Information Discovery on Electronic Medical Records[1]

Vagelis Hristidis* Fernando Farfán* Redmond P. Burke+ Anthony F. Rossi+ Jeffrey A. White†
*School of Computing and Information Sciences, Florida International University
+Miami Children's Hospital
†Teges Corporation
{vagelis,ffarfan}@cis.fiu.edu, redmond111@aol.com, anthony.rossi@mch.com, jwhite@teges.com

Abstract

As the use of Electronic Medical Records (EMRs) becomes more widespread, so does the need to search and provide effective information discovery on them. Information discovery methods will allow practitioners and other healthcare stakeholders to locate relevant pieces of information in the growing corpus of available EMRs. The success of Web search engines has shown that keyword queries are a useful tool for locating relevant information in an intuitive and effective manner. However, questions arise of the form: What are the semantics of keyword queries on EMRs? What is a meaningful result? What is the role of medical and clinical ontologies and dictionaries like SNOMED (Systematized Nomenclature of Human and Veterinary Medicine) in answering such queries?

In this position paper we introduce the problem of keyword-based information discovery on EMRs and enumerate the salient challenges that must be addressed to facilitate quality information discovery. The objective is to create interest in new medical information management research initiatives, and potentially create new paradigms for using medical data. The primary focus of the paper is the newest XML-based EMR standard created by the Health Level Seven (HL7) group, the Clinical Document Architecture (CDA) Release 2.0, although the same issues arise for any other standard hierarchical format.

1. Introduction

The National Health Information Network (NHIN) and its data-sharing building blocks, RHIOs (Regional Health Information Organizations), are encouraging the widespread adoption of *Electronic Medical Records (EMR)* for all hospitals within the next five years. To date, there has been little or no effort to define methods or approaches to rapidly search such documents and return meaningful results. One of the most promising standards for EMR manipulation and exchange is the XML-based Health Level 7's [19] Clinical Document Architecture (CDA) [8].

The definition and adoption of this standard presents new challenges to related computer science disciplines like data management, data mining and information retrieval. In this position paper we study the problem of facilitating information discovery on a corpus of CDA documents, i.e., given a question (query) and a set of CDA 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 mainly focus on keyword queries in this paper, although some issues going beyond plain keyword queries are also examined.

As an example, consider the usual scenario where a doctor wants to check possible conflicts between two drugs. Keyword query "drug-A drug-B death" could be submitted to discover cases where a patient who took both drugs died. Note that the word "death" can be specified in many different elements of a CDA document, and also synonyms or related terms like "mortality" can be used instead. The latter can be tackled by leveraging appropriate medical ontologies like SNOMED Clinical Terminology (SNOMED CT) [27] as discussed below.

The key ranking criteria found in current systems as well as the bibliography [26, 7, 14] are (a) relevance, (b) quality (authority) and (c) specificity. It is challenging to define the information discovery semantics for CDA documents such that the three aforementioned key ranking criteria are considered, given the hierarchical structure and specific semantics of CDA, and the common references to outside entities like dictionaries, ontologies, separate text, or multimedia patient data. Medical dictionaries and ontologies typically used in CDA are SNOMED CT [27] and LOINC [22]. We also study how previous work on information discovery on XML data (Section 2.2) can be leveraged, and what limitations might exist in this unique domain. We note that our study does not discuss the important privacy issues involved in accessing patient information, as required by HIPAA [18].

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

The extended version [15] of this work describes more challenges and discusses more related work.

The rest of this paper is organized as follows: Section 2 presents a background exposition of current clinical information standards and a brief survey on information discovery on XML data. Section 3 addresses the challenges that we have identified to execute information discovery on a corpus of EMR documents. Our concluding remarks are presented in Section 4.

2. Background

In this section we review key standards used to represent clinical data and EMRs and present previous work on information discovery on general XML documents.

