An RDF Based Semantic Approach to Model Temporal Relations in Health Records

Oya Beyan[1,2], Stefan Decker[1,2]
[1] RWTH Aachen University, Informatik 5, Ahornstr. 55,
52056 Aachen, Germany
{ beyan, decker }@dbis-rwth-aachen.de
[2] Fraunhofer Institute for Applied Information Technology FIT
DE-53754 Sankt Augustin, Germany

Abstract. Progression of diseases may vary for each patient due to genetic make-up, life style, or previous health history. Even for well-known medical conditions, temporal signatures can be different for specific genotypes. Secondary use of health records can help us to identify these signatures. We propose an RDF based approach for modelling the temporal relations in health records. RDF graphs compared to relational data representations provide advantages with their inherent notion of a hierarchy and a temporal model. In this work, we suggest a new approach to representing temporal relations in RDF graphs. The proposed approach will help to improve the efficiency of data mining by including a more relevant set of patient attributes.

Keywords: RDF, graphs, temporal relations, secondary use of health data

1 Introduction

This work is focused on presenting the challenges of temporal data mining of Electronic Health Records (EHRs) and discussing how RDF data model representation may help to address some of the shortcomings of temporal modelling and abstraction of health data. With the widespread use of EHRs, the secondary use of this rich data source for discovering new knowledge becomes a predominant research question. EHRs contain longitudinal health information, including demographics, laboratory test results, medication orders, medical diagnosis procedures, and progress notes [1]. They naturally contain multiple series of clinical variables and medical events. Therefore, effective mining of EHRs incorporates the temporal dimension. Although temporal data mining promises better understanding of disease prognosis and individual pathways, due to the longitudinal and heterogeneous properties of EHRs, temporal analysis is an inherently difficult challenge. Most of the temporal pattern mining approaches such as times series classification methods, times series similarity measures, and time series feature extraction methods cannot be directly applied to complex EHR data [2]. Considering the trade-off between exploring a large enough time span to discover patterns and reducing the computational cost with smaller window size, the selection of relevant and non-redundant features remains a challenge. Therefore defining a language that can adequately represent the temporal dimensions of data becomes a key issue. The basic

properties of EHR data can be described as: (i) Multivariate – a large number of clinical variables are measured; (ii) Heterogeneous – contains multiple types of events; (iii) Irregular in time – variables are measured asynchronously; and (iv) Sparse – contains many unknown and missing values [1]. We argue that RDF as a data model is capable of satisfying the requirements to represent the temporal dimensions of health data. Firstly, the RDF data model does not follow a fixed schema. Therefore heterogeneous and highly interconnected data can be easily represented. Secondly, RDF graphs can be nested as well as chained, and so complex objects can be modelled. Thirdly, RDF resources are identified by unique international resource identifiers (IRI's), which makes it easier to add additional information by creating references between two different RDF graphs.

2 Semantic Modelling of Temporal Health Data Graphs

The RDF data mapping approach has been applied to integrate health records from heterogeneous resources and to generate integrated data in different non-RDF data formats or semantics to support various clinical research applications [3]. Although the most important part of the medical data is stored as narrative notes, new approaches such as high-throughput phenotyping promise to generate thousands of phenotypes with minimal human intervention [4]. Despite the progress in semantic modelling of health records, less attention has been given to defining complex temporal relations. In this section, we suggest three extensions to the current state of the art, namely: introducing new temporal relations for the construction of semantic health graphs; a flexible window size selection approach based on the introduced semantic temporal relations; and the use of contextual information to abstract events in a time point.

A. Semantic Construction of Temporal Graphs: Longitudinal and heterogeneous properties of EHRs increase the complexity and results in the pattern explosion problem in sequential pattern mining. In other words, the improper setting of thresholds leads to the detection of a huge number of patterns. In a recent study, Liu, Chuanren, et al proposed temporal graph representation for event sequences to address this challenge [5]. In their model, patient EHRs were represented as temporal graphs wherein the nodes are medical events and the edges indicate the temporal relations among those events. A weight is also associated with each edge, which encodes the average duration between two EHR events. The result is a directed and weighted graph, in which is assigned a smaller edge weight for larger intervals, when time interval is smaller than given threshold. Although time directed temporal graphs provide a practical solution to pattern explosion, they have two main shortcomings. Firstly, the hard threshold with a certain cut-off value ignores events and laboratory measurements that happened in the far past. For example, today we know that genetic makeup has an impact on the occurrence of many diseases as well as reactions to treatment. When we have genetic profiling as a laboratory value, its impact on other nodes should be time independent. Similarly, the same events, even though they occurred in the far past, might have an impact on current acute diseases more than recent events. For example in some rheumatic fever cases, the inflammation may cause long-term complications. Damage to the mitral valve, other heart valves, or other heart tissues can cause problems with the heart later in life. Resulting conditions may include atrial fibrillation and heart failure

The RDF data model helps us to overcome the limitations of directed temporal graphs which represent time only as a weight in terms of days or hours. The rich semantic of the RDF graphs facilitates the creating of edges between medical events, not only with time interval weights, but by representing complex relations such as etiological associations between comorbid diseases. Prior knowledge accumulated in curated research data repositories, such as NIH dbGAP or Cosmic, can be utilized to define possible associations.

