Journal of Computer Science and Cybernetics, V.31, N.4 (2015), 305–321
DOI: 10.15625/1813-9663/31/4/5742

A PROBABILISTIC RELATIONAL DATABASE MODEL AND ALGEBRA

NGUYEN HOA

Department of Information Technology, Saigon University;
nguyenhoa@sgu.edu.vn

Abstract. This paper introduces a probabilistic relational database model, called PRDB, for representing and querying uncertain information of objects in practice. To develop the PRDB model, first, we represent the relational attribute value as a pair of probabilistic distributions on a set for modeling the possibility that the attribute can take one of the values of the set with a probability belonging to the interval which is inferred from the pair of probabilistic distributions. Next, on the basis representing such attribute values, we formally define the notions as the schema, relation, probabilistic functional dependency and probabilistic relational algebraic operations for PRDB. In addition, a set of the properties of the probabilistic relational algebraic operations in PRDB also are formulated and proven.

Keywords. Probability distribution, probabilistic triple, probabilistic relation, probabilistic functional dependency, probabilistic relational algebraic operation

1. INTRODUCTION

As we all know, the classical relational database model is very useful for modeling, designing and implementing large-scale systems. However, this model is restricted for representing and handling uncertain and imperfect information of objects in the real world [1, 2]. For example, applications of the classical relational database model cannot deal with queries as find all players that are 80-90% likely to be the top scorers of English Premier League, in year 2015; nor find all patients who are at least 70% likely to catch a cirrhosis or hepatitis, etc.

So far, there have been many relational database models studied, developed and built based on the probability theory for modeling objects about which information may be uncertain and imperfect to overcome the limitation of the classical relational database model. Such models are called probabilistic relational database models [3–6].

Some models were built by extending each classical relation to a probabilistic relation as in [7, 8]. That is, each tuple in a probabilistic relation has an uncertainty degree, measured by a probability value for it belonging to the relation.

Some models like [5, 9], assigning a probability to an attribute value to represent the uncertain level for the attribute could take the value. Some models in [10–12] allowed the value of each attribute associated with a probability interval to represent the uncertainty degree of both the probability and the value that the attribute could take. More flexibly, the model in [13] represented the value of each attribute as a probability distribution on a set. It means that each attribute associated with a set of values and a probability distribution expressing possibility that the attribute can take one of values of the set with a probability computed from the distribution. The models mentioned above

are extensions with probability of the classical relational database model in different levels to represent uncertain information of objects in practice. However, these models still have the restriction. Particularly, the probability value that is assigned to each tuple or each attribute value in the models [5, 7–9, 13] is not always determined exactly in practice. The models in [11, 12] overcame the shortcoming by estimating a probability interval for each attribute value of the relations. However, in [11, 12], each attribute was only assigned to a definite value with a respective probability interval, but in the real world, there are situations in which we do not know exactly the value of each attribute whereas we know that the attribute may take one of the values of a certain set. In addition, the probabilistic functional dependencies were not defined in models mentioned above. In [14] the probabilistic functional dependent notion were presented, however, the limitations of representing the probability value for a tuple belonging to a relation also as in [7].

In this paper, using the probabilistic triple concept in the probabilistic object base model [15], we build a new probabilistic relational database model (PRDB) with all of the basic probabilistic relational algebraic operations that can overcome the mentioned shortcomings of the models in [11–13] to represent and manipulate uncertain information in practice. PRDB model is also a next developmental step of the model proposed in [4].

Basic probability definitions as a mathematical foundation for PRDB are presented in Section 2. The schema, relation and probabilistic functional dependency in PRDB are introduced in Section 3. Section 4, 5 and 6 present probabilistic relational algebraic operations and their properties in PRDB. Finally, Section 7 concludes the paper and outlines further research directions in the future.

2. PROBABILITY AND PROBABILISTIC COMBINATION STRATEGIES

In this section, some probability definitions and probabilistic combination strategies are presented as the basis for representing and handling uncertain information in PRDB.

2.1. Probability distribution functions and probabilistic triples

For representing uncertain attribute values in PRDB, we use probability distribution functions and probabilistic triples in [15]. Concepts of the probability distribution function and probabilistic triple respectively are defined as below.

Definition 1. Let X be a finite set, *a probability distribution function* α over X is a mapping $\alpha : X \to [0, 1]$ such that $\sum_{x \in X} \alpha(x) \leq 1$.

An important probability distribution function which often encountered in practice is the uniform distribution $u(x) = 1/|X|, \forall x \in X$. For example, if $X = \{24, 48, 72\}$, the uniform distribution u over X is $u(x) = 1/3,\ \forall x \in \{24, 48, 72\}$.

Definition 2. A *probabilistic triple* $\langle X, \alpha, \beta \rangle$ consists of a finite set X, a probability distribution function α over X, and a function $\beta : X \to [0, 1]$ such that $\alpha(x) \leq \beta(x),\ \forall x \in X$ and $\sum\limits_{x \in X} \beta(x) \geq 1$ hold.

Informally, a probabilistic triple $\langle X, \alpha, \beta \rangle$ assigns each element $x \in X$ a probability interval $[\alpha(x), \beta(x)]$ to express the uncertainty degree of x in X. This assignment is consistent in the sense that each element $x \in X$ is assigned a probability $p(x) \in [\alpha(x), \beta(x)]$ such that $\sum_{x \in X} p(x) = 1$.

The probabilistic triple is a tool to represent uncertain information of objects in practice. For example, when examining a patient, a doctor may be unsure about what disease the patient is

suffered from. However, if the doctor is sure that the patient's disease is *hepatitis* or *cirrhosis* with a probability between 40% and 60%, then this knowledge may be encoded by the probabilistic triple $\langle\{hepatitis, cirrhosis\}, 0.8u, 1.2u\rangle$. Here, u is the uniform distribution function over $\{hepatitis, cirrhosis\}$, $0.8u$ and $1.2u$ are probability distribution functions α and β respectively with $\alpha(x) = 0.8u(x) = 0.8(1/2) = 0.4$ and $\beta(x) = 1.2u(x) = 1.2(1/2) = 0.6$, $\forall x \in \{hepatitis, cirrhosis\}$.

2.2. Probabilistic combination strategies

Given two events e_1 and e_2 having probabilities in the intervals $[L_1, U_1]$ and $[L_2, U_2]$, one may need to compute the probability intervals of the conjunction event $e_1 \wedge e_2$, disjunction event $e_1 \vee e_2$, or difference event $e_1 \wedge \neg e_2$. In this paper, we employ the conjunction, disjunction, and difference strategies given in [15, 16] as presented in Table1, where $\otimes$, $\oplus$, and $\ominus$ denote the conjunction, disjunction, and difference operators, respectively.

