Efficient Medical Information Retrieval in Encrypted Electronic Health Records

Cédric PRUSKI1 and François WISNIEWSKI
CR SANTEC – AHI
Centre de Recherche Public Henri Tudor
2A rue Kalchegrück L-1852
Luxembourg

Abstract. The recent development of eHealth platforms across the world, whose main objective is to centralize patient’s healthcare information to ensure the best continuity of care, requires the development of advanced tools and techniques for supporting health professionals in retrieving relevant information in this vast quantity of data. However, for preserving patient’s privacy, some countries decided to de-identify and encrypt data contained in the shared Electronic Health Records, which reinforces the complexity of proposing efficient medical information retrieval approach. In this paper, we describe an original approach exploiting standards metadata as well as knowledge organizing systems to overcome the barriers of data encryption for improving the results of medical information retrieval in centralized and encrypted Electronic Health Records. This is done through the exploitation of semantic properties provided by knowledge organizing systems, which enable query expansion. Furthermore, we provide an overview of the approach together with illustrating examples and a discussion on the advantages and limitations of the provided framework.

Keywords. Information Retrieval; Ontology; EHR; Encrypted Data

Introduction

Over the last decade, several countries have decided to develop a so-called eHealth platform that offers a central access to citizens’ Electronic Health Records (EHR) for health professionals in order to reduce global costs and optimize cares. Nevertheless, the exploitation of the stored data is usually constrained by a set of restrictions mainly from legal or technical nature. This is for instance the case in Luxembourg regarding the security of the stored data [1]. Actually, in order to protect patient privacy and to prevent illegal use of personal health information, the data located on the platform must remain de-identified and encrypted [2]. This strong constraint is a clear barrier to implement efficient information retrieval mechanisms to support health professionals (HP) when searching for relevant information since submitted queries cannot be evaluated without the decryption of the data causing security breaches. This is why existing information retrieval techniques must be adapted in order to cope with this data encryption issue.

1 Corresponding Author.
For the description of EHR's content, the use of metadata has been pointed out as promising [3] in order to promote interoperability and it has shown convincing capabilities for various purposes like semantic search [4] or data integration [5]. It is therefore natural that standard metadata has been defined and integrated in well-accepted initiative for eHealth (e.g. HL7, Integrating the Healthcare Enterprise). Besides, the emergence of ontologies, part of the Semantic Web paradigm, and the use of well-known biomedical knowledge organizing systems like terminologies, taxonomies or thesauri, offer properties for improving metadata capabilities. Such properties allow to make the semantics of the metadata explicit for both human and computer agents, which, in turn, facilitated the development of advanced techniques like query expansion [6] in order to make the query more accurate and therefore improving, in terms of relevance, the results of a search process. Nevertheless, the selection of the metadata must be done in a way that the identity of the patient is not betrayed.

In this paper, we introduce an original approach for searching relevant information in a shared EHR composed of encrypted Clinical Document Architecture (CDA) documents. The proposed framework is based on the exploitation of standard XDS metadata for CDA document description together with ontologies and standard medical terminologies like the UMLS, SNOMED CT or LOINC. We briefly introduce through a concrete example a query expansion mechanism exploiting semantics properties of the metadata in order to enhance the results of a search. The remainder of the paper is structured as follows: Methods section presents the various components of our approach for retrieving medical information. We explain how standard metadata and ontologies are combined to improve the result of a search. Then we discuss the advantages and limitations of the framework. We finally wrap up with concluding remarks and outline future work.

1. Methods

Our approach for efficient information retrieval in encrypted EHR relies on two main ideas. The first one consists in the use of relevant metadata for the description of the content of the CDA documents sent to the platform. The definition of such metadata is twofold. On one hand, medical documents provided by data provider (see Figure 1) are enriched with a set of XDS metadata (e.g. authorPerson, creationTime, classCode, etc) as recommended by IHE for interoperability reasons. Nevertheless, such standard metadata is not tailored enough for the precise description of the documents’ content. This is the reason why additional metadata must be associated with documents to better describe their content and facilitate their retrieval. In consequence, the connector software, application in charge of bridging the interoperability gap between actors (see Figure 1), is able to extract, using natural language processing techniques combined with the properties of the UMLS metathesaurus (vocabulary of reference in Figure 1), additional metadata (called meta-information) by analyzing the content of the document before being encrypted. The so-extracted meta-information consists of a set of concept codes coming from the UMLS. It provides an overview of the content of the document without revealing information values such as the result of a medical exam and the identity of the patient. The encryption and pseudonymization of the data is ensured by a mechanism implementing a trusted third party in charge of maintaining the association between patient identity and pseudonyms used to index the documents.
but this aspect goes beyond the scope of this paper. As a result, three kinds of
document or information are available. The encrypted document itself (stored in the
repository of the eHealth platform), the reference of the document associated with the
XDS metadata (located in the registry) and, finally, the reference of the document
associated with a set of meta-information (stored in the meta-information index) that
contains much precise information that tags the content of the encrypted document.

