ARTICLE IN PRESS
JID: KNOSYS [m5G;April 15, 2017;13:18]
Knowledge-Based Systems 000 (2017) 1–12
Contents lists available at ScienceDirect
Knowledge-Based Systems
journal homepage: www.elsevier.com/locate/knosys

A temporal model in Electronic Health Record search

Jiayue Zhang*, Weiran Xu, Jun Guo, Sheng Gao

Beijing University of Posts and Telecommunications, Beijing, PR China

ARTICLE INFO

Article history:
Received 17 October 2016
Revised 29 March 2017
Accepted 30 March 2017
Available online xxx

Keywords:
Temporal model
Dynamic time warping
Electronic Health Record search
Clustering

ABSTRACT

Electronic Health Records (EHRs) refer to a collection of patient data, including diagnosis, medical history, medication, allergies, etc., mostly contained in the form of unstructured text. EHRs are designed to capture the state of a patient over time, thus the temporal information is crucial. Most previous works processing time in EHRs narrative focused on temporal expression extraction, using textual dimension to embody the temporal dimension. In this paper, we propose to model the textual and the temporal dimension of EHRs narrative jointly. To meet the challenge, we propose to model the EHRs narrative as temporal sequential data. A novel representation framework is designed to model the clinical narrative text as document sequence, where the textual and temporal dimension are modeled simultaneously. In the framework, a dynamic time warping based measure is proposed to quantify the similarity between EHRs of different patients. To verify the effectiveness of the model, the proposed model is applied in EHRs search via clustering algorithm. Experiments on real-world EHRs data set demonstrate that the proposed model sufficiently expresses the temporal feature of the EHRs and provides an effective solution for measuring the temporal similarity between EHRs of different patients.

1. Introduction

With the prevalence of Electronic Health Records (EHRs) systems in modern healthcare, massive amount of EHRs data is becoming a rich resource for data mining. EHRs refer to a collection of documents that are designed to record the changing conditions of a patient over time, where most of the information in each record is kept in the form of unstructured text, usually called clinical narrative. Examples of a person's EHRs collection are shown in the following figures. Fig. 1. shows that a person's EHRs is composed of a series of records created over time. Fig. 2 shows the details of an EHR narrative, where different types of information are indicated with different colors. From the perspective of a doctor, it only makes sense when a person's medical records are reviewed chronologically, because the temporal order of the occurrences of symptoms is an essential feature. Consider two patients having hypertension and diabetes: one patient is diagnosed to have hypertension then diabetes after a period of time, while the other is diagnosed to have the two in the opposite order. From the view of lexical matching, the two patients belong to one group because they both have the terms of 'hypertension' and 'diabetes'. But when taking the temporal order into account, the two patients may fall into two different categories of diabetes. This suggests that when analyzing EHRs, besides textual content similarity, the temporal similarity is an important factor as well.

It is challenging to process time in clinical narrative data. Previous work dealing with temporal information in medical records includes finding textual temporal expressions [24] and temporal relationships among medical events [29]. They both embody time in text and represent it as textual expressions. It is impossible to measure the temporal similarity between records of patients in this way, since when represented by textual expressions, time becomes static. In fact, a medical history depicts a dynamic process of a patient's changing conditions. It is more appropriate to model time as an individual dimension. Our intuition is to model the textual and temporal dimension jointly, then the temporal similarity can be measured.

Such a model faces several challenges. First, most current works processing time in EHRs narrative embody temporal information in text form. Modeling both textual content and temporal dimension simultaneously remains unexploited. The temporal distribution of records poses the second challenge. We extract the temporal intervals of two person's EHRs and put them into sequences as shown in Fig. 3. It shows two temporal characteristics of EHRs: i) the two sequences have different numbers of intervals, ii) the intervals have different lengths. These characteristics imply that content alignment is required when comparing two EHR sequences which leads to another challenge. It is obvious that the conventional similarity measures such as Euclidean distance and Cosine Similarity are not applicable for EHRs due to the fact that they require

* Corresponding author.
E-mail addresses: zhangjiayue@bupt.edu.cn (J. Zhang), xuweiran@bupt.edu.cn (W. Xu), guojun@bupt.edu.cn (J. Guo), gaosheng@bupt.edu.cn (S. Gao).

http://dx.doi.org/10.1016/j.knosys.2017.03.029
0950-7051/© 2017 Elsevier B.V. All rights reserved.

