MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190199

A Concept for Graph-Based Temporal Similarity of Patient Data

Matthias Ganzinger, Jens Schrodt, Petra Knaup

Institute of Medical Biometry and Informatics, Heidelberg University, Heidelberg, Germany

Abstract

Computer-based decision support systems are often used for dedicated tasks such as the detection of sepsis. However, positive predictive values for sepsis detection are reported to achieve only around 46%. In this paper we describe a novel approach to use temporal data of electronic patient records based on similarity measures. We apply the concept of case-based reasoning, which is well-established in many fields of medical informatics. Temporal patient data are organized in a time-graph structure. For the quantification of similarity between cases, we exploit graph theory based approaches. For development and evaluation of our time-graph similarity frame we use the open MIMIC III dataset. In a later phase, we envision to transfer our concept from sepsis to other diseases.

Keywords:

Medical Informatics, Computer-Assisted Decision Making.

Introduction

Computer-based clinical decision support systems (CDS) are important elements for improving the quality of patient care. A relevant clinical example is the early detection and management of sepsis. Sepsis is a systemic reaction of the human body against inflammatory infections with approximately 154,000 diagnosed cases in Germany per year leading to as many as 56,000 deaths per year [1]. Early detection and start of treatment are crucial factors for improving survival and complete recovery [2].

CDS have been a field of research in medical informatics for several decades. In recent years, the field was complemented by data science approaches analyzing vast amounts of patient-related and disease-related data in order to build computer models of diseases. Terms commonly used for such concepts are systems medicine, personalized medicine, or precision medicine. One of the ideas that are currently under discussion in systems medicine is to identify similar patients or cohorts within comprehensive patient-related databases such as electronic health record (EHR) systems [3]. However, the concept of providing clinical decision support on the basis of treatment experiences made with past patient cohorts has been established in medical informatics as case-based reasoning (CBR) already in the 1980s based on previous works in cognitive sciences [4, 5].

Currently, temporal data are rarely used in similarity-based decision support systems. To change this, it would be necessary to bring existing temporal data into a form that supports reasoning on these data. For CBR, the case of a patient would have to be described considering all relevant temporal events. Such a temporal profile or network of different event values over time could be used to find a similar patient in the case base, whose case shows a similar sequence of temporal events in addition to the similarity of attribute values. This could lead to improved clinical conclusions for diagnosis and therapy or medication. For our example of sepsis, this would mean to compare the course of disease of current patients with those who had developed a sepsis in the past in order to detect sepsis in an early stage to save time until treatment is started.

Since there are no established approaches for temporal similarity in CBR, we describe a possible model for the development of such a similarity measure in the course of this manuscript. This includes potential forms of representing temporal data for similarity and reference data sets for the development process.

Patient Similarity and CBR

Technically, historic treatment information is represented as a patient case. Each case is described by a set of attributes, often represented as structured fields in the database of an EHR system. To quantify similarity of two individual cases, a so-called local similarity measure is defined to compare the values of a certain attribute for the two cases. Global similarity between two cases is gained by computing and weighting the importance of attributes and merging local similarity measures into a single value. Great care has to be taken to choose an appropriate similarity measure that takes into account the characteristics of the respective attributes. For example, the similarity of the body weights of two patients might be calculated as the Euclidian distance between the values. For attributes with discrete value domains, such as ABO blood groups, the context for which similarity is being calculated has to be considered. For example, when looking at the risk of developing coronary heart disease (CHD), blood group values of A, B, and AB might be considered more similar to each other than compared to O. The reason is that bearers of blood group O seem to have a significantly lower risk to develop CHD than bearers of the other groups [6]. For other diseases, a different assumption of similarity might be more conclusive. In these cases, value distributions or expert knowledge can be used to establish a numeric similarity value.

Due to the development of big data analysis initiatives, more and more patient related data sources are expected to become available in a structured way for research on clinical decision support systems. Using these sources for finding similar patients is only possible if suitable similarity measures exist and are correctly chosen for the respective attributes.

