SEMCARE - SEMantic Data Platform for HealthCARE

P. Daumke, C. Haid, C. Riede, M. Ihle, L. Mertens

Averbis GmbH, Freiburg

Introduction: The definition of patient cohorts by combining clinical patient data is a precondition for different scenarios like feasibility studies, patient enrolment in clinical studies, quality assurance within the hospital, selection of cases tor teaching purposes and identification of undiagnosed patients [1]. As large parts of patient-level data in electronic health records (EHRs) are only available as free text, text mining tools are a prerequisite for this process [2]. The research project SEMCARE „SEMantic Data Platform for HealthCARE", funded by the EU for a two-year period, builds a semantic data platform to identify patient cohorts. This platform combines the power of full text search with text analytics and semantic web technologies for a hybrid semantic full-text search.

Use cases: The three participating European health centres (Erasmus Universitair Medisch Centrum Rotterdam, Medical University of Graz and Saint George's University of London) have agreed on one initial, general use case on which they will focus during the project. The use case is called 'Risk Stratification and Differential Diagnosis of Patients suffering from transient loss of consciousness'. A number of phenotypic features can help to stratify patients under risk, most of which are available from routine assessment and investigation. Using a semantic data platform, high-risk patient cohorts will be identified based on these criteria scattered in heterogeneous clinical data contained in electronic healthcare records (EHRs).

Methods: The information extraction methodology we develop and describe is generic and could be adapted to other patient cohorts and medical inquiries and allow an interaction with third parties tools (e.g. tranSMART, QlikView). This will be enabled by providing a common data model (e.g. the i2b2 star schema) that can easily be used by third party applications using the same data model.

The Extraction-Transfer-Load (ETL) tool Talend aggregates the patient data from different clinical data sources and stores them into one data store, a PostgreSQL database with the I2B2 schema.

In addition, Talend transfers the unstructured patient data to the Averbis Extraction Platform (AEP) for further text analysis. The AEP consists of a number of modular text analysis components, so called Analysis Engines (AEs), connected together with the Apache UIMA[1] framework. Rule-based as well as statistical methods, in combination with medical terminologies, are used to detect relevant diagnoses, symptoms, laboratory values or procedures in the patient data.

The obtained, structured data is then stored in both the relational PostgreSQL database and a free text search engine. For the Averbis Search Platform the data is indexed to guarantee fast access in the query process and to permit faceted search, which means that the search results are organized according to a faceted classification system. The user is so enabled to narrow down the search results by filtering with independent categories.
By combining indexing of data in the database and search platform the user benefits from the power of both areas. In the SEMCARE platform, both complex queries on structured data, and search engine like queries on free text is possible.

[1] http://opennlp.apache.org

Results: Although the software is still in development a demonstration of a prototype during the GMDS/IMIA Workshop "Research Databases“ is possible. Results based on the use cases are to be expected in 2015.

[1] C. Gallenmüller, S. Ullherr, M. Greeff, V. Stümpflen, T. Klopstock; Semantisches Textmining als neuartiger Ansatz zur Identifikation bisher undiagnostizierter Patienten mit Niemann-Pick-Erkrankung Typ C. 85. Kongress der DGN mit Fortbildungsakademie; Sept. 2012
[2] F. Köpcke, B. Trinczek, R. W. Majeed, B. Schreiweis, J. Wenk, H.-U. Prokosch et al. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. BMC Medical Informatics and Decision Making 2013, 13:37