Text Mining Applied to Electronic Medical Records: A Literature Review

Luís Pereira, School of Technology and Management, Polytechnic Institute of Leiria, Leiria, Portugal

Rui Rijo, School of Technology and Management, Polytechnic Institute of Leiria, Leiria, Portugal

Catarina Silva, School of Technology and Management, Polytechnic Institute of Leiria, Leiria, Portugal

Ricardo Martinho, School of Technology and Management, Polytechnic Institute of Leiria, Leiria, Portugal

ABSTRACT

The analysis of medical records is a major challenge, considering they are generally presented in plain text, have a very specific technical vocabulary and are nearly always unstructured. It is an interdisciplinary work that requires knowledge from several fields. The analysis may have several goals, such as assistance on clinical decision, classification of medical procedures, and to support hospital management decisions. This work presents the concepts involved, the relevant existent related work, and the main open issues for future research within the analysis of electronic medical records, using data and text mining techniques. It provides a comprehensive contextualization to all those who wish to perform an analytical work of medical records, enabling the identification of fruitful research fields. With the digitalization of medical records and the large amount of medical data available, this is an area of wide research potential.

Keywords: Data Mining, Electronic Medical Records, Healthcare Informatics, ICD Codes, Machine Learning, Text Mining

INTRODUCTION

Analysis of medical records using text mining is a complex field demanding considerable time and effort (Brown & Trimble, 2000). Many sources of data such as symptoms, exams, patient history, procedures, treatments, and medications need to be taken into account for a correct scrutiny. Additionally, this analysis requires knowledge in several different fields, namely, the clinical specific area, data mining, text mining, medical records, hospital and clinical procedures.
Developing a process to mine medical records can be troublesome. Usually these records are in free text, have an unstructured format and a specific and complex domain. In fact, each physician has usually his own approach of describing events or symptoms, depending on his previous learning experiences and medical practices (Weiner, Swain, Wolf, & Gottlieb, 2001). Medical doctors use a particular language, which usually demands additional tools to interpret the designated terms and symptoms and to extract semantic information from medical records (Fernandez et al., 2004).

Another challenge is the huge amount of data. Online medical information is increasing each day, generating massive amounts of electronic data. Medical Literature Analysis and Retrieval System Online, or MEDLARS Online (MEDLINE) databases contain more than 12.5 million records, growing at the rate of 500,000 new citations each year (Cohen & Hersh, 2004). The increase of electronic medical records provides a huge quantity of information to process and, at the same time, offers an opportunity for developing research in order to reduce the time and effort to find and classify a correct diagnosis for a particular patient.

There are several types of medical data, e.g., patient demographics, medical history, and lab tests that can be used to different ends, such as, ordering, managing, scheduling, and billing. This data helps physicians diagnosing and treating patients and the correct management of resources (Ludwick & Doucette, 2008).

Furthermore, physicians use standard codes to describe diagnoses, procedures or treatments, which may include features such as symptoms, diseases or disorders. International Classification of Diseases (ICD) and Systematized Nomenclature of Medicine Clinical Terms (SNOMED) are two examples of standard codes (Schulz, Rector, Rodrigues, & Spackman, 2012). This classification is critical for sharing information among clinicians and also for billing. The classification process is not straightforward, representing another challenge in the area.

In the following sections, we introduce the relevant concepts of data mining, text mining, medical records, and international standard classifications for diagnose, clinical procedures and treatments, indicating important techniques, procedures and tools used. We carry on by reviewing techniques, main challenges and tools in text mining applied to medical records, decision support, health management and classification systems. Finally, relevant open issues are discussed presenting promising future research trends.

**DATA MINING BACKGROUND**

Text mining is a variation of data mining (Navathe, B., & Ramez, 2000). Thus, this section introduces the concept of data mining, its conception, techniques, learning methods, and processes that can be used and its automation.

Data mining is the process of understanding and discovering patterns in large data sets to retrieve useful knowledge (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Regarding healthcare, this knowledge can be useful to improve the quality of treatments, increasing revenues or lower costs of healthcare organization such as a hospital. Data mining is widely used in the fields of computer science, economics, communication and marketing, allowing finding important patterns on information that are, otherwise, difficult to analyse.

It is hard to say when this concept appeared. Data mining use several of Bayes algorithms created in the XVIII century (Pawlak, 2002), but only in 1989-1991 was the term introduced by researcher Gregory Piatetsky-Shapiro (Piatetsky-Shapiro, 1990).
Alerts in Healthcare Applications: Process and Data Integration

www.igi-global.com/article/alerts-healthcare-applications/2247?camid=4v1a