Data-Centric Probabilistic Process: A PhD. Symposium Paper for ICSOC

Haizhou Li

Universit Blaise Pascal Laboratoire LIMOS - UMR 6158 Complexe Scientifique des
Czeaux, 63177 AUBIERE cedex, France
li@isima.fr

Abstract. The theory of business process management system actively plays a principle role in the domain of semantic web service and business process management. This paper briefly described my current Ph.D work result and the perspective about specification and verification of data-centric probabilistic business process system which integrating traditional probabilistic business process and probabilistic database. The probabilistic business process system is described by probabilistic automaton which is an abstracted framework for the transition system. By integrating probabilistic database, probabilistic automaton would be provided greater expressive power to handle the probabilistic event and high volume probabilistic data.

Keywords: Data-centric probabilistic process system, probabilistic automata, probabilistic database.

1 Introduction

This paper is a description of my Ph.D. work. My work focus on data-centric business process management system which integrates the theory of business process management described by transition system methodology (ex. probabilistic business process modelled by probabilistic automata)and extending class of incomplete database(incomplete database,probabilistic database and uncertain database). My current research focus on a specification and verification of data-centric probabilistic business process system which are capable to represent probabilistic transition system and uncertain data with the help of probabilistic automata and probabilistic database. This paper organized as follows. Section 2 states the problem and the main challenge of my research. Section 3 provides a coarse model of data-centric business process system depicted by probabilistic automata integrating probabilistic database. In Section 5, I tries to compare the envisioned research outcome with current literature. Last section is a conclusion.

2 Problem Statement

The research of business process management methodology is blooming and researched by numerous researchers in disparate way. Number of techniques and

A. Ghose et al. (Eds.): ICSOC 2012, LNCS 7759, pp. 406–411, 2013.

models are integrated in the domain of business process management. My work mainly focus on the process system and uncertain data. Due to the main task of business process management, the model and the methodology are served to improve the business efficiency and aid the business people to manage their business. So we would like to propose updated framework to enhance the expressive power of existing models. Supply chain management is always a suitable entry point as an experimental field for a novel method of business process management. In the traditional supply chain management system, researchers are fond of process system or called transition system to describe the work flows of supply chain system. Theoretically, a transition system could be interpreted as a finite automaton (In this paper, we focus on finite transition system. Moreover in reality, there would be infinite process occurred due to errors, deadlocks, design problem,etc.). Normally, an automaton consists of four components as follows:

- A set of finite states $\{s_0, s_1, ..., s_n\}$ which stand for the current state of the system, denoted as *S*.
- A non-empty start states s_0.
- A set of actions $\{a_0, a_1 ..., a_n\}$denoted as *Act*.
- Transition relations(or called steps) denoted as $\rightarrow \subseteq S \times Act \times S$ present the transition from one state to another assigned by actions

The behaviours of automata are described by traces denoted as *trace(S)*or execution fragments which are a sequence of alternating states and actions starting with a state. A trace would be exampled as follows: $trace(S)_i = \{s_0 a_0 s_1 a_1 ... s_n a_n\}$. Figure 1 shows a common work flow (described in finite state automaton). We would like to know if we selected different product, how the process would be. Imaging if we would like to query data from high volume dataset such as database and also show this data in the automata. There are several ways to introduce data into the process, converting the actions of the transitions is a kind of expressive method and we could find out clearly how the data flow and how the data influence the transitions in the automata at the same time. So we expand the notion of actions to guarding formulas which could query or update the database to return a boolean value *true* of $false$ such that we capture a kind of notion of data-centric process model which could be described by finite state automata ([6] discussed a similar web service model).

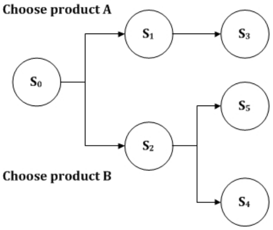

Fig. 1. An example work flow

2.1 Probabilistic Database Centric

After introducing the notion of data-centric service, integrating deterministic database is not a innovative domain. Nowadays, in the wake of information explosion, the system is facing to the problem of selecting duplicate data from the different sources, managing uncertain data etc. Probabilistic database is a ideal tool to manage the uncertain data.

