35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

A Temporal Model for Clinical Data Analytics Language

Leila Safari, Jon D. Patrick

Health Language Laboratories, School of Information Technology
The University of Sydney, NSW, Australia
lsaf7301@uni.sydney.edu.au, jonpat@it.usyd.edu.au

Abstract- **The proposal of a special purpose language for Clinical Data Analytics (CliniDAL) is presented along with a general model for expressing temporal events in the language. The temporal dimension of clinical data needs to be addressed from at least five different points of view. Firstly, how to attach the knowledge of time based constraints to queries; secondly, how to mine temporal data in different CISs with various data models; thirdly, how to deal with both relative time and absolute time in the query language; fourthly, how to tackle internal time-event dependencies in queries, and finally, how to manage historical time events preserved in the patient's narrative. The temporal elements of the language are defined in Bachus Naur Form (BNF) along with a UML schema. Its use in a designed taxonomy of a five class hierarchy of data analytics tasks shows the solution to problems of time event dependencies in a highly complex cascade of queries needed to evaluate scientific experiments. The issues in using the model in a practical way are discussed as well.**

I. Introduction

The description of and reasoning with temporal data has attracted the attention of many researchers from different backgrounds including philosophy, artificial intelligence, database management, computational linguistics, and biomedical informatics[1]. Temporal information is crucial in the clinical domain especially in clinical research. For instance, investigating disease progression is practical only by definition of a time line; or possible causes of a clinical condition have to be found by referring to a patient's past clinical history. In [1], the basic concepts of temporal representation in the medical domain has been described to include: category of time (natural, conventional, logical), structure of time (line, branch, circular, parallel), instant of time vs. interval, and, absolute time vs. relative time. However, this is still a challenging and active subject of research.

Our substantive goal is to create a special purpose query language for clinical data analytics (CliniDAL) to place in any clinical information system (CIS) and answer any answerable question from the CIS. A category scheme of five classes of increasing complexity, including point-of-care retrieval queries, descriptive statistics, statistical hypothesis testing, complex hypotheses of scientific studies and semantic record retrieval have been designed to capture the scope encompassed by CliniDAL's objectives[2].

The temporal dimension is of prime significance to our primary objective. Reasoning and analytics cannot be validated without considering temporal information. So, the different approaches for dealing with time in the clinical domain have been reviewed in the next section. In the method section the basic requirements for dealing with the temporal dimension in clinical data analytics and our method of addressing them are explained. Finally, a proposed temporal model is presented.

II. Background

A review of temporal query languages reflects that the importance of time has led to the development of custom temporal management solutions, which are mostly built to extend relational database systems (e.g. TSQL2[3] and T4SQL[4]), while there are some approaches which focus on ontologies to address time dimensions such as T-SPARQL[5] and SQWRL[6]. Efforts in the relational database field have led to developing expressive temporal query languages but they still suffer from two issues: firstly, they are only applicable to structural relational databases; secondly, it is difficult for hospital staff with poor IT skills to apply them. On the other hand, in most ontology based approaches composing queries can be difficult due to a complex underlying model representation and lack of expressivity.

Recently, OWL has gained attention from researchers wishing to manage time in the clinical domain such as CNTRO[7], an OWL based ontology for temporal inference in clinical narratives. It provides the ability to annotate temporal expressions and relations in clinical narratives. But it is far from being a complete system for temporal querying. In addition, the capabilities of CNTRO have only been evaluated on 5 clinical records and its reasoning and inference capabilities need to be proven on temporal information.

While the ontology based approaches mainly focus on extracting time oriented data from structured data bases, some other approaches have been proposed to deal with temporal issues by using natural language based techniques on clinical notes. In comparison to ontology based approaches these approaches are less mature in querying facilities especially in dealing with complex queries. This shortcoming is mainly found in the complexity of dealing with natural language based clinical narratives instead of structured databases.

In TimeText[8] a temporal reasoning system is designed to represent, extract and reason about temporal information in clinical text. This system has been evaluated with discharge

summaries. Despite all the effort it is only capable of answering very simple questions.