Our group has conducted research on similarity measures in a systems medicine research project on multiple myeloma. For this project we developed an IT architecture for systems medicine applications that rely on the concept of patient similarity. As one result we established a novel similarity measure that considers survival data (e.g., progression free survival, or overall survival) as they are measured in clinical trials for the computation of similarity. However, in the course of our research we found additional aspects of patient similarity that have not been taken into consideration sufficiently so far

due to the lack of appropriate similarity measures. Specifically, these additional aspects are temporal relations among symptoms and events like diagnostic and therapeutic measures. In the everyday life of a physician it is often essential to know the sequence of such events to get an appropriate diagnosis [7].

Decision Support for Sepsis Diagnosis

Supporting early diagnosis and management of sepsis has been subject to previous studies. Corfield et al. describe a single, not computer-based, score as an early warning sign for sepsis [8]. In terms of computer-based CDS, one approach is to generate sepsis alerts if a number of criteria is met by entries in the electronic medical record of a patient [9]. While this system had no false negative results, its positive predictive value was only 44.7%. Amland et al. report on a similar system with a likewise positive predictive value of 46% [10]. Such results suggest the necessity for further research to improve accuracy of computer-based CDS for sepsis. Considering temporal aspects might help to fill this gap [11].

Temporal Abstraction

Temporal abstractions are qualitative representations of temporal intervals with respect to a specific context [12]. Such abstractions are used to derive higher level concepts from raw, quantitative data. For example, individual measurements of body temperature could be aggregated to intervals and labelled with qualitative descriptors like “normal temperature” or “hyperthermia”. A widely used set of temporal abstractions has been defined by Allen [13, 14]. He describes a system of thirteen basic relations such as “before”, “overlaps”, or “during”, covering all relations that are possible between two temporal intervals. Temporal abstractions can be used to reduce complexity when working with temporal dependencies.

Temporal Data Structures

Temporal relations among health related events could be an important source of information for diagnosis and prognosis of patients. Thus, such relationships in patient data are in focus of many research activities such as temporal mining. A basic question for the development of a temporal similarity measure is how data should be organized and stored for efficient processing. Traditionally, medical events are recorded in the form of tables either as paper charts or electronic database tables. To support temporal information, the 2001 edition of the SQL standard query language for relational databases supports temporal tables. However, this extension does not focus on working with temporal relations among entities, but provides means to model the time when entries are valid in the real-world context. This is similar to conserving the state of a database into a snapshot which makes it possible to reproduce the data for time points in the past. So the word ‘temporal’ has a different meaning in the SQL standard. As a consequence, temporal relations have to be implemented outside the database management system into an application with the potential requirement to transfer large amounts of data. The goal of such an application should not be to save different states of database tables over time but to save and compare temporal data like patient events in a timeline.

Time-Graphs

An intuitive representation of temporal relationships among medical data might be graphs or networks. Medical events could be presented as nodes while edges show the temporal relations among them. Interestingly, the term *temporal network* is widely used in computer science, but usually not in the sense that the network shows temporal sequences. Instead, the term describes the research on the development of networks over the time. In research on social networks, for example, it is investigated how the contacts of a person change by comparing the respective network at different points in time [15].

To avoid confusion with such temporal developments of graphs, the term *time-graph* is used for sequential graphs throughout this proposal.

A graph-based representation of temporal EHR data has been shown to work by Liu et al. in a paper on temporal phenotyping [16]. They introduce temporal phenotypes as a set of basis graphs that are combined to construct an observed temporal graph. The authors describe the application of similarity-based regularization for calculating the temporal phenotypes. However, the aim of their work is not to identify similar patients in the sense of CBR but to predict medical events such as the onset risk of heart failure.

Thus, additional research is necessary to establish time-graphs for the use of finding patients with similar temporal disease development. This could be achieved by calculating the similarity of time-graphs each representing one patient.

