TELKOMNIKA, Vol.11, No.11, November 2013, pp. 6686~6692
e-ISSN: 2087-278X

An Indeterminacy Temporal Data Model based on Probability

Ren Shuxia[1,2], Zhao Zheng*[1], Zou Xiaojian[3]
[1]College of Computer Science and Technology, Tianjin University, China
[2]College of Computer Science and Soft, Tianjin Polytechnic University, Tianjin 300072, China
[3]Military Transportation University, China
*Corresponding author, e-mail: zhengzh@tju.edu.cn

Abstract

There are many kinds of indeterminacy temporal data in temporal database. Therefore, many researchers have focused on building indeterminacy temporal data models. Unfortunately, established models can't adequately address the challenges posed by indeterminacy temporal information, and can't adapt to all sorts of involved applications. In this paper, we propose a temporal data model, named BPTM (Temporal model based on probability), to manage the indeterminacy temporal semantics of indeterminacy data. Firstly, we present our tuple-timestamp method to represent and store these temporal data including determinacy and indeterminacy data. Then we introduce the temporal primitives to process temporal relations needed in BPTM. A new probability method is brought forward to get potential information among these indeterminacy data. At last a query example based on CPR (Computer-based Patient Record) is given to show that our method is effective and feasible.

Keywords: *indeterminacy, a temporal data model, probability, CPR (Computer-based Patient Record)*

1. Introduction

Many applications such as AI, database management, multimedia system, history management system, medical informatics, etc., inevitably encounter indeterminacy temporal data because of the dynamic changes of the real world. They can not be efficiently represented and stored in database, especially the valid time of some incidents and their temporal relations can not be accurately determined [1, 2]. Therefore, many researchers have focused on building indeterminacy temporal data models. There are two broad categories of approaches emerged in the previous research. One is point-based semantics model and the other is interval-based semantics model. The typical point-based model is C. Combi proposed model based on time point, which is suitable to deal with a variety of medical data. But the model has some limitations in coping with data based on interval. The typical interval-based model is HAMP in which users can define time point and time interval with indeterminacy [3]. HAMP is focused on querying information about natural language expressions, while it can not express a finite union of intervals and represent relative time. NLTM (Temporal model of Natural language) model is also a interval-based model which has overcome HAMP's limitations, but NLTM still exists some faults [4]. For example, date elements and time-of-day are represented separately, so space cost is very high than models that store them unity.

Above-mentioned models, HAMP and NLTM all can express determinacy information and indeterminacy information. But query results are only "uncertain" when users query indeterminacy information. Potential information among these indeterminacy data can not been get when users are querying in HAMP or NLTM model. In order to overcome these shortcomings, we propose an indeterminacy temporal data model based on probability, named BPTM (Temporal model based on probability), to manage the indeterminacy temporal semantics of medical data.

Firstly, Section 2 and section 3 present our tuple-timestamp method to represent and store these temporal data including determinacy and indeterminacy data. In section 4, we introduce the temporal primitives to process temporal relations needed in BPTM. A new probability method is brought forward to get potential information among these indeterminacy

Received May 11, 2013; Revised June 23, 2013; Accepted July 25, 2013

data. At last a query example based on CPR (Computer-based Patient Record) is given in section 5 to show that our method is effective and feasible.

2. Temporal Concepts and Terms

Temporal DBMS has three kinds' styles' of time:

a) Valid-time [1, 5]: a period time in which a real event remains true.

b) Transaction-time: the time when a database object happens.

c) User-defined time [6]: the time that users input according to their needs.

Events are always associated with valid time and transaction time in temporal database. We only deal with valid time because the main purpose of the latter one is validating database, and moreover, it brings a powerful cost in terms of computing complexity, storage capacity and performance.

The timestamp types for representing BPTM are time points, intervals, duration and temporal elements. In general, the standard Gregorian calendar is adopted, which allows timestamps to be declared at any of the following collection of granularity: year, month, day, hour, minute and second.

Temporal DBMS has five kinds styles of temporal data [7-10]:

a) Chronon: We assume time domain TD is a non-empty, finite, totally ordered set. TD with the corresponding domain ELEM, models the time domain. Its elements are termed Chronon. Chronon is a non-decomposable time interval of some fixed minimal duration. Second is utilized as the Chronon in this paper.

b) Instant: a fixed time point on time axis. It relates to instantaneous situations.

c) Interval [7]: an movable and continuous time period.

d) Period: an immovable interval

A period is an immovable interval and very useful data style, but it is not supported by business service DBMS and SQL92. This paper introduces a method to simulate period with both instant, one instant means the begin of the period, the other means the end of the period. In addition, periods are not entirely orderly and have seven temporal relations [5, 11]. Figure 1 shows the seven temporal relations of two periods.

