Ontology-Aware Search on XML-based Electronic Medical Records

Fernando Farfán[*1], Vagelis Hristidis[*2], Anand Ranganathan[†3], Redmond P. Burke[‡4]

[*] *School of Computing and Information Sciences, Florida International University*
Miami, Florida, United States of America
[1]ffarfan@cis.fiu.edu
[2]vagelis@cis.fiu.edu

[†]*IBM T.J. Watson Research*
Yorktown Heights, New York, United States of America
[3]arangana@us.ibm.com

[‡]*Miami Children's Hospital*
Miami, Florida, United States of America
[4]redmond111@aol.com

***Abstract*— As the use of Electronic Medical Records (EMRs) becomes more widespread, so does the need for effective information discovery on them. Recently proposed EMR standards are XML-based, having as a key characteristic the frequent use of ontological references, i.e., ontological concept codes appear as XML elements and are used to associate portions of the EMR document with concepts defined in a domain ontology. In this paper we present the XOntoRank system which tackles the problem of ontology-aware keyword search on XML documents with a particular focus on EMR XML documents. Our running examples and experiments use the Health Level Seven (HL7) Clinical Document Architecture (CDA) Release 2.0 standard of EMR representation and the Systematized Nomenclature of Human and Veterinary Medicine (SNOMED) ontology, although the presented techniques and results are applicable to any EMR hierarchical format and any ontology that defines concepts and relationships.**

I. Overview

In this paper we present the XOntoRank system, which tackles the problem of facilitating ontology-aware information discovery on a corpus of XML-based EMR documents, i.e., given a question (query) and a set of EMRs, find the entities (typically subtrees) that are "good" for the query, and rank them according to their "goodness" with respect to the query. The success of Web search engines has shown that keyword queries are a useful and intuitive information discovery approach. Therefore, we focus on keyword queries in this paper.

There is a large corpus of work on keyword search on XML documents [3], [4], [6], where the query keywords are matched to XML nodes and a minimal tree containing these nodes is returned. A variety of ranking techniques are used ranging from the size of the result-trees to Information Retrieval (IR) scoring adaptations. There is also work on exploiting ontologies in XML querying [10], [9], [7]; however they are not appropriate for plain keyword queries since they are based on query expansion which would lead to non-minimal results.

The use of ontological definitions allows us to perform semantic search on the XML documents. We no longer

```
1 <? xml version="1.0" ?>
2 <ClinicalDocument>
3   <recordTarget>
4    <patientRole>
5     <id extension="12345"
       root="2.16.840.1.113883.3.933"/>
6      <patientPatient>
7       <name>
8        <given>SampleName</given>
9        <family>SamplePatient</family>
10      </name>
11      <genderCode code="M" codeSystem="2.16.840.1.5.1"/>
12      <birthTime value="20020924"/>
13     </patientPatient>
14  </patientRole></recordTarget>
15  <component>
16   <StructuredBody>
17    <component>
18     <section>
19      <code code="10160-0"
         codeSystem="2.16.840.1.113883.6.1"
         codeSystemName="LOINC"/>
20      <title>Medications</title>
21      <entry>
22       <Observation>
23        <code code="84100007"
           codeSystem="2.16.840.1.113883.6.96"
           codeSystemName="SNOMED CT"
           displayName="Medications "/>
24        <value xsi:type="CD" code="195967001"
           codeSystem="2.16.840.1.113883.6.96"
           codeSystemName="SNOMED CT"
           displayName="Asthma">
25      </value></Observation></entry>
26      <entry>
27       <SubstanceAdministration>
28        <consumable>
29         <manufacturedProduct>
30          <manufacturedLabeledDrug>
31           <code code="66493003"
             codeSystem="2.16.840.1.113883.6.96"
             codeSystemName="SNOMED CT"
             displayName="Theophylline"/>
32          </manufacturedLabeledDrug>
33         </manufacturedProduct>
34        </consumable>
35       </SubstanceAdministration>
36      </entry>
37 </section></component></StructuredBody></component>
38</ClinicalDocument>
```

Fig. 1 Sample CDA Document

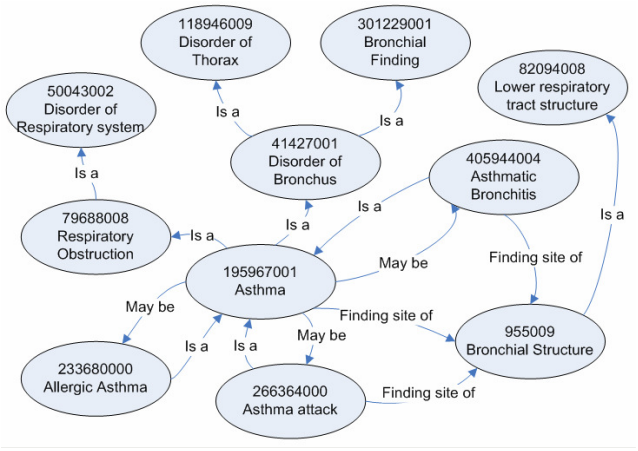

Fig. 2 Subgraph of SNOMED Ontology.

require an exact match between keywords in the query and in the document, but we can make use of the domain ontology to infer a semantic relationship between keywords in the query and terms in the document. This allows returning more results than would otherwise be returned with an exact-match requirement.