Strategy	Operators
Ignorance	$([L_1, U_1] \otimes_{ig} [L_2, U_2]) \equiv [\max(0, L_1 + L_2 - 1), \min(U_1, U_2)]$ $([L_1, U_1] \oplus_{ig} [L_2, U_2]) \equiv [\max(L_1, L_2), \min(1, U_1 + U_2)]$ $([L_1, U_1] \ominus_{ig} [L_2, U_2]) \equiv [\max(0, L_1 - U_2), \min(U_1, 1 - L_2)]$
Independence	$([L_1, U_1] \otimes_{in} [L_2, U_2]) \equiv [L_1 \cdot L_2, U_1 \cdot U_2]$ $([L_1, U_1] \oplus_{in} [L_2, U_2]) \equiv [L_1 + L_2 - (L_1 \cdot L_2), U_1 + U_2 - (U_1 \cdot U_2)]$ $([L_1, U_1] \ominus_{in} [L_2, U_2]) \equiv [L_1 \cdot (1 - U_2), U_1 \cdot (1 - L_2)]$
Positive correlation (when e_1 implies e_2, or e_2 implies e_1)	$([L_1, U_1] \otimes_{pc} [L_2, U_2]) \equiv [\min(L_1, L_2), \min(U_1, U_2)]$ $([L_1, U_1] \oplus pc[L_2, U_2]) \equiv [\max(L_1, L_2), \max(U_1, U_2)]$ $([L_1, U_1] \ominus_{pc} [L_2, U_2]) \equiv [\max(0, L_1 - U_2), \max(0, U_1 - L_2)]$
Mutual exclusion (when e_1 and e_2 are mutually exclusive)	$([L_1, U_1] \otimes_{me} [L_2, U_2]) \equiv [0, 0]$ $([L_1, U_1] \oplus_{me} [L_2, U_2]) \equiv [\min(1, L_1 + L_2), \min(1, U_1 + U_2)]$ $([L_1, U_1] \ominus_{me} [L_2, U_2]) \equiv [L_1, \min(U_1, 1 - L_2)]$

Table 1: Examples of probabilistic combination strategies

In following sections, the notation $[L_1, U_1] \leq [L_2, U_2]$ is used to replace $L_1 \leq L_2$ and $U_1 \leq U_2$ whereas the notation $[L_1, U_1] \subseteq [L_2, U_2]$ is used to replace for $L_2 \leq L_1$ and $U_1 \leq U_2$.

2.3. Conjunction, disjunction and difference of probabilistic triples

For building algebraic operations such as the join, intersection, union and difference of probabilistic relations in PRDB, the conjunction, disjunction and difference of probabilistic triples in [15] are used as the basis for combining the probability of attribute values in outcome relations of the operations. First, the conjunction of probabilistic triples is defined as follows.

Definition 3. Let $pt_1 = \langle V_1, \alpha_1, \beta_1\rangle$ and $pt_2 = \langle V_2, \alpha_2, \beta_2\rangle$ be two probabilistic triples, and $\otimes$ be a probabilistic conjunction strategy. The *conjunction* of pt_1 and pt_2 under $\otimes$, denoted by $pt_1 \otimes pt_2$, is the probabilistic triple $pt = \langle V, \alpha, \beta\rangle$, such that:

1. $V = \{v \in V_1 \cap V_2 | [\alpha_1(v), \beta_1(v)] \otimes [\alpha_2(v), \beta_2(v)] \neq [0, 0]\}$, and
2. $[\alpha(v), \beta(v)] = [\alpha_1(v), \beta_1(v)] \otimes [\alpha_2(v), \beta_2(v)],\ \forall v \in V$.

Example 1. Let $pt_1 = \langle\{hepatitis, cirrhosis\}, 0.8u, 1.2u\rangle$ and $pt_2 = \langle\{hepatitis\},\ u,\ u\rangle$ be probabilistic triples, then $pt_1 \otimes_{in} pt_2$ with the independence probabilistic conjunction strategy is the probabilistic triple $pt = \langle\{hepatitis\},\ 0.4u,\ 0.6u\rangle$.

Next, the disjunction and difference of probabilistic triples in turn are defined as below.

Definition 4. *Let $pt_1 = \langle V_1, \alpha_1, \beta_1\rangle$ and $pt_2 = \langle V_2, \alpha_2, \beta_2\rangle$ be two probabilistic triples, and $\oplus$ be a probabilistic disjunction strategy. The disjunction of pt_1 and pt_2 under $\oplus$, denoted by $pt_1 \oplus pt_2$, is the probabilistic triple $pt = \langle V, \alpha, \beta\rangle$, such that:*

1. $V = V_1 \cup V_2$, and
2. $[\alpha(v), \beta(v)] = \begin{cases} [\alpha_1(v), \beta_1(v)],\ \forall v \in V_1 - V_2 \\ [\alpha_2(v), \beta_2(v)],\ \forall v \in V_2 - V_1 \\ [\alpha_1(v), \beta_1(v)] \oplus [\alpha_2(v), \beta_2(v)],\ \forall v \in V_1 \cap V_2 \end{cases}$

Definition 5. Let $pt_1 = \langle V_1, \alpha_1, \beta_1\rangle$ and $pt_2 = \langle V_2, \alpha_2, \beta_2\rangle$ be two probabilistic triples, and $\ominus$ be a probabilistic difference strategy. The *difference* of pt_1 and pt_2 under $\ominus$, denoted by $pt_1 \ominus pt_2$, is the probabilistic triple $pt = \langle V, \alpha, \beta\rangle$, such that:

1. $V = V_1 - \{v \in V_1 \cap V_2 | [\alpha_1(v), \beta_1(v)] \ominus [\alpha_2(v), \beta_2(v)] = [0, 0]\}$, and
2. $[\alpha(v), \beta(v)] = \begin{cases} [\alpha_1(v), \beta_1(v)], v \in V_1 - V_2 \\ [\alpha_1(v), \beta_1(v)] \ominus [\alpha_2(v), \beta_2(v)], \forall v \in V_1 \cap V_2. \end{cases}$

3. SCHEMA AND PROBABILISTIC RELATIONS

3.1. Probabilistic relational schemas

A probabilistic relational schema in PRDB describes a set of attributes of a set of certain objects of which each attribute is associated with probabilistic triples as the following definition.

Definition 6. A *probabilistic relational schema* is a pair $R = (\boldsymbol{U}, \wp)$, where $\boldsymbol{U} = \{A_1, A_2 \ldots, A_k\}$ is a set of pairwise different attributes $\wp$ is a function that maps each attribute $A \in U$ to a non-empty set of probabilistic triples f whose each element has the form $\langle V, \alpha, \beta\rangle$ where V is a subset of the domain of A.

Note that as in the classical relational database, for simplicity, the notations $R(\boldsymbol{U}, \wp)$ and R can be used to replace $R = (\boldsymbol{U}, \wp)$. In addition, the domain of each attribute A is denoted by $dom(A)$.

3.2. Probabilistic relations

A probabilistic relation is an instance of a probabilistic relational schema in which each attribute may be take uncertain values represented by a probabilistic triple as the following definition

Definition 7. Let $\boldsymbol{U} = \{A_1, A_2, \ldots, A_k\}$ be a set of k pairwise different attributes A *probabilistic relation* r over the probabilistic relational schema $R(\boldsymbol{U}, \wp)$, is a finite set $\{t|t = (\langle V_1, \alpha_1, \beta_1\rangle, \langle V_2, \alpha_2, \beta_2\rangle, \ldots, \langle V_k, \alpha_k, \beta_k\rangle)\}$ in which each element t is a list of k probabilistic triples such that $\langle V_i, \alpha_i, \beta_i\rangle$ belongs to the set $f_i = \wp(A_i)$, for every $i = 1, 2, \ldots, k$

For simplicity, each element $t = (\langle V_1, \alpha_1, \beta_1\rangle, \langle V_2, \alpha_2, \beta_2\rangle, \ldots, \langle V_k, \alpha_k, \beta_k\rangle)$ in a probabilistic relation is also called a *tuple* t as in a classical relation. Each probabilistic triple $\langle V_i, \alpha_i, \beta_i\rangle$ represents the uncertain value of the attribute A_i of the tuple t, the notation $t.A_i$ denotes the probabilistic triple, that is $t.A_i = \langle V_i, \alpha_i, \beta_i\rangle$. Each tuple t in the relation r over $R(\boldsymbol{U}, \wp)$ is called a tuple over the set of the attributes $\boldsymbol{U}$. For each set of attributes $X \subseteq \{A_1, A_2, \ldots, A_k\}$, the notation $t[X]$ is used to denote the rest of t after eliminating the value of attributes not belonging to X.