Graph Similarity

Graphs have been studied for a long time in graph theory. One problem related to comparing two graphs is the task of determining their structural identity or the existence of an isomorphism between them. In addition, quantifying the similarity of graphs and subgraphs has been of interest in many application areas like analyzing the world wide web, business process models (BPM), timetables, schema matching, or chemical structures [17–21]. Comparable results have not yet been researched for time-graphs. No specific time-graph similarity approach is known to the authors. Nevertheless, there are some approaches that are related to graph similarity, but it has not been investigated yet, if they can be applied for medical time-graphs.

One generic approach that has been demonstrated to work with BPM, for example, is the *graph edit distance*. It is capable of measuring similarity between two graphs by assessing the number of modifications necessary to transform one graph into the other [22]. A different method for quantifying similarity among two graphs is using an iterative framework based on the idea that two nodes or edges are similar if their neighbors are similar. Since similarity is propagated to adjacent elements during iterations, this method is also called similarity flooding [20]. Another method is *NetSimile*, which calculates different node-specific values, like the number of neighbors, the clustering coefficient of a node, and others. These attributes are finally aggregated into one value that describes the similarity of two graphs using the Canberra Distance [23]. These graph similarity approaches are candidates for being transferred to time-graphs.

Graph-Based Temporal Similarity

Taking up the concept of case-based reasoning, we develop a computer-based CDS for early detection of sepsis aiming for high accuracy. Specifically, we will develop a novel measure for calculating temporal similarities of disease progress among patients. This measure relies on graph-theory concepts. Later, we plan to extend our approach to other disease patterns.

Methods

The main research question of our project is as follows: How could temporal medical data of different patients be used to identify patients with a similar temporal profile for clinical decision support systems?

This question is divided into three sub-questions:

1. How can medical data be represented as a time-graph?

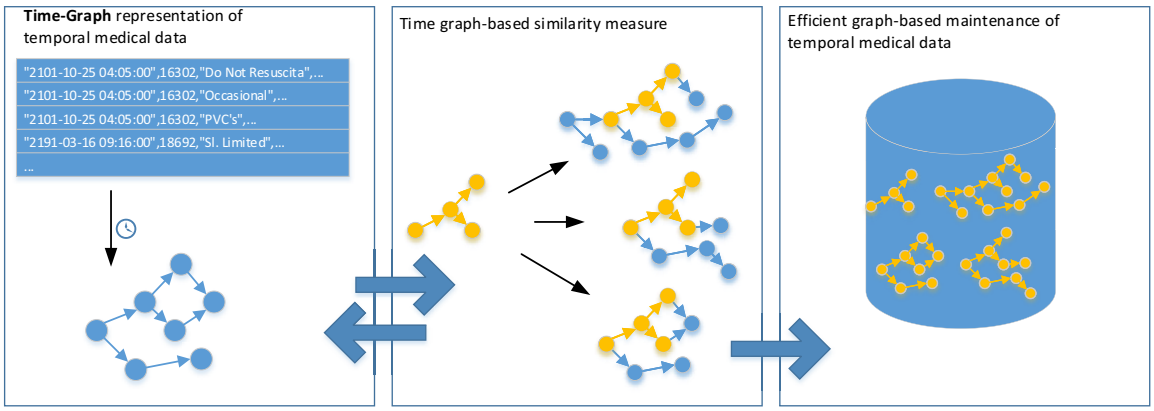

Figure 1 – Main Parts of the Time-Graph Similarity Framework.

