Temporal clinical events clustering and visualization

Mohcine Madkour, Jingcheng Du, Hsing-Yi Song, Cui Tao

Abstract—Events in clinical narratives are naturally associated with medical trials such as surgery, vaccination, lab test, medication, medical procedure, diagnosis, and they are interrelated with many temporal relations, however it is difficult to define these events quantitatively or consistently in coarse time-bins (e.g. before vaccination, after admission). The grouping of medical events onto temporal clusters is a key to applications such as longitudinal studies, clinical question answering, and information retrieval. In this paper, we developed two algorithms based on Min-conflicts and K-means to enable labeling a sequence of medical events with predefined time-bins. The computation is based solely on temporal similarity and integrated with a timeline visualization tool.

Index Terms—Temporal event visualization, Clinical narratives, Temporal event ordering, Temporal event clustering

Introduction

Visual analytics is a new interdisciplinary field of study that combine concepts from data mining, machine learning, human computing interaction, and human cognition. One of the interesting fields of applications is the cluster analysis of temporal events from clinical narratives. Clinical narratives, indeed, contain large amounts of events and temporal expressions, such as dates and time, and temporal relations between events. These temporal information are very important for many clinical tasks, for example analysing disease progression pathways in terms of observed events could provide important insights into how diseases evolve over time and can help clinicians understand how certain progression paths may lead to better or worse outcomes. However temporal structure of a clinical narrative is not always coherent (i.e. lack of time expressions and temporal relations, etc.) and the order of temporal events sometimes is difficult to articulate. In order to solve this problem, we use temporal information of an ontology-based annotated file to assign values to lacked expressions of temporal closures of events using a constraint satisfaction problem algorithm, thereafter we label sequence of events with highly probable sequence of time-bins using K-Means clustering based on a temporal similarity measure.

1 Previous work

Recently, a large body of research had interest in classifying medical events into a temporal timeline. [1], [2]. For instance, A. Dehghan [3] uses NLP methods to extract and order clinical events based on a priori defined classifier (i.e., Problem - Treatment -Test) using a temporal ordering that is based completely on lexical features and NLP. P. Raghavan et.al.[2] presents a Natural Language Processing (NLP) based tool to extract sequence of medical events from across medical narratives and anchor them with predefined time-bins. The classification is made based on lexical, section (document-level structure), and temporal features of medical events. In another work, [1] the authors represent a medical event as a time duration with a corresponding start and stop, and learn to rank the starts/stops based on their proximity to the admission date. This approach allows to infer Allen's temporal relations [4] between medical events using a rule-based temporal reasoning. Overall, most of the methodologies used were either classification based (i.e. NLP extraction) or relationships based (i.e. representing and reasoning). Our approach combined both techniques and integrated both reasoning over temporal relations and heuristics and machine learning algorithms for temporal clustering of events.

- *Mohcine Madkour is with Ontology Research Group at the School of Biomedical Informatics, UTHealth. E-mail:mohcine.madkour@uth.tmc.edu*
- *Jingcheng Du is with Ontology Research Group at the School of Biomedical Informatics, UTHealth. E-mail: Jingcheng.Du@uth.tmc.edu*
- *Hsing-Yi Song is with Ontology Research Group at the School of Biomedical Informatics, UTHealth. E-mail: Hsingyi.Song@uth.tmc.edu*
- *Cui Tao is with Ontology Research Group at the School of Biomedical Informatics, UTHealth. E-mail:cui.tao@uth.tmc.edu*

Proceedings of the IEEE VIS 2016 Workshop on Temporal & Sequential Event Analysis. Available online at: http://eventevent.github.io

2 Representation of unstructured temporal records

Recently the ordering of events has been at the primary interest of Clinical NLP research community which has investigated many tasks for this purpose [3]. One of the goals behind temporal ordering is to allow the clustering of events onto a temporal visualization. For this purpose, we found that there is a gap in current literature of methods for ordering of temporally associated events based solely on temporal information. To fill this gap, we developed a tool that uses qualitative temporal relations to infer temporal expressions. This tool uses the Java OWL-API to parse and extract axioms and annotations from the RDF data file as the annotation was done through ontology. The tool has 4 functionalities: import RDF data file, order events, cluster event sequence, and generate output format for timeline visualization. The second part of this paper shows the visualization of the generated JSON data set into a timeline using the Exhibit Smile framework tool [5]. Figure 1 is a screen shot of our data extraction and processing framework.