From Definition 2, it is noted that, each attribute A_i of a tuple t in the relation r over $R(\boldsymbol{U}, \wp)$ only takes one of the values $v_i \in V_i$ with a probability $p(v_i) \in [\alpha_i(v_i), \beta_i(v_i)]$. Therefore, each probabilistic relation r corresponds with a set of classical relations $w(r)$ such that each tuple t of the relation $r_w \in w(r)$ has the form $t = (v_1, v_2, \ldots v_k)$, where $v_i \in V_i$. As in [13,17], the model PRDB adopts the closed world assumption (CWA). It means, for each tuple t, every value $v \in dom(A_i) - V_i$ has the probability 0.

Now, the notion of a probabilistic relational database is defined as follows.

Definition 8. A *probabilistic relational database* over a set of attributes is a set of probabilistic relations corresponding with the set of their probabilistic relational schemas.

Note that, if we only care about a unique relation over a schema then we can unify its symbol name with its schema's name.

Example 2. A simple probabilistic relational database about patients at the clinic of a hospital can be structured as Tables 2, 3 and 4. In the database, the attributes PATIENT_NAME, WEIGHT MEDICAL_HISTORY and DISEASE describe the information about the name, weight, medical history and disease of each patient. Some other attributes can be DURATION, COST that define the treatment duration and treatment cost per day of each patient. In reality, while diagnosing the disease of each patient is not always determined certainly by the physicians. Similarly, the treatment duration, treatment cost for patients are also not known accurately even as the patients know about their diseases. Here, the conventional units for treatment duration and treatment cost are established as date and 1000 (VND). The unit for the physicians' experience is year. We note that, in the database, the name of each relation and the name of its schema are identical, the set of probabilistic triples $\wp(A)$ for each attribute A in the schemas of the relations consists of all probabilistic triples $\langle X, \alpha, \beta\rangle$ such that X is a subset of the domain of A. Some attributes have been removed (for simplicity) and they do not affect the illustration of the probabilistic relational database model. In addition, each probabilistic triple $\langle V, u, u\rangle$ with $V = \{v\}$, will be represented as a single value v Because if the attribute takes such a probabilistic triple, then actually it only takes a value v with the probability is 1 (Definition 2). In other words, the attribute certainly takes the value v. At that time, the attribute and its value have the same meaning

as those in the classical relational database. Therefore, we can say that the the classical relational database model is a particular case of PRDB.

PATIENT_ID	PATIENT_NAME	WEIGHT	MEDICAL_HISTORY
PT0421	*N.V. An*	$\langle\{71, 72\}, u, u\rangle$	$\langle\{bronchitis\}, u, u\rangle$
PT3829	*L.T. Huong*	$\langle\{50\}, u, u\rangle$	$\langle\{cholecystitis, gall\text{-}stone\}, 0.8u, u\rangle$
PT2938	*T.V.Hung*	$\langle\{60\}, u, u\rangle$	$\langle\{cholecystitis\}, u, u\rangle$

Table 2: Relation PATIENT

PHYSICIAN_ID	PHYSICIAN_NAME	EXPERIENCE
DT005	*L.V Cuong*	$\langle\{30, 31\}, u, u\rangle$
DT093	*N.V. Son*	$\langle\{25, 26\}, u, u\rangle$
DT102	*N.T.L Huong*	$\langle\{6\}, u, u\rangle$

Table 3: Relation PHYSICIAN

PATIENT_ID	PHYSICIAN_ID	DISEASE	DURATION	COST
PT0421	*DT005*	$\langle\{lung\ cancer, tuberculosis\}, 0.8u, 1.2u\rangle$	$\langle\{400, 500\}, u, u\rangle$	$\langle\{300, 350\}, u, u\rangle$
PT3829	*DT093*	$\langle\{hepatitis, cirrhosis\}, u, u\rangle$	$\langle\{30, 40\}, u, u\rangle$	$\langle\{60, 70\}, u, u\rangle$
PT2938	*DT102*	$\langle\{hepatitis\}, u, u\rangle$	$\langle\{30\}, u, u\rangle$	$\langle\{60\}, u, u\rangle$

Table 4: Relation DIAGNOSE

For defining the probabilistic functional dependent concept in PRDB we first propose a probability measure to determine the equal degree of two values of the same attribute for two different tuples in a relation

Definition 9. Let t_1 and t_2 be two tuples in a probabilistic relation r, A be an attribute of r and $\otimes$ be a probabilistic conjunction strategy. The *probability interval* for the values of the attribute A of two tuples t_1 and t_2 respectively are equal under $\otimes$ is

$$p(t_1A =_{\otimes} t_2A) = \begin{cases} [\sum\limits_{v\in W} \alpha(v), \min(1, \sum\limits_{v\in W} \beta(v))], & \text{if } W \neq \emptyset \\ [0, 0], & \text{otherwise} \end{cases}$$

where $t_1.A = \langle V_1, \alpha_1, \beta_1\rangle$, $t_2.A = \langle V_2, \alpha_2, \beta_2\rangle$, $W = \{(v_1, v_2) \in V_1 \times V_2 | v_1 = v_2\}$ and $[\alpha(v), \beta(v)] = [\alpha_1(v_1), \beta_1(v_1)] \otimes [\alpha_2(v_2), \beta_2(v_2)]$, $\forall v = (v_1, v_2) \in W$.

For example, in relation DIAGNOSE above, if t_1DISEASE $= \langle\{hepatitis\}, u, u\rangle$ and t_2DISEASE $= \langle\{hepatitis, cirrhosis\}, u, u\rangle$, then the probability interval for two patients represented by t_1 and t_2 has the same disease, under $\otimes_{in}$, is $p(t_1\text{DISEASE} =_{\otimes in} t_2\text{DISEASE}) = [0.5, 0.5]$

Now, the probabilistic functional dependency in PRDB is extended from the functional dependency in classical relational database as below.

Definition 10. Let $\boldsymbol{U} = \{A_1, A_2, \ldots, A_k\}$ be a set of k pairwise different attributes $R(\boldsymbol{U}, \wp)$ be a probabilistic relational schema, r be any probabilistic relation over R, $\otimes$ be a probabilistic conjunction strategy, $X = \{A_i, \ldots, A_l\}$ and $Y = \{A_j, \ldots, A_m\}$ be two subsets of $\boldsymbol{U}$. A *probabilistic functional dependency* of Y on X under $\otimes$ over R, denoted by $X \rightarrow_{\otimes} Y$, if and only if

$$\forall t_1, t_2 \in r, p(t_1[X] =_{\otimes} t_2[X]) \leq p(t_1[Y] =_{\otimes} t_2[Y]),$$

where $p(t_1[X] =_{\otimes} t_2[X]) = p(t_1.A_i =_{\otimes} t_2.A_i) \otimes \ldots \otimes p(t_1.A_l =_{\otimes} t_2.A_l)$ and $p(t_1[Y] =_{\otimes} t_2[Y]) = p(t_1.A_j =_{\otimes} t_2.A_j) \otimes \ldots \otimes p(t_1.A_m =_{\otimes} t_2.A_m)$.