Finally, some information visualization systems have been developed to extract and visualize temporal knowledge in the clinical domain. A comprehensive study was conducted on 12 information visualization tools to investigate their capabilities from different points of view[9]. However, in our review we mainly focus on temporal issues. Among the selected tools for study [10-13], only Lifeline2, Similan and VISITORS have provided considerable support for temporal querying in clinical research, however, they suffer from issues such as dependency on the underlying data model of the CIS or the need for domain knowledge in composing time oriented queries.

III. METHOD

Our ideal goal is to propose a special purpose query language for performing clinical data analytics on any CIS so as to extract information for analysis. This is essential in many areas such as investigating disease progress or individualized treatments. Hence, the temporal dimension of clinical data needs to be addressed from at least five different points of view which are explained in the following sub-sections.

A. *Attaching knowledge of time based constraints to queries*

An investigation was performed on some temporal expressions which are used in clinical information systems so as to identify a means for expressing temporal constraints in the query language. Table 1 summarizes some of the most frequent temporal expressions extracted from the literature to query various CISs. Temporal expressions and related temporal terms are highlighted.

It is clear from the table that the temporal expressions are free form in natural language but they are not very wide in scope, so some general patterns can be established to describe them. In addition, although there are some abbreviated forms of temporal expressions (like "post op day" , "po qd", "10/12" , "3/7", etc.) they are not ambiguous like other abbreviated forms in clinical NLP tasks. For instance "10/12" means 10 months, "3/7" means 3 days. In addition, following Zhou's comprehensive work in categorization of temporal expressions in the medical domain[14], any temporal constraints can be expressed with rules which cover all Zhou's categories except fuzzy time. Zhou's, temporal categories and subcategories which were mainly designed by the analysis of 200 records and evaluated on 100 clinical records are shown in table 2.

TABLE 1. A SAMPLE OF TEMPORAL EXPRESSIONS [7, 12]

Temporal Expressions and Queries
Patients who receive paracetamol **after their temperature recorded >38** Patients who **within 12 hours of admission** attain a temperature>38 and are **then** administered paracetamol
Find **days during** which the WBC value was less than "normal" Find patients who have had a "very-low" Hemoglobin-state value **starting anywhere between January 1st, 2006 and January 7th, 2006** Patient's INR value is below normal **today**
The second cycle of chemotherapy was **on June 10, 2004** Monitor patient's heart rate for 72 hours **starting from today** Take antibiotics **every 8 hours for 10 days starting from today** Patient's bilirubin is elevated **2 weeks after the 2nd cycle of chemotherapy**
Has the patient had a headache **since the lumbar puncture?** Has the patient had a fever **since the administration of vincristine?** Has the patient's blood pressure improved **since the blood transfusion?** Has the patient passed urine **since the dose of Lasix?**

TABLE 2. CATEGORIES OF TEMPORAL EXPRESSIONS[14]

Categories of Temporal Exp.	Subcategories
Date and time	Date, Time of day, Part of day, Time Range, Conjunction, Disjunction
Relative date and time	Yesterday, today, tomorrow , Past/next time unit, A period of time Ago/later, In/within a period of time
Duration	Treatment duration
Key events	Admission, Specific hospital day, Hospital stay, Discharge, Other key events
Fuzzy time	Past, Present, Future, This time/that time, Non-specific time
Recurring time	Every day

Based on the above points we propose a set of rules for temporal expressions and define a context free grammar to formalize them to be attached to the core language (CliniDAL). A parser for the grammar has been successfully built with popular tools of python versions of Lex and Yacc (PLY) to validate instances of temporal expressions. Table 3 shows the Bachus-Naur Form (BNF) of temporal expressions.

Any temporal expression in CliniDAL can be represented in three basic forms. A simple comparison, a parenthesized comparison and disjunction or conjunction of various comparisons. The "TemporalComparison" consists of a temporal relation as a subset of Allen's[15] temporal relations and a comparative. Temporal relations include 4 concepts: *Equal* (keywords: AT, IN, ON in the BNF), *Before* (keywords: BEFORE, PRIOR), *After* (keyword: AFTER) and *During* (keywords: DURING, BETWEEN, WITHIN + PAST, LAST, NEXT or WITHIN + BEFORE, AFTER, PRIOR).

