ICIC Express Letters
Volume **16**, Number **3**, March **2022**

ICIC International ©2022 ISSN 1881-803X
pp. **241–248**

A PROBABILISTIC RELATIONAL DATABASE MODEL WITH UNCERTAIN MULTIVALUED ATTRIBUTES

Hoa Nguyen, Thanh-Nam Nguyen and Thi-Thu-Nga Tran

Information Technology Faculty
Saigon University
No. 273, An Duong Vuong, District 5, Ho Chi Minh City 72710, Vietnam
{ nguyenhoa; ntnam.ncs; tttnga }@sgu.edu.vn

Received July 2021; accepted September 2021

Abstract. *In this paper, we introduce a new probabilistic relational database model whose relational attributes may take a set of values associated with a probability for representing and handling uncertain information. A probabilistic interpretation of binary relations on sets is proposed to compute uncertain degree of relations on attribute values. Probabilistic triples over a set are extended to probabilistic triples over a set of sets to represent multivalued relational attributes. Then, fundamental concepts as schemas, probabilistic relations, probabilistic relational database and selection operation are defined coherently and consistently for the new model.*

Keywords: Uncertain multivalued attribute, Probabilistic interpretation, Probabilistic triple, Probabilistic relation, Selection operation

1. Introduction.

It is true that the classical relational database model (CRDB) is very useful for modeling, designing and implementing large-scale systems [1,2], but it is restricted for representing and handling uncertain information in practice. Currently, there have been many non-classical database models, including probabilistic relational database models, studied and built to overcome the limitation of CRDB (e.g., [4-13,16-21,24,25]). However, no model would be so universal that could include all measures and tackle all aspects of uncertainty of information in the real world.

Probabilistic relational database models are developed and built as extensions of CRDB based on the probability theory. There are two main approaches corresponding to two levels to extend CRDB to a probabilistic relational database model [22]: 1) at the relation level, each relation is defined by a set of tuples that each tuple is associated with a probability to represent the uncertainty degree of it in the relation; or 2) at the attribute level, each attribute in a relation is associated with a probability to define the uncertainty degree of the values that it may take.

For the first approach, as the works in [3-9,24], each tuple of a relation was associated with a probability in the interval $[0,1]$ to express the uncertainty membership degree of that tuple for the relation. The uncertainty degree of the attribute values of a tuple was inferred from the uncertainty membership degree of that tuple. However, in many real situations, we do not know exactly the probability as a number in the interval $[0,1]$ but only can estimate it as an approximate number in a subinterval of $[0,1]$. The models in [10-13,16,22] were extended with probability intervals associated with each tuple to overcome the shortcoming.

For the second approach, as in [14,15], each value of an attribute was assigned to a probability in the interval $[0,1]$ to represent the uncertain level for that attribute taking the value. More flexibly, the model in [19] represented the value of each attribute as a probability distribution on a set. It means that each attribute associated with a set

DOI: 10.24507/icicel.16.03.241

of values and a probability distribution expressing possibility that the attribute might take one of values of the set with a probability computed from the distribution. The models in [20,21] extended more the model in [19], where a pair of lower and upper bound probability distributions is used instead of a probability distribution as in [19].

In above mentioned works including both approaches, the attribute of a tuple or an object only took a single, unique value in a set of values with some probability. For instance, the attribute DISEASE in [20] represented by DISEASE: $\langle\{$*hepatitis*, *cirrhosis*, *cholecystitis*$\}$, $0.9u$, $1.5u\rangle$ said that the patient's disease might be *hepatitis* or *cirrhosis* or *cholecystitis* with a probability in the interval $[0.3, 0.5]$. However, in practice, a patient may have both hepatitis and cirrhosis or cholecystitis with a determined probability interval, and then the model in [20] cannot express. The shortcoming of above mentioned models including the models in [20,21] is that those models cannot represent multivalued relational attributes.

In this paper, we propose a new relational database model for uncertain information, denoted (URDB), as an extension of the models in [20,21] with multivalued attributes to overcome the limitation of above mentioned models. This extension is also consistent with the classical relational database model in [1] and the decision making support system in [23], where tuples and objects can have multivalued attributes.