An obvious example of the probabilistic functional dependency is every attribute A_i depending on the set $\{A_1 A_2 \ldots A_k\}$ that consists of all attributes of the schema R. It is noted that in the classical database, one can consider the probability for two values of an attribute are equal is only 0 or 1, so the functional dependency in the classical relational database is a particular case of the probabilistic functional dependency in this definition.

As for the classical relational database, the keys of a schema in PRDB are the basis for recognising a tuple in a probabilistic relation. In the model and management systems of the classical relational database, key attributes are constrained not to take the value NULL [2]. Similarly, in PRDB, we assume that the value of each key attribute is always certain and definite. The key concept of probabilistic relational schema is defined using the probabilistic functional dependency as follows.

Definition 11. Let $R(\boldsymbol{U}, \wp)$ be a probabilistic relational schema, r be any relation over R and $\otimes$ be a probabilistic conjunction strategy, a set of attributes $K \subseteq \boldsymbol{U}$ is called a *key of* R under $\otimes$ if the value of each attribute of K is always certain in r and there is a probabilistic functional dependency $K \rightarrow_{\otimes} \boldsymbol{U}$ such that not to exist any subset of K has the properties.

In the patient database above, if assuming that each patient has a unique identifier corresponding with the value of the attribute PATIENT_ID and the identifier differs from every identifier of other patients, then by the definition, PATIENT_ID is a key of the schema **PATIENT** under any probabilistic conjunction strategy.

4. SELECTION OPERATION ON A PROBABILISTIC RELATION

4.1. Syntax of selection conditions

As for the classical relational database, the selection is a basic algebraic operation in PRDB. The result of a selection query over a probabilistic relation r of a schema R is a probabilistic relation r' over R such that tuples of r' have attribute values satisfying the selection condition of this query.

Before defining the selection operation, we present the formal syntax and semantics of selection conditions by extending those definitions of the classical relational database for PRDB taking into account probability intervals to satisfy the selection conditions. We start with the syntax of selection expressions as follows.

Definition 12. Let R be a schema in PRDB and χ be a set of relational tuple variables, θ be a binary relation of $\{ =, \neq, \leq, <, >, \geq \}$. Then *selection expressions* are inductively defined and have one of the following forms:

1. $x.A\theta v$, where $x \in \chi$, A is an attribute in R and v is a value.

2. $x.A_1\theta_{\otimes}x.A_2$, where $x \in \chi$, A_1 and A_2 are two different attributes in R and $\otimes$ is a probabilistic conjunction strategy.

3. $E_1 \otimes E_2$, where E_1 and E_2 are selection expressions on the same relational tuple variable, $\otimes$ is a probabilistic conjunction strategy.

4. $E_1 \oplus E_2$, where E_1 and E_2 are selection expressions on the same relational tuple variable, $\oplus$ is a probabilistic disjunction strategy.

It is noted that, the selection expression $x.A_1\theta_{\otimes}x.A_2$ in the definition is a general form of the selection expression $xA_1 =_{\otimes} x.A_2$ in [18].

Example 3. Consider the relational schema **DIAGNOSE** in the patient database above, the selection of "all patients who have treatment duration less than 40 days or pay the treatment cost not less than 60 (thousand VND/day)", can be expressed by the selection expression $x.\text{DURATION} < 40 \oplus x.\text{COST} \geq 60$.

In PRDB, each selection condition is a logical combination of selection expressions with probability intervals needed to be satisfied as the following definition.

Definition 13. Let R be a schema in PRDB, *selection conditions* are inductively defined as follows:

1. If E is a selection expression and $[L, U]$ is a subinterval of $[0, 1]$, then $(E)[L, U]$ is a selection condition

2. If ϕ and ψ are selection conditions on the same tuple variable, then $\neg\phi$, $(\phi \wedge \psi)$, $(\phi \vee \psi)$ are selection conditions.

Example 4. Consider the patient database in Example 2, with schema **DIAGNOSE**, then the selection of "all patients who have hepatitis and pay treatment cost not less than 70 (thousand VND/day) with a probability from 0.4 to 0.6 or have treatment duration not less than 30 days with a probability of at least 0.9", can be done by using the selection condition $(x.\text{DISEASE} = \textit{hepatitis} \otimes x.\text{COST} \geq 70)[0.4, 0.6] \vee (x.\text{DURATION} \geq 30)[0.9, 1.0]$.

4.2. Semantics of selection conditions

For defining the semantics of selection conditions, the probabilistic interpretations of selection expressions are first presented as in the definition below.

Definition 14. Let R be a relational schema in PRDB, r be a relation over R, x be a tuple variable and t be a tuple in r. The *probabilistic interpretation* of selection expressions with respect to R, r and t, denoted by $prob_{Rr,t}$, is the partial mapping from the set of all selection expressions to the set of all closed subintervals of $[0, 1]$ that is inductively defined as follows:

1. $$prob_{R,r,t}(x.A\theta\, v) = \begin{cases} \left[\sum\limits_{c\in W} \alpha(c), \min(1, \sum\limits_{c\in W} \beta(c))\right], & \textit{if } W \neq \emptyset \\ [0,\, 0], & \text{otherwise} \end{cases},$$

   where $t.A = \langle V, \alpha, \beta\rangle$ and $W = \{c \in V | c\theta v\}$.

2. $prob_{R,r,t}(x.A_1\,\theta_{\otimes}\,x.A_2) = \begin{cases} \left[\sum\limits_{c\in W}\alpha(c), \min(1, \sum\limits_{c\in W}\beta(c))\right], & \text{if } W \neq \emptyset \\ [0,\,0], & \text{otherwise}, \end{cases}$

   where $t.A_1 = \langle V_1, \alpha_1, \beta_1\rangle, t.A_2 = \langle V_2, \alpha_2, \beta_2\rangle$ and
   $[\alpha(c), \beta(c)] = [\alpha_1(c_1), \beta_1(c_1)] \otimes [\alpha_2(c_2), \beta_2(c_2)], \forall c \in W = \{(c_1, c_2) \in V_1 \times V_2 | c_1\theta c_2\}$

3. $prob_{R,r,t}(E_1 \otimes E_2) = prob_{R,r,t}(E_1) \otimes prob_{R,r,t}(E_2)$.

4. $prob_{R,r,t}(E_1 \oplus E_2) = prob_{R,r,t}(E_1) \oplus prob_{R,r,t}(E_2)$.

Intuitively, $prob_{R,r,t}(x.A\theta v)$ is the probability interval for the attribute A of the tuple t having a value c such that $c\theta v$ and $prob_{R,r,t}(x.A_1\theta_{\otimes}x.A_2)$ is the probability interval for the attributes A_1 and A_2 of the tuple t having values c_1 and c_2, respectively such that $c_1\theta c_2$.

Example 5. Let r denote the relation DIAGNOSE in the patient database in Example 2 and R denote the schema of DIAGNOSE, consider the tuple t_2 (the second tuple) in r, then $prob_{R,r,t2}(x.\text{DISEASE} = \textit{hepatitis}) = [0.5, 0.5]$ and $prob_{R,r,t2}(x.\text{COST} \geq 70) = [0.5, 0.5]$. So, the probabilistic interpretation of the selection expression $\psi_{in} = x.\text{DISEASE} = \textit{hepatitis} \otimes_{in} x.\text{COST} \geq 70$ (under the independent probabilistic conjunction strategy $\otimes_{in}$) with respect to t_2 is $prob_{R,r,t2}(\psi_{in}) = [0.5, 0.5] \otimes_{in} [0.5, 0.5] = [0.25, 0.25]$.

