ABSTRACT

While the digitization of medical documents has greatly expanded during the past decade, health information retrieval has become a great challenge to address many issues in medical research. Information retrieval in electronic health records (EHRs) should also reduce the difficult tasks of manual information retrieval from records in paper format or computer. The aim of this article was to present the features of a semantic search engine implemented in EHRs. A flexible, scalable and entity-oriented query language tool is proposed. The program is designed to retrieve and visualize data which can support any Conceptual Data Model (CDM). The search engine deals with structured and unstructured data, for a sole patient from a caregiver perspective, and for a number of patients (e.g. epidemiology). Several types of queries on a test database containing 2,000 anonymized patients EHRs (i.e. approximately 200,000 records) were tested. These queries were able to accurately treat symbolic, textual, numerical and chronological data.

CCS Concepts

• Information systems → Specialized information retrieval; Query representation;

Keywords

Automatic Indexing, Electronic Health Record, Information Retrieval, Search Engine

1. INTRODUCTION

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR ’16 July 21, 2016, Pisa, Italy
© 2016 ACM. ISBN:
DOI:
and analysis with database tools. Data semantics is particularly important as it derives from the concrete healthcare providing process in hospitals. EHR data is mainly composed of several key entities semantically related to one another: (a) patient, (b) hospital, (c) stay and then (d) the "classical" and more basic levels (procedures, diagnosis related group (DRG) coding, lab tests, reports, metadata from reports etc.). As a consequence, IR from EHR is more difficult and different when compared to the "classical" IR.

In this context, the aim of this study was twofold. First, describe a conceptual model which represents the conceptual and intuitive representation that non-IT medical provider users can have of EHR data. Secondly, describe a query language (QL) used to query those data and providing users the possibility to build queries accessing the entire set of EHR entities by taking advantage of the semantic network of entities. This search engine was funded by the French National Agency (TecSan program) in the Retrieval and Visualization In Electronic Health records (RAVEL) project.

2. MATERIALS

2.1 EHR Data Sources

A corpus of 2,000 anonymized patients and 200,000 reports from Rouen University Hospital (RUIH) 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, natural language processing tools developed by the Vidal and Lille teams of the RAVEL project were also used to partially re-structure the unstructured data via multiterminological automatic indexing (AI) using more than 65 terminologies partially or totally translated into French.

2.2 EHR Conceptual schema and data model

The underlying database of the system is based on a generic Entity-Attribute-Value (EAV) physical data model [3]. This data model is able to integrate all types of data without structural changes to the data model (e.g. without adding columns or tables). This physical database structure enable to store any kinds of data in only a few tables. This helps to optimize IR, maintain the database and manage heterogeneous data types. As described in the thesis by A.D. Dirieh Dibad [2], EHRs are structured and organized in the four key concepts: (i) patient, (ii) hospital, (iii) stay, (iv) medical procedures, laboratory tests. A dedicated CDM using concepts i, iii and iv was designed to abstract the EHR data contained in the physical database data model (Figure 1). The query language syntax is patterned on that CDM instead of the physical database schema which provides the search engine with semantic features and capabilities.

3. METHODS

3.1 Query Language Description

The specific QL syntax is based on the CDM. Hence, building a query only requires real-life knowledge of existing entities in the database, their properties and their relationships with each other (e.g. "a patient undergo a medical test"). This query language has three main characteristics:

Semantic Information Retrieval capabilities: The QL is built with an entity-oriented vision. It enables semantic information retrieval since it provides the ability to display and query EHRs semantically related entities on any level (patient, stay, procedure, biological test etc.). It can also deal with multiple terminologies and hierarchical relationships.

Scalability: It is a scalable and flexible QL. It can automatically handle modifications on the CDM. More technically, new conceptual entities, attributes or relationships between entities are automatically taken into account directly from the database and without any search engine modification. This enabled an easy and rapid extension to omics data (genomics, metabolomics, proteomics, methylation etc.) [1].

Comprehensive querying capabilities: The full scope of entities can be queried using constraints built upon several types of data:

- Textual and symbolic data (e.g. the query patient(gender="M") target male patients).
- Numerical data (e.g. the query medicalTest(numericResult > 6 AND numericResult <= 6.25) targets lab tests with a result value between 6 and 6.25).
- Chronological (e.g. the query stay(entryDate = 2010-03-10) targets stays, which entry date is 2010-03-10).

All comparators and operators available are specified in Table 1.  