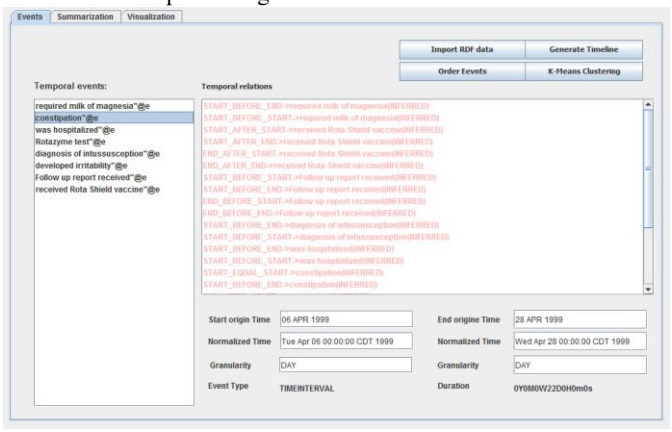

Fig. 1. Extraction and mining of temporal events from RDF data set

3 Temporal ordering of events

In this section we show how we use the annotated and inferred temporal relations in RDF file as a homogeneous collection of finite constraints to formulate a Constraint Satisfaction Problem (CSP)[6]. Also we demonstrate the Min-conflict algorithm, a CSP solving approach, to find a consistent order of all events and to estimate lacked events' endpoints' values

The CSP consists of a triple <V,D,C>, where V is a set of variables and D is a set of variables' domains and C is a set of constraints for assigning values to variables V from the domain D.

Each constraint is represented by a scope which is a subset of variables V and relation which is the set of assignments to the scope. Each constraint can involve a subset of variables and restricts the values these variables can simultaneously take. Solving the CSP consists of finding the full assignments (if one exists) to current variables from elements of defined domains in such a way that assigning constraints are all satisfied. By analogy to the temporal ordering problem, we consider events' endpoints as the set of variables, and the number of variables, which is twice the number of events, as the variables domain, and temporal relations as the constraints. We create three types of constraints, the before constraint to represent all basic relations that include the term before such as start-before-start and start-before-end, and the constraints equal and after to represent the basic temporal relations that contain the words equal and after, respectively.

Once the elements of CSP are created, we apply a heuristic search algorithm called Min-conflicts [7]. This algorithm is guided by an ordering heuristic, that we call conflict function, which attempts to minimize the number of constraint violations after each step. Given an initial assignment of values to all the variables of a CSP, the algorithm randomly selects a variable from the set of variables with conflicts violating one or more constraints of the CSP [8] . Then it assigns to this variable the value that minimizes the number of conflicts. If there is more than one value with a minimum number of conflicts, it chooses one randomly. This process of random variable selection and min-conflict value assignment is iterated until a solution is found or a pre-selected maximum number of iterations is reached. To demonstrate our solution we use a narrative from a Vaccine Adverse Event Reporting System (VAERS) report. VAERS is a national passive vaccine safety surveillance program that was established to help assess the safety of vaccines and gather adverse event (AE) data to serve as post-approval surveillance systems.

4 Temporal Clustering of Event

The previous section demonstrated finding crisp assignments of events' temporal endpoints. In this section, we investigate the task of tagging a sequence of events using a clustering algorithm. For this purpose we assume that each medical note can be associated with a predefined set of coarse of times that we refer to as time bins. For our example of VAERS note, the potential time-bins are: "before vaccination", "soon after vaccination", and "way after vaccination". The time-bin "before vaccination" is intended to capture past medical history of the patient including the medical state of the patient on time of vaccination; "soon after vaccination" captures medical events that occurred immediately after the vaccination; and "way after vaccination" captures medical events that occurred after an extended duration from the vaccination. The issue in clustering events in predefined time-bins is that the time duration of each time-bin varies based on the patient. For instance, the coarse of time "soon after vaccination" could be the first few hours after or a few days after depending on the general conditions. For that we consider that related events happen in relatively close proximity of time. We use a non-hierarchical clustering to classify the set of events. We consider the temporal distance between events as the measure of similarity between events of same clusters and dissimilarity between events of different clusters.