The probability base for URDB is presented in Section 2. The proposed URDB model including fundamental concepts as the schema, relation, database and the query of uncertain information is introduced in Section 3. Finally, Section 4 concludes the paper and outlines further research directions in the future.

2. **Probability Base.**

The URDB model is built on a probability base including probability notions and definitions extended and proposed.

For representing multivalued attributes in URDB, probabilistic triples over a set in [17,20] are extended to probabilistic triples over a set of sets as follows.

Definition 2.1. *Let X be a finite set, a probabilistic triple $\langle V, \alpha, \beta\rangle$ over X consists of a subset V of the set 2^X (i.e., the set of all subsets of X) whose elements are disjointed, a probability distribution function α: $V \to [0, 1]$, and a function β: $V \to [0, 1]$ such that $\alpha(x) \leq \beta(x)$, $\forall x \in V$ and $\sum_{x\in V} \beta(x) \geq 1$ hold.*

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

Example 2.1. *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 diseases are hepatitis and cirrhosis or cholecystitis with a probability between 40% and 60%, then this knowledge may be encoded by the extended probabilistic triple $\langle\{\{$hepatitis, cirrhosis$\}$, $\{$cholecystitis$\}\}, 0.8u, 1.2u\rangle$. Here, u is the uniform distribution function over $\{\{$hepatitis, cirrhosis$\}, \{$cholecystitis$\}\}$, $0.8u$ 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$\}, \{$cholecystitis$\}\}$.*

We note that an element e in X is also considered as a special set $\{e\}$ on X; thus, a probabilistic triple $\langle\{\{e_1\}, \{e_2\}, \ldots, \{e_k\}\}, \alpha, \beta\rangle$ can be written as $\langle\{e_1, e_2, \ldots, e_k\}, \alpha, \beta\rangle$ for simplicity. Also, "an extended probabilistic triple" is called "a probabilistic triple".

For computing uncertain degree of relations on attribute values in URDB, we propose a probabilistic interpretation of binary relations on sets as below.

Definition 2.2. *Let A and B be sets, U and V be value domains, and θ be a binary relation from $\{=, \neq, \leq, \geq, <, >, \subseteq, \supseteq\}$. The probabilistic interpretation of the relation $A\ \theta\ B$, denoted $Pr(A\ \theta\ B)$, is a value in $[0, 1]$ that is defined by*

1) $Pr(A\ \theta\ B) = p(u\ \theta\ v|u \in A, v \in B)$, where A is a subset of U, B is a subset of V and $\theta \in \{=, \neq, \leq, <, \geq, >\}$ assumed to be valid on $(U \times V)$, $p(u\ \theta\ v|u \in A, v \in B)$ is the conditional probability of $u\ \theta\ v$ given $u \in A$ and $v \in B$.

2) $Pr(A\ \theta\ B) = \begin{cases} p(u \in B|u \in A), & \theta \text{ is the relation } \subseteq \\ p(u \in A|u \in B), & \theta \text{ is the relation } \supseteq \end{cases}$ where A and B are two subsets of U, $p(u \in B|u \in A)$ is the conditional probability for $u \in B$ given $u \in A$ and $p(u \in A|u \in B)$ is the conditional probability for $u \in A$ given $u \in B$.

We note that, the probabilistic interpretation of binary relations on sets defined here is an extension of that in [21] with relations "$\subseteq$" and "$\supseteq$"; meanwhile, no probabilistic interpretation of binary relations on sets was proposed in [20].