A probabilistic database (in [2], [3]) is a set of finite possible complete database which distributed by probabilities denoted as $P-DB : \{W, Pr\}$, $W = \{D_1, D_2, ...Dn\}$, Pr is a function $Pr : W \rightarrow [0, 1]$ such that $\sum_{D \in W} Pr(D) = 1$. Formally, a probabilistic instance consists of a collection of tuples which are distributed by probabilities and relation between tuples are disjoint (or called mutually exclusive) or independent. Table 1 shows an example of a probabilistic instance in which t_1 and t_2 are disjoint (capable to be considered as a cluster) and t_2 and t_3 are independent, as well as the possible world of this probabilistic instance.

Table 1. Example of probabilistic database and its possible world

	ID	pr
t_1	A	70%
t_2	B	30%
t_3	c	50%

PW	tuples	Pr
W_1	$\{t_1\}$	35%
W_2	$\{t_2\}$	15%
W_3	$\{t_1, t_3\}$	35%
W_4	$\{t_2, t_3\}$	15%

As a result, we would like to integrate probabilistic database as the data source in the data-centric process model described above. Figure 2 shows a data-centric probabilistic process model by reusing the example in figure 1 and formal definition of data-centric probabilistic process will be provided in section 3. In the state of art of investigating the probability and processes, the methodology of probabilistic process is a critical field. It is valuable to review the methodology of probabilistic process.

2.2 Probabilistic Process

The methodology of automata has been widely utilized to illustrate the structure and the behaviours of a system but researchers found that it was not adequate for more complicated utilizations. In order to increase the expressive power of the automata, researches considered the probability of the events and the branches in the automata. Probabilistic business process model (introduced in [4], [8])which described by probabilistic automata (depicted in [5]) is an attractive innovation to the theory of process system. As its name described, probabilistic automata reuse the main components of the traditional automata in additional to provide probability on the transitions. The transitions of probabilistic automata are assigned not only actions but also probability distributions $\{\mu_0, \mu_1, ..., \mu_n\}$ denoted as *Distri(S)* such that the transition relation of probabilistic automata is denoted as $\rightarrow \subseteq S \times Act \times Distri(S) \times S$.

2.3 Research Problems and Challenge

Our approach attempts to integrate probabilistic process and probabilistic database in the form of data-centric probabilistic automata. It would be a combination of two asynchronous semantics of probability methodology. In the theory of probabilistic process, the probability distribution from different branches in one step are mutually exclusive but in contrast, due to the guarding formulas querying independent tuples from probabilistic database, the transition relation by considering guarding formulas would be independent transitions. So the research problems and challenge of my work are as follows:

- Modelling data-centric probabilistic business process especially finding a proper way to present these two asynchronous semantics of probability as above described.
- Verifying if the data-centric probabilistic process model could be reduced to the model of probabilistic process and reuse some algorithms which have been proposed in the probabilistic process.
- Verifying the simulation relation between two data-centric probabilistic processes if their guarding formulas queried one certain probabilistic instance, one uncertain probabilistic instance or two different probabilistic instances and compare with probabilistic process if the simulation relation algorithm of probabilistic process could be reused.
- Programming a prototype system to make this model into reality.

3 Proposed Solution

In this section, I would like to introduce this data-centric process model which captures the stochastic transitions and probabilistic data. Data-centric probabilistic process system has a collection of probabilistic distributions on the transitions and data. This system could provide an ability to manage uncertain data and support probability branching choices in the domain of business process management.

In the model of data-centric probabilistic process system, there are five principle components as follows.

- A set of finite states $\{s_0, s_1, ..., s_n\}$ which stand for the current state of the system, denoted as *S*.
- A nonempty start states s_0.
- The real world is modelled by probabilistic database which denoted as *p-D*.
- A set of actions $\{a_0, a_1 ..., a_n\}$denoted as *Act*. An action a_n may be a guarding formula f_n. A guarding formula queries probabilistic database *p-D* returning a boolean value *true* or *false* or update the database (insertion, deletion or modification) returning a boolean value *true* or *false*..
- Transition relations which are signed by actions and accompany with probabilistic distributions $\{\mu_0, \mu_1, ..., \mu_n\}$ which denoted as *Distri(S)*. Transition relations denoted as $\rightarrow \subseteq S \times Act \times Distri(S) \times S$