K-means is one of the simplest algorithms for solving the clustering problem [9]. Clustering is an unsupervised learning problem whereby we aim to group subsets of entities with one another based on a temporal distance similarity. The idea is to define k centroids for the k assumed clusters and to associate each point belonging to a given data set to the nearest center. A point represents the time instant of the event or the center of interval if its time interval event. When no point is pending, the first step is completed and an early group age is done. At this point we re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, we re-bind the same data set points to their nearest new center. A loop has been generated. As a result of this loop the k centers change their location step by step until no more changes are done or in other words centres do not move any more.

Finally, we visualize the results using the Exhibit dashboard solution [5]. The timeline dashboard enables intuitive cluster analysis by user interactions. Also our visualization allows summarizing by the various types of events information. Figure 2 shows a screen shot of this visualization. The visualization of the example used in this paper is available at: https://mohcinemadkour.github.io/TEvent/.

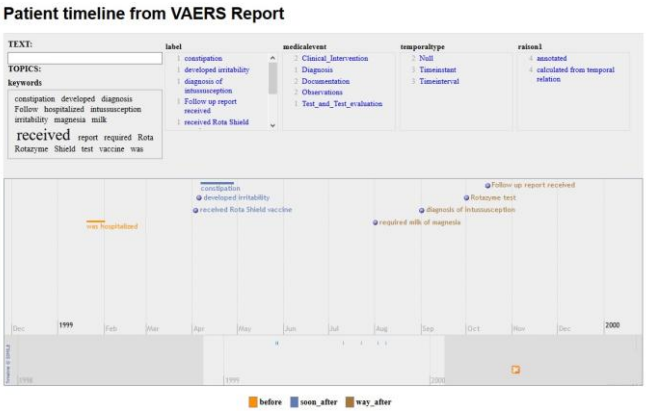

Fig. 2. Exhibit-based visualization dashboard

5 Conclusion

This paper presented a cluster analysis visualization integration approach to time binning events from narrative clinical records. The main innovation was the assignment of medical events to time-bins with the help of the K-means clustering and temporal events ordering. In the future, we plan to explore other features of events in temporal clustering along with other temporal similarity measures.

Acknowledgments

The research was partially supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM011829

References

[1] P. Raghavan, E. Fosler-Lussier, and A. M. Lai, "Learning to temporally order medical events in clinical text," Proceedings of ACL 2016. pp. Vol 2, 70–74, 2012.

[2] P. Raghavan, E. Fosler-Lussier, and A. M. Lai, "Temporal classification of medical events," Proceedings of ACL 2012. pp. 29–37, 2012.

[3] A. Dehghan, "Temporal ordering of clinical events," arXiv Prepr. arXiv1504.03659, no. December 2014, pp. 1–35, Apr. 2015.

[4] J. F. Allen, "An Interval-Based Representation of Temporal Knowledge.," in IJCAI, 1981, vol. 81, pp. 221–226.

[5] D. F. Huynh, D. R. Karger, and R. C. Miller, "Exhibit," in Proceedings of the 16th international conference on World Wide Web - WWW '07, 2007, p. 737.

[6] A. K. Mackworth, "Constraint satisfaction problems," Encycl. AI, vol. 285, p. 293, 1992.

[7] S. Minton, M. D. Johnston, A. B. Philips, and P. Laird, "Minimizing conflicts: a heuristic repair method for constraint satisfaction and scheduling problems," Artif. Intell., vol. 58, no. 1–3, pp. 161–205, Dec. 1992.

[8] S. Minton, M. Johnston, A. Philips, and P. Laird, "Solving Large-Scale Constraint-Satisfaction and Scheduling Problems Using a Heuristic Repair Method.," AAAI, 1990.

[9] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," J. R. Stat. Soc. Ser. C (Applied Stat., vol. 28, no. 1, pp. 100–108, 1979.