The second part of our approach consists in exploiting the XDS metadata and the
meta-information previously identified and linked to the documents for improving the
results of a search. Our strategy is based on the implementation of an advanced query
expansion mechanism making the most of the properties of the ontology (see Figure 1),
the submitted queries that are keywords based and the existing metadata and the meta-
information. The goal is to make explicit information that are obvious for the user and
are usually never mentioned in the queries. The proposed mechanism has been defined
based on the observation made about the queries that HPs (or data consumer in
Figure 1) are likely to submit. These are usually short (i.e. composed of one or two keywords)
because HPs are not supposed to be familiar with existing tools for information
retrieval that implement expressive query language. Furthermore, queries are built on
very specific keywords highly representative of the kind of information the HP is really
interested in (e.g. glucose or cholesterol). Starting from such queries, the system must
be intelligent enough to understand the demand of the user in order to decide the type
of document that contains the requested information. This is where the ontology comes
into play. Actually, the ontology constitutes the heart of the proposed approach for
information retrieval. It is constructed in a way that enables to connect the XDS
metadata model with the UMLS metathesaurus. It serves as support for extracting
additional keywords to put in the initial query in order to make it more precise
regarding the information the user is really interested in. These keywords can be seen
as constraints whose role is to restrict as much as possible the search space in order to
improve both precision and recall of the search. The combination of these constraints
acts as a filter to skim many irrelevant documents.

In our context, the designed query expansion mechanism [6] exploits the structure
of the initial query (i.e. how the keywords are combined) as well as the keywords that

Figure 1: The Framework
compose it, but also the semantic information provided by the ontology including the metadata and the meta-information and the semantic relation between the keywords. Actually, the first objective is to clearly target the type of documents that is supposed to contain the relevant information. This is done through the analysis of the relations between the keywords composing the query and their position in the ontology. For instance, if the entered query refers to the glucose rate of a patient, the system must be able to search only in laboratory reports stored in the platform and neglect all other kind of documents like medical images for instance. Therefore, thanks to the ontology, the system is able to infer such information. With respect to the previous example, glucose is located in the subpart of the ontology attached to the typeCode, part of XDS metadata model, reporting about the type of the considered documents (e.g. only laboratory reports). Moreover, in order to restrict even more the search space, we exploit additional information provided by the UMLS and more precisely by the contained terminologies and mainly their structure relying on the subsumption relation (i.e. is-a). According to the previous example, the glucose concept has been decided as part of the chemistry class in existing codification schemes. It is therefore useful for the system to enrich the initial query with this information since it can filter many unwanted documents. In consequence documents not annotated with the LOINC code referring to chemistry as eventCodeList of the XDS metadata model can be excluded. According to the architecture of the platform (see Figure 1), the system must be able to interpret the enriched queries on the logic structure of the registry and meta-information index components and return the result of the search to the user.

2. Preliminary Results and Discussion

The approach presented in this paper has been only evaluated on a small set of data as a proof of concept. Although being extremely promising, it is irrelevant to mention in this paper the obtained results in term of precision and recall. This aspect will be the subject of our future work in this field. Nevertheless, even if data encryption is a serious barrier to overcome towards the development of efficient information retrieval techniques, our experiences show that the combined use of metadata and knowledge organizing systems including ontologies, terminologies, taxonomies and thesauri offers interesting properties to improve, in terms of relevance, the results of a search. Moreover, the use of standard vocabularies facilitate the construction and interpretation of the queries and, in a long term perspective, will simplify the connection with cross-border systems in order for all HPs to retrieve relevant information concerning their patients for optimizing care. However, there are some restrictions. Actually, in order to be efficient in query interpretation and enrichment, the keywords composing the submitted queries must be labels of existing concepts in the termino-ontological resources used otherwise the system will not be able to make the link between queries and ontologies. To overcome this lack, we coupled a lexical analyzer to normalize terms of the query. The idea is to substitute terms that are misspelled or divergent from syntactic but not semantic point of view from the normalized terms contained in the vocabulary of reference to reinforce the membership of the terms as label of concept. It consists basically in replacing terms written under their plural form with the singular one (e.g. feet/foot). The use of a lexical analyzer coupled with the various versions of the standard termino-ontological resources also limits issues induced by the multilingualism problem. As it is the case in Luxembourg, some health professionals
are likely to be fluent in French other rather in German, which directly impacts the way queries are written (i.e. in French, German or English).

Concerning security issues, the proposed software architecture as well as the implemented concepts (i.e. meta information index) does not weaken the secure aspect of the system, which relies on the pseudonymization of the data, because the information stored in the meta information index only contains an association of tags (i.e. meta-information) with documents ID (e.g. glucose information is contains in doc1, doc3, doc7, etc.). In consequence, if hackers have access to this element, they will never be able to establish the correspondence between a document ID and the identity of the patient without the exploitation of the registry. Moreover, the idea to clearly separate meta-information and metadata simplifies the treatment of queries that can be submitted in the context of vast study like for instance analyzing public health issues. Such queries that are usually devoted to compute statistics over the whole (or part of the) population, will be evaluated only on the logic structure of the meta-information index and not that of the registry which again will hide patient identity to the institution interested in public health. The real needs of HPs, in terms of medical information, can give place to standard queries to optimize IR. For instance, physicians in charge of patient with diabetes would require only document dealing with this disease therefore preformatted queries should be available to both reduce errors in queries building and optimize their interpretation by the system.

Although promising, the work introduced in this paper deserves several improvements. First of all, the involvement of the end users in the validation of the query expansion mechanism will be done. This will allow the identification of new enrichment rules based on users’ feedbacks, which, in turn, will strengthen the overall approach. To achieve this, we will have to test our framework on all kinds of health professionals having access to the platform ranging from physicians to pharmacists including nurses and hospital staff. The maintenance of the used ontologies and vocabulary of reference will also be the subject of future investigations. It will also be the case for the management of the evolution of the meta-information index required as soon as new meta-information will be identify. In a long-term perspective we will integrate this approach in a broader context involving Luxembourg’s neighbors in order to reinforce the validity of our concepts for treating multilingualism issues. Further research should also show how the HPs want to retrieve the information.

References