2.1. Clinical Information Model and Ontologies

Reference Information Model (RIM): HL7 is a language, and every language has a grammar. The HL7 RIM [25] specifies the grammar of HL7 messages and the basic building blocks of the language and their permitted relationships. For more details see [15].

Systematized Nomenclature of Medicine (SNOMED): SNOMED [27] has grown up into a comprehensive set of over 150,000 records in twelve different chapters or axes. *SNOMED Clinical Terms (SNOMED CT)* is a universal health care terminology and infrastructure. Figure 1 shows a sub-graph of the SNOMED ontology graph. For more details see [15].

Clinical Document Architecture: The Clinical Document Architecture (CDA) is an XML-based document markup standard that specifies the structure and semantics of clinical documents, such as discharge summaries and progress notes, for the purpose of exchange. It is an American National Standards (ANSI) approved HL7 standard, intended to become the de facto electronic medical record. Figure 2 depicts a sample CDA document D_1, which is wrapped by the "ClinicalDocument" element, as it appears in Line 2 of this figure. For more details see [15].

2.2. Searching XML Documents

In this section we present an overview of previous work on searching XML documents. This corpus of work will be viewed as the starting point to present the challenges of information discovery on CDA XML documents in Section 3. XRANK [14] ranks the XML elements by generalizing the Page-Rank algorithm [6], combining the ranking of elements with keyword proximity. XSEarch [10] ranks the results taking into consideration both the degrees of the semantic relationship and the relevance of the keyword. Cohen et al. [9] present an extended framework to specify the semantic relationship of XML elements. XIRQL [13] utilizes a different strategy to compute its ranking, defining index units, specific entity types that can be indexed and used for tf-idf computation. For more details see [15].

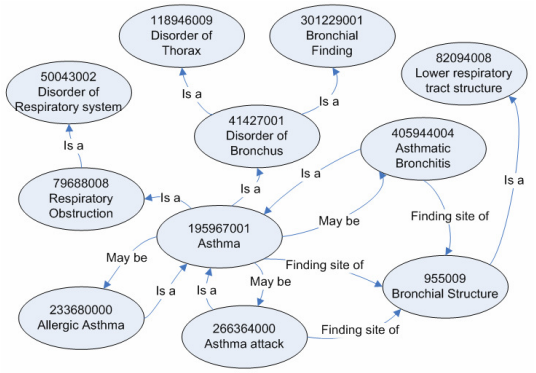

Figure 1: Partial SNOMED ontology for the term "Asthma".

3. Challenges of Information Discovery on CDA Documents

In this section we present a series of challenges that have to be addressed to effectively perform information discovery on a corpus of CDA documents. For simplicity we focus on plain keyword queries, although the same challenges are valid for semi-structured queries as well a semi-structured query is a query where partial information about the structure of the results is provided. Detailed discussion and examples for each of the challenges are presented in the extended version [15].

We discuss why the general work on searching on XML documents (Section 2.2) is not adequate to provide quality information discovery on CDA XML documents. The key reasons are the complex and domain-specific semantics and the frequent references to external information sources like dictionaries and ontologies. We use Document D_1 depicted in Figure 2 as our running example.