RDF descriptive properties can model etiological association between medical conditions, e.g. direct causation: the presence of disease A is directly responsible for another; associated risk factors: for disease A are correlated with the risk factor for another disease; heterogeneity: disease risk factors are not correlated but each is capable of causing disease associated with other risk factors [6].

Allen's temporal logic describes 13 possible relations of any pair of states [7]. From the medical point of view, four of them are meaningful for representing co-morbidities, risk factors, and disease aetiologies as temporal dimensions. Each type of relation can be represented as a predicate which connects medical events, including diseases and risk factors. Table 1 presents these relations and example medical cases.

Table 1. Subset of Allen's temporal relations and example medical cases.

E1----- E2----	*E1 before E2:* Subacute sclerosing panencephalitis (SSPE) is a very rare but fatal disease of the central nervous system that results from a measles virus infection acquired earlier in life. SSPE generally develops 7 to 10 years after a person has measles, even though the person seems to have fully recovered from the illness.
E1------- E2----	*E1 overlaps E2:* Concurrent damage to different organs and systems, which is caused by a singular pathological agent (for example due to alcoholism in patients suffering from chronic alcohol intoxication); Diabetic nephropathy (Kimmelstiel-Wilson disease) in patients with type 2 diabetes.
E1---------- E2----	*E1 contains E2:* Development of cerebrovascular accident resulting from complications due to hypertensive crisis in patients suffering from hypertension; Development of cataract as a diabetes complication.
E1------ E2----------	*E1 starts E2:* Neurofibromatosis in early life may cause learning and behaviour problems, and individuals might have light brown dermatological spots (café-au-lait spots), neurofibromas, growths on the eye's iris, and abnormal growth of the spine (scoliosis).

Figure 1 presents time relations in a prostate cancer case. Patient diagnosed with prostate cancer at time t_i, 16 days before prostate-specific antigen (PSA) test ordered, 4 days before the PSA test a digital rectal exam (DRE) performed. Obesity diagnosed 8 years ago linked to time graph with "overlaps" predicate. Similarly genetic profile sequenced in early childhood linked with "contain" predicate.

B. Flexible Window Size Selection: In temporal data mining, the discovery process usually includes sliding time windows or time constraints [8]. Specification of window size defines the maximum pattern time periods between adjacent elements of the sequential pattern and set them as a fixed value. This means that for every patient for a time point t_i only temporal patterns within the window size can be observed. This approach assumes events very far away from each other are not of interest for explaining the current state. However, for medical histories, this assumption is not always valid. RDF representation of temporal relations will enable us to set flexible window size for different medical contexts. The following features for flexible window size selection

can be supported: (i) *Selective inclusion:* Genetic, environmental, and life style factors jointly influence the risk of developing disease. The multifactorial risk factors, independent from their occurrence time stamp, can be selectively included in the temporal analysis. (ii) *Repetitive events:* Chronic disease monitoring requires continuous measurement of certain laboratory and physiological parameters. However in patient records, these measurements may not be complete. In these cases, the required measurements can be included in the time window, even though they are far away from the current medical event. Similar conditions can be valid for recovery periods or follow up. (iii) *Compelling events:* Some diseases have risk factors which may date back to early childhood, such as starting menstrual periods at a young age being a risk factor for breast cancer. Similarly some diseases may impact a later stage of life and increase the likelihood of developing other diseases. The semantic representation of temporal health graphs in RDF support flexible window size selection by querying repetitive and compelling events as well as the selective inclusion of risk factors. These extended flexible window sizes provide relevant attributes for constructing models in machine learning and improving the success of temporal data mining algorithms.