Problem Definition: We view an *XML document* as a labeled tree T_i in which each node v_i has a textual description, concatenating its tag name, attribute names and values, and text content, and an optional ontological reference typically consisting of an integer code for the ontological system and an integer code for the specific concept. We view the *ontology* as a graph where the nodes x_i represent concepts, and the edges represent relationships between concepts.

A *keyword query* is a set of keywords w_i. Previous works, which ignore ontological references, have generally defined the results as subtrees of the XML documents that contain all query keywords. We adopt the result semantics of XRANK [6] and extend it to account for ontological references; instead of requiring keywords to be contained in the nodes of the result subtree, we require that the result subtree has nodes associated with every query keyword. We measure the association degree of a node v with respect to a keyword w which is directly contained in v or is associated to v through an ontology. To encompass this modification we generate XOnto-DIL lists, similar to the DIL lists generated by XRANK, but instead of storing ElemRank we compute *NodeScore NS(v,w)*, that is, the relevance score of node v with respect to keyword w given the XML documents and the ontological systems.

Example: Let's consider a query *"Bronchial Structure Theophylline"* on a CDA document (such as the one in Fig. 1). The phrase *Bronchial Structure* does not appear anywhere in this document. Hence, most traditional XML-based keyword search systems will not return any results. However, this document contains an ontological reference to an *"Asthma"* concept defined in SNOMED (in Line 24, Fig. 1). The SNOMED ontology further defines a *"finding-site-of"* relationship between *"Asthma"* and *"Bronchial Structure"* (as shown in Fig. 2). Hence, based on the definitions in the ontology, a result tree connecting the *"Asthma"* node of Line 24 and the *"Theophylline"* node of Line 31 can be output.

II. Algorithms

A key component of XOntoRank is the derivation of semantic relevance of a concept v in the ontology to a query keyword w. Since nodes in an XML document may refer to concepts in the ontology, this derivation essentially quantifies the semantic relevance of an XML node to a query keyword based on terminological definitions in the ontology. Our approach for calculating the semantic relevance of a concept to a query keyword is inspired by the idea of authority flow (e.g. as used in ObjectRank [2]). Initially, each concept in the ontology is granted a certain authority based on how strongly it is related to w, as measured by its IR score. Authority then flows from these concepts to other concepts in the ontology based on the following rules:

Ontology as Undirected Graph (Graph): This strategy treats the ontology as an undirected graph, with no distinction among the different kinds of relationships between concepts.

Ontology as Taxonomy (Taxonomy): This strategy only considers the taxonomic portion of the ontology, i.e. we only consider *is-a* links between concepts. The *is-a* links form a Directed Acyclic Graph *(DAG)*, since cycles are not permitted based on subclass relationships. It follows two cases:

(i) Concept x is a superclass of concept v: Since x is a superclass of v, any query for x is completely and logically satisfied by v.

(ii) Concept x is a subclass of concept v: Since x is a subclass of v, any query for x is partially satisfied by v. Our heuristic for calculating the extent of the partial satisfaction is based on the number of subclasses of v, similarly to the authority flow distribution in [2].

Ontology as Collection of Relationships (Relationships): In order to handle different kinds of relationships, we interpret concepts and relationships in SNOMED using description logics [1], allowing us to describe every concept as a subclass of a set of atomic concepts or existential role restrictions. Hence, we can reduce a graph with different kinds of relationships into one that has only subclass or (*is-a*) relationships.

III. Results

A. Impact of Leveraging Ontology

To evaluate the impact of the four proposed approaches of incorporating ontological knowledge, we performed a series of two-keyword queries and compared their top-k result lists. The four approaches are denoted as *XRANK* (baseline, no use of ontology), *Graph*, *Taxonomy* and *Relationships*. We executed a series of keyword queries obtained from domain expert collaborators.

TABLE I

Normalized Kendall Tau Values for four Approaches.

	XRANK	*Graph*	*Taxonomy*	*Relationships*
XRANK	0.000	0.362	0.577	0.600
Graph	0.362	0.000	0.401	0.748
Taxonomy	0.577	0.401	0.000	0.193
Relationships	0.600	0.748	0.193	0.000

We use the top-k Kendall Tau [5] to determine the distance between the lists. Table 1 reports the average normalized Kendall Tau values for *k=10*. We observe the largest distance between the result of *XRANK* and the *Relationships* algorithm. As expected, the shortest distance between the lists occurs between the *Taxonomy* and *Relationships* lists, since the latter uses the first one as its base set.

