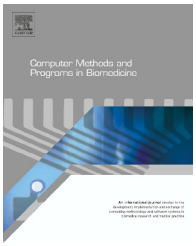
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 128 (2016) 52–68

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



Temporal data representation, normalization, extraction, and reasoning: A review from clinical domain

Mohcine Madkour[a], *Driss Benhaddou*[b], *Cui Tao*[a,*]

[a] *School of Biomedical Informatics, University of Texas Health Science Center at Houston, 7000 Fannin St, Houston, TX 77030, United States*

[b] *Department of Engineering Technology, University of Houston, 4800 Calhoun Rd, Houston, TX 77004, United States*

ARTICLE INFO

Article history:
Received 18 August 2015
Accepted 16 February 2016

Keywords:
Clinical temporal information
Temporal representation
Temporal extraction
Ontologies of time
Medical NLP

ABSTRACT

Background and objective: We live our lives by the calendar and the clock, but time is also an abstraction, even an illusion. The sense of time can be both domain-specific and complex, and is often left implicit, requiring significant domain knowledge to accurately recognize and harness. In the clinical domain, the momentum gained from recent advances in infrastructure and governance practices has enabled the collection of tremendous amount of data at each moment in time. Electronic health records (EHRs) have paved the way to making these data available for practitioners and researchers. However, temporal data representation, normalization, extraction and reasoning are very important in order to mine such massive data and therefore for constructing the clinical timeline. The objective of this work is to provide an overview of the problem of constructing a timeline at the clinical point of care and to summarize the state-of-the-art in processing temporal information of clinical narratives.

Methods: This review surveys the methods used in three important area: modeling and representing of time, medical NLP methods for extracting time, and methods of time reasoning and processing. The review emphasis on the current existing gap between present methods and the semantic web technologies and catch up with the possible combinations.

Results: The main findings of this review are revealing the importance of time processing not only in constructing timelines and clinical decision support systems but also as a vital component of EHR data models and operations.

Conclusions: Extracting temporal information in clinical narratives is a challenging task. The inclusion of ontologies and semantic web will lead to better assessment of the annotation task and, together with medical NLP techniques, will help resolving granularity and co-reference resolution problems.

* *Corresponding author.*
E-mail addresses: mohcine.madkour@uth.tmc.edu (M. Madkour), dbenhadd@Central.uh.edu (D. Benhaddou), cui.tao@uth.tmc.edu (C. Tao).

http://dx.doi.org/10.1016/j.cmpb.2016.02.007
0169-2607/

Contents

1. Introduction

Time is a universal phenomenon that has interested many disciplines of science for many years. It provides basic elements for understanding the world in its dynamics: (a) in mining actions and changes to recognize pattern evolution, and (b) describing time-oriented relations for intelligent decision-making. Similarly, time plays a major role in the clinical domain by helping understanding chronological development of clinical procedures such as diagnosis (e.g. the order in which symptoms develop), treatment (e.g. time of taking medications), and prevention (e.g., signals for pre-disease). Remarkably, researchers have avidly studied time concepts and their representations. Mathematicians formulate time theories in order to abstract elements of time and temporal entities; philosophers contest changes and dynamics since ancient time; physicists, in both Newton's and Einstein's physics, debate the notions of special time and the dimension of time.

In the clinical domain, patients' data have been collected over time and recorded in electronic health record (EHR) systems; the ever growing complexity of such patterns of data reveals challenges in handling its high dimensionality taking into consideration complex parameters such as amplitude modifications, time warping and noise. In addition, clinical data differs from other time series by the fact that observations are made at irregular time intervals, and some of them may be missed or disrupted [1]. Therefore, temporal mining will need to provide solutions and innovations in both, theoretical view, such as time parameterization and abstraction, and methodological view, such as temporal relations extraction and event calculation. Moreover, as we are heading toward the "Big Clinical Data" era, we are faced with a torrent of data generated and captured in digital form as a result of the advancement of sciences, engineering and information technology. Consequently, there is a great potential of new waves of innovation to be aroused on detailed trend analyses by taking advantage of this large-scale and high-resolution data sets. Nevertheless, the heterogeneity and the complex nature of big data make it challenging to be leveraged directly by any algorithms without intensive and manual analysis.

Additionally, approximately 80% of EHR is unstructured [2,3]. Correspondingly, temporal reasoning and interpretation will confront additional challenges as reasoning about time requires so-called "common knowledge" which can be notoriously difficult to establish (e.g. the yesterday in "yesterday she experienced some pain" could be implemented rigidly as a microsecond after midnight). Furthermore, idiosyncratically structured and disparate health information suitable for large-scale analyses as well as inference, needs to be dynamically transformed into standards. Many grounded researches have been established for better extraction of temporal information and providing guidelines for the standardization of temporal statements. In this connection, there is yet a gap between semantic inference technologies research and clinical and medical NLP approaches. We posit that the fundamental challenges that hinder the secondary use of EHR temporal data include: (1) temporal information exists in different formats (structured, semi-structured and non-structured) [4]; (2) mechanisms to harvest time for specific purposes are not formalized or readily available; and (3) Medical Natural Language Processing (Medical NLP) tools are of variable quality and completeness with respect to a given extraction purpose.

In this survey, we provide the background approaches of representation and reasoning about time-oriented aspects. It is devised in a way that even a non-specialized reader will be able to grasp. The current state of this research is inspired