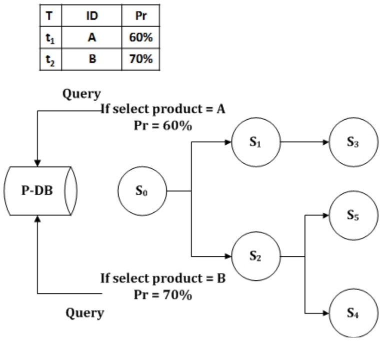

Fig. 2. An example of data-centric probabilistic business process

The behaviour of this model is described by traces. For a given set of state S,there are several traces $\{trace(S)_0, trace(S)_1, ..., trace(S)_n\}$ denoted as *Trace(S)* corresponding to the transition relations *Tr(S)*. A possible trace includes principle components of the model and depict the execution order of the states inside. A trace is just like the follows $trace(S)_0 = \{s_0a_0\mu_0s_1a_1\mu_1s_3a_3\mu3...s_na_n\mu_n\}$. In the context of finite process model, we just consider complete trace which is definitely finite. Figure 2 shows a data-centric probabilistic process and its possible world modelled by an instance of probabilistic database.

4 Comparison between the Envisioned Research Outcome and Current Literature

The data-centric probabilistic process model integrated probabilistic process and probabilistic database ([2] and [3] described the semantic of probabilistic database and the query method of probabilistic database in details). [4] provided a specification model of probabilistic business process and [9] described a mathematical methodology about probabilistic process. [8] discussed the method of testing probabilistic pre-order which could be a infrastructure if we discussed the simulation relation of data-centric probabilistic business model. Comparing with traditional probabilistic business process model, data-centric probabilistic process model introduced the accessing ability to probabilistic database and accommodated massive data manipulation. The theory of probabilistic database extends from incomplete database (Specifically presented in [1] and [7]) and provide a tool to manage and store enormous stochastic data. The idea of atomic process accessing database is inspired by [6] which was introduced "Colombo" model. "Colombo" model is a data-centric process model without considering probabilistic transitions or probabilistic data and it is specific in composition of web service. Our data-centric probabilistic process model introduced probabilistic business process and utilized probabilistic database to focus on stochastic data management and random event occurrence in the field of process system.

Meanwhile, our model described a framework of data-centric probabilistic process system and it could be utilized not only in web service but also other business process management system by adding specification features.

5 Conclusion

This paper coarsely stated the principle problem of my Ph.D work and introduced a model of data-centric probabilistic process system which integrate probabilistic process and probabilistic database. In future, we would like to improve this model and discuss the verification of this model. Following the theoretical work, the prototype of this model would be designed to attempt to handle the realistic applications.

References

1. Imielinkski, T., Lipski Jr., W.: Incomplete information and dependencies in relational databases. In: Proc. ACM-SIGMOD International Conference on Management of Data, pp. 178–184 (1983)
2. Dalvi, N., Suciu, D.: Efficient query evaluation on probabilistic databases. University of Washington Technical Report (TR 04-03-04) (2004)
3. Dalvi, N., Re, C., Suciu, D.: Query evaluation on probabilistic database. IEEE Data Engineering Bulletin 29(1), 25–31 (2006)
4. Deutch, D., Milo, T.: On models and query languages for probabilistic processes. SIGMOD Record 39(2), 27–38 (2010)
5. Stoelinga, M.: An introduction to probabilistic automata. Alea jacta est: verification of probabilistic, real-time and parametric systems. PhD thesis, University of Nijmegen, the Netherlands, ch. 2 (2002)
6. Berardi, D., Calvanese, D., De Giacomo, G., Hull, R., Mecella, M.: Automatic composition of transition-based semantic web services with messaging. In: Proceedings of the 31st VLDB Conference, pp. 613–624 (2005)
7. Grahne, G.: The problem of incomplete information in relational database. University of Helsinki (1991)
8. Jonsson, B., Wang, Y.: Testing pre-orders for probabilistic processes can be characterized by simulations. Theoretical Computer Science 282, 33–51 (2002)
9. Lustig, Y., Nain, S., Vardi, M.Y.: Synthesis from probabilistic components. In: Proc. CSL 2011. LIPICS, vol. 12, pp. 412–427 (2011)