3.1. Structure and Scope of Results

In contrast to traditional Web search where whole HTML documents are returned as query results, in the case of XML documents and particularly CDA documents, we need to define what a meaningful query result is. Previous work has studied different approaches to define the structure of results. A corpus of works [1, 13, 14] consider a whole subtree as result, that is, a result is unambiguously defined by the lowest common ancestor (LCA) node of the keyword nodes. We refer to this approach as *subtree-*

```
1  <? xml version="1.0" ?>
2  <ClinicalDocument xmlns="urn:hl7-org:v3" xmlns:voc="urn:hl7-org:v3/voc"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
     xsi:schemaLocation="urn:hl7-org:v3 CDA.ReleaseTwo.CommitteeBallot03.Aug.2004.xsd"
     templateId="2.16.840.1.113883.3.27.1776">
3   <id extension="c266" root="2.16.840.1.113883.3.933"/>
4   <confidentialityCode code="N" codeSystem="2.16.840.1.113883.5.25"/>
5   <author>
6    <time value="20000407"/>
7    <assignedAuthor>
8     <id extension="KP00017" root="2.16.840.1.113883.3.933"/>
9     <assignedPerson>
10     <name>
11      <given>Robert</given>
12      <family>Dolin</family>
13      <suffix>MD</suffix>
14  </name></assignedPerson></assignedAuthor></author>
15  <recordTarget>
16   <patientRole>
17    <id extension="12345" root="2.16.840.1.113883.3.933"/>
18    <patientPatient>
19     <name>
20      <given>Henry</given>
21      <family>Levin</family>
22      <suffix>the 7th</suffix>
23     </name>
24     <administrativeGenderCode code="M" codeSystem="2.16.840.1.113883.5.1"/>
25     <birthTime value="19320924"/>
26    </patientPatient>
27    <providerOrganization>
28     <id extension="M345" root="2.16.840.1.113883.3.933"/>
29  </providerOrganization></patientRole></recordTarget>
30  <component>
31   <StructuredBody>
32    <component>
33     <section>
34      <code code="10160-0" codeSystem="2.16.840.1.113883.6.1" codeSystemName="LOINC"/>
35      <title>Medications</title>
36      <entry>
37        <Observation>
38          <code code="84100007" codeSystem="2.16.840.1.113883.6.96"
              codeSystemName="SNOMED CT" displayName="history taking (medication)"/>
39          <value xsi:type="CD" code="195967001" codeSystem="2.16.840.1.113883.6.96"
              codeSystemName="SNOMED CT" displayName="Asthma">
40           <originalText><reference value="m1"/></originalText>
41      </value></Observation></entry>
42      <entry>
43        <Observation>
44          <code code="84100007" codeSystem="2.16.840.1.113883.6.96"
              codeSystemName="SNOMED CT" displayName="history taking (medication)"/>
45          <value xsi:type="CD" code="32398004" codeSystem="2.16.840.1.113883.6.96"
              codeSystemName="SNOMED CT" displayName="Bronchitis">
46           <value xsi:type="CD" code="91143003" codeSystem="2.16.840.1.113883.6.96"
              codeSystemName="SNOMED CT" displayName="Albuterol" />
47      </value></Observation></entry>
48      <entry>
49       <SubstanceAdministration>
50        <text><content ID="m1">Theophylline</content> 20 mg every other day, alternating
           with 18 mg every other day. Stop if temperature is above 103F.</text>
51        <consumable>
52         <manufacturedProduct>
53          <manufacturedLabeledDrug>
54           <code code="66493003" codeSystem="2.16.840.1.113883.6.96"
              codeSystemName="SNOMED CT" displayName="Theophylline"/>
55        </manufacturedLabeledDrug></manufacturedProduct></consumable>
57    </SubstanceAdministration></entry></section></component>
58    <component>
59     <section>
60      <code code="11384-5" codeSystem="2.16.840.1.113883.6.1" codeSystemName="LOINC"/>
61      <title>Physical Examination</title>
62      <component>
63       <section>
64        <code code="8716-3" codeSystem="2.16.840.1.113883.6.1" codeSystemName="LOINC"/>
65        <title>Vital Signs</title>
66        <text>
67          <table>
68            <tr>
69              <th>Temperature</th>
70              <td>36.9 C (98.5 F)</td>
71            </tr>
72            <tr>
73              <th>Pulse</th>
74              <td>86 / minute</td>
75        </tr></table></text>
76        <entry>
77         <Observation>
78          <code code="50373000" codeSystem="2.16.840.1.113883.6.96"
              codeSystemName="SNOMED CT" displayName="Body height"/>
79          <effectiveTime value="200004071430"/>
80          <value xsi:type="PQ" value="1.77" unit="m" />
81         </Observation></entry></section></component></section></component>
82 </StructuredBody></component></ClinicalDocument>
```

