

Information Retrieval in Electronic Health Record: a feasibility study

Ahmed-Diouf DIRIEH DIBAD^a, Nicolas GRIFFON^a, Saoussen SAKJI^a,
Suzanne PEREIRA^b, Philippe MASSARI^a, Stéfan DARMONI^{a1}
^a*CISMeF, Rouen University Hospital & TIBS, LITIS EA 4108,
Institute of Biomedical Research, University of Rouen, France*
^b*VIDAL, Issy les Moulineaux, France*

Abstract. Background: To allow Electronic Health Records (EHRs) being useful for medical decision making or research, the information must be easily found in it, even in voluminous EHRs. This requires to develop search capabilities for information retrieval (IR). Methods: To perform this, we propose an adapted concise model to IR. The data analysis of EHRs of Rouen University Hospital has led us to consider EHRs as being a set of events linked by conceptual relationships. After implementation, we have evaluated the capacity of the adapted concise model to take into account all data from the EHRs and its accommodation to IR. Results: We performed 31 queries on EHR. The results in 22 cases were considered successful, although mistakes are avoidable. These results confirm the ability of the adapted concise model to take into account all relevant data of EHR in IR. Conclusion: Based on the preliminary evaluation of the adapted concise model, we have demonstrated its accommodation to IR in EHR. Nevertheless, further work on larger sets is required to confirm our preliminary results.

Keywords. information retrieval; electronic health record; modeling

1. Introduction

With the spread of information and communication technologies in the medical domain, an increasingly amount of health information is computerized into the Electronic Health Record (EHR). To make this information available for clinical use, information retrieval (IR) tools are needed. We present here a new model and an evaluation of its capacity to retrieve information.

2. Method

The Rouen University Hospital information system (IS) gathers clinical data in dozens of table. Only trained end-users are able to query EHR using specific codes namely CCAM and ICD10.

Based on this IS, we built a model in which: one table gathered all the events (e.g. patients, hospital visits, surgical procedure, biology test), one table summed up all the relations between events (surgical procedure *A* takes place during stay 1), one table

¹ Corresponding Author: Stefan J. Darmoni, CISMeF, Rouen University Hospital, 1 rue de Germont, 76031 Rouen Cedex, France; E-mail: stefan.darmoni@chu-rouen.fr.

contains events' attributes (biology test B was performed at date T , found value V), one table contained all the indexing data (diagnosis D was indexed by ICD-10 code I20.0).

To ensure the model effectiveness for IR, a physician (PM) created 31 test cases on 20 anonymous EHR, totalizing 2,075 hospital visits and 2,377 procedures. These test cases are clinical queries concerning only structured data (procedures, diagnoses and biology) in order to evaluate specific issues: mono vs. multi-patient search, chronology events, specific diagnoses or kind of pathology, procedures.

The results of the test cases were manually judged by an expert (PM) and classified as *relevant* when all pertinent information was restituted. In the other case, test cases results were classified as *irrelevant* and were explored manually to understand IS limits and to improve it.

3. Results

Results were relevant for 71% ($IC_{95\%} = [55\%-87\%]$) of queries. Six irrelevant results were due to query interpretation: terms used for querying did not belong to terminologies used for indexing and the IS did not match them with the correct controlled terms. Two irrelevant results were due to data incompleteness introduced in the model and one irrelevant result was explained by manual indexing error.

4. Discussion & Conclusion

We proposed a new model to optimize information retrieval in EHR for individual or cohort data. Its structure allows simplified querying that is conceptually better for scalability e.g. less computing response time.

Manual exploration of results showed that six of the irrelevant results should be avoided using natural language processing tools. They have not yet been integrated in our model. Therefore, the end-user has to be very cautious in writing the queries. However, using tools developed elsewhere, i.e. predefine queries [1], super concept [2] stemming and lemmatization and synonymies will facilitate querying for end user and improve information retrieval performance of our model.

As RUH EHR follows HISA norms, this may allow the use of our adapted concise model into other EHR respecting this norm. The next step is to make our model HL7 compliant to allow its use for most EHR commercial solutions. We described here an adapted concise model of EHR. Evaluation showed its efficacy for IR. However, future series will have to corroborate this and specify limits which remain unrecognized.

References

- [1] Douyère M, Soualmia L, Névéol A, et al. Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. *Health Info Libr J* 2004; 21(4):253-261
- [2] Massari P, Pereira S, Thirion B, Derville A, Darmoni SJ. Use of Super-Concepts to Customize Electronic Medical Records Data Display. In: *Proc. MIE-2008. Göteborg, Sweden, May, Studies in Health Technology and Informatics*, 2008: 136:845-50.