TABLE 3. BNF OF TEMPORAL EXPRESSIONS IN CLINIDAL

TemporalExpression: TemporalComparison | '('TemporalExpression')'
| TemporalExpression JUNCTION TemporalExpression
TemporalComparison: AT | IN | ON | BY | InstantReference
| [Temporaloffset](BEFORE | PRIOR | AFTER) InstantReference
| [Frequency] FOR duration (BEFORE | PRIOR | AFTER)
InstantReference
| [Frequency] (FOR | WITHIN) (PAST | LAST | NEXT) OTHER
GRANULARITY
| WITHIN [TemporalOffset] (BEFORE | PRIOR | AFTER)
InstantReference | (DURING | BETWEEN) IntervalReference
InstantReference: Expression | Date | DateTime
| BoundaryPoint OF [POSITION_NUMBER] Expression
IntervalReference: Expression | '(' Date ',' Date ')' |
'(' DateTime ',' DateTime ')' | Date BOOL_AND Date
| DateTime BOOL_AND DateTime
TemporalOffset: OTHER GRANULARITY
Duration: OTHER GRANULARITY
Frequency: EVERY OTHER GRANULARITY
BoundaryPoint: START | END

A temporal comparative can be an "InstantReference" or "IntervalReference". InstantReference can be an explicit date/time or relative time, such as today, yesterday, etc. It also can be a start or an end time point of an event. We suppose any event happens during a time interval so it has a start and end time. IntervalReference in the BNF reflects the time interval between two explicit date/times, or an implicit time interval like "in April", which means the time interval between the beginning and end of April of the reference year. It also can be the whole time interval of the duration of an event. The non-terminal "Expression" in the BNF syntactically represents relative time and events in our temporal expression.

An important issue to be considered is that there has been a debate among researchers on whether to set the temporal primitive unit as an instant or an interval[1]. Both instants and intervals have been considered in representing temporal information for medical information systems. In general, a point-based representation is desirable if every event is assigned a date. However, in real applications many events cannot be assigned a precise date. In such cases, time intervals are convenient[1]. Consequently, we have set the interval as our primitive time unit in CliniDAL which is modeled by its endpoints. So, an interval is an ordered pair of points where the second point is never before the first. Moreover, the proposed grammar consists of a rich set of different features for relative times or events which are derived from the temporal model (Fig. 1).

In the model, every RAStatement (a retrieval aggregation statement in CliniDAL is representative of a descriptive statistics query) can have a temporal relation with an absolute time (actual date/time) or a relative time or event. Relative times such as "today" or "yesterday" are recognized by the string text while for relative events more attributes have been provided in the model. These attributes mainly include "PositionNumber" and "BoundaryPoint". The former, has been described by regular expressions in the lexical rules of the query language to identify the ordering number of any event (e.g. in "before 2nd Chemotherapy", 2nd is the PositionNumber). The latter, specifies the start or end time point of an event to be used as a reference time (in "after the end of Chemotherapy", "end" is the BoundaryPoint).

The "TemporalRelation" class specifies a relation between a RAStatement and temporal constraint in the form of an instant or interval time point extracted from the "TemporalExpression". It then has to be applied in CliniDAL's constraint clauses to describe a temporal dimension for them. In addition to the "Relation" property of this class which comes from Allen's temporal relations (like >, <) and "RelationLabel" which is a text form of the relation (like "before", "after"), the "TemporalRelation" class encapsulates more features including "TemporalOffset", "Frequency", and "Duration". For instance, in "every 8 hours 5 days after admission", "5 days" is a TemporalOffset, "every 8 hours" is Frequency and in "for 5 days after Chemotherapy", "for 5 days" is Duration. These 3 classes have attributes of Granularity which reflects the time unit to be used in later computations and analytics. Its value can be set to one of second, minute, hour, day, week, month and year time units.

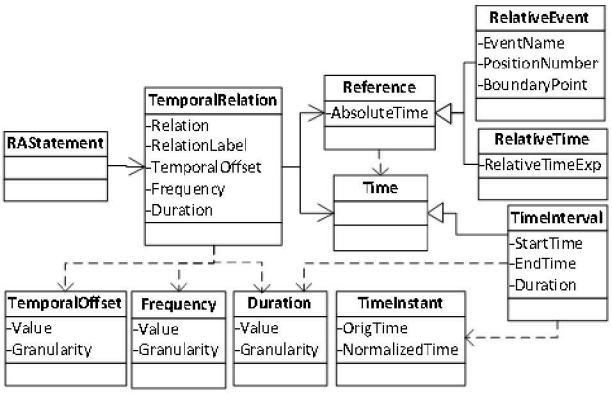

Figure1. CliniDAL's temporal model

B. *Mining temporal data from different CISs*

The next challenge in this work is mining temporal data from different CISs with various data models. Data models of interest in the work include (1) Entity Relationship (ER) design model or n-ary relational model which is usually implemented as an ordinary relational database or a star schema; (2) Entity Attribute Value (EAV) model which is mostly implemented as binary relationships in ordinary relational databases, and (3) Document or Forms design model which can be implemented as a XML database form or a NOSQL representation.

