Information Retrieval in Electronic Health Records Using a Multiple Layer Query Language

Romain Lelong, Chloé Cabot, Tayeb Merabti, Julien Grosjean, Nicolas Griffon, Badisse Dahamna, Philippe Massari, Stéfan Darmoni

Information Retrieval in Electronic Health Records Using a Multiple Layer Query Language

Romain Lelong[1], Chloé Cabot[1], Tayeb Merabti[1], Julien Grosjean[1], Mher B. Joulakian[1], Nicolas Griffon[1,2], Badisse Dahamna[1], Philippe Massari[1], Stéfan J. Darmoni*[1,2]

[1] Service d'Informatique Biomédicale, Rouen University Hospital, Haute-Normandie & TIBS, LITIS EA 4108, France.
[2] LIMICS, INSERM, U1142, Paris, France.
* Corresponding author, stefan.darmoni@chu-rouen.fr.

Abstract *- **Information Retrieval (IR) in Electronic Health Records (EHRs) should provide healthcare professionals with accurate information to the right person at the right time and place. It should also reduce the difficult tasks of manual information retrieval from records in a paper format or computers. In this paper, a flexible, scalable and multi-layer object-oriented query language is described. It is designed for retrieving and viewing data which support any conceptual schema. The search engine deals with structured and unstructured data. Several types of queries on test databases containing some 200,000 anonymized records from 2,000 patients were tested. These queries were able to accuratly treat with symbolic, textual, numerical and chronological data.***

Index Terms *- **Medical Informatics***

I. INTRODUCTION

An Electronic Health Record (EHR) or an Electronic Medical Record is defined as "an electronic version of the traditional record used by the healthcare provider" [3]. EHR plays a central role since it includes a long-term record of care and a record of events from different types of care, including instructions, prospective information such as plans, orders and evaluations [1]. In this context, the goal of the Information Retrieval (IR) System on EHR is to provide physicians with the correct information at the right place for the right person [2].

Several tools and frameworks for searching in EHRs for one patient have been proposed. These tools are adapted according to each data format: structured, not structured or mixed. CISearch has been developed and implemented in the Columbia University Hospital EHR. The CISearch end-user may query all the textual reports (imaging, pathology, discharge summaries, etc.), using certain Lucene tools. Medical Information Retrieval System (MIRS) is also based on Lucene tools. The objective of the LERUDI project was to perform IR from a projected French EHR model in emergency management, using a domain ontology. There are also various IR Systems (IRS) that are available based on health data warehouse for several patients. The main system is Informatics for Integrating Biology and the Bedside (I2B2), an open source platform developed in the USA and dedicated to translational research.

In this context, the main objective of this paper is to present the functionnalities of a specific query language dedicated to information retrieval in EHRs. Since EHR data types are multiple, on several levels (patient, hospital, stays) and linked directly to patient care, IR from EHR is more difficult and different when compared to the "classical" IR. The conceptual schema used in this study allowed to define a language close to the medical representation of care management. This search engine is under development in the Retrieval and Visualization In Electronic Health records (RAVEL) project [4] funded by the French National Agency (TecSan program).

II. MATERIALS

A corpus of 2,000 anonymized patients and 200,000 reports was used in this study, approved by the French National Commission on Computers and Liberty. Almost any clinical information available in the EHR is integrated in the RAVEL model, e.g. Diagnosis related group codes (ICD10), patient data (age, gender), lab tests and all medical reports. Moreover, NLP tools developed by the Vidal and Lille teams of the RAVEL project were also used to partially re-structure the unstructured data via multi-terminological automatic indexing.

III. METHODS

The goal of developing a new query language is to ease the consultation and the information retrieval in the EHR by health professionals and to propose an alternative to query languages such as SQL.The query language syntax is currently patterned on the conceptual schema. This query language has three main characteristics: (i) structured information retrieval capabilities, (ii) scalability and (iii) comprehensive querying capabilities.