In PRDB, each selection condition is a logical combination of selection expressions with probability intervals needed to be satisfied. In other words, the logical satisfaction of a selection condition is the satisfaction of probabilistic bounds associated with selection expressions in this selection condition. The satisfaction or semantics of a selection condition under a probabilistic interpretation is defined as below

Definition 15. Let R be a relational schema in PRDB, r be a relation over R and $t \in r$. The *satisfaction* of selection conditions under $prob_{R,r,t}$ is defined as follows:

1. $prob_{R,r,t} \models (E)[L, U]$ if and only if (iff) $prob_{R,r,t}(E) \subseteq [L, U]$.
2. $prob_{R,r,t} \models \neg\phi$ iff $prob_{R,r,t} \models \phi$ does not hold.
3. $prob_{R,r,t} \models \phi \wedge \psi$ iff $prob_{R,r,t} \models \phi$ and $prob_{R,r,t} \models \psi$
4. $prob_{R,r,t} \models \phi \vee \psi$ iff $prob_{R,r,t} \models \phi$ or $prob_{R,r,t} \models \psi$

Note that, in the classical database, the concepts of selection expression and selection condition are identical and we can consider probability intervals $[L, U]$ in selection conditions being always equal to $[1.0, 1.0]$. This also means that the concept of satisfaction of selection conditions in the classical relational database model is a particular case of the concept of satisfaction of selection conditions in PRDB.

Example 6. Consider tuple t_2 in the relation r over the schema R as in Example 5. It is easy to see that $prob_{R,r,t2} \models (x.\text{DISEASE} = \textit{hepatitis} \otimes_{in} x.\text{COST} \geq 70)[0.2, 0.8]$, because $prob_{R,r,t2}(x.\text{DISEASE} = \textit{hepatitis} \otimes_{in} x.\text{COST} \geq 70) = [0.25, 0.25] \subseteq [0.2, 0.8]$.

Now, the selection operation on a relation in PRDB is defined as follows.

Definition 16. Let R be a relational schema in PRDB, r be a relation over R and ϕ be a selection condition over a tuple variable x. The *selection* on r with respect to ϕ, denoted by $\sigma_\phi(r)$, is the relation r' over R, including all satisfied tuples of the selection condition ϕ

$$r' = \{t \in r | prob_{R,r,t} \models \phi\}$$

Example 7. Consider the relation DIAGNOSE in the patient database in Example 2. Then, the query "Find all patients who have hepatitis and pay treatment cost not less than 70 (thousand VND/day) with a probability of at least 0.25" can be done by the selection operation $r' = \sigma_\phi(\text{DIAGNOSE})$ with $\phi = (x.\text{DISEASE} = \textit{hepatitis} \otimes_{in} x.\text{COST} \geq 70)[0.25, 1.0]$.

The selection is implemented by checking the satisfaction of all tuples in DIAGNOSE for the selection condition ϕ. From Example 5 and 6, we can easily see only the tuple t_2 satisfies ϕ because $prob_{R,r,t2}(\phi) = [0.25, 0.25] \subseteq [0.2, 1.0]$. So, the result of the selection is the relation r' as in Table 5.

PATIENT_ID	PHYSICIAN_ID	DISEASE	DURATION	COST
PT3829	*DT093*	$\langle\{\textit{hepatitis}, \textit{cirrhosis}\}, u, u\rangle$	$\langle\{30, 40\}, u, u\rangle$	$\langle\{60, 70\}, u, u\rangle$

Table 5: Relation r'$=\sigma_\phi$ (DIAGNOSE)

5. OTHER ALGEBRAIC OPERATIONS ON PROBABILISTIC RELATIONS

As for the classical relational database, other basic operations on probabilistic relations are the projection, Cartesian product, join, intersection, union, and difference. We now extend those operations of the classical relational database for PRDB taking into account uncertain value of relational attributes.

5.1. Projection

A projection of a probabilistic relation on a set of attributes is a new probabilistic relation in which only the attributes in that set are considered for each tuple of the new relation as the following definition.

Definition 17. Let $R = (\boldsymbol{U}, \wp)$ be a probabilistic relational schema, r be a relation over R and $\boldsymbol{L}$ be a subset of attributes of $\boldsymbol{U}$. The *p rojection* of r on $\boldsymbol{L}$ denoted by $\Pi_L(r)$, is the probabilistic relation r' over the schema R', determined by:

1. $R' = (\boldsymbol{L}, \wp')$ and $\wp'(A) = \wp(A)$, $\forall A \in \boldsymbol{L}$.
2. $r' = \{t' = t[\boldsymbol{L}] | t \in r\}$, i.e., r' consists of tuples t' to achieve from the tuples $t = (\langle V_1, \alpha_1, \beta_1\rangle, \ldots, \langle V_k, \alpha_k, \beta_k\rangle) \in r$ by eliminating every $\langle V_j, \alpha_j, \beta_j\rangle$ that $A_j = \langle V_j, \alpha_j, \beta_j\rangle$ and $A_j \notin \boldsymbol{L}$.

5.2. Cartesian product

For the Cartesian product of two probabilistic relations, as in the classical relational database, assuming the set of attributes of their schemas is disjoint. Also, for the operation being commutative, we assume every k-tuple $t = (\langle V_1, \alpha_1, \beta_1\rangle, \ldots, \langle V_k, \alpha_k, \beta_k\rangle)$ is an un-ordered list.

Definition 18. The probabilistic relational schemas $R_1(\boldsymbol{U}_1, \wp_1)$ and $R_2(\boldsymbol{U}_2, \wp_2)$ are *Cartesian product-compatible* if and only if $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ have not any common attribute.

Note that, any schemas $R_1(\boldsymbol{U}_1, \wp_1)$ and $R_2(\boldsymbol{U}_2, \wp_2)$ can be made Cartesian product-compatible by renaming of attributes in $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$.

Now, the Cartesian product of two probabilistic relations in PRDB is extended from that operation of the classical relational database as follows.

Definition 19. Let r_1 and r_2 be two probabilistic relations over the Cartesian product-compatible schemas $R_1 = (\boldsymbol{U}_1, \wp_1)$ and $R_2 = (\boldsymbol{U}_2, \wp_2)$, respectively. The *Cartesian product* of r_1 andr_2, denoted by $r_1 \times r_2$, is the probabilistic relation r over R determined by:

1. $R = (\boldsymbol{U}, \wp)$, where $\boldsymbol{U} = \boldsymbol{U}_1 \cup \boldsymbol{U}_2, \wp(A) = \wp_1(A)$ if $A \in \boldsymbol{U}_1$ and $\wp(A) = \wp_2(A)$ if $A \in \boldsymbol{U}_2$.
2. $$r = \Big\{ t = (\langle V_1, \alpha_1, \beta_1\rangle, \ldots, \langle V_k, \alpha_k, \beta_k\rangle, \langle V_{k+1}, \alpha_{k+1}, \beta_{k+1}\rangle, \ldots, \langle V_{k+m}, \alpha_{k+m}, \beta_{k+m}\rangle) \Big| \\ t_1 = (\langle V_1, \alpha_1, \beta_1\rangle, \ldots, \langle V_k, \alpha_k, \beta_k\rangle) \text{ and} \\ t_2 = (\langle V_{k+1}, \alpha_{k+1}, \beta_{k+1}\rangle, \ldots, \langle V_{k+m}, \alpha_{k+m}, \beta_{k+m}\rangle),\ t_1 \in r_1 \text{ and } t_2 \in r_2 \Big\}.$$

5.3. Join

The join of two probabilistic relations in PRDB is extended from the natural join of two relations in the classical relational database. The join only works with relations whose schemas are join-compatible as the definition below

Definition 20. The probabilistic relational schemas $R_1(\boldsymbol{U}_1, \wp_1)$ and $R_2(\boldsymbol{U}_2, \wp_2)$ are *join -compatible* if and only if the domains of two attributes of the same name A in $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$, respectively are identical

From Definition 6, we can see that for two attributes of the same name A in $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ of two join–compatible schemas $R_1(\boldsymbol{U}_1, \wp_1)$ and $R_2(\boldsymbol{U}_2, \wp_2)$ then $\wp_1(A) = \wp_2(A)$. For building the join of two probabilistic relations, we first define the join of two tuples in PRDB as follows.