Example 2.2. *Let $A = \{3, 4\}$ and $B = \{4, 5\}$ be two sets on the domain $\{1, 2, 3, 4, 5, 6\}$. Then $Pr(A = B) = p(u = v|u \in A, v \in B) = p(u = v|u \in \{3, 4\}, v \in \{4, 5\}) = 0.25$ and $Pr(A \subseteq B) = p(u \in B|u \in A) = p(u \in \{4, 5\}|u \in \{3, 4\}) = 0.5$.*

In this work, we use the combination strategies of probability intervals in [17] to compute the probability intervals of the conjunction, disjunction or difference event of two events. Let two events e_1 and e_2 have probabilities in the intervals $[L_1, U_1]$ and $[L_2, U_2]$, respectively. Then the *probability intervals* of the conjunction event $e_1 \wedge e_2$, disjunction event $e_1 \vee e_2$, or difference event $e_1 \wedge \neg e_2$ can be computed by alternative strategies as follows:

1) Independence conjunction, disjunction, and difference strategies, denoted $\otimes_{in}$, $\oplus_{in}$, and $\ominus_{in}$ respectively, are determined by
   - $[L_1, U_1] \otimes_{in} [L_2, U_2] \equiv [L_1.L_2, U_1.U_2]$
   - $[L_1, U_1] \oplus_{in} [L_2, U_2] \equiv [L_1 + L_2 - (L_1.L_2), U_1 + U_2 - (U_1.U_2)]$
   - $[L_1, U_1] \ominus_{in} [L_2, U_2] \equiv [L_1.(1 - U_2), U_1.(1 - L_2)]$
2) Mutual exclusion conjunction, disjunction, and difference strategies (when e_1 and e_2 are mutually exclusive), denoted $\otimes_{me}$, $\oplus_{me}$, and $\ominus_{me}$ respectively, are determined by
   - $[L_1, U_1] \otimes_{me} [L_2, U_2] \equiv [0, 0]$
   - $[L_1, U_1] \oplus_{me} [L_2, U_2] \equiv [\min(1, L_1 + L_2), \min(1, U_1 + U_2)]$
   - $[L_1, U_1] \ominus_{me} [L_2, U_2] \equiv [L_1, \min(U_1, 1 - L_2)]$
3) Positive correlation conjunction, disjunction, and difference strategies (when e_1 implies e_2, or e_2 implies e_1), denoted $\otimes_{pc}$, $\oplus_{pc}$, and $\ominus_{pc}$ respectively, are determined by
   - $[L_1, U_1] \otimes_{pc} [L_2, U_2] \equiv [\min(L_1, L_2), \min(U_1, U_2)]$
   - $[L_1, U_1] \oplus_{pc} [L_2, U_2] \equiv [\max(L_1, L_2), \max(U_1, U_2)]$
   - $[L_1, U_1] \ominus_{pc} [L_2, U_2] \equiv [\max(0, L_1 - U_2), \max(0, U_1 - L_2)]$
4) Ignorance conjunction, disjunction, and difference strategies, denoted $\otimes_{ig}$, $\oplus_{ig}$, and $\ominus_{ig}$ respectively, are determined by
   - $[L_1, U_1] \otimes_{ig} [L_2, U_2] \equiv [\max(0, L_1 + L_2 - 1), \min(U_1, U_2)]$
   - $[L_1, U_1] \oplus_{ig} [L_2, U_2] \equiv [\max(L_1, L_2), \min(1, U_1 + U_2)]$
   - $[L_1, U_1] \ominus_{ig} [L_2, U_2] \equiv [\max(0, L_1 - U_2), \min(U_1, 1 - L_2)]$

3. **Proposed URDB Model.** As in CRDB, fundamental concepts in URDB are the schema, relation and database. A URDB schema consists of a set of relational attributes respectively associated with domains that define probabilistic triples representing uncertain values of those attributes. The URDB schema is extended from that of the models [20,21] with uncertain multivalued attributes as follows.

Definition 3.1. *A URDB schema is a pair $R = (\boldsymbol{U}, P)$, where*

1) $\boldsymbol{U} = \{A_1, A_2, \ldots, A_k\}$ is a set of pairwise different attributes.

2) P is a function that maps each attribute $A \in \boldsymbol{U}$ to the set of all probabilistic triples on the value domain of A.

For simplicity, the notation $R(\boldsymbol{U}, P)$ and R can be used to denote $R = (\boldsymbol{U}, P)$.

