User Centred Networked Health Care
A. Moen et al. (Eds.)
IOS Press, 2011
© 2011 European Federation for Medical Informatics. All rights reserved.
doi:10.3233/978-1-60750-806-9-584

Roogle: An Information Retrieval Engine for Clinical Data Warehouse

Marc CUGGIA[a], Nicolas GARCELON[a], Boris CAMPILLO-GIMENEZ[a], Thomas BERNICOT[a], Jean-François LAURENT[b], Etienne GARIN[b], André HAPPE[c], and Régis DUVAUFERRIER[a]
[a]*UMR 936 Inserm, Faculté de médicine de Rennes. France*
[b]*CRLCC Centre Eugène Marquis, Rennes, France*
[c] *Intermède – Guignen. France*

Abstract. High amount of relevant information is contained in reports stored in the electronic patient records and associated metadata. R-oogle is a project aiming at developing information retrieval engines adapted to these reports and designed for clinicians. The system consists in a data warehouse (full-text reports and structured data) imported from two different hospital information systems. Information retrieval is performed using metadata-based semantic and full-text search methods (as Google). Applications may be biomarkers identification in a translational approach, search of specific cases, and constitution of cohorts, professional practice evaluation, and quality control assessment.

Keywords. Information retrieval, electronic patient record, ontology, indexing

1. Introduction

As of today, medical informatics is going through a strong shift of paradigm. Data from medical reports are most of the time not structured. They are however of high medical value as they correspond to the expert interpretation of the clinician. These information sources consist therefore in a data repository potentially highly relevant for scientific research. From this perspective, the combined exploitation of metadata and information contained in exam reports on a large data corpus by tailored search engines becomes very relevant. The main objective of the R-oogle project is to implement a system aimed at offering to researchers the possibility to exploit, for a scientific goal, the huge amount of medical data that are metadata and exam reports with the ease of Google[tm] search engine, combining search methods for structured data elements and full text. The objective of R-oogle is to implement a platform consisting of: (i) A Clinical data warehouse (CDW) containing a large collection of patient data coming from different hospitals (ii) A search engine combining semantic search and full text search (with semantic enrichment) exploiting information contained in the exam reports.

2. Background

Building and exploiting a multi-domain medical CDW built from Electronic Health Records (EHR) is currently an active research topic, e.g. Rubin et al [1] or the open

source platform (I2B2) [5] developed by Harvard. Information retrieval (IR) is also a very active research domain and the medical field appears to be very suitable for such techniques, despite medical documents were, as for semantic ambiguity, more suitable for indexing than documents from other domains [2]. Ehrler, Ruch et al proposed in 2007 [3] an approach based on the full text indexing of medical reports, exploiting the context in which can be found information on the report's structure (motive of the exam, description, conclusion). In this vein, Spat and Cadonna [4] described a system for document retrieval based on the metadata enriched by automatically extracted concepts from the reports indexed in German. A literature review done in 2008 [6] on 174 publications showed the increased scientific activity around IR on EHRs, despite significant storage and processing limitations to implement systems outside of experimental context.

In a previous publication, we evaluated contribution of full-text search versus encoding data with the Diagnosis Related Groups (DRG)) in an epidemiological study context [7]. This study highlighted the contribution of full-text search to the DRG database search. In this work, we added semantic enrichment to the search engine, and we compared document retrieval with or without semantic enrichment. Shultz and al. developed MorphoSaurus, a German concept-based document search engine, connected to hospital information system in order to support search across the whole corpus of patient discharge letters and other clinically relevant documents [8].

3. Material and Methods

Building the biomedical data warehouse: A "star" data base schema is used. CDW includes patients, full-text documents and structured documents linked to thesauri. Each document is defined as a medical production by a medical department on one patient, at a specific date. Data come from different software, so common kind of information have been defined to all documents (i.e. patient's identification number, date, author, title, type of document, or text). Into some full-text documents, zones of specific texts can be extracted: motive, results, conclusion, technique, exam, and medical issues. These metadata have been added in the document description to help targeting a search. Structured data have been recorded into a separate table with a scalable architecture to allow integration of heterogeneous data. Each data is attached to a document as a data element, wherever applicable a thesaurus (such as the Logical Observation Identifiers Names and Codes for laboratory data (LOINC)), combined with a data value (such as a date, a number or text), wherever applicable a thesaurus for the data value (such as the French procedure classification (CCAM) used for encoding the DRG data).

The CDW was implemented with the Oracle® database management system. Scripts using the Open Source Talend ETL were programmed to feed the database. The first step consisted in loading documents produced between 2005 and July 2010 from different sources. These scripts will be used to feed the CDW "on the fly".