2. What similarity measure can be used on time-graphs?
3. How can time-graphs be stored efficiently?

To answer these questions, our objective is to develop a new approach of making temporal aspects of medical data available in decision support systems, especially with respect to similarity-based approaches like case-based reasoning. We aim to establish graphs of clinical events as an innovative representation of temporal relations among these events. Here, we will consider concepts like networks, directed/undirected graphs, weights of edges, trees, and the significance of cyclic, i.e. reoccurring parts. The challenge is to find a graph-based representation of temporal medical data that provides an optimized basis for similarity quantification. To support case-based reasoning, a novel graph-based similarity measure for time-graphs in the form of a framework of methods is part of this approach. For this framework we investigate graph-theoretical approaches for the assessment of graph similarity. A set of criteria to classify these methods for applicability in the medical CBR context is established. As part of this development, graph similarity methods are tested in various disease-specific scenarios.

Finding the most similar patient with respect to his time-graph in a big case base can be very time-consuming, making efficient management of graph-based data structures essential. Thus, we investigate which kind of database management systems could form an appropriate basis for storing time-graphs. Specifically, we compare relational data base systems with temporal index structures with dedicated graph-based database systems. The exact requirements towards such database systems depend on the characteristics of the graph-based similarity measure.

Results

Our concept is comprised of three major blocks, corresponding to our research questions. First, research on the graph-based representation of temporal medical data is performed. Second, a novel similarity measure for temporal data in a graph representation is developed. In a third block, existing database technologies for efficient storage and similarity retrieval of graph-based temporal data are reviewed.

In addition, a user interface for temporal queries on the database is part of the concept. For research and testing of our approach, we use the MIMIC III database [24]. This dataset provides data of approximately 40,000 patients who were treated in an intensive care unit (ICU). All data are annotated with timestamps making them ideal for temporal research in clinical context.

Time-Graph Representation of Temporal Medical Data

As the first step, the MIMIC III database is analyzed with data mining tools for cases that can be used as blue prints for the development of a time-graph representation of the case documentation. In this phase, we consider cases suitable, if the temporal relation between events is known *a priori*. Medical treatment guidelines are helpful to identify such cases. For example, the guidelines for diagnosis and treatment of sepsis list a number of conditions that have to be met within a certain timespan. Thus, we will mine the database for patients who are diagnosed for sepsis and other diseases commonly occurring in an ICU context with temporal aspects. This will result in a set of reference cases with typical temporal disease progress is identified in the MIMIC III dataset.

Further, an appropriate graph-based representation for patient cases is established. Existing approaches will be reviewed: The most straight-forward approach is to form a timeline where all data entries are lined up in order of the respective timestamps. In this approach, each node has one edge pointing to the subsequent node. A timeline already allows to compare the temporal developments of different cases. However, in this naïve approach no difference is made between different classes of events. As a consequence, we will investigate more complex representations that also consider different classes of nodes. Nodes within each class have a linear temporal order, but are interwoven with nodes of other classes. In contrast to a timeline, such additional edges might help to better represent event-specific temporal relations. For example, known cause-effect relations such as the administration of antipyretics and the consecutive decrease of body temperature are modelled into the graphs.

Further research is made regarding the most suitable topology of the resulting graph. Possible topologies include directed graphs with or without cycles, weighted graphs, and trees.

Time-Graph-Based Similarity Measure

We investigate methods described in the literature for calculating the similarity between two graphs for their applicability on time-graphs. Aspects to be considered are the topology of the graphs and the possibility to work with sub-graphs. This is important, since the time-graph of medical events of a newly diagnosed patient is likely not as extensive as that of a patient who has finished treatment. We investigate similarity methods of different categories such as graph

isomorphism (e.g., edit distance based measures [22]), iterative methods like similarity flooding [20], and feature extraction methods like NetSimile [23].

The base methods for graph-based similarity are adapted to fulfil requirements of temporal graphs. For this development process, cases of the MIMIC III database are used as models and test cases. Criteria for the successful development of the similarity measure are defined during this task. They are based on plausibility checks and correlation analysis of the cases determined as similar by the new similarity measure. The graph-based temporal similarity measure developed in this work packages can be used by itself, but also in combination with conventional similarity measures based on attribute values. In this task, we investigate, how such different types of similarity measures can be combined to calculate a comprehensive similarity value. The approach will be based on the paradigm that global similarity is comprised as a combination of weighted local values. In this task, a suitable algorithm for combining the different similarity measures into one global value will be researched.

