COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 112 (2013) 529–552

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

A new tool for sharing and querying of clinical documents modeled using HL7 Version 3 standard

Vasil Slavov[a], *Praveen Rao*[a,*], *Srivenu Paturi*[a], *Tivakar Komara Swami*[a], *Michael Barnes*[a], *Deepthi Rao*[b], *Raghuvarun Palvai*[a]

[a] *Department of Computer Science Electrical Engineering, University of Missouri-Kansas City, United States*
[b] *Department of Pathology and Laboratory Medicine, University of Kansas Medical Center, United States*

ARTICLE INFO

Article history:
Received 10 September 2012
Received in revised form
16 April 2013
Accepted 2 July 2013

Keywords:
HL7 v3
P2P
XML
Query processing
Security

ABSTRACT

We present a new software tool called CDN (Collaborative Data Network) for sharing and querying of clinical documents modeled using HL7 v3 standard (e.g., Clinical Document Architecture (CDA), Continuity of Care Document (CCD)). Similar to the caBIG initiative, CDN aims to foster innovations in cancer treatment and diagnosis through large-scale, sharing of clinical data. We focus on cancer because it is the second leading cause of deaths in the US. CDN is based on the synergistic combination of peer-to-peer technology and the extensible markup language XML and XQuery. Using CDN, a user can pose both structured queries and keyword queries on the HL7 v3 documents hosted by data providers. CDN is unique in its design – it supports *location oblivious queries* in a large-scale, network wherein a user does not explicitly provide the location of the data for a query. A location service in CDN discovers data of interest in the network at query time. CDN uses standard cryptographic techniques to provide security to data providers and protect the privacy of patients. Using CDN, a user can pose clinical queries pertaining to cancer containing aggregations and joins across data hosted by multiple data providers. CDN is implemented with open-source software for web application development and XML query processing. We ran CDN in a distributed environment using Amazon EC2 as a testbed. We report its performance on real and synthetic datasets of discharge summaries. We show that CDN can achieve good performance in a setup with large number of data providers and documents.

1. Introduction

Today, it is well agreed upon that through effective use of Information Technology (IT), health care costs can be reduced and better quality care can be delivered to patients. The US government is spending billions of dollars to promote the adoption of electronic health records and to develop Health Information Exchanges (HIEs) [46]. HIEs aim to enable "the electronic movement of health-related information across organizations according to nationally recognized standards" [46]. They are considered to be the building blocks for

* *Corresponding author at*: Department of Computer Science Electrical Engineering, University of Missouri-Kansas City, United States. Tel.: +1 816 235 2705.
E-mail addresses: vgslavov@mail.umkc.edu (V. Slavov), raopr@umkc.edu (P. Rao), sp895@mail.umkc.edu (S. Paturi), tk27f@mail.umkc.edu (T.K. Swami), mjb5a6@umkc.edu (M. Barnes), drao@kumc.edu (D. Rao), rp7x5@mail.umkc.edu (R. Palvai).
0169-2607/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2013.07.002

Nationwide Health Information Network (NHIN) initiative [45] and are designed to achieving Institute of Medicine's (IOM) vision of a learning healthcare system [19]. Some of the established HIEs such as HealthBridge, CareSpark, Indiana Health Information Exchange, and MedVirginia serve up to few million patients and few thousand physicians, thereby, hosting large volumes of patient data [47].

Recently, "data sharing and collaboration" and "large scale management of health care data" have been identified as the key IT challenges to advance the nation's healthcare system [80]. This is because vast amounts of health-related information remain untapped due to the lack of suitable IT solutions. Personal health information resides in digital silos and healthcare systems do not easily share information with each other. However, by tearing down these silos, health-related information can be utilized by medical practitioners and researchers to provide efficient, quality, timely, and cost–effective care to patients.

The National Cancer Institute's caBIG is a nation-wide initiative, whose vision is to advance research on cancer and improve clinical outcomes for patients by connecting the members of the cancer community to share knowledge and data [32]. The caBIG community has more than 190 organizations [13]. Today, there are 124 participating institutes connected to caGrid – the underlying network infrastructure of caBIG. The community has shown great interest in sharing large amounts for biospecimen annotations, microarray data, cancer genome data (e.g., tissue samples), and so forth [14]. Such large-scale sharing of biomedical and clinical data is the first step towards collaborative e-science in the 21st century.