A URDB relation is an instance of a URDB schema, where each relational attribute may take more than one uncertain value represented by a probabilistic triple. The URDB relation is extended from that of the models in [20,21] with uncertain multivalued attributes as the following definition.

Definition 3.2. *Let $\boldsymbol{U} = \{A_1, A_2, \dots, A_k\}$ be a set of k pairwise different attributes. A URDB relation r over the schema $R(\boldsymbol{U}, P)$ is a finite set of elements $\{t_1, t_2, \dots, t_n\}$, where each element $t_i = (\langle V_{i1}, \alpha_{i1}, \beta_{i1}\rangle, \langle V_{i2}, \alpha_{i2}, \beta_{i2}\rangle, \dots, \langle V_{ik}, \alpha_{ik}, \beta_{ik}\rangle)$ is a list of k probabilistic triples such that $\langle V_{ij}, \alpha_{ij}, \beta_{ij}\rangle$ belongs to the set $P(A_j)$ and $V_{ij} \neq \emptyset$, for every $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, k$.*

Each element t in the relation r over $R(\boldsymbol{U}, P)$ is called a tuple on $\boldsymbol{U}$. For each tuple t_i, the probabilistic triple $\langle V_{ij}, \alpha_{ij}, \beta_{ij}\rangle$ represents an uncertain valued set of the attribute A_i of the tuple t_i. We write $t_i.A_i$ or $t_i[A_i]$ to denote $\langle V_{ij}, \alpha_{ij}, \beta_{ij}\rangle$.

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 3.1. *In the database about patients at the clinic of a hospital, a simple URDB relation, named PATIENT, over the URDB schema **PATIENT**$(\{NAME, AGE, DISEASE, D\_COST\}, P)$ can be given as Table 1.*

Table 1. Relation PATIENT

NAME	AGE	DISEASE	D_COST
$\langle\{Oliver\}, u, u\rangle$	$\langle\{65\}, u, u\rangle$	$\langle\{lung\,cancer, tuberculosis\}, 0.6u, 1.2u\rangle$	$\langle\{30, 35\}, 0.7u, 1.3u\rangle$
$\langle\{Blair\}, u, u\rangle$	$\langle\{43, 44\}, u, u\rangle$	$\langle\{\{hepatitis, cirrhosis\}, \{cholecystitis\}\}, 0.9u, 1.3u\rangle$	$\langle\{6, 7\}, 0.8u, 1.4u\rangle$
$\langle\{Alice\}, u, u\rangle$	$\langle\{36\}, u, u\rangle$	$\langle\{cholecystitis\}, u, u\rangle$	$\langle\{8\}, u, u\rangle$
$\langle\{Anne\}, u, u\rangle$	$\langle\{15\}, u, u\rangle$	$\langle\{\{bronchitis, angina\}\}, u, u\rangle$	$\langle\{7\}, u, u\rangle$

In the relation, the attributes NAME, AGE, DISEASE and D_COST describe the information about the name, age, disease and daily treatment cost of each patient, respectively. In reality, while diagnosing, the disease of each patient is not always determined certainly by the physicians. Similarly, the daily treatment cost for patients is also not known definitely even as the patients know about their diseases. Here, the conventional unit for the daily treatment cost is 1 (USD), u is the uniform distribution function as presented in Example 2.1. Note that, for each attribute A in the schema **PATIENT**, $P(A)$ includes all probabilistic triples on the domain of A (Definition 3.1).

The URDB relational database is defined as an extension of CRDB and the probabilistic relational databases in [20,21] as follows.

Definition 3.3. *A URDB relational database over a set of attributes is a set of URDB relations corresponding to the set of their URDB schemas.*

As in CRDB model, the selection is a basic algebraic operation in URDB model for querying data on relations of databases. The selection operation in URDB is extended from that of the models in [20,21] taking account of uncertain multivalued relational attributes. Before defining the selection operation, we present the formal syntax and semantics of selection expressions and conditions as below.