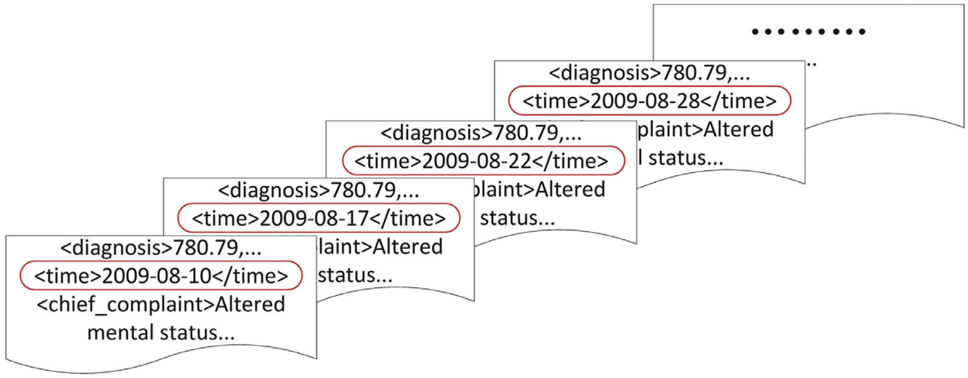

Fig. 1. An example of a person's EHR collection.

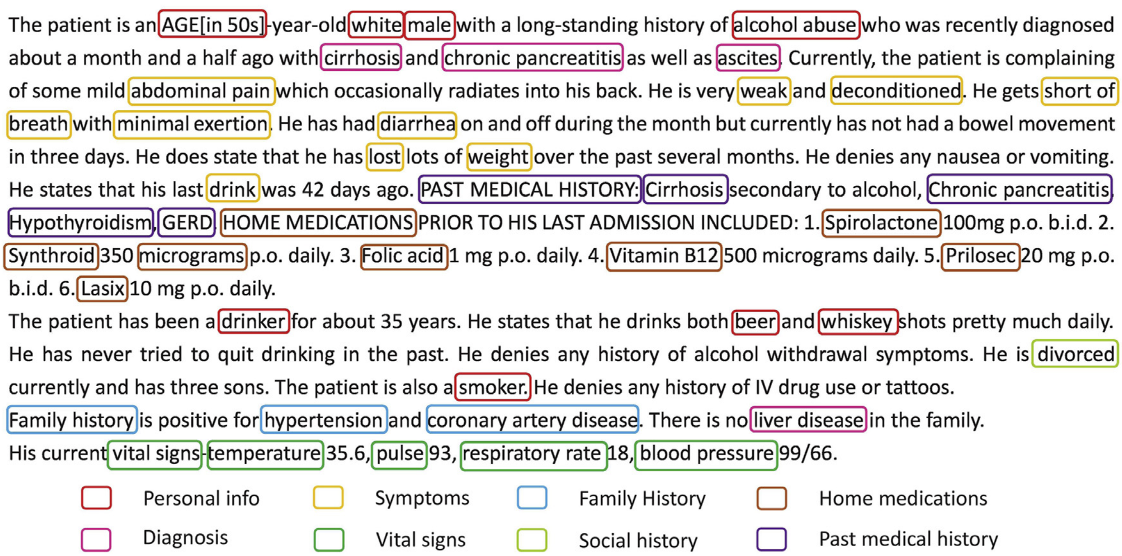

Fig. 2. Narrative of an EHR.

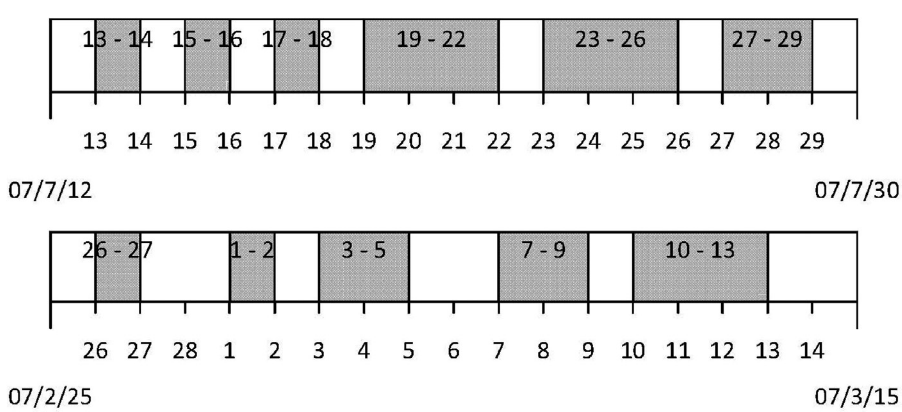

Fig. 3. Temporal distributions of two visits. A white space denotes a date on which there are records, and a dark space denotes a date on which there is no record. The dark spaces between every two adjacent white spaces are called intervals.

identical length of sequences. Therefore, to measure the temporal similarity, a method that makes content alignment and allows different lengths between two sequences needs to be developed.

In this paper, we propose a temporal model (TM) to address the challenges above. We consider a series of EHRs of a patient as a whole, and take them as a document sequence with temporal order retained. By extending the feature space from vector to vector sequence, we achieve the purpose of modeling textual and temporal dimension jointly. With the temporal framework, we modify the Dynamic Time Warping algorithm (DTW) [27] to make alignment of two vector sequences and compute the similarity between them. While processing time in EHRs has been proposed in some past work [3,4,13], our work differs from previous ones in several important aspects. First, compared to finding textual