Figure 2: HL7 CDA Sample Document.

as-result. For example, XRANK favors deeply nested elements, returning the deepest node containing the keywords as the most specific one, having more context information. In contrast, a path as the result is proposed by [2, 4, 20, 10, 21]; where a minimal path of XML nodes is returned that collectively contain all the query keywords. Note that we use the term "path" loosely to differentiate it from the subtree-as-result approach, because it can be a collection of meeting paths (a tree) for more than two query keywords. We refer to this approach as *path-as-result*. It is unclear whether the subtree-as-result or the path-as-result is a better fit for searching CDA documents. The discussion on minimal information unit below sheds more light to this aspect. Another issue is the *scope* of a result, in particular, whether results spanning across EMRs should be produced. Finally, doctors would like to be able to specify the results' schema in some cases, which in turn limits the types of elements searched for the query keywords.

3.2. Minimal Information Unit (MIU)

It is challenging to define the granularity of a piece of information in a way that it is self-contained and meaningful, but at the same time specific. For example, in Document D_1 returning the "value" element of Line 45 without the preceding "code" element is not meaningful for the user. Hence, the "value" element is not an appropriate MIU, whereas the enclosing "Observation" element could be.

Furthermore, for some queries it is required to include into the result some elements that do not contribute in connecting the query keywords or are part of the MIU of such a connecting node. For instance, the "patientPatient" element should be included in the result of query "Asthma Theophylline" if a practitioner submits the query, but not if a researcher does. Such personalization issues are further discussed in Section 3.11.

3.3. Semantics of Node and Edge Type

It is challenging to incorporate the rich semantic information available for the clinical domain, and particularly for the elements of a CDA document, in the results' ranking process. At the most basic, a domain expert statically assigns a weight to each node and edge type, as in BANKS [4]. In addition to that, we can assign a relevance to whole paths on the schema as explained below. Furthermore, it is desirable that the degrees of semantic association are adjusted dynamically exploiting relevance feedback and learning techniques.

3.4. Access to Dictionaries and Ontologies

CDA documents routinely contain references to external dictionary and ontology sources through numeric codes. As an example, document D_1 includes references to LOINC [22] and SNOMED

CT [27] in Lines 34 and 38 respectively. Hence, it is no longer enough to answer a query considering the CDA document in isolation, as is done by all previous work on information discovery on XML documents (Section 2.3). In this setting, the query keywords may refer to text in the CDA document or an ontology that is connected to the CDA document through a code reference. For example, the query keyword "appendicitis" may not be present in the document but its code might be present, so we need to go to the ontology and search for the query keyword there.

3.5. Access to Dictionaries and Ontologies

We need to assign an appropriate value to each of the relations present in the ontologies. SNOMED CT, for example, has four different types of relationships: (1) Defining characteristics, (2) Qualifying characteristics, (3) Historical relationships and (4) Other relationships. Figure 1 includes relations such as "May be", "Finding site of" and "Has finding site" in addition to the most common "Is a" relationship. Stricter and stronger relations in the ontology should intuitively have a higher weight. Furthermore, we need to take into consideration the direction of the edges. A possible approach to measure the degree of association between nodes of an ontology graph is to execute ObjectRank [5] on the ontology graph, as described by Hwang et al. [17].

3.6. Arbitrary Levels of Nesting

We can find an arbitrary number of levels of nesting and recursion in the definition of components and sections, as shown in the path *component.section. component.section* in Lines 58-63 of Figure 2. Taking into consideration the semantics of the document, the interconnection relationship rule of XSEarch [10], where the same tag may not appear twice in internal nodes of a result path, cannot be applied since the same tag can appear twice in a vertical path (top-to-bottom).