B. Performance Results

Preprocessing Phase: Building the XOnto-DIL lists for all keywords in the SNOMED ontology was not practical given they are in the order of millions and they cannot be extracted from the provided SNOMED API. Hence, we built XOnto-DIL lists for all the keywords in the CDA documents and for all keywords contained in a concept up to 2 relationships away from a concept referenced in a CDA document. Fig. 3 presents the times to build a keyword's Onto-DIL list for the two involved stages: (*OntoScore Computation* and *Combination*) for a set of keywords. We measured separately the time spent to navigate the ontology graph, reported as *SNOMED Navigations*.

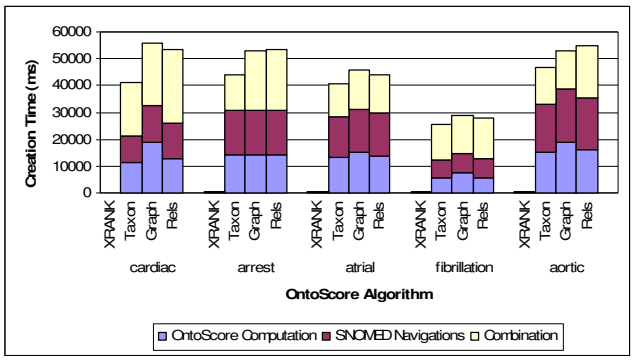

Fig. 3 Xonto-DIL Lists Creation Time.

Query Phase: Fig. 4 presents the average execution times of keyword queries for varying the number of keywords, for *k=10*. Note that the time for the Relationship-based approach is higher due to the larger number of nodes in the XML document that are ontologically related to the query keywords.

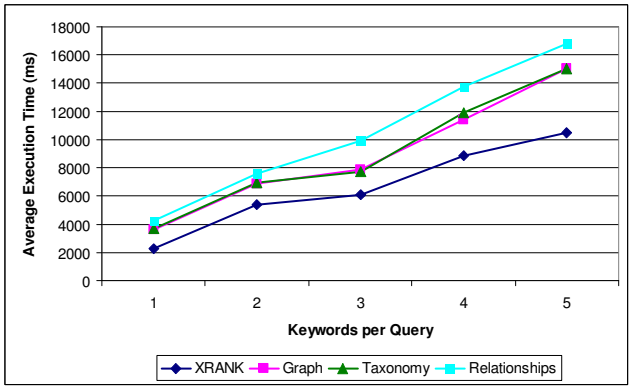

Fig. 4 Average Execution Time for Keyword Queries of different lengths.

However, none of the approaches suffers significant overhead compared to XRANK, with the largest overhead being about 60% for 5 keywords.

IV. Conclusions

We have introduced the problem of ontology-aware keyword search on XML-based EMR documents, which contain references to clinical ontological concepts. We defined semantics for this problem, where the ontological references, as well as the relationships within the ontology are used in creating and ranking the query results. Alternative views of the ontology were considered. We created efficient algorithms, building on previous work, to generate the top-k query results. The algorithms were evaluated experimentally.

A critical future direction is the optimization of the index creation process. Our current index creation approach relies on the API and data provided by [8], which are based on flat files. Implementing approximation and early pruning techniques may prove useful in scaling to larger ontologies and datasets.

Acknowledgment

This project was supported in part by the National Science Foundation Grant IIS-0534530.

References

[1] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, Eds. 2003 *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press.

[2] A. Balmin, V. Hristidis, and Y. Papakonstantinou. ObjectRank: authority-based keyword search in databases. In Proceedings of the International Conference on Very Large Databases (VLDB), August 2004.

[3] S. Cohen, J. Mamou, Y. Kanza and Y. Sagiv. XSEarch: A semantic search engine for XML. In VLDB, 2003.

[4] N. Fuhr and K. Großjohann. XIRQL: a query language for information retrieval in XML documents. In Proc. 24th International ACM SIGIR Conference on Research and Development in Information Retreival, pages 172–180, New Orleans (Louisiana, USA), Sept. 2001. ACM Press.

[5] R Fagin, R Kumar and D Sivakumar. Comparing Top *k* Lists. In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 2003.

[6] L. Guo, F. Shao, C. Botev, and J. Shanmugasundaram. XRANK: Ranked keyword search over XML documents. In Proc. ACM SIGMOD International Conference on Management of Data, San Diego (California), June 2003.

[7] M. S. Kim, Y. H. Kong, and C. W. Jeon. Remote-Specific XML Query Mobile Agents. In DEECS 2006, June 2006.

[8] SNOMED Clinical Terms (SNOMED CT). http://www.snomed.org/snomedct/index.html. 2007.

[9] R. Schenkel, A. Theobald, and G. Weikum. Semantic similarity search on semistructured data with the XXL search engine. Information Retrieval, 8(4):521–545, December 2005.

[10] A. Theobald. An Ontology for Domain-oriented Semantic Similarity Search on XML Data. Datenbanksysteme für Business, Technologie und Web (BTW) (2003) 217-226.