Definition 21. Let t_1 and t_2 be two tuples on two sets of attributes $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ respectively, and $\otimes$ be a probabilistic conjunction strategy. The *join* of t_1 and t_2 under $\otimes$, denoted by $t_1 \bowtie_\otimes t_2$, is the tuple t on $\boldsymbol{U}_1 \cup \boldsymbol{U}_2$ defined by:

1. $t.A = t_1.A,\ \forall A \in \boldsymbol{U}_1 - \boldsymbol{U}_2$.
2. $t.A = t_2.A,\ \forall A \in \boldsymbol{U}_2 - \boldsymbol{U}_1$.
3. $t.A = t_1.A \otimes t_2.A,\ \forall A \in \boldsymbol{U}_1 \cap \boldsymbol{U}_2$.

It is easy to see that $t_1 \bowtie_\otimes t_2 = t_2 \bowtie_\otimes t_1$ for every probabilistic conjunction strategy, i.e., the join of two tuples is commutative.

Definition 22. Let r_1 and r_2 be two probabilistic relations over the join-compatible schemas $R_1 = (\boldsymbol{U}_1, \wp_1)$ and $R_2 = (\boldsymbol{U}_2, \wp_2)$, respectively and let $\otimes$ be a probabilistic conjunction strategy. The *join* of r_1 and r_2 under $\otimes$, denoted by $r_1 \bowtie_\otimes r_2$, is the probabilistic relation r over the schema R, determined by:

1. $R = (\boldsymbol{U}, \wp)$ where $\boldsymbol{U} = \boldsymbol{U}_1 \cup \boldsymbol{U}_2, \wp(A) = \wp_1(A)$ if $A \in \boldsymbol{U}_1 - \boldsymbol{U}_2, \wp(A) = \wp_2(A)$ if $A \in \boldsymbol{U}_2 - \boldsymbol{U}_1$ and $\wp(A) = \wp_1(A) = \wp_2(A)$ if $A \in \boldsymbol{U}_1 \cap \boldsymbol{U}_2$ (because $\wp_1(A) = \wp_2(A)$ Definition 20).

2. $r = \{t = t_1 \bowtie_{\otimes} t_2 | t_1 \in r_1 \text{ and } t_2 \in r_2\}$.

Example 8. Given two relations PATIENT_1 and PATIENT_2 as in Tables 6 and 7, then the result of the join of them under the probabilistic conjunction strategy $\otimes_{in}$ is the relation PATIENT computed as in Table 8.

PATIENT_ID	MEDICAL_HISTORY
PT0421	$\langle\{bronchitis\}, u, u\rangle$
PT3829	$\langle\{cholecystitis,\ gall\text{-}stone\}, 0.8u, u\rangle$

Table 6: Relation PATIENT_1

PATIENT_NAME	MEDICAL_HISTORY
N.V. An	$\langle\{bronchitis\}, u, u\rangle$
L.T. Huong	$\langle\{cholecystitis,\ cirrhosis\}, 0.8u, u\rangle$

Table 7: Relation PATIENT_2

PATIENT_ID	PATIENT_NAME	MEDICAL_HISTORY
PT0421	*N.V. An*	$\langle\{bronchitis\}, u, u\rangle$
PT3829	*L.T. Huong*	$\langle\{cholecystitis\}, 0.16u, 0.25u\rangle$

Table 8: Relation $\text{PATIENT} = \text{PATIENT}_1 \bowtie_{\otimes in} \text{PATIENT}_2$

Here, the name of each relation and its schema is identical, the set of probabilistic triples $\wp(A)$ for each attribute A in the schemas consists of all probabilistic triples $\langle X, \alpha, \beta\rangle$ such that $X \subseteq dom(A)$.

5.4. Intersection, union, and difference

Intersection, union and difference of two probabilistic relations, respectively, over the same schema is a probabilistic relation over that schema, in which the value of attributes in common tuples of those two relations associated by a probabilistic combination strategy. A common tuple of two probabilistic relations over the same schema is the tuple whose key attributes' values are identical in both relations. It is due to the uncertainty of attribute values, a common tuple of two probabilistic relations is not completely identical as that of two relations in the classical relational database.

First, the intersection of two tuples as the basis for the intersection of two probabilistic relations is defined as follows.

Definition 23. Let t_1 and t_2 be two tuples on the same set of attributes $\boldsymbol{U}$ and $\otimes$ be a probabilistic conjunction strategy. The *intersection* of t_1 and t_2 under $\otimes$, denoted by $t_1 \cap_{\otimes} t_2$, is the tuple t on $\boldsymbol{U}$ defined by $t.A = t_1.A \otimes t_2.A$ for every $A \in \boldsymbol{U}$.

Definition 24. Let r_1 and r_2 be two probabilistic relations over the same schema $R(\boldsymbol{U}, \wp)$, K be a key of R and $\otimes$ be a probabilistic conjunction strategy. The *intersection* of r_1 and r_2 under $\otimes$, denoted by $r_1 \cap_{\otimes} r_2$, is the probabilistic relation r over R defined by $r = \{t = t_1 \cap_{\otimes} t_2 | t_1 \in r_1, t_2 \in r_2 \text{ such that } t_1[K] = t_2[K]\}$.

It is noted that, the notation $t_1[K] = t_2[K]$ is used in the definition due to the value of each key attribute assumed is certain and definite as in the Definition 11.

Example 9. Consider two relations DIAGNOSE_1 and DIAGNOSE_2 over the same schema **DIAGNOSE**$(\boldsymbol{U}, \wp)$ as in Tables 9 and 10 where $\boldsymbol{U} = \{\underline{\text{PATIENT\_ID}}, \text{DISEASE}, \text{COST}\}$ and PATIENT_ID is a key, then the intersection of them under $\otimes_{in}$ is the relation DIAGNOSE computed as in Table 11.

PATIENT_ID	DISEASE	COST
PT0421	$\langle\{$*lung cancer*, *tuberculosis*$\}$, $0.8u$, $1.2u\rangle$	$\langle\{300, 350\}, u, u\rangle$
PT3829	$\langle\{$*hepatitis*, *cirrhosis*$\}, u, u\rangle$	$\langle\{60, 70\}, u, u\rangle$

Table 9: Relation DIAGNOSE_1

PATIENT_ID	DISEASE	COST
PT3830	$\langle\{$*lung cancers*$\}, u, u\rangle$	$\langle\{350, 400\}, u, u\rangle$
PT3829	$\langle\{$*hepatitis*, *gall-stone*$\}, u, u\rangle$	$\langle\{60, 70\}, u, u\rangle$
PT2938	$\langle\{$*hepatitis*$\}, u, u\rangle$	$\langle\{60\}, u, u\rangle$

Table 10: Relation DIAGNOSE_2

PATIENT_ID	DISEASE	COST
PT3829	$\langle\{$*hepatitis*$\}$, $0.25u$, $0.25u\rangle$	$\langle\{60, 70\}, 0.5u, 0.5u\rangle$

Table 11: Relation DIAGNOSE = $\text{DIAGNOSE}_1 \cap_{\otimes \text{in}} \text{DIAGNOSE}_2$

The union of two probabilistic relations over the same schema in PRDB are based on the union of tuples as below.