Patient identity mapping: Data sources come from two distinct medical hospitals, so we implemented an algorithm mapping the different identification numbers of the two establishments. The mapping is successively realized following four methods, from the most accurate (identical surname or maiden name, first name, sex and birth date) to the least one (surname or maiden name, sex, date of birth and the first four letters of the first name). The accuracy criterion is applied when mapping so that

manual mapping of patient identification numbers is allowed for the least accurate method. The identification numbers of the Rennes Cancer Institute are stored to ease patients mapping in future imports and to provide a way to search the data warehouse by its own identification numbers.

Semantic enrichment and indexing documents: Medical concepts are extracted from reports using NOMINDEX [9], a concept extraction tool based on the ADM tool (Aide au Diagnostic Médical). Medical concepts associated to each document are stored in the structured part of the data warehouse. Concepts are restricted to the MeSH thesaurus. Lucene, an Open Source full-text engine written in Java by Apache, performed document indexing with tri-grams [10]. Full-text search is optimized and very elaborate queries may be composed (such as boolean, fuzzy, joker) on the entire information contained in a document or on part of the metadata (e.g.: *+(mediator benfluorex) +conclusion:valvulop*

All French synonyms issued from the UMLS (Unified Medical Language System Metathesaurus) and the hierarchical parents of concepts are also integrated in indexing. Subsomption is therefore taken care of during the indexing of the document and not at the query time, the whole Lucene capabilities remaining intact for allowing complex queries.

Developing multi-criteria search engine: The first part of the engine consists in a high-level search: patient's sex, age range and medical department producing the document. The second part consists in a full-text search using Lucene syntax and the third part consists in the structured search. This way the user will have the capability to build structured queries on whole disjoined documents or only on specific parts of documents (e.g. motive, conclusion).

Information retrieval assessment: We placed the evaluation in a non-habit user perspective, without complex structured queries. The corpus of assessed documents consisted of the textual part of multidisciplinary prostate cancer reviews. This one was selected for the natural language properties of disease descriptions contained into this part, because our main objective was to assess the potential benefit of documents' semantic enrichment in a full text search problematic. The assessment was performed with two contextual designs, one where search terms were presumed to be found with a high level of occurrence in the whole documents (i.e. prostatic adenocarcinoma), and one with a low level of search term occurrence (i.e. heart failure). Four types of search process have been conducted on the corpus of documents: one with and one without semantic enrichment by the search engine, one exact match term search by a human medical expert, and one textual search with clinical interpretation of each document by a human medical expert too. We used recall, precision with their 95% confidence interval and f-measure to describe the whole assessment results.

4. Results

Status of the data warehouse and the search engine: The CDW was fed by six sources of heterogeneous data, five sources managed by Rennes hospital (pathology reports, radiology reports, hospitalization/consultation reports, gastrointestinal endoscopy reports, DRG data) and one source managed by Rennes Cancer Institute (imaging reports). We are not yet reaching completeness, and the CDW contains as of today 2 115 581 documents. The results of a query are displayed as a table of the retrieved documents. When the user opens a document, the text searched through the full-text

search engine is highlighted to ease validating the document (like Google cache). The user can then display all the documents on the patient as a sortable list or according to a temporal representation modeled as a Gantt diagram. The user has then the possibility to add the selected patient in a "cohort" he/she is building or, on the contrary, rule out this patient. All the documents of enrolled or ruled-out patients for the current cohort will be displayed as attached to patients already enrolled or ruled out, and won't appear to to be checked again. Other functionalities will provide another general view of the CDW.

Table 1. Results of the evaluation in the high term prevalence context

	TP	FP	TN	FN	Recall [95% CI]	Precision [95% CI]	F-measure [95% CI]
*nWES search engine * HEM search*	142	1	100	15	0,90 [0,86-0,95]	0,99 [0,98-1,00]	0,95
*WES search engine * HEM search*	157	26	75	0	1,00	0,86 [0,81-0,91]	0,92
*nWES search engine * HCI search*	141	2	45	70	0,67 [0,60-0,73]	0,99 [0,97-1,00]	0,80
*WES search engine * HCI search*	180	3	44	31	0,85 [0,81-0,90]	0,98 [0,97-1,00]	0,91

nWES : without semantic enrichment ; WES : with semantic enrichment; HEM : human exact match; HCI: human clinical interpretation; TP: true positive; FP: false positive; TN: true negative; FN: false negative; CI: confidence interval.