Definition 3.4. *Let R be a URDB schema and X be a set of relational tuple variables. Then selection expressions are inductively defined and have one of the following forms.*

1) $x.A\ \theta\ c$, where $x \in X$, A is an attribute in R, θ is a binary relation from $\{=, \neq, \leq, \geq, <, >, \subseteq, \supseteq\}$, and c is a single value or a set of values.

2) $x.A_1 =_\otimes x.A_2$, where $x \in X$, 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, and $\otimes$ 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, and $\oplus$ is a probabilistic disjunction strategy.

Example 3.2. *Consider the schema **PATIENT** in Example 3.1, the selection of "all patients who get hepatitis and pay the daily treatment cost over 6 USD" can be expressed by the selection expression* $x.\mathrm{DISEASE} = \textit{hepatitis} \otimes x.\mathrm{D\_COST} > 6$.

Now, selection conditions in URDB are formally defined based on selection expressions as follows.

Definition 3.5. *Let R be a URDB schema. Then 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 3.3. *Given the schema **PATIENT** in Example 3.1, the selection of "all patients who are over 40 years old with a probability of at least 0.8 or have tuberculosis and pay the daily treatment cost not less than 30 USD with a probability from 0.5 to 0.6" can be done using the selection condition* $(x.\mathrm{AGE} > 40)[0.8, 1] \vee (x.\mathrm{DISEASE} = \mathrm{tuberculosis} \otimes x.\mathrm{D\_COST} \geq 30)[0.5, 0.6]$.

The probabilistic interpretation (i.e., semantics) of selection expressions in URDB is extended from that of the models [20,21] with the probabilistic interpretation of binary relations on sets as below.

Definition 3.6. *Let R be a URDB schema, 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 $\mathit{Prob}_{R,r,t}$, 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) $\mathit{Prob}_{R,r,t}(x.A\ \theta\ c) = \left[\sum_{v\in V} \alpha(v).Pr(v\ \theta\ c), \min\left(1, \sum_{v\in V} \beta(v).Pr(v\ \theta\ c)\right)\right]$, where $t.A = \langle V, \alpha, \beta\rangle$.

2) $\mathit{Prob}_{R,r,t}(x.A_1 =_\otimes x.A_2) = \left[\sum_{v\in V} \alpha(v).Pr(v_1 = v_2), \min\left(1, \sum_{v\in V} \beta(v).Pr(v_1 = v_2)\right)\right]$, where $t.A_1 = \langle V_1, \alpha_1, \beta_1\rangle$, $t.A_2 = \langle V_2, \alpha_2, \beta_2\rangle$ and $[\alpha(v), \beta(v)] = [\alpha_1(v_1), \beta_1(v_1)] \otimes [\alpha_2(v_2), \beta_2(v_2)]$, $\forall v = (v_1, v_2) \in V = V_1 \times V_2$.

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

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

Intuitively, $\mathit{Prob}_{R,r,t}(x.A\ \theta\ c)$ is the probability interval for the attribute A of the tuple t having a value v such that $v\ \theta\ c$, while $\mathit{Prob}_{R,r,t}(x.A_1 =_\otimes x.A_2)$ is the probability interval for the attributes A_1 and A_2 of the tuple t having values v_1 and v_2, respectively, such that $v_1 = v_2$.

Example 3.4. *Let R denote the schema **PATIENT** and r denote the relation PATIENT in Example 3.1. Consider the second tuple in r, denoted by t_2. We have*

$$\begin{aligned}
&\mathit{Prob}_{R,r,t_2}(x.\mathrm{DISEASE} \supseteq \{\mathit{hepatitis}, \mathit{cirrhosis}\})\\
&= [0.9u(\{\mathit{hepatitis}, \mathit{cirrhosis}\}).Pr(\{\mathit{hepatitis}, \mathit{cirrhosis}\} \supseteq \{\mathit{hepatitis}, \mathit{cirrhosis}\})\\
&\quad + 0.9u(\{\mathit{cholecystitis}\}).Pr(\{\mathit{cholecystitis}\} \supseteq \{\mathit{hepatitis}, \mathit{cirrhosis}\}),\\
&\quad \min(1, 1.3u(\{\mathit{hepatitis}, \mathit{cirrhosis}\}).Pr(\{\mathit{hepatitis}, \mathit{cirrhosis}\} \supseteq \{\mathit{hepatitis}, \mathit{cirrhosis}\})
\end{aligned}$$