Definition 25. Let t_1 and t_2 be two tuples on the same set of attributes $\boldsymbol{U}$ and $\oplus$ be a probabilistic disjunction strategy. The *union* of t_1 andt_2 under $\oplus$, denoted by $t_1 \cup_{\oplus} t_2$, is the tuple t on $\boldsymbol{U}$ defined by $t.A = t_1.A \oplus t_2.A$ for every $A \in \boldsymbol{U}$.

Definition 26. Let r_1 and r_2 be two probabilistic relations over the same schema $R(\boldsymbol{U}, \wp)$, K be a key of R and $\oplus$ be a probabilistic disjunction strategy. The *union* of r_1 and r_2 under $\oplus$, denoted by $r_1 \cup_{\oplus} r_2$, is the probabilistic relation r over R defined by $r = \{t_1 \in r_1|$ there is not any tuple $t_2 \in r_2$ such that $t_1[K] = t_2[K]\} \cup \{t_2 \in r_2|$ there is not any tuple $t_1 \in r_1$ such that $t_2[K] = t_1[K]\} \cup \{t = t_1 \cup_{\oplus} t_2 | t_1 \in r_1, t_2 \in r_2$ such that $t_1[K] = t_2[K]\}$.

As for the intersection and union operations, for defining the difference operation of two probabilistic relations, we first define the difference operation of two tuples as follows.

Definition 27. Let t_1 and t_2 be two tuples on the same set of attributes $\boldsymbol{U}$ and $\ominus$ be a probabilistic difference strategy. The *difference* of t_1 and t_2 under $\ominus$, denoted by $t_1 -_{\ominus} t_2$, is the tuple t on $\boldsymbol{U}$ defined by $t.A = t_1.A \ominus t_2.A$ for every $A \in \boldsymbol{U}$.

Definition 28. Let r_1 and r_2 be two probabilistic relations over the same schema $R(\boldsymbol{U}, \wp)$, K be a key of R and $\ominus$ be a probabilistic difference strategy. The *difference* of r_1 andr_2 under $\ominus$, denoted by $r_1 -_{\ominus} r_2$, is the probabilistic relation r over R defined by $r = \{t_1 \in r_1|$ there is not any tuple $t_2 \in r_2$ such that $t_1[K] = t_2[K]\} \cup \{t = t_1 -_{\ominus} t_2 | t_1 \in r_1, t_2 \in r_2$ such that $t_1[K] = t_2[K]\}$.

6. PROPERTY OF ALGEBRAIC OPERATIONS

In this section, we propose some properties of the probabilistic relational algebraic operations in PRDB which are extended from those in the classical relational database. Clearly, these properties say that PRDB model is sound.

Theorem 1. *Let r be a probabilistic relation over the schema R in PRDB, ϕ_1 and ϕ_2 be two selection conditions. Then*

$$\sigma_{\phi 1}(\sigma_{\phi 2}(r)) = \sigma_{\phi 2}(\sigma_{\phi 1}(r)) = \sigma_{\phi 1 \wedge \phi 2}(r) \tag{1}$$

where, the last expression assumes that ϕ_1 and ϕ_2 have the same tuple variable.

The first property shows that the selections may be reordered.

Proof. Let $r_1 = \sigma_{\phi 1}(r)$, $r_2 = \sigma_{\phi 2}(r)$ and $r_{1\wedge 2} = \sigma_{\phi 1 \wedge \phi 2}(r)$. Then for each $t \in r$, it yields

$$\begin{aligned}
\sigma_{\phi 1}(\sigma_{\phi 2}(r)) =& \{t \in r_2 | prob_{R,r2,t} \models \phi_1\} \\
=& \{t \in r | (prob_{R,r,t} \models \phi_2) \wedge (prob_{R,r2,t} \models \phi_1)\} \\
=& \{t \in r | (prob_{R,r,t} \models \phi_2) \wedge (prob_{R,r,t} \models \phi_1)\} (\text{ because of } r_2 \subseteq r) \\
=& \{t \in r | prob_{R,r,t} \models \phi_1 \wedge \phi_2\} (\text{Definition 15}) \\
=& \sigma_{\phi 1 \wedge \phi 2}(r).
\end{aligned}$$

So, $\sigma_{\phi 1}(\sigma_{\phi 2}(r)) = \sigma_{\phi 1 \wedge \phi 2}(r)$ is proven. Equation $\sigma_{\phi 2}(\sigma_{\phi 1}(r)) = \sigma_{\phi 2 \wedge \phi 1}(r)$ is proven similarly. Since $\phi_1 \wedge \phi_2 \Leftrightarrow \phi_2 \wedge \phi_1$(the logical conjunction of selection conditions are commutative), hence $\sigma_{\phi 1 \wedge \phi 2}(r) = \sigma_{\phi 2 \wedge \phi 1}(r)$. Therefore, it results in $\sigma_{\phi 1}(\sigma_{\phi 2}(r)) = \sigma_{\phi 2}(\sigma_{\phi 1}(r))$ and so $\sigma_{\phi 1}(\sigma_{\phi 2}(r)) = \sigma_{\phi 2}(\sigma_{\phi 1}(r)) = \sigma_{\phi 1 \wedge \phi 2}(r)$. Thus, Theorem 1 is proven. □

Theorem 2. *Let r be a probabilistic relation over the schema R in PRDB, $\boldsymbol{A}$ and $\boldsymbol{B}$ be two subsets of attributes of$R$ and $\boldsymbol{A} \subseteq \boldsymbol{B}$. Then*

$$\Pi_A(\Pi_B(r)) = \Pi_A(r) \tag{2}$$

Proof. Because $\boldsymbol{A} \subseteq \boldsymbol{B}$, so $\boldsymbol{A} \cap \boldsymbol{B} = \boldsymbol{A}$. From Definition 17, it is easy to see $\Pi_A(\Pi_B(r)) = \Pi_{A \cap B}(r) = \Pi_A(r)$. Thus, equation (2) is proven. □

Theorem 3. *Let R_1, R_2 and R_3 be pairwise join-compatible schemas in PRDB, r_1, r_2 and r_3 be relations over R_1, R_2 and R_3 respectively. Let $\otimes$ be a probabilistic conjunction strategy. Then*

$$r_1 \bowtie_{\otimes} r_2 = r_2 \bowtie_{\otimes} r_1 \tag{3}$$

$$(r_1 \bowtie_{\otimes} r_2) \bowtie_{\otimes} r_3 = r_1 \bowtie_{\otimes} (r_2 \bowtie_{\otimes} r_3) \tag{4}$$

Equation (3) and (4) say that the join operation of probabilistic relations is commutative and associative.

Proof. Clearly, $r_1 \bowtie_{\otimes} r_2$ and $r_2 \bowtie_{\otimes} r_1$ are two relations over the same schema. By Definition 3, the conjunction of probabilistic triples is commutative (due to the commutativity of probabilistic conjunction strategies and the intersection of sets). Consequently, the join of tuples is commutative (by Definition 21). So, by Definition 22, it yields $r_1 \bowtie_{\otimes} r_2 = r_2 \bowtie_{\otimes} r_1$.