Efficient Graph-Based Maintenance of Temporal Medical Data

Electronic health record data are usually stored in relational databases with no specific focus on temporal data retrieval. Graph databases are potentially good candidates for high efficiency, especially in combination with a time-graph data model. In addition, in this kind of databases not only the raw data is saved like it is the case in traditional relational databases but also the connections between the different entities of the raw data are saved. This aspect makes graph databases potentially more efficient than relational databases but it still has to be investigated if this is still true for temporal data like it is used in our case.

We investigate if relational databases such as PostgreSQL can be used as an efficient storage for time-graphs. Specifically, we assess if existing SQL standard extensions support our graph-oriented similarity measure. We also evaluate proprietary temporal extensions that exist for some open-source database management systems. Further we investigate if dedicated index tables can support queries for similarity retrieval. The idea for such index tables is to pre-calculate temporal relations (respecting temporal abstractions) within cases to allow for fast retrieval without the need of calculating temporal relations at query time.

In contrast to relational databases, graph databases are designed to store graph data structures. In this task, we will investigate open source graph databases such as neo4j for their applicability in our context. Special focus lays on the efficient retrieval of similar cases. We will compare the respective query possibilities for retrieving cases based on subgraphs. One criterion for the assessment of the database management systems is the availability of features for efficient similarity retrieval. Further, the query times are compared for the different approaches.

For using the database system selected in tasks 3.1 and 3.2, data have to be prepared first. To do this, we will follow the extract, transform, and load (ETL) approach that is well-established for example in the context of data warehouses. Thus, we will develop ETL processes to make a subset of MIMIC III data available in our temporal database. Depending on the specific database management system, we will use open-source ETL software such as Talend Open Studio or KNIME to automate this process.

Discussion

In this manuscript, we present a comprehensive approach for the establishment of a graph-based similarity measure temporal patient data. The first step will be the development of a time-graph representation for temporal medical data as basis for the remaining parts. Further, efficient storage approaches and graph-based temporal mining are investigated.

In this phase, the development will be performed on the well-established open data set MIMIC III for patients suffering from sepsis. Being ICU derived data, this data set has a rather fine-grained temporal resolution compared to other fields of EHR data. After establishing this primary application, we are going to extend our research to other ICU related disease patterns as well as to other clinical areas, where documentation frequency and data types differ from ICU data. In addition, we are going to validate our findings with other clinical data sets.

References

[1] C. Engel, F.M. Brunkhorst, H.G. Bone, R. Brunkhorst, H. Gerlach, S. Grond et al., Epidemiology of sepsis in Germany: results from a national prospective multicenter study, *Intensive Care Med* **33** (2007), 606–618.

[2] A. Rhodes, L.E. Evans, W. Alhazzani, M.M. Levy, M. Antonelli, R. Ferrer et al., Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016, *Intensive Care Med* **43** (2017), 304–377.

[3] S.A. Brown, Patient Similarity: Emerging Concepts in Systems and Precision Medicine, *Front Physiol* **7** (2016), 561.

[4] J.L. Kolodner, Reconstructive memory: A computer model, *Cogn Sci* **7** (1983), 281–328.

[5] A. Aamodt, E. Plaza, Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches, *AI Communications* **7** (1994), 39–59.

[6] M. He, B. Wolpin, K. Rexrode, J.E. Manson, E. Rimm, F.B. Hu et al., ABO blood group and risk of coronary heart disease in two prospective cohort studies, *Arterioscler Thromb Vasc Biol* **32** (2012), 2314–2320.

[7] J.C. Augusto, Temporal reasoning for decision support in medicine, *Artif Intell Med* **33** (2005), 1–24

[8] A.R. Corfield, F. Lees, I. Zealley, G. Houston, S. Dickie, K. Ward et al., Utility of a single early warning score in patients with sepsis in the emergency department, *Emerg Med J* **31** (2014), 482–487.