Duration: Duration is the length between two time points. Duration of uncertain interval is uncertainty and it has the minimum and maximal values. The expression of duration is an ordered sequence of time value on time domain. It can be different granularities such as year, month, day and hour and so on.

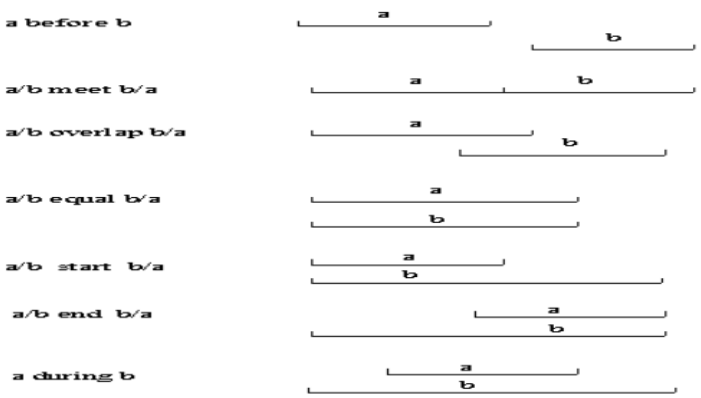

Figure 1. Seven Temporal Relations of Periods

3. Indeterminacy Temporal Information

Valid-time can be expressed by a or many points, an interval, duration and a period. We adopt a period to show a valid time. For example, Disease attribute in CPR (Computer-based Patient Record) is a temporal one (change over time), its valid time is sometime determinacy,

sometime indeterminacy (Disease begins and ends at a movable time). We have resolved the problem by providing upper bound and lower bound for the begin and end time of disease.

a. The representation of indeterminacy data

Attributes set of temporal relationship is composed of non-temporal attributes (Values do not change over time) and temporal attributes (values change with time) [8]. We express a temporal attribute as a couple of attribute value and period, namely (v, <t1, t2>), v is a temporal attribute name, <t1,t2> is a couple of begin time and end time of temporal attribute, namely tuple-timestamp. If the valid time of temporal attribute is determinacy, t1 and t2 become time points and begin time and end time is equal. If one of t1 and t2 is a period at least, valid time of temporal attribute is indeterminacy [9].

In the BPTM, all types of time are represented as a tuple form <t1, t2>, t1 or t2 is expressed as a period to show a valid time. When determinacy temporal information is a instant, the instant is denoted as <t1, t2>, t1 is equal to t2. If the instant is indeterminacy, it is denoted as <t1, t2>, t1 is not equal to t2. When the indeterminacy temporal information is a period, the period is still denoted as <t1, t2>, but t1 and t2 are all two-tuples, namely t1 or t2 has start time and end time. Compared with HATM and NLTM in flexibility, BPTM model has a great advantage in the unified form of expression.

b. The storage structure of indeterminacy data

Indeterminacy and determinacy valid time are all stored in the same table structure. Although Oracle do not provide data style about periods, we can stimulate period with Date style. If valid time is determinacy, we need two date fields to imitate. If valid time is indeterminacy (at least one of t1 and t2 is a period), we use four Date fields to determine upper and lower bound of begin and end point of valid time. In the Oracle DBMS, a field of date style can denote year-month-day and hour-minute-second simultaneously.

If indeterminacy temporal data is <t1, Duration > or <Duration, t2>, we adopt three kinds of database schemas to solve the Duration problem. The forward mode can solve <t1, Duration >, Backward mode can solve <Duration, t2>. The two modes can figure out the other unknown t1 or t2 by means of adding and subtraction.

4. Temporal Relations of Indeterminacy Temporal Information

The temporal query language and processing are key contents in temporal management and have close connection. This paper introduces a temporal model BPTM based on relations, so the temporal query language is also relational query one including extension of SQL. Relational data model is able to process the added valid-time and transaction time.