$$+ 1.3u(\{cholecystitis\}).Pr(\{cholecystitis\} \supseteq \{hepatitis, cirrhosis\}))]$$
$$= [0.9 \times 0.5 \times 1.0 + 0.9 \times 0.5 \times 0.0, \min(1, 1.3 \times 0.5 \times 1.0 + 1.3 \times 0.5 \times 0.0)]$$
$$= [0.45, 0.65].$$

The satisfaction (i.e., semantics) of selection conditions in URDB is defined as below.

Definition 3.7. *Let R be a URDB schema, 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 CRDB, the concepts of selection expression and selection condition are identical, where probability intervals $[L, U]$ in selection conditions are always equal to $[1.0, 1.0]$. This also means that the satisfaction of selection conditions in CRDB is a special case of that in URDB.

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

Definition 3.8. *Let R be a URDB schema, 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 $r^* = \{t \in r | Prob_{R,r,t} \models \phi\}$ over R, including all satisfied tuples of the selection condition ϕ.*

Example 3.5. *Let r denote the relation PATIENT in Example 3.1 and R denote its schema. The query "Find all patients who are over 40 years old with a probability of at least 0.9, and have both hepatitis and cirrhosis and pay the daily treatment cost not less than 6 USD with a probability between 0.3 and 0.7" can be done by the selection operation $\sigma_\phi(PATIENT)$, where $\phi = (x.\text{AGE} > 40)[0.9, 1] \wedge (x.\text{DISEASE} \supseteq \{hepatitis, cirrhosis\}$ $\otimes_{in} x.\text{D\_COST} \geq 6)[0.3, 0.7]$.*

Only the second tuple t_2 of the relation PATIENT in Example 3.1 satisfies ϕ, because $Prob_{R,r,t_2}(x.\text{AGE} > 40) = [u(43) \times Pr(43 > 40) + u(44) \times Pr(44 > 40), \min(1, u(43) \times Pr(43 > 40) + u(44) \times Pr(44 > 40))] = [0.5 \times 1 + 0.5 \times 1, \min(1, 0.5 \times 1 + 0.5 \times 1)] = [1, 1] \subseteq [0.9, 1]$, $Prob_{R,r,t_2}(x.\text{D\_COST} \geq 6) = [0.8u \times Pr(6 \geq 6) + 0.8u \times Pr(7 \geq 6), \min(1, 1.4u \times Pr(6 \geq 6) + 1.4u \times Pr(7 \geq 6))] = [0.8 \times 0.5 \times 1 + 0.8 \times 0.5 \times 1, \min(1, 1.4 \times 0.5 \times 1 + 1.4 \times 0.5 \times 1)] = [0.8, 1]$.

From the result of the computation in Example 3.4, we have $Prob_{R,r,t_2}(x.\text{DISEASE} \supseteq \{hepatitis, cirrhosis\} \otimes_{in} x.\text{D\_COST} \geq 6) = [0.45, 0.65] \otimes_{in} [0.8, 1] = [0.36, 0.65] \subseteq [0.3, 0.7]$.

For the other tuples, one has $Prob_{R,r,t_i}(x.\text{DISEASE} \supseteq \{hepatitis, cirrhosis\} \otimes_{in} x.$ $\text{D\_COST} \geq 6) = [0, 0] \not\subseteq [0.3, 0.7]$, $\forall i \neq 4$. Thus, the result of the query is as Table 2.

Table 2. Relation σ_ϕ (PATIENT)

NAME	AGE	DISEASE	D_COST
$\langle\{Blair\}, u, u\rangle$	$\langle\{43, 44\}, u, u\rangle$	$\langle\{\{hepatitis, cirrhosis\},$ $\{cholecystitis\}\}, 0.9u, 1.3u\rangle$	$\langle\{6, 7\}, 0.8u, 1.4u\rangle$

As for CRDB, the selection operation in URDB is not dependent on the order of selection conditions as the following theorem.