Among these 3 forms of data models, the EAV model appears to be more complicated to mine for any data but also for temporal data. On the one hand it is a more popular data model in clinical information systems due to its advantages of flexibility in data storage, easy schema design, and effective storage for sparse data. On the other hand EAV brings difficulties of complexity in data extraction and mining. An EAV design is less efficient than a conventional structure for the bulk retrieval of numerous objects occurring at the same time. The process of performing complex attribute-centric queries, which are based on values of attributes, and returning a set of objects is both significantly less efficient as well as technically more difficult. Although for schemas that are relatively static or simple (e.g., databases for business applications, such as inventory or accounting), the overhead of an EAV design exceeds its advantages but in the clinical domain we are dealing with very dynamic data and a schema which makes EAV more suitable.

In addition, documents of a forms-centric design model XML have some common features with the EAV model. The so called entities or tables in an EAV model are equivalent to the highly hierarchical structure of an XML schema. The attributes are equivalent to tags in the XML hierarchy and the values of attributes are inter-connected to tags in the XML hierarchy.

So as to mine temporal data from the EAV model, the time attribute is indexed during the automatic mapping of the query terms. For instance, in the ICIP CIS (Philips), after automatically mapping "temperature" to a specific field in the "ptAssessment" table, we know that related temporal information can be found in the same table in the field "chart time". Hence, the temporal information can be extracted to make it ready for later analytics to answer a time based CliniDAL query.

The formal evaluation of the proposed model can be done only after finalizing CliniDAL's implementation. The evaluation will clarify that in what extent the model fits with user requirements and covers the problems in the literature which needs extensive literature survey and discussion with clinicians to accomplish. Collecting the user feedback by field testing will be another part of the evaluation.

C. Dealing with relative time and absolute time

Based on the proposed context free grammar for temporal expressions in CliniDAL, there are two kinds of temporal references in the queries: instant reference and interval reference. Instant reference can be either a specific date/time as absolute time or start or end time point of any specific event in the patient's history, such as, operation, admission, etc. In addition, it can be a relative time like "today", "yesterday" for which an equivalent absolute time is constructed using the combination of a dictionary of relative time expressions and some computation. In this solution the user cannot use fuzzy time expressions, so the equivalent absolute time is found by computation. As mentioned before, these relative times and events syntactically, are modeled as an "Expression" in the grammar, which during the mapping process in CliniDAL are mapped to internal values of the underlying data model of the CIS and the absolute time of those events can be extracted from the stored time field of that internal value.

Interval reference reflects two instant time points as start and end absolute times. It also can reflect the duration of occurrence of any event. For instance the time period of starting BMT (Bone Marrow Transplantation) until the end of BMT can be referred to as an interval reference of the BMT event. Finally, a relative time event can be referred to as an interval reference as well. For instance, we may want to find all patients who received a particular medication in "April". So the whole interval from the beginning of April to the end of April is considered as a relative interval time event, and the equivalent absolute time can be computed in a similar way to the relative time instant.

D. Internal time-event dependencies in queries

Most of the research questions posed by hospital staff carry internal time-event dependencies to some extent. These dependencies occur when it is desired to retrieve some content that is defined to be dependent on the occurrence of some other data item which has a time or event relationship with the first event. A simple form of query with a temporal constraint is "retrieve all records after the 20th July 2006". This query does NOT have an internal dependency although it does have a temporal constraint. A query "retrieve all patients who receive paracetamol after their temperature reached >38" has the internal dependency in that all patients with "temperature>38" must be found before the paracetamol condition is valid. A query "retrieve all patients who within 12 hours of admission attain a temperature >38 and are then administered paracetamol" has an internal time-event dependency and an explicit trigger event defining the starting point for valid data. An analysis of the structure of these queries, has led us to propose a cascaded query model in CliniDAL with sequences of up to five CliniDAL queries in a way that each query can refer to the retrieved attributes of the previous query with temporal attributes included. Hence, the mapping strategy which led to identifying temporal information related to query terms in a CliniDAL query, together with the cascaded query model resolves the difficulties of handling time-event dependencies.