Snodgrass brought forward probability method to solve the relations of indeterminacy temporal data by defining the “Before” relation of indeterminacy instant. Based on which, the paper extends this method by adding two new temporal primitive definitions of indeterminacy instant as well as indeterminacy period: “BeforeI” and “simultaneity”. By introducing an argument called “Fuzziness” to describe the degree of indeterminacy, meanwhile the “Fuzziness” is modified by NiaveBayes classifier to ensure accuracy of the mined indeterminacy temporal data.

a. Temporal relations of indeterminacy period.

Seven relations of periods can be summed up to “<”or “<<” relation between a kind of time points. If two end-points of a period are determinacy, its temporal relation can be achieved by using temporal relation of time points. If two end-points of a period are indeterminacy, the query result may be ambiguity. The paper adopts probabilistic approach to resolve the problem about indeterminacy temporal information.

(1) Probabilistic approach [12]

There is a prerequisite before using probabilistic approach that any incident happens at any time instant during the period with equal probability. In addition, one incident and the other incident has no any relations.

Probabilistic Ordering [7, 12]: Given there are two indeterminacy time points α and β, the probability of α before β, that is probability of α≤β, namely Before (α,β), can be defined as:

$$\Pr[\alpha \le \beta] = \sum_{i \in A \wedge j \in B \wedge i \le j} \Pr[\alpha = i] \times \Pr[\beta = j] \quad (1)$$

(2) “Simultaneity” temporal primitive

In order to obtain more flexibility to query indeterminacy temporal information, we define two new temporal primitives based on seven temporal relations in Figure 1: “simultaneity” and “non-simultaneity”.

Given incident A occurs at a and finishes at b, incident B occurs at c and finishes at d. Obviously, “a< b’ and “c< d” are all true. a, b, c and d can be time points and also time periods.

Definition 1 “simultaneity”: Given “a≤c< b” or “c< a< d” is true, then incident A and incident B will have simultaneity relation. “simultaneity” definition includes five of all the relations in Figure 1 (overlap, during, equal, start, end).

Definition 2 “non-simultaneity”: Given “b≤c” or “d≤ a” is true, then incident A and incident B will have non-simultaneity relation. “Non-simultaneity” definition includes two relations in Figure1 (meet, before).

(3) New temporal primitive “BeforeI” and Fuzziness

The query of indeterminacy temporal information often has not an explicit answer. If the computed probability is too small, it is no good help users’ decision-makings. In order to resolve the problem, we introduce an argument “γ”, which is probable value that can be achieved at least according to user’s experience.

Definition 3 Fuzziness:

We call “γ” as “fuzziness”, which the value of “γ” is between ‘0’ to ‘1’. If the value of “γ” is bigger, and then the query result of indeterminacy temporal information is more meaningful for users.

In SQL, “Before [6, 8]” is algebra relation “≤”, the “Before” relation of any both time points can be described as Before (α,β)=α≤β. But for indeterminacy temporal information can not use “Before” operation [12]. So we define a new operation “BeforeI” which includes three arguments such as fuzziness “γ”, incident A and incident B with indeterminacy valid-time.

Definition 4 “BeforeI” operation:

$$\text{BeforeI}(\alpha, \beta, \gamma) = \{\text{True} \mid Pr[\alpha \le \beta] \ge \gamma\} \cup \{\text{False} \mid Pr[\beta < \alpha] \ge \gamma\} \quad (2)$$

The result set of BeforeI (a, b, γ) includes four elements such as {}, {True}, {False}, {True, False}. If BeforeI (a, b, γ) ={True}, the relation of ”a≤b” is true. If BeforeI (a, b, γ)={False}或{}, the relation of “a ≤b” is false. If BeforeI (a, b, γ) = {True, False}, the query result is indeterminacy, but the value of probability, which result is true or false, must not be smaller than fuzziness “γ”.

b. Modifying fuzziness “γ”

Before mining the indeterminacy temporal data, the argument “γ” as fuzziness is inputted by users. But different “γ” by different users offered all has some deviation which will influence the accuracy of the mining results. NiaveBayes classifier is used to modify fuzziness “γ” for ensuring better accuracy of the mined indeterminacy temporal data.

The modifying procedure is as follows:

(1) Initializing arguments “γ” and “p” which is step width.
(2) Preparing data set for an evaluation function---fit(x), which can get classification accuracy given a “γ” by NiaveBayes classifier training.