[9] S.Q. Nguyen, E. Mwakalindile, J.S. Booth, V. Hogan, J. Morgan, C.T. Prickett et al., Automated electronic medical record sepsis detection in the emergency department, *PeerJ* **2** (2014), e343.

[10] R.C. Amland, K.E. Hahn-Cover, Clinical Decision Support for Early Recognition of Sepsis, *Am J Med Qual* **31** (2016), 103–110.

[11] J. Kim, J.M. Blum, C.D. Scott, *Temporal features and kernel methods for predicting sepsis in postoperative patients: Technical report*, University of Michigan, USA, 2010.

[12] M. Stacey, C. McGregor, Temporal abstraction in intelligent clinical data analysis: A survey, *Artif Intell Med* **39** (2007), 1–24.

[13] J.F. Allen, Towards a general theory of action and time, *Artif Intell* **23** (1984), 123–154.

[14] J.F. Allen, G. Ferguson, Actions and Events in Interval Temporal Logic, *J Logic Computation* **4** (1994), 531–579.

[15] Holme P, Saramäki J, editors, *Temporal Networks*, Berlin, Heidelberg, Springer Berlin Heidelberg, 2013.

[16] C. Liu, F. Wang, J. Hu, H. Xiong, Temporal Phenotyping from Longitudinal Electronic Health Records. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15. New York, New York, USA: ACM Press; 2015. p. 705–14.

[17] V.D. Blondel, A. Gajardo, M. Heymans, P. Senellart, P. van Dooren, A measure of similarity between graph vertices: Applications to synonym extraction and web searching, *SIAM Rev Soc Ind Appl Math* **46** (2004), 647–666.

[18] R. Dijkman, M. Dumas, L. García-Bañuelos, Graph Matching Algorithms for Business Process Model Similarity Search. In: Dayal U, Eder J, Koehler J, Reijers HA, editors. Business Process Management: 7th International Conference, BPM 2009, Ulm, Germany, September 8-10, 2009. Proceedings. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009. p. 48–63 Available from: URL:http://dx.doi.org/10.1007/978-3-642-03848-8_5

[19] E.K. Burke, B. MacCarthy, S. Petrovic, R. Qu, Case-Based Reasoning in Course Timetabling: An Attribute Graph Approach. In: Goos G, Hartmanis J, van Leeuwen J, Aha DW, Watson I, editors. Case-Based Reasoning Research and Development. Berlin, Heidelberg: Springer Berlin Heidelberg; 2001. p. 90–104 (Lecture Notes in Computer Science).

[20] S. Melnik, H. Garcia-Molina, E. Rahm, Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In: Proceedings 18th International Conference on Data Engineering: IEEE Comput. Soc; 2002. p. 117–28.

[21] J.W. Raymond, RASCAL: Calculation of Graph Similarity using Maximum Common Edge Subgraphs, *Comput J* **45** (2002), 631–644.

[22] H. Bunke, On a relation between graph edit distance and maximum common subgraph, *Pattern Recognit Lett* **18** (1997), 689–694.

[23] M. Berlingerio, D. Koutra, T. Eliassi-Rad, C. Faloutsos, NetSimile: A Scalable Approach to Size-Independent Network Similarity, Available from: URL:http://arxiv.org/pdf/1209.2684v1

[24] A.E.W. Johnson, T.J. Pollard, L. Shen, L.W.H. Lehman, M. Feng, M. Ghassemi et al., MIMIC-III, a freely accessible critical care database, *Sci Data* **3** (2016), 160035.

Address for Correspondence

Matthias Ganzinger
Institut für Medizinische Biometrie und Informatik
Im Neuenheimer Feld 130.3
69120 Heidelberg
Germany
Email: matthias.ganzinger@med.uni-heidelberg.de