Since R_1, R_2 and R_3 are pairwise join-compatible, so the results of two sides of (4) are the relations over the same schema. Moreover, the intersection of sets has the associativity, by Definition 3, it follows that the conjunction of probabilistic triples is associative. From associativity of the classical relational join and by Definition 21, it is easy to see that the join of tuples which are based on the conjunction of probabilistic triples is associative. Thus, by Definition 22, it results in $(r_1 \bowtie_{\otimes} r_2) \bowtie_{\otimes} r_3 = r_1 \bowtie_{\otimes} (r_2 \bowtie_{\otimes} r_3)$. □

Because the Cartesian product is a particular case of the join, it yields the corollary straight of Theorem 3 below.

Corollary 1. *Let R_1, R_2 and R_3 be pairwise Cartesian product-compatible schemas in PRDB, r_1, r_2 and r_3 be relations over R_1, R_2 and R_3 respectively. Then*

$$r_1 \times r_2 = r_2 \times r_1 \tag{5}$$

$$(r_1 \times r_2) \times r_3 = r_1 \times (r_2 \times r_3) \tag{6}$$

Theorem 4. *Let r_1, r_2 and r_3 be probabilistic relations over the same schema R. Let $\otimes/\oplus$ be a probabilistic conjunction/disjunction strategy. Then*

$$r_1 \cap_{\otimes} r_2 = r_2 \cap_{\otimes} r_1 \tag{7}$$

$$(r_1 \cap_{\otimes} r_2) \cap_{\otimes} r_3 = r_1 \cap_{\otimes} (r_2 \cap_{\otimes} r_3) \tag{8}$$

$$r_1 \cup_{\oplus} r_2 = r_2 \cup_{\oplus} r_1 \tag{9}$$

$$(r_1 \cup_{\oplus} r_2) \cup_{\oplus} r_3 = r_1 \cup_{\oplus} (r_2 \cup_{\oplus} r_3) \tag{10}$$

Equations of (7), (8), (9) and (10) say that the intersection, union and difference of relations in PRDB are commutative and associative.

Proof. Equations in the theorem are proven respectively as follows:

Equations (7) and (8): From commutativity and associativity of the intersection of sets, it follows the conjunction of probabilistic triples has commutativity and associativity. Thus, the intersection of tuples, by Definition 23, has commutativity and associativity. So, the intersection of tuples that have the same key value in r_1, r_2 and r_3 respectively are commutative and associative. From that, it follows Equations (7) and (8).

Equations (9) and (10): From commutativity of the union, intersection of sets, the union of probabilistic triples (Definition 4) and the union of tuples (Definition 25), by Definition 26 it results in Equation (9).

For Equation (10), let K be the key used to determine common tuples of r_1, r_2 and r_3. Without loss of generality, we may assume that each tuple t belongs to one of the three relations r_1, r_2 and r_3 then there exists two tuples belonging to the two remaining relations, respectively such that the value of the key K of the three tuples respectively are always the same. This can be done by adding t to the relations in which it is missing and resetting $\alpha(v) = \beta(v) = 0$ for every v in the set V of values of each attribute $A \notin K$of the tuple t. Under this technical assumption, the result of each expression in Equation (10) is not changed and only the union case of two tuples in Definition 26 is relevant. Now, from associativity of the disjunction of probabilistic triples (Definition 4) and the union of tuples, Equation (10) obviously holds. □

7. CONCLUSION

In this paper, we have proposed a probabilistic relational database model, abbreviated to PRDB, as a development following the model built in [4]. PRDB is an extension of the classical relational database model with integrating uncertain values into the relational attributes. Such uncertain values are represented by a probabilistic triple. From that, the notions of schema, relation, probabilistic functional dependency and probabilistic relational algebraic operations have been defined formally and consistently to allow querying and manipulating uncertain information on relations of database. A set of basic properties of the algebraic operations in PRDB is also proposed as theorems and proven completely.

Toward applying PRDB for modeling and handling uncertain information in the real world, we will build a management system for PRDB with the familiar querying and manipulating language like SQL by developing more language and management system that had been built in [4]. Next work will be to integrate fuzzy set values into relational attributes for building a fuzzy and probabilistic relational database model to represent and manipulate objects about which information may be uncertain and imprecise.

REFERENCES

[1] E. F. Codd, "A relational model of data for large shared data banks," *Communications of the ACM*, vol. 13, no. 6, pp. 377–387, 1970.

[2] C. J. Date, *An introduction to database systems*, 8th ed. Addison-Wesley publ., 2008.

[3] Y. Li, J. Chen, and L. Feng, "Dealing with uncertainty: a survey of theories and practices," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 11, pp. 2463–2482, 2013.

[4] N. Hoa and T. D. Hieu, "A probabilistic relational data model for uncertain information," in *Information Science and Technology (ICIST), 2013 International Conference on.* IEEE, 2013, pp. 607–613.

[5] D. Suciu, D. Olteanu, C. Ré, and C. Koch, "Probabilistic databases," *Synthesis Lectures on Data Management*, vol. 3, no. 2, pp. 1–180, 2011.

[6] Z. Ma and L. Yan, *Advances in Probabilistic Databases for Uncertain Information Management.* Springer, 2013, vol. 304.

[7] N. Fuhr and T. Rölleke, "A probabilistic relational algebra for the integration of information retrieval and database systems," *ACM Transactions on Information Systems (TOIS)*, vol. 15, no. 1, pp. 32–66, 1997.

[8] R. Tang, R. Cheng, H. Wu, and S. Bressan, "A framework for conditioning uncertain relational data," in *Database and Expert Systems Applications.* Springer, 2012, pp. 71–87.

[9] M. Pittarelli, "An algebra for probabilistic databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 6, no. 2, pp. 293–303, 1994.

[10] R. Ross, V. Subrahmanian, and J. Grant, "Aggregate operators in probabilistic databases," *Journal of the ACM (JACM)*, vol. 52, no. 1, pp. 54–101, 2005.

[11] W. Zhao, A. Dekhtyar, and J. Goldsmith, "Databases for interval probabilities," *International journal of intelligent systems*, vol. 19, no. 9, pp. 789–815, 2004.

[12] ——, "Query algebra operations for interval probabilities," in *Database and Expert Systems Applications.* Springer, 2003, pp. 527–536.

[13] T. Eiter, T. Lukasiewicz, and M. Walter, "Extension of the relational algebra to probabilistic complex values," in *Foundations of Information and Knowledge Systems.* Springer, 2000, pp. 94–115.

[14] D. Dey and S. Sarkar, "Generalized normal forms for probabilistic relational data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 3, pp. 485–497, 2002.

[15] T. Eiter, J. J. Lu, T. Lukasiewicz, and V. Subrahmanian, "Probabilistic object bases," *ACM Transactions on Database Systems (TODS)*, vol. 26, no. 3, pp. 264–312, 2001.

[16] L. V. Lakshmanan, N. Leone, R. Ross, and V. S. Subrahmanian, "Probview: A flexible probabilistic database system," *ACM Transactions on Database Systems (TODS)*, vol. 22, no. 3, pp. 419–469, 1997.

[17] N. N. Dalvi and D. Suciu, "Efficient query evaluation on probabilistic databases," in *VLDB*, 2004, pp. 864–875.

[18] H. Nguyen and T. H. Cao, "Extending probabilistic object bases with uncertain applicability and imprecise values of class properties," in *Proc. 5th IEEE International Conf. on Fuzzy Systems, London, England.* IEEE, 2007, pp. 487–492.

Received on December 15 - 2014
Revised on November 11 - 2015