3.7. Free Text Embedded in CDA Documents

In some cases, plain text descriptions are added to certain sections to enrich the information about the record or to express a real life property not codified in dictionaries or ontologies. As a first measure, traditional text-based Information Retrieval techniques should be included in the architecture to support such cases. Another technique to address the coexistence of semi-structured and unstructured data is presented in [16], where IR and proximity rankings are combined.

3.8. Time and Location Attributes

After discussing with medical researchers and practitioners, we found that time and location are critical attributes in most queries. For instance, for the query "drug-A drug-B" the doctor is probably looking for any conflict between these drugs, and hence the time distance between the prescriptions of these drugs for a patient is a critical piece of information. Location is also important since two patients located in nearby beds in the hospital should be viewed as associated because infections tend to transmit to neighboring beds. Clearly, it is challenging to standardize the representation of such location information within an EMR.

Furthermore, time and location can lead to the definition of metrics similar to the inverse document frequency (idf) in Information Retrieval [26]. For instance, asthma is more common in summer; hence a patient who has asthma in winter should be ranked higher for the query "asthma". Similarly, a patient who has the flu in a town where no one else has it should be ranked higher for the query "flu". These associations are too complex since time can be used to define time, distance, or periodicity. Similarly, location relationships can be specified either within a hospital or across towns.

3.9. EMR Document-as-Query

An alternative query type to the plain keyword query is using a whole (or part of) EMR (CDA) document as the query. This approach can be used to find similar CDA documents, that is, CDA documents of patients with similar history, demographic information, treatments, and so on. The user should be able to customize and personalize such an information discovery tool to fit her needs. For instance, a researcher may not consider the physician's (author of CDA document) name when matching CDA documents, and could specify that a generic medication should be viewed as identical to the non-generic equivalent. Previous work on document content similarity [3] and XML document structural similarity [23] can be leveraged to solve this problem.

Furthermore, such document-as-query queries can be used to locate medical literature relevant to the current patient. In this scenario, the EMR application could have a button named "relevant literature" that invokes an information discovery algorithm on PubMed or other medical sources. Price et al. [24] present a first attempt towards this direction, where they extract all MeSH terms (MeSH refers to the U.S. National Library of Medicine's controlled vocabulary used for indexing articles for MEDLINE/PubMed) from an EMR (not specific to CDA) and then query MEDLINE using these terms. The structured format of CDA documents can potentially

allow more elaborate searching algorithms where multiple terms that are structurally correlated can construct a single and more focused query on medical literature sources.

3.10. Handle Negative Statements

A substantial fraction of the clinical observations entered into patient records are expressed by means of negation. Elkin et al. [12] found SNOMED-CT to provide coverage for 14,792 concepts in 41 health records from Johns Hopkins University, of which 1,823 (12.3%) were identified as negative by human review. Today, one has to examine the terms preceding a diagnosis to determine if this diagnosis was excluded or not. Ceusters and Smith [11] propose new ontological relationships to express "negative findings". It is challenging to handle such negative statements for an information discovery query in a way that the user can specify whether negated concepts should be excluded or not from the search process.

3.11. Personalization

The information discovery engine should provide personalized results depending on the preferences of each individual user. For example, for different doctors, different entities and relationships in the CDA components are more important. For some healthcare providers, the medication may be more relevant than the observation, or the medication may be more relevant than the doctor name. Also the relationships in ontologies may be viewed differently. Furthermore, depending on whether a user is a nurse, a pharmacist, a technician or a physician, the system could automatically assign different weights on edges and nodes of the CDA Object Model to facilitate the information needs of the users.

4. Concluding Remarks

We have introduced the problem of Information Discovery on Electronic Medical Records (EMR), enumerating a series of challenges that must be addressed to provide quality information discovery services on EMRs, specifically on HL7 CDA documents. The key challenges are related to the semantics of the architecture, the XML structure of CDA documents, and the interconnection of EMR documents with ontologies and dictionaries. Additional challenges include the incorporation of time and location semantics, as well as handling negative statements. We hope that this work will spawn new research directions to address these challenges. The successful realization of information discovery on EMRs is expected to have a great impact on the quality of healthcare.