S = load('Data.mat');
x = S.Data;
DecA: a decision-making attribute（The second column of X, DecA = x(:,2)）;
ConA: a condition attribute（The first column of X, ConA = x(:,1)）;

The method of 10 fold cross validation is used to compute the classification accuracy of sample. The implementation procedure of the function fit(x) is as follows:

```
indices = crossvalind('Kfold',DecA,10);
cp = classperf(DecA);
for k = 1:10
test = (indices == k);
train = ~test;
nb = NaiveBayes.fit(ConA(train,:),DecA(train,:));
class = nb.predict(ConA(test,:));
classperf(cp,class,test);
end
```

```
        Fit = cp.CorrectRate;
    end
(3) For a given fuzziness "γ", computing probability value and its class mark.
        for j = 1:length(x)
        if x(j,1) >= r
          x(j,2) = 1;
        else
        x(j,2) = 0;
        end
        end
(4) Initializing the best classification accuracy. pbest = fit(x);
(5) Through many times iterations to find the optimal fuzziness "γ". Specific process is as
    follows:
        for i = 1:m
        r = r + step;
        for j = 1:length(x)
        if x(j,1) >= r
          x(j,2) = 1;
       else
          x(j,2) = 0;
       end
     end
     pref = fit(x);
     if pref < pbest
        pbest = pref;
        rbest = r;
     end
   end
   P = pbest;
   R = rbest;
End
```

Figure 2 is a program flow chart for Modifying fuzziness "γ".

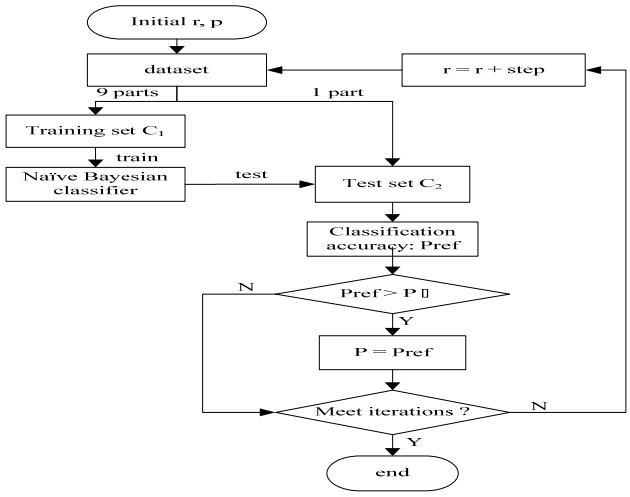

Figure 2. The Program Flow Chart for Modifying Fuzziness "γ"

5. The Mining of Indeterminacy Temporal Information

Using above-mentioned method, indeterminacy temporal information in CPR is mined. First, some data about hypertension and arteriosclerosis has been filtered from Database, and then we want to continue analyzing "simultaneity" relation when we find time of illness attack

and recovery is indeterminacy. For example, disease A and B have valid-time as follows in Table 1, ES is the earliest time of onset of illness, LS is the latest time of onset of illness, EF is the earliest time of illness recovery, and LF is the latest time of illness recovery.

Table 1. Stimulation of Valid-time

Name	LS	ES	EF	LF
A	1999-5-30	1999-6-18	2000-2-10	2000-10-1
B	1999-6-8	1999-6-29	2002-10-10	2002-12-3

Given time of disease A attack is ‘a’ and the time of recovery is ‘b’, the time of disease B attack is ‘c’ and the time of recovery is ‘d’, then temporal relations of 1999-5-30≤a≤1999-6-18, 2000-2-10≤b≤2000-10-1, 1999-6-8≤c≤1999-6-29 and 2002-10-10≤d≤2002-12-3 are all true, the modified fuzziness “γ” is equal to 0.6 which is obtained by NiaveBayes classifier training. Now, we need to confirm the relations of a, b, c and d by estimating if the relation of “a≤c< b” or “c< a< d” is true. Because a, b, c and d are all indeterminacy, we need probability method to compute their relations.