The query language is basically composed of nested syntactical units, with the following syntax: `ENTITY(CONSTRAINTS_CLAUSE)`, where: `ENTITY` corresponds to any kind of entity of the conceptual data model (e.g. patient, stay, medicalUnit, etc.) and

Example	Description
`stay(patient(id="DM_PAT_1736"))`	all stays linked to patient with id "DM_PAT_1736" entity, ie. all patient 1736 stays
`stay(icd10SC(label="Burns involving less than 10% of body surface"))`	all stays with a diagnosis of "Burns involving less than 10% of the body surface" using ICD10
`patient(medicalTest(exe( label="Sodium") AND numericResult> upperBound))`	all patients linked (via stays entity) to a biological test coded on the "sodium" and with a result greater than normal, ie. all patients with hypernatremia.

Table 1: Examples of basic semantic querying

`CONSTRAINTS_CLAUSE` corresponds to a constraint applied to the specified entity using Boolean operators (AND, OR, NOT) and parenthesis to logically link unitary constraints. Boolean operators, parenthesis and comparators (e.g. `=`) are explicitly defined in the grammar of the language whereas entities (e.g. `patient`) and unitary constraint keyword such as `birthDate` and `gender` keywords are automatically deduced from the database. It is also possible to combine two attributes in one constraint for instance: `stay(leavingDate - entryDate >= 10)` (stays with a duration of 10 days or more), `medicalTest(numericResult > upperBound)` (lab test result higher than normal level). Minimum and maximum can be searched on numerical and date data.
Thanks to its nesting capabilities, this query language is able to explore the relationships between entities and thereby enables to build a query based on the full semantic network.

IV. RESULTS

All the presented queries (see Table 1) have been tested with a database containing the data extracted from 2,000, including 200,000 reports from RUH.
Using this database, we have successfully answered several use cases: (i) visualize over time the neutrophil rate of a patient with rheumatoid arthritis, (ii) produce all the medical reports containing the "concept of metastasis", and (iii) retrieve all stays where "REMICADE" (infliximab) was used. For each of them, an EHR prototype displays each stay, medical procedure, medical test and report available. The neutrophil rate is displayed with a timeline in this same interface.
To extract all the medical reports containing the "concept of metastasis", a plain text query such as `compteRendu(FILE.f_html()="concept of metastasis")` was used.
Response times can be very variable depending on the query specificities (wildcards, hierarchical level, clauses number, etc.). Most queries are very fast (less than 3 seconds) but in some cases, response time could be improved.

V. DISCUSSION-CONCLUSION

In this work the main functionalities of the RAVEL project search engine were presented. We mainly focused on its query language that is generic and powerful, from our point of view. The presented search engine is able to deal with numerical, textual and chronological data. Furthermore, a comparative evaluation of this query language with I2B2 will be launched. I2B2 is a de facto standard in the field of EHRs data information retrieval as it is used in several French hospitals. The broad use of this solution should provide the necessary materials to make a comparison of search capabilities both in terms of precision and querying scope of the query languages.
The search engine is currently being tested outside the RUH: at Bordeaux University Hospital, Aquitaine, France. However, the current model still does not operate on the establishment level. This lack will be fulfilled in the near future. A scaling up study is ongoing at RUH with all the patients with at least one stay (in or out patient) in the dermatology department since 1992 ($n = 65,000$).

ACKNOWLEDGMENTS

This study was supported by a grant from the French National Research Agency (Tecsan ANR-11-TECS-012).

REFERENCES

[1] S. Garde, P. Knaup, E. Hovenga, and S. Heard. Towards semantic interoperability for electronic health records. *Methods Inf Med*, 46(3):332–43, 2007.

[2] K. Ondo, J. Wagner, and K. Gale. The electronic medical record: Hype or reality. *Journal of Healthcare Information Management*, 17(4):2, 2002.

[3] J. Sewell and L. Thede. Informatics and nursing: Opportunities and challenges. online glossary of terms, 2012.

[4] F. Thiessard, F. Mougin, G. Diallo, V. Jouhet, S. Cossin, N. Garcelon, B. Campillo, W. Jouini, J. Grosjean, P. Massari, N. Griffon, M. Dupuch, F. Tayalati, E. Dugas, A. Balvet, N. Grabar, S. Pereira, B. Frandji, S. Darmoni, and M. Cuggia. Ravel: retrieval and visualization in electronic health records. *Stud Health Technol Inform*, 180:194–8, 2012.