Health Policy and Technology (▮▮▮▮) ▮, ▮▮▮–▮▮▮

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/hlpt

Organizing standardized electronic healthcare records data for mining

Shivani Batra, Shelly Sachdeva*

Jaypee Institute of Information and Technology University, Sector-128 Jaypee Wish Town, Village Sultanpur, Noida 201304, Uttar Pradesh, India

KEYWORDS
Entity attribute value model;
Electronic healthcare records;
Mining;
Preprocessing

Abstract
Digitized and standard health data is highly desirable in healthcare domain. Current research started with the aim of mining standardized electronic health records (EHRs) data. It is observed that the available column-modeled database lacks in volatility and ability to deal with sparseness. Due to highly sparse nature of EHRs, there is need of organized standard based EHRs data to ascertain the benefit of mining and also to improve storage capacity. This paper proposes an architecture for preprocessing of data that deals with the issues of volatility and sparseness, and then transforming the data in a form which is desirable for data mining. The architecture proposed in this paper can be used by the data administrator to organize standardized EHRs for the application of any type of mining which can reveal knowledge to benefit various healthcare users, such as, doctors, patient and insurance company users. Simulation of proposed architecture is done on a sample set of 6700 instances of EHRs. Results demonstrate benefits of our approach in terms of storage required and time consumed (reduced by a factor of 2.5 approximately) to retrieve the desired piece of information. Consequently, data mining techniques capable of automatically extracting useful knowledge to provide clinical decision support may be applied. The decision support thus developed can be used by stakeholders, policy makers and other persons involved directly or indirectly with EHRs to take accurate and calculative decisions.

Introduction

Electronic health records (EHRs) refers to the healthcare records of an individual which are stored electronically. It is a systematic documentation of a patient's medical history and care. Nowadays usability of EHRs is limited not only to doctors, it has a great impact on decision making of stakeholders and policy makers too. EHRs integrate all the data (administration, nursing, lab, clinical, radiology and pharmacy) related to a patient in single record. EHRs have eliminated the need of carrying a large number of papers on which health records were maintained. EHRs data have

*Corresponding author. Tel.: +91 9810490704.
E-mail address: shelly.sachdeva@jiit.ac.in (S. Sachdeva).

http://dx.doi.org/10.1016/j.hlpt.2016.03.006
2211-8837/

Please cite this article as: Batra S, Sachdeva S. Organizing standardized electronic healthcare records data for mining. Health Policy and Technology (2016), http://dx.doi.org/10.1016/j.hlpt.2016.03.006

multitude of representations [1,2]. The contents may be structured, semi-structured, unstructured, or a mixture of all three. These may be plain text, coded text, paragraphs, measured quantities (with values and units); date, time, date-time (and partial date/time); encapsulated data (multimedia, parsable content); basic types (such as Boolean, state variable); container types (list, set); and uniform resource identifiers (URI). EHRs serve as a basis for planning patient care. The healthcare providers can access the medical history of a patient. EHRs collection can be mixed to extract interesting patterns, group entities classes, or decide whether an EHR satisfies some given property. Mining EHRs reveal knowledge which can benefit mankind in several ways, such as evaluation of treatment effectiveness; management of healthcare; customer relationship management; and detection of fraud and abuse [3]. The main concerns of EHRs considering mining are lack of standardization, sparseness and volatility.

Need of standardization

With an advancement in technology, nowadays all organizations maintain their records electronically. However, there is a difference in the schema being followed by all organizations. For example, a health organization dedicated for treating eyes will maintain records related to eyes only and not about heart. Moreover if two organizations are maintaining records of heart only, there may be a difference in the attributes, order of attributes or name of attributes (i.e. ambiguous terms). So, a standardized method is required to store the data in a common format. Other aspects such as acceptability, interoperability, usability, accuracy and space requirement are also required along with standardization [2,4]. Moreover, the original semantics of data is required to retain irrespective of the point of access. Therefore, the current research focus on standards related to semantic interoperability of EHRs.

Sparseness

Ideally, when data of a patient is maintained in an EHR, all attribute fields should be recorded, while in a real scenario only some details are recorded and others are kept null. For example, if a disease is diagnosed by the symptoms, the patient does not require to undergo any lab or radiology test. Consequently, the fields corresponding to lab and radiology test are kept null. This leads to sparseness. Due to presence of sparseness, a huge amount of memory is wasted. Authors in the current research have aimed to organize standardized EHRs in a storage efficient manner to achieve benefits of mining.

Volatility

Traditional method of organizing data in various columns is the simplest way of modeling data. However, column-modeled approach has limited number of columns which can be defined. For example, the maximum number of columns in DB2 and Oracle is 1012 [5]. Whenever a new concept arises in medical field, clinical information system needs to be modified to accommodate that concept, which might not be possible in column-modeled approach. This is the problem of volatility. Authors propose an architecture in the current research which provides the flexibility of adding any number of attributes to the existing EHRs database which is to be organized for the purpose of mining.

The research started with the objective of mining data/clinical content used for treating patients (based on semantic interoperable standards). But considering the issues of sparseness and volatility along with the lack of standardization, organization of EHRs in a storage efficient manner is observed as a main challenge. Thus, there is a need to first prepare and organize the data before achieving the benefits of data mining. The current research focus on organizing standard based EHRs to be used effectively for the purpose of mining. This paper is organized as follows. Details of storing standardized EHRs in a storage efficient manner are described in Section 2. Section 3 provides details about the present scenario of mining EHRs database. Section 4 gives the details about how standardized EHRs are stored in current research. In Section 5, an architecture is proposed to mine the data stored as per the storage mechanism adopted. Section 6 presents the comparative results of the architecture proposed earlier in [2] and the architecture proposed in the current research. Section 7 throws some light on related work. Section 8 finally concludes the research work with future scope.

Standardization and data storage for EHRs

In today's scenario, every organization maintains their records, but schema followed by every organization may not be same. Absence of common standard attribute set hurdles in mining EHRs database and in other issue like interoperability. Several standard development organizations (openEHR, CEN, ISO and HL7) [6-10] are working to have guidelines which can be used to build a common format for the EHRs data. Standard defines constraints on attributes for each and every medical concept. For example, BP concept consists of five attributes namely systolic, diastolic, mean arterial, pulse pressure and comments (among which only comments is of 'TEXT' (textual) type and other attributes are of 'QUANTITY' (quantifiable) type).

Data mining (DM) can be applied to various sectors. However, application of DM to healthcare domain can benefit mankind a lot as good health is very important for every living being [4]. Providing an efficient decision support can help doctors in making decision about human life which is very critical such as, in case of emergency and epidemics. The complex and voluminous data generated by healthcare transactions are transformed into useful information for decision making through DM. Tomar et. al. presented a survey on various data mining techniques like Association, Clustering and Classification in [11]. Various advantages and disadvantages of mining techniques are explored in the study. Various data mining approaches used in healthcare by different researchers along with the accuracy achieved are also summarized.

Approaching standardization – dual model approach

Three methodologies to build EHR system described in [12] are the unstructured approach, the "BIG" model approach and the generic approach. Unstructured approach follows no rules to

Please cite this article as: Batra S, Sachdeva S. Organizing standardized electronic healthcare records data for mining. Health Policy and Technology (2016), http://dx.doi.org/10.1016/j.hlpt.2016.03.006