Chapter 2
Pre-Processing Highly Sparse and Frequently Evolving Standardized Electronic Health Records for Mining

Shivani Batra
KIET Group of Institutions, Delhi-NCR, India

Shelly Sachdeva
https://orcid.org/0000-0003-4088-1271
National Institute of Technology, Delhi, India

ABSTRACT

EHRs aid in maintaining longitudinal (lifelong) health records constituting a multitude of representations in order to make health related information accessible. However, storing EHRs data is non-trivial due to the issues of semantic interoperability, sparseness, and frequent evolution. Standard-based EHRs are recommended to attain semantic interoperability. However, standard-based EHRs possess challenges (in terms of sparseness and frequent evolution) that need to be handled through a suitable data model. The traditional RDBMS is not well-suited for standardized EHRs (due to sparseness and frequent evolution). Thus, modifications to the existing relational model is required. One such widely adopted data model for EHRs is entity attribute value (EAV) model. However, EAV representation is not compatible with mining tools available in the market. To style the representation of EAV, as per the requirement of mining tools, pivoting is required. The chapter explains the architecture to organize EAV for the purpose of preparing the dataset for use by existing mining tools.

DOI: 10.4018/978-1-7998-2742-9.ch002

INTRODUCTION

Electronic Health Records (EHRs) provide a digital support to the healthcare industry. A database of EHRs assembles health data of a patient from various departments of a healthcare organization including administration, pharmacy, clinical, radiology, laboratory and nursing. Contents within EHRs can be structured, semi-structured, unstructured, or a hybridization of these. For example, the contents of EHRs can be in the form of plain text, basic types (such as state variable and Boolean), time, date, date-time (including partial date/time), paragraphs, coded text, encapsulated data (such as parsable and multimedia content), measured quantities (providing units with values), uniform resource identifiers (URI) and container types (such as set and list) (Sachdeva S. & Bhalla S., 2012). EHRs aid in exchanging patients' health information electronically from one hospital to another. This electronic exchange of EHRs diminishes the burden of patients to carry reports printed on papers and other health related documents. However, exchange of EHRs needs to be semantic interoperable i.e. communicating parties must depict the same meaning of the exchanged EHRs data without any ambiguity.

Semantic Interoperability

To attain semantic interoperability, distinguished standard organizations, such as ISO (ISO 13606-1. 2008. Health informatics -- Electronic health record communication -- Part 1: Reference Model,.; ISO/DIS 13606-2 - Health informatics -- Electronic health record communication -- Part 2: Archetype interchange specification), openEHR (Beale T., Heard S., Kalra D., & Lloyd D., OpenEHR architecture overview, 2006), and HL7 (Health Level Seven International - Homepage) suggest adopting a dual model approach for the management of EHRs in an information system. Dual model approach segregates the information regarding the structure of various medical concepts (such as blood pressure, thyroid, and body mass index) from the knowledge about constraints on different attributes (that belong to an underlying medical concept). The first layer of the dual model approach, i.e., Reference Model (RM) (Beale T., Heard S., Kalra D., & Lloyd D., The OpenEHR Reference Model, 2007) is defined by various IT experts in the form of numerous classes that describe all possible data type and data structures that an EHR can use. RM aids in capturing the complex structure of EHRs. The second layer of the dual model approach, i.e., Archetype Model (AM) (Beale T., The openEHR archetype model-archetype object model, 2008) enables various clinical experts to portray their knowledge regarding various medical concepts in terms of participating attributes, their data types, and any other constraints such as, ranges and cardinality. AM delivers the maximal definition of a medical concept in form of an artefact known as archetype. Archetypes are released on a standard online library after a rigorous review process. Further, any revision in the definition of an existing archetype is released as a new version. Thus, adoption of archetypes enables semantic interoperability and capturing any future evolution. However, evolution within the medical concept also needs to be reflected to the database level. Moreover, EHRs are characterized by highly sparse behavior. Thus, there is a need of database that can efficiently store standardized EHRs data considering sparseness and frequent evolution. Moreover, database used for storing EHRs must possess the compatibility with mining tools for successful implementation of analytics.