is normalized into a normal form in terms of the normalization rule for the EL$^+$ ontology. Each type of EL$^+$ normalized axioms corresponds to one table in the database. The RDF triples are stored into two tables where TYPEOF(ind, concept) is designed for RDF membership triples and RELATIONSHIP(ind, role, ind') stores all relationship triples. Secondly, we define Datalog rules for each EL$^+$ normalized axiom and then evaluate these rules iteratively using a bottom-up strategy until no new inferred triples are generated. Please refer to [5] for detailed information of the reasoner.

However, the expressiveness of the existential axiom $A \subseteq \exists R.B$ from EL$^+$ is beyond Datalog rules. The underlining meaning of $A \subseteq \exists R.B$ is $A(x) \rightarrow \exists y, s.t.R(x, y)$ *and* $B(y)$, while the Datalog rule is not powerful enough to do the existential individual generation. Even if we extend Datalog with a generation function, an infinite sequence of individuals might be generated. To solve this problem, we proposed the idea of *canonical individual*. For each existential axiom $A \subseteq \exists R.B$, we upfront generated one canonical individual ind' with respect to the role R and the class B, and then stored all canonical individuals in a table as CANONIND(ind', R, B). In this way, we are allowed to define a Datalog rule that if an individual *ind* is typed of a concept *sub* where $sub \subseteq \exists role.sup$, and there is a canonical individual ind' with respect to *role* and *sup* in the table CANONIND, then triples of RELATIONSHIP(ind, role, ind') and TYPEOF(ind', sup) are inferred out.

Semplore. We use an IR based engine, i.e. Semplore [6], to index the RDF triples and a hybrid query that integrates keyword search and structure navigation is supported. Semplore leverages the inverted lists to store the keywords, the type information of resources and the structure information of triples. A facet search interface is provided where users could start from the keyword searching followed by the relationship or type constraints navigation.

4. EVALUATION

Our iSMART system is evaluated by an experimental study where 100, 900, 9000 clinical documents are uploaded into iSMART. These documents are collected from a large hospital with protected privacy information. The upper table in Table 1 lists the statistic information about the number of triples extracted from CDA documents and the number of triples generated by the EL+ reasoner. It is observed that the number of triples increases five folds after the inference. The additional inferred triples provide more comprehensive search results. The bottom table in Table 1 shows the running time for each step. This is to measure the time taken to make a batch of documents ready for semantic searching. It is observable that the time increases linearly with the number of triples.

#document	#triples from CDAs	#triples from reasoner
100	56,451	289,342
900	426,251	2,137,702
9000	4,550,055	23,129,897

#document	X2R	Reasoning	Indexing
100	105	100	278
900	749	628	2228
9000	8001	4550	25800

Table 1: Statistic Information and Results

5. DEMONSTRATION SCENARIO

We plan to demonstrate the following two aspects of the iSMART system : (1) the procedure to create indexes on CDAs for semantic retrieval; (2) the effectiveness of the reasoning using domain ontology to improve the recall of the search results.

The IBM IHE XDS is used as the EHR repository. The above real documents are used to demonstrate the system. Users could select a set of documents to upload into the system, and perform the *extract*, *infer* and *index* steps to add these documents for semantic search. We could build indexes on both the original set of triples from X2R and the inferred set of triples from the reasoner. Users could search on both dataset to understand the usefulness of the inference. Let us consider a sample query "*Find documents about hearing normal and the finding site is entire left ear*". On the dataset with 9000 documents, no results are returned from triples without inference while four documents are returned with inference. The reason is that these four documents have the information about hearing normal with finding site of entire ear. As entire left ear is subclass of entire ear, we know that these four documents should be the answer.

6. REFERENCES

[1] iSMART. http://sites.google.com/site/crlsemantics/files/ismart.swf.

[2] College of american pathologists. systematized nomenclature of medicine-clinical terms (SNOMED CT). http://www.ihtsdo.org/snomed-ct/, 2007.

[3] W3c recommendation. gleaning resource descriptions from dialects of languages (GRDDL). http://www.w3.org/2004/01/rdxh/spec, 2007.

[4] R. H. Dolin, L. Alschuler, S. Boyer, C. Beebe, F. M. Behlen, P. V. Biron, and A. Shabo Shvo. HL7 Clinical Document Architecture, Release 2. *J Am Med Inform Assoc*, 13(1):30–39, 2006.

[5] J. Mei, S. Liu, G. Xie, A. Kalyanpur, A. Fokoue, Y. Ni, H. Li, and Y. Pan. A practical approach for scalable conjunctive query answering on acyclic EL+ knowledge base. *To Appear in ISWC*, 2009.

[6] L. Zhang, Q. L. Liu, J. Zhang, H. F. Wang, Y. Pan, and Y. Yu. Semplore : an IR approach to scalable hybrid query of semantic web data. In *ISWC*, 2007.