First, we compute Pr(a ≤c) by using formula 1. Length of ‘a’ and ‘c’ is 20 and 22 days respectively as granularity with a day. ‘a’ and ‘c’ have equal probability in each own period. So the probability is computed as follows:

Pr(a ≤c) =(1/20×1/22)×12+(1/20×1/22)×13+ ……+(1/20×1/22)×21+[(1/20×1/22)×22] ×10=0.875

Using probability approach, other probability results of indeterminacy temporal information are computed as follows:

Pr[a>c]=0.125 Pr[c<b]=1 Pr[c ≥ b]=0

Compared with γ=0.6, Pr[a ≤ c]=0.875> γ Pr[a>c]=0.125< γ Pr[c<b]=1>γ Pr[c ≥ b]=0<γ our goal will be to obtain Before I (a, c, γ)={True} and Before I (c, b, γ)={True}. This shows that ‘a ≤c’ and ‘c< b’ are all true, namely ‘a≤c< b’ is true. The concluded results are determinacy, namely disease A and B is “simultaneity” relation. If one result is {True}, the other is {True, False}, and then we need to compute “Pr[a≤c<b]” or “Pr[c<a<d]” and show their indeterminacy results for all users. For computing “Pr[a≤c<b]” or “Pr[c<a<d]” is very complex, we introduce a simple method for a case of “Pr[a≤c<b]” as follows:

$$Pr[a \le c < b] = Pr[c < b] + Pr[a \le c] - 1 \quad (3)$$

Using above-mentioned methods, a temporal DBMS based on CPR is constructed to implement the mining of indeterminacy medical temporal data. Figure 3 is probability computing results of CPRS.

Figure 3. Probability Computing Results

6. Conclusion

In order to overcome some shortcomings of current temporal models, an optimization model is proposed [13], named BPTM (Temporal model based on probability), to manage the indeterminacy temporal semantics of indeterminacy data. Firstly, we present our tuple-timestamp method to represent and store these temporal data including determinacy and indeterminacy data. Then we introduce the temporal primitives to process temporal relations needed in BPTM. A new probability method is brought forward to get potential information among these indeterminacy data. At last a temporal CPRS is constructed to implement the mining of indeterminacy medical temporal data. This CPRS is uncertain system to supply a good help for doctors to make a clinical diagnosis [14]. This CPRS example related to query also shows that our method is effective and feasible.

However, our model still has some faults. For example, the query speed will slow when data records exceed one hundred thousand or more. And therefore we shall optimize the query algorithm on temporal information and the extension of temporal index and join operator in the future.

Acknowledgment

This work was supported by Tianjin Natural Science Foundation (Grant No.07JCZDJC06700).

References

[1] Zhang Shichao, Yan Xiaowei, Nie Wenlong. A few problems intemporal database. *Journal of Guangxi Normal University*.1995; 13(4): 10-14.
[2] Zhou Xiaoning. Research on CPR. *Medical information.* 1998; 11(1): 6-8.
[3] C Combi, G Pozzi. HMAP - A temporal data model managing intervals with different granularities and indeterminacy. *The VLDB Journal*. 2008; l9(4): 294–311.
[4] Xiaowei ZHANG. A Temporal Data Model for Handling Uncertain Temporal Medical information. *Journal of Computational Information Systems*. 2012; 8(10): 3971–3978.
[5] Allen JF. Maintaining Knowledge about Temporal Intervals. CACM. 1983; 6(11): 832-843.
[6] Dyresson CE, RT Snodgrass. Timestamp Semantics and Representation. *Information Systems*. 1993; 18(3): 143-166.
[7] Jensen CS, L Mark. Temporal Specialization and Generalization. *IEEE Transactions on Knowledge and Data Engineering*. 1994; 6(6): 954-974.
[8] Myrach T, GF Knolmayer, R Barnert. On Ensuring Keys and Referential Integrity in the Temporal DataBase Language TSQL2. In HM Haav, B Thalheim. *editors*. DataBase and Information systems. Proceeding of the Second International Baltic Workshop, Tallinn. Reaserch Track, Tampere University of Technology Press, 1996; 1: 171-181.
[9] Richard T Snodgrass. *The temporal Query Language TQuel*. ACM Transactions on DataBase System. 1987; 12(2): 247-298.
[10] Richard T Snodgrass. *Developing Time-Oriented Database Application In SQL*. Morgan Kanfmann Publishers. 2000.
[11] Jia Chao. *Processing extension of indeterminacy time interval with temporal query language*. Computer development. 2002; 21-25.
[12] Curtis E Dyreson, Richard T Snodgrass. *Supporting Valid-Time Indeterminacy*. ACM Transactions on Database Systems. 1998; 23(1):1-57.
[13] Lei Zhao, Yihua Zhong, Yilin Wan. Bi-Level Multi-criteria Multiple Constraint Level Optimization MODELS and Its Application. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 11(5): 3.
[14] Y Zhu, Q Feng, J Wang. Neural network-based adaptive passive output feedback control for MIMO uncertain system. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2012; 10(6): 1263-1272.