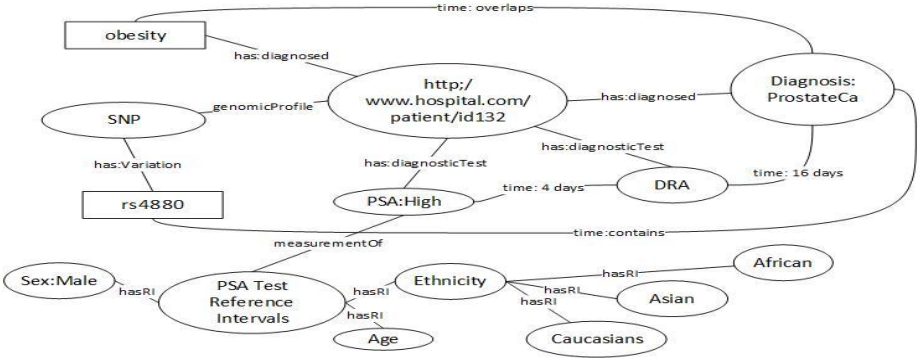

Fig. 1. Temporal model of a patient data diagnosed with Prostate Cancer. Obesity as a comorbidity overlaps in time with the diagnosis. Patient's genetic profile has SNPs associated with Prostate Cancer. PSA test reference interval values differ for age and ethnicity groups.

C. Context Sensitive Temporal Abstraction: Temporal abstraction can be defined as a generic interpretation task that interprets states and trends for a given set of goals [9]. Temporal abstraction transforms raw numeric time series variables for clinical variables into a high-level qualitative form [2]. Sets of clinical variables and lab measurements, such as blood glucose level, transformed into interval based representation$(v_1\{b_1, a_1\}, .. v_n\{b_n, a_n\})$, where $v_i \in \Sigma$ and is a finite set of all permitted abstractions that holds from time a_i to time b_i. The value abstraction Σ finite set includes values such as very low [VL], low [L], normal [N], high [H], and very high [VH]; whereas in trend abstraction, values such as decreasing [D], increasing [I], and steady [S]. Each laboratory and physiological measurement in health records with time stamp can be represented as a time point event. The abstraction of data should be based on a prior domain knowledge. Normal and abnormal values for physiological and laboratory measurements are based on reference intervals. The reference intervals may vary by age, sex, ethnicity, genetic profile, or accompanying diseases. Conditions like pregnancy, delivery, and the postpartum period are other specific cases as physiological

changes in human life [10]. The RDF data model provides us the opportunity to represent, acquire, maintain, use, share, and reuse this knowledge effectively. Hence boundaries between healthy and pathological states are influenced by many biological factors, and so it is misleading to abstract time point events with a single threshold. Rather, reference interval prior knowledge can be represented with a graph as a collection of classes corresponding varying properties for measurement. Figure 1 presents a reference interval graph for PSA test. Health data of the patient is represented as a separate graph, and the required knowledge for interpretation of time point events is interpreted by referring to another graph based on varying properties of case. Separating the reference interval model and patient data, and late binding for interpretation, will enable us to utilize all the individual characteristics, e.g. age and gender, in the abstraction stage for precise modelling of patient states. Moreover, the proposed architecture will support the re-abstraction of data over time for different medical contexts or goals by simply replacing the reference interval model with the one designed for the required context.

3 Conclusion and Future Work

In this work we have presented advantages of the RDF data model in overcoming the shortcomings of the feature selection and data abstraction in temporal data mining for health care data. We have summarized the distinct features of medical data and we propose the RDF as a viable solution to target challenges of complex EHRs. The main obstacle to the implementation of the suggested model is the absence of the rich phenotype data. Most of the information in EHRs is buried in free text format, and semi structural representation of this data requires natural language processing. Another challenge is discovering the relevant risk factors and comorbidities in EHRs. This can be overcome by linking more knowledge bases, including publications, and exploiting clinical research results for creating semantic links between temporal events.

References

[1] Batal, I., and C. A. San Ramon: Temporal data mining for healthcare data. Healthcare Data analytics. Boca Raton, FL: Chapman and Hall/CRC (2015): 379-402.

[2] Batal, I., et al. : A temporal pattern mining approach for classifying electronic health record data. ACM Transactions on Intelligent Systems and Technology (TIST) 4.4 (2013): 63.

[3] Sun, H., et al. : Semantic processing of EHR data for clinical research. Journal of biomedical informatics 58 (2015): 247-259.

[4] Hripcsak, G., and David J. A.. :Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association 20.1 (2013): 117-121.

[5] Liu, C., et al.: Temporal phenotyping from longitudinal electronic health records: A graph based framework. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015.

[6] Valderas, J. M., et al. : Defining comorbidity: implications for understanding health and health services. The Annals of Family Medicine 7.4 (2009): 357-363.

[7] Allen, J.F.: Towards a general theory of action and time. Artificial intelligence23.2(1984):123-154.

[8] Ali, A. B. M., ed. Dynamic and Advanced Data Mining for Progressing Technological Development: Innovations and Systemic Approaches: Innovations and Systemic Approaches. IGI Global, 2009.

[9] Shahar, Y. :A framework for knowledge-based temporal abstraction." Artificial intelligence 90.1 (1997): 79-133.

[10] Siest, G., et al. The theory of reference values: an unfinished symphony. Clinical chemistry and laboratory medicine 51.1 (2013): 47-64.