MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Index and Query of Clinical Documents Complied with HL7 CDA Documents Using Weighted Navigational Expression

Zhijing Liu PhD[a], Sam Brandt MD[a], Floyd Eisenberg MD[a]

[a] *Department of Clinical Informatics, Siemens Medical Solution, United States*

Abstract

The Health Level 7 (HL7) Clinic Document Architecture (CDA) is an American National Standards Institute (ANSI)–approved, document markup standard that specifies the hierarchical structure and semantics of clinical document (such as a discharge summary, progress note, procedure report) for the purpose of information exchange. In this paper, a method is developed for index and query of CDA documents using weighted navigational expression. The method is based on templates that define the weighted navigational expression to guide the processes of index and query of clinical documents, so it is configurable. With the method, the index and query of clinical documents is fast and reliable. The method can be easily implemented in a distributed, heterogeneous environment such as a web application and adopted to different clinical settings.

Keywords:

clinical documents, CDA, retrieval, index, query

Introduction

HL7 CDA documents are encoded in Extensible Markup Language (XML). CDA documents derive their machine processable meaning from the HL7 Reference Information Model (RIM) and use the HL7 V3 Data Types. The CDA specification is richly expressive and flexible. Document-level, section-level and entry-level templates can be used to constrain the generic CDA specification. A CDA document can be transferred within a message, and can exist independently, outside the transferring message, with the characteristics of persistence, context, and wholeness. To be able to fully exploit the potential the CDA documents, one has to be able to effectively index, archive, and retrieve such clinical documents and more, query in a manner that takes advantage of the structured nature of such documents.

In the presentation, a weighted navigational expression method is developed to index and retrieve the clinical document, which precisely addresses this concern. The method is based on templates that define the weighted navigational expression to guide the processes of index and query clinical documents.

Methods

The first step of the weighted navigational expression method is document analysis, which analyzes the document *structures,* metadata, XML fragment and value, the filter terms of index and query, and stop context. A XML fragment is structure data represented and a part of a clinical document. The paths of the data fragments are specified in the index template. The filter terms are the content or attribute of atomic elements of XML data fragments. The textual contents of metadata, data fragments, and filter terms are analyzed based on linguists, thesauri or ontology. There are some dependencies between analysis steps because some results of the one analysis are helpfully or even necessary for the other analysis. Term position must be determined before stop word elimination because some terms are not counted and some phrase query may fail. Stop-word elimination should be processed before stemming because stemming is expensive depending on the number of words. Stop words such as ('the', 'a' 'if' etc) that increase the overhead but add no value during query are removed here.

After CDA documents are analyzed, an index process extracts metadata, XML data fragments, ranking terms, filter terms, and their relations to the content. The weighted data to be ranked according to the importance of data and the structure relationship among different levels of data fragments in the clinical documents are profiled in an index template. Figure 1 illustrates some data fragments and its tree structure. The structure relationships are used with the method as paths to locate and extract data from the clinical documents. The context of an individual CDA document, a set of the documents, or multiple sets of the documents are navigated and extracted to generate an index dictionary with different granularities to facilitate the flexibility of the query of the clinical documents. The indexes of clinical documents, their collection, their context, and the clinical setting are stored dictionary which is stored as a set of files. The indexes can be in-memory or map to a RMDS database specified by the index template.