E. Historical time events preserved in the patient's narrative

Although our main focus in CliniDAL is working with CISs with three structural data models ER, EAV and XML, there is some valuable information buried in text fields of those data models. Especially, pertinent events that have occurred during the patient's lifetime and the time of starting, ending or duration of such events is effectively managed in CliniDAL's analytical approach. Thus, we need to extract those events and their temporal dimension and integrate them into the mapping process of CliniDAL.

REFERENCES

[1] L. Zhou, & G. Hripcsak, "Temporal reasoning with medical data--a review with emphasis on medical natural language processing". *Journal of Biomedical Informatics*. vol. 40, pp. 183-203, 2007.

[2] J. D. Patrick, L. Safari, & Y. Cheng, "Knowledge Discovery and Knowledge Reuse in Clinical Information Systems", in *Proc. The 10th IASTED International Conference on Biomedical Engineering (BioMed 2013)*, Innsbruck, Austria, 2013,

[3] R. T. Snodgrass, "The TSQL2 temporal query language": Springer. Vol. 330. 1995.

[4] C. Combi, A. Montanari, & G. Pozzi, "The t4sql temporal query language", in *the sixteenth ACM conference on information and knowledge management*, ACM: Lisbon, Portugal. pp. 193-202, 2007.

[5] J. Tappolet, & A. Bernstein, "Applied temporal RDF: Efficient temporal querying of RDF data with SPARQL". *The Semantic Web: Research and Applications*. vol., pp. 308-322, 2009.

[6] M. J. O'Connor, & A. Das, "SQWRL: a Query Language for OWL", in *Proc. the 6th OWL: Experiences and Directions Workshop (OWLED2009)*, 2009,

[7] C. Tao, W. Q. Wei, H. R. Solbrig, G. Savova, & C. G. Chute, "CNTRO: A semantic web ontology for temporal relation inferencing in clinical narratives", in *Proc. AMIA Annual Symposium Proceedings*, American Medical Informatics Association, 2010, pp. 787-792.

[8] L. Zhou, C. Friedman, S. Parsons, & G. Hripcsak, "System architecture for temporal information extraction, representation and reasoning in clinical narrative reports", in *Proc. AMIA Annual Symposium Proceedings*, American Medical Informatics Association, 2005, pp. 869-873.

[9] A. Rind, T. Wang, W. Aigner, S. Miksh, K. Wongsuphasawat, C. Plaisant, & B. Shneiderman, "Interactive information visualization for exploring and querying electronic health records: A systematic review", Technical Report HCIL-2010 2010.

[10] A. Inokuchi, K. Takeda, N. Inaoka, & F. Wakao, "MedTAKMI-CDI: interactive knowledge discovery for clinical decision intelligence". *IBM Systems Journal*. vol. 46, pp. 115-133, 2007.

[11] K. Wongsuphasawat, & B. Shneiderman, "Finding comparable temporal categorical records: A similarity measure with an interactive visualization", in *Proc. Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, IEEE, 2009, pp. 27-34.

[12] D. Klimov, Y. Shahar, & M. Taieb-Maimon, "Intelligent visualization and exploration of time-oriented data of multiple patients". *Artificial intelligence in medicine (Tecklenburg, Germany)*. vol. 49, pp. 11-31, 2010.

[13] T. D. Wang, K. Wongsuphasawat, C. Plaisant, & B. Shneiderman, "Visual information seeking in multiple electronic health records: design recommendations and a process model", in *Proc. the 1st ACM International Health Informatics Symposium*, ACM, 2010, pp. 46-55.

[14] L. Zhou, G. B. Melton, S. Parsons, & G. Hripcsak, "A temporal constraint structure for extracting temporal information from clinical narrative". *Journal of Biomedical Informatics*. vol. 39, pp. 424-439, 2006.

[15] J. F. Allen, "Maintaining knowledge about temporal intervals". *Communications of the ACM*. vol. 26, pp. 832-843, 1983.