3.2 Basic Querying

The query language is basically composed of nested syntactical units with the following syntax ENTITY(CONSTRAINTS_CLAUSE). ENTITY can correspond to any kind of entity of the CDM (e.g. patient, stay, medicalUnit etc.) and specify the type of object that the search engine should return (or target when nested). For instance, the queries patient() and medicalUnit() would respectively return all the patients and all the medical units of the database. The CONSTRAINTS_CLAUSE is a boolean expression enabling to apply constraints to the targeted ENTITY. For instance, in the
query patient(birthDate=1937-01-01 AND gender = "M") uses the two attributes birthDate and gender of the patient entity to return all male patients born on 1937-01-01. stay(leavingDate-entryDate>=10) will return stays with a duration of 10 days or more.

### 3.3 Semantic querying

The strength of the query language originates from its ability to deal with nested syntactical units. For instance the query stay(patient(id = "DM_PAT_42")) targets all stays associated with at least one relationship to the patient number 42. More complex queries can be performed by using the relationships between these entities. Some example queries are given in Table 2. This nesting functionality allows the exploitation of the relationships between entities and thereby enables to build queries based on the full semantic network. TheQL has other querying capabilities: full text search, minimum and maximum on numerical data, hierarchical expansion, chronological and temporal queries.

### 3.4 Search Engine Process

The internal process of the search engine is composed of three main stages. Stages 1 and 2 are dedicated to build a comprehensive and computer-processable representation of the input string query. Stage 3 stands for the core of the search engine and consists of the precise querying of the EHRs data to return a list of entities.

**Stage 1:** Query parsing: A parser was designed to comprehensively define the query language syntax requirements. The parser matches and extracts this syntax respectively

**Stage 2:** Tree representation of the query: Stage 2 provides a computer-processable representation of a) the Boolean logic and b) the nested structure of the query. Stage 2 is particularly important as semantic search capabilities rely on it. A tree representation is an optimal computer processable structure to achieve that goal.

**Stage 3:** SQL query building: A SQL query is generated recursively from root nodes to leaf nodes of the tree built in Stage 2 and executed to return the list of entities.

### 4. RESULTS

#### 4.1 RAVEL project use cases

| Table 1: Types of data handled by the search engine |
|-------------------------------|-----------------|-----------------|
| Data type | Available operators | Available comparators |
| character string data | None | = (equal), != (not equal), * (wildcard) |
| Numerical data | + (add), - (subtract), * (multiply), / (divide) | <= (less or equal), => (greater or equal) |
| Chronological data | +, - | >= (greater or equal) |

#### 4.2 Comparison to I2B2 workbench

Several use cases were successfully answered in the RAVEL project:

**Use case 1:** Visualize over time the neutrophil rate of a patient with rheumatoid arthritis

**Use case 2:** Produce all the medical reports containing the concept of metastasis

**Use case 3:** Retrieve all stays where "REMICADE" (infliximab) was used.

The use cases resolution required to use: AI in medical records, full text search, and multiple terminological resources. Some of the queries used to answer these three use cases are shown in Table 3.

### 4.2 Comparison to I2B2 workbench

| Table 3: Example of RAVEL search engine queries |
|----------------|-----------------|
| Example | Description |
| stay(patient(id="DM_PAT_21") AND procedure(label="BLOOD SAMPLE")) | Patient 21 stays in which a blood sample procedure was taken. |
| medicalUnit(stay(patient(id="DM_PAT_21") AND procedure(label="BLOOD SAMPLE"))) | Medical units of the patient 21 stays in which a blood sample was taken. |
| biologicalTest(patient(id="DM_PAT_1078") AND exe(label="Platelets") AND 10*numericResult<lowerBound) | Patient 1078 platelet tests with a result more than 10 times lower than normal level. |
| procedure(ccamMP(id="CCA_\_AM_EQQM006") AND procedure=Date="MAX") | The last procedure coded with EQQM006 |
In this study, a search tool dedicated to retrieving health information into an EHR has been presented. This search tool is able to adapt to any CDM and thus address a large variety of issues. Its specific query language provides practical and flexible querying capabilities but remains difficult to grasp for health professionals.

6. CONCLUSION

In this study, a search tool dedicated to retrieving health information into an EHR has been presented. This search tool is able to adapt to any CDM and thus address a large variety of issues. Its specific query language provides practical and flexible querying capabilities but remains difficult to grasp for health professionals.

7. REFERENCES