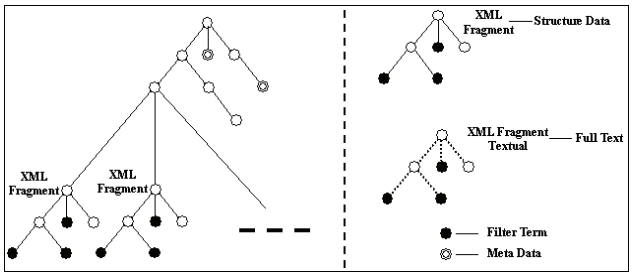

Figure 1 - Filter Term Nodes and XML Data Fragment Nodes

After the index dictionary is generated, users can query and retrieve the documents and can select the query context. If a single collection is selected, users can either specify an individual, multiple, or all the clinical document(s) in the collection. A query agent can perform query parse, query evaluation, pattern matching, and then present the query results. The query agent can perform both information retrieval (text query) and structure query (XML data fragments). The text query expressions can be constructed with simple keyword search, combining keywords within Boolean expressions, or querying for phrases, sentences etc. It can query the metadata, the textual content of XML data fragments, the element values or attribute values of filter terms. The XML Path Language is used to query structure data. The path expressions are embedded in the query template. Path expressions are also used to specify the return document fragments since often not the whole XML document should be referenced in the query result but only a certain fragment. By default, references to the root nodes of the matching documents or matching XML data fragments are returned depending on which how results are presented.

Results

The method has been successfully implemented as web application as shown in Figure 2. The implementation includes two major modules – index module and query module. The index module has an index user interface (UI) and an index agent. The query module has a query UI and a query agent.

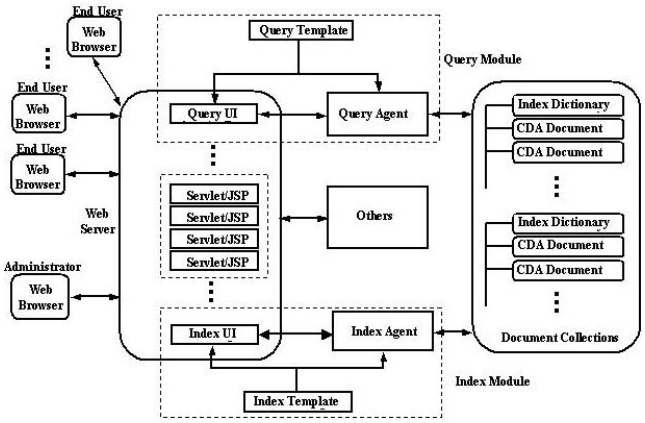

Figure 2 - Implementation of the weighted navigational expression method

Indexer agent performs document analysis, extracts data from the clinical documents, and stores index information to the index files in the server. Both index UI and index agent is configurable with the index template.

The query UI allows users to specify the query range. Users can query information from keywords and can be scoped down with filter terms. The query agent queries data from indexes and performs query parse, query evaluation, pattern matching, and present the query results. The query results are presented to users with two ways. All resulting data fragments are organized together and send back from query agent to query UI, where the results can be displayed either in a HTML page or PDF document depending upon the preferred style.

All matched clinical document names are listed in query UI. Users can select and look their interested document in detail. The matched contents are highlighted in the clinical document. The view of the document can also be displayed either as a HTML page or PDF document depending upon the preferred style

Discussion

There are two kinds of information retrieval IR approaches that deal with the retrieval of structured documents:

The structural approaches enrich text query by conditions relating to the document structure, e.g. that words should occur in certain parts of a document, or that a condition should be fulfilled in a document part preceding the part satisfying another condition, but generally do not support weighting or ranking [3].

Content-based approaches aim at the retrieval of the most relevant part of a document with respect to a given query, but do not allow for structural conditions [4].

Only a few researchers have dealt with the combination of explicit structural information and content-based retrieval. Mayan and colleague [5] use belief networks for determining the most relevant part of structural documents, but allows only for plain text queries, without structural conditions. The FERMI multimedia model [6] presents a general framework for relevance- based retrieval of documents.

In the paper, a weighted navigational expression method is developed to index and retrieve the clinical document, which precisely addresses this concern. The method is based on templates that define the weighted navigational expression to guide the processes of index and query clinical documents. The method can be easily implemented in a distributed, heterogeneous environment such as a web application and adopted to different clinical settings.

Conclusion

The weighted navigational expression method is an easy, effective, and configurable method of index and query of CDA documents. The method is implemented as a web application. The method can also be implemented as a stand-alone application or can be integrated with other web application.