Achieving interoperability among applications processing clinical data has been a topic of interest for several years. Many advances have been made in developing standards for clinical data with regard to exchange/messaging, terminology, application, architecture, and so forth [53]. The standards from Health Level Seven International (HL7) have become popular for the exchange, integration, sharing and retrieval of electronic health information. HL7 standards are used by 90% of the hospitals in the US.[1] More recently, HL7 Version 3 standard was developed to enable *semantic interoperability* in healthcare data interchange [63]. (XML is used to encode the data.) The documents in HL7 v3 are derived from the Reference Information Model (RIM) and use terminologies such as SNOMED CT, LOINC, CPT and ICD-9. Software tools are available for modeling data using HL7 v3 standards (e.g., Model-Driven Health Tools [67], caAdapter [15], HL7 Tooling [66]).

We present a new software tool called CDN (Collaborative Data Network) for sharing and querying of clinical data modeled in HL7 v3 standard. Of particular interest to us are the HL7 CDA (Clinical Document Architecture) and CCD (Continuity of Care Document) standards. CDN is ideal tool for data providers (e.g., clinic, hospital, research lab) who wish to selectively enable data sharing and querying of HL7 v3 documents. While CDN is not restricted to a particular health condition, the GUI of CDN is designed for posing clinical queries related to cancer diagnosis and treatment. Cancer is the second most leading cause of deaths in the US. CDN *differs from the aim of HIEs in the sense that it is not designed for the electronic movement of health-related information across organizations.*

The remainder of the paper is organized as follows. Section 2 provides the background and motivations. Section 3 describes the novel architecture of CDN, the query processing approach, and security schemes in CDN. Section 4 describes the implementation and evaluation of CDN. We provide a discussion in Section 5 and conclude in Section 6.

This work has previously appeared in the conference proceedings of the 2010 and 2012 ACM SIGHIT International Health Informatics Symposium [68,71]).

2. Background and motivations

2.1. The peer-to-peer model of computing

We have witnessed a huge success of the P2P model of computing in the last decade. This has culminated in the development of Internet-scale applications such as Kazaa, BitTorrent, and Skype. P2P computing has also become popular in ecommerce and ebusiness and has lead to the development of many Internet-scale systems. Innovations in P2P computing, most notably the concept of Distributed Hash Table (DHT) (e.g., Chord [81], Pastry [76], CAN [72], Tapestry [90], Kademlia [62]), has been embraced by key-value stores of production quality such as Dynamo [24], Cassandra [56], and Voldemort [58]. DHT-based systems have good scalability, fault-tolerance, and load balancing properties. Because of these useful properties, CDN employs a DHT for indexing HL7 v3 documents and locating relevant documents during query processing.

2.2. XML and distributed XQuery

The extensible markup language XML has become the de facto standard for information representation and interchange on the Internet. It is widely adopted in a variety of domains ranging from ecommerce to health informatics. XQuery is a popular query language for XML and is recommended by the W3C. It is a functional language that subsumes XPath – a query language for selecting qualifying nodes in an XML document. XQuery allows for the creation of new elements and attributes and the specification of their contents and relationships. Queries in XQuery can contain `for`, `let`, `where`, `order by`, and `return` clauses and are frequently called FLWOR expressions.

A great deal of work has been done in the area of distributed query processing [55]. There are three well-known approaches to processing a distributed query, namely, pure data shipping, pure query shipping, and hybrid shipping. Neither pure data shipping nor pure query shipping are the best choices in all scenarios in a distributed setting and a hybrid approach has shown to perform better [55]. Distributed XQuery processing [73,33,89,34,29,25,38] has been studied in recent years. The underlying principle is to ship portions of a query to remote servers which then execute them. Locations of remote servers are specified in the query. These previous solutions were not designed for a P2P network, where the locations of relevant data of interest may not be known apriori. In contrast, CDN

[1] http://www.itl.nist.gov/div897/docs/Message_Maker.html.