5. References

[1] S. Amer-Yahia, C. Botev and J. Shanmugasundaram. TeXQuery: A Full-Text Search Extension to XQuery. In WWW 2004.

[2] S. Agrawal, S. Chaudhuri, and G. Das. DBXplorer: A System For Keyword-Based Search Over Relational Databases. In ICDE, 2002.

[3] R. K. Ando. Latent Semantic Space: Iterative scaling improves precision of inter-document similarity measurement. SIGIR 2000.

[4] G. Bhalotia, A. Hulgeri, C. Nakhey, et al. Keyword searching and browsing in databases using BANKS. In ICDE, 2002.

[5] A. Balmin, V. Hristidis, and Y. Papakonstantinou. ObjectRank: authority-based keyword search in databases. In VLDB 2004.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1-7), 1998.

[7] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. New York, ACM Press, 1999

[8] HL7 Clinical Document Architecture, Release 2.0. http://lists.hl7.org/read/attachment/61225/1/CDA-doc%20version.pdf. 2007.

[9] S. Cohen, Y. Kanza, and B. Kimelfeld. Interconnection semantics for keyword search in XML. In CIKM, 2005.

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

[11] W. Ceusters and B. Smith. Tracking Referents in Electronic Health Records. Medical Informatics Europe (MIE 2005), Geneva, Stud Health Technol Inform. 2005;116:71–76.

[12] PL Elkin at al. A controlled trial of automated classification of negation from clinical notes. BMC Medical Informatics and Decision Making 2005;5:13.

[13] N. Fuhr and K. Großjohann. XIRQL: a query language for information retrieval in XML documents. In ACM SIGIR, 2001.

[14] L. Guo, F. Shao, C. Botev, and J. Shanmugasundaram. XRANK: Ranked Keyword Search over XML documents. In ACM SIGMOD, 2003.

[15] V. Hristidis, F. Farfán, R. Burke, A. Rossi, J. White. Challenges for Information Discovery on Electronic Medical Records. Florida International University Technical Report. Feb 2007. http://www.cis.fiu.edu/~vagelis/publications/TR_2007_02_02.pdf

[16] V. Hristidis, L. Gravano, and Y. Papakonstantinou. Effcient IR-style keyword search over relational databases. In VLDB, 2003.

[17] H. Hwang, V. Hristidis and Y. Papakonstantinou. ObjectRank: A System for Authority-based Search on Databases. Demo paper, ACM SIGMOD 2006.

[18] Health Insurance Portability and Accountability Act. http://www.hipaa.org/. 2007.

[19] Health Level Seven Group. http://www.hl7.org/. 2007.

[20] V. Hristidis and Y. Papakonstantinou. DISCOVER: Keyword Search in Relational Databases. In VLDB, 2002.

[21] V. Hristidis, Y. Papakonstantinou, and A. Balmin. Keyword proximity search on XML graphs. In ICDE, 2003.

[22] Logical Observation Identifiers Names and Codes (LOINC). http://www.regenstrief.org/medinformatics/loinc/. 2006.

[23] A. Nierman, H.V. Jagadish. Evaluating Structural Similarity in XML Documents. WebDB 2002.

[24] SL Price, WR Hersh, DD Olson, PJ Embi, SmartQuery: context-sensitive links to medical knowledge sources from the electronic patient record, Proceedings of the 2002 Annual AMIA Symposium, 2002, 627-631

[25] HL7 Reference Information Model. http://www.hl7.org/library/data-model/RIM/C30204/rim.htm. 2007.

[26] G. Salton. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison Wesley, 1989.

[27] SNOMED Clinical Terms (SNOMED CT). http://www.snomed.org/snomedct/index.html. 2006.