Evaluation results: We evaluated the contribution of semantic enrichment of the document database for the search engine. Two hundred and fifty eight notices of prostatic cancer multidisciplinary meetings have been analyzed. Related to the high term prevalence context using the combined query: "adenocarcinoma" AND "prostatic", the search without semantic enrichment retrieved 143 documents, the search with semantic enrichment retrieved 183 documents, the human exact match search retrieved 157 documents and the human search with clinical interpretation retrieved 211 documents. Related to the low term prevalence context using the combined query: "heart" AND "failure", the search without semantic enrichment retrieved 0 documents, the search with semantic enrichment retrieved 7 documents, the human exact match search retrieved 0 documents and the human search with clinical interpretation retrieved 8 documents. To compare complete results of the search engine and the human evaluation, see table 1 and table 2.

Table 2. Results of the evaluation in the low term prevalence context

	TP	FP	TN	FN	Recall [95% CI]	Precision [95% CI]	F-measure [95% CI]
*nWES search engine * HEM search*	0	0	258	0	-	-	-
*WES search engine * HEM search*	0	7	251	0	-	0	-
*nWES search engine * HCI search*	0	0	250	8	0	-	-
*WES search engine * HCI search*	4	3	247	4	0,50 [0,15-0,85]	0,57 [0,21-0,94]	0,53

nWES : without semantic enrichment ; WES : with semantic enrichment; HEM : human exact match; HCI: human clinical interpretation; TP: true positive; FP: false positive; TN: true negative; FN: false negative; CI: confidence interval.

5. Discussion - Conclusion

In this paper we demonstrated the feasibility and applicability of a CDW that benefits from full-text search capabilities, as opposed to I2B2 that is mainly based on a structured data approach and does not address French NLP specificities. We however applied NLP methods for annotating reports using relevant concepts as well as synonyms and ancestors. This enrichment permitted to deal with subsumption issues using full-text search, and also to cluster cases by projecting reports on a MeSH hierarchy. Results show that semantic enrichment provides a better recall while precision stays quite stable. This is the best situation for prescreening patient for clinical trials. Prescreening aims at spotting patients better too often than too seldom. Knowing that each document returned by the searching engine, can be quickly checked and validated by an end-user (with the keyword highlighting feature), it is then really easy to rule out un-relevant documents. Even though temporality is taken care of through the Gantt diagram representation and the multi-documents search within a single hospitalization is possible, the search engine could be improved to address specific situations, e.g. retrieving hospital-acquired infection cases would require the detection of a positive bacteremia at 48h or later after the admission. We encountered some issues related to reference terminologies we used to encode patient data in the CDW (e.g concerning the integration of lab test, as for Cormont et al [11], we were confronted with the lack of coverage and the missing French translations of LOINC). As perspective, we are currently working a web portal intended to research technicians and investigators. This portal aims at managing the workflow to access to the CDW.

Acknowledgement: We would like to thank the CRITT Santé Bretagne for their financial support and Delphine Rossille.

References

[1] Rubin DL, et al. A data warehouse for integrating radiologic and pathologic data,. *J Am Coll Radiol*, 2008. 5(3): p. 210-7.
[2] Ruch P, et al. Comparing general and medical texts for information retrieval based on natural language processing: an inquiry into lexical disambiguation. *Stud Health Technol Inform*, 2001. 84(Pt 1): p. 261-5.
[3] Ehrler F, et al. Challenges and methodology for indexing the computerized patient record, *Stud Health Technol Inform*, 2007. 129(Pt 1): p. 417-21.
[4] Spat S, et al. Enhanced information retrieval from narrative German-language clinical text documents using automated document classification, *Stud Health Technol Inform*, 2008. 136: p. 473-8.
[5] Murphy SN, et al. Integration of clinical and genetic data in the i2b2 architecture. *AMIA Annu Symp Proc*. 2006:1040. Available at: Consulté novembre 30, 2010.
[6] Meystre SM, et al. Extracting information from textual documents in the electronic health record: a review of recent research, *Yearb Med Inform*, 2008: p. 128-44.
[7] Cuggia M, et al. A full-text information retrieval system for an epidemiological registry, *Studies in Health Technology and Informatics*, vol. 160, n°. 1, p. 491-495, 2010
[8] S. Schulz, Daumke P, Fischer P, Müller ML. « Evaluation of a document search engine in a clinical department system », *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, p. 647-651, 2008.
[9] Happe, A, et al. Automatic concept extraction from spoken medical reports, *Int J Med Inform*, 2003. 70(2-3): p. 255-63.
[10] Hatcher E, et al. *Lucene in action,* Action series. Manning Publications Co., Greenwich, CT, 2004.
[11] Cormont S, et al. Construction of a dictionary of laboratory tests mapped to LOINC at AP-HP, *In Actes AMIA Annual Fall Symposium* 2008, page 1200, Washington, DC, novembre 2008. AMIA