Theorem 3.1. *Let r be a relation over the schema R in URDB, ϕ_1 and ϕ_2 be two selection conditions. Then*

$$\sigma_{\phi_1}(\sigma_{\phi_2}(r)) = \sigma_{\phi_2}(\sigma_{\phi_1}(r)) \tag{1}$$

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 $t \in r$, we have

$$\begin{aligned}
\sigma_{\phi_1}(\sigma_{\phi_2}(r)) &= \{t \in r_2 | Prob_{R,r_2,t} \models \phi_1\} \\
&= \{t \in r | (Prob_{R,r,t} \models \phi_2) \wedge (Prob_{R,r_2,t} \models \phi_1)\} \\
&= \{t \in r | (Prob_{R,r,t} \models \phi_2) \wedge (Prob_{R,r,t} \models \phi_1)\} \text{ (because of } r_2 \subseteq r) \\
&= \{t \in r | (Prob_{R,r,t} \models \phi_1 \wedge \phi_2)\} \text{ (Definition 3.7)} \\
&= \sigma_{\phi_1 \wedge \phi_2}(r).
\end{aligned}$$

Thus, the equation $\sigma_{\phi_1}(\sigma_{\phi_2}(r)) = \sigma_{\phi_1 \wedge \phi_2}(r)$ is proven. The equation $\sigma_{\phi_2}(\sigma_{\phi_1}(r)) = \sigma_{\phi_2 \wedge \phi_1}(r)$ is similarly proven. Since $\phi_1 \wedge \phi_2 \Leftrightarrow \phi_2 \wedge \phi_1$. So, Theorem 3.1 is proven.

4. **Conclusions.** We have presented a new relational database model whose relational attributes may take more than one uncertain value represented by an extended probabilistic triple. A probabilistic interpretation of binary relations on sets has proposed for defining the selection operation to query uncertain information expressed by relations of this model.

In the next steps, we will extend algebraic operations in the classical relational database model as projection, Cartesian product, join, intersection, union, difference for the new model and build a management system with the language like SQL for querying and manipulating uncertain information in the real world applications.

Acknowledgment. This work is funded by Saigon University in the scientific research project of 2021.

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] R. Elmasri and S. B. Navathe, *Fundamentals of Database Systems*, 6th Edition, Addison-Wesley, 2011.

[3] A. Ali, S. Talpur and S. Narejo, Detecting faulty sensors by analyzing the uncertain data using probabilistic database, *Proc. of the 3rd International Conference on Computing, Mathematics and Engineering Technologies*, Sukkur, Pakistan, pp.143-150, 2020.

[4] I. I. Ceylan, S. Borgwardt and T. Lukasiewicz, Most probable explanations for probabilistic database queries, *Proc. of the 26th International Joint Conference on Artificial Intelligence*, Melbourne, Australia, pp.950-956, 2017.

[5] I. I. Ceylan, A. Darwiche and G. V. D Broeck, Open-world probabilistic databases: Semantics, algorithms, complexity, *Journal of Artificial Intelligence*, vol.295, no.11, pp.103474-103513, 2021.

[6] D. Dey and S. Sarkar, A probabilistic relational model and algebra, *ACM Trans. Database Systems*, vol.21, no.3, pp.339-369, 1996.

[7] N. Fuhr and T. Rolleke, A probabilistic relational algebra for the integration of information retrieval and database systems, *ACM Trans. Information Systems*, vol.15, no.1, pp.32-66, 1997.

[8] Y. Li, J. Chen and L. Feng, Dealing with uncertainty: A survey of theories and practices, *IEEE Trans. Knowledge and Data Engineering*, vol.25, no.11, pp.2463-2482, 2013.

[9] S. Zhang and C. Zhang, A probabilistic data model and its semantics, *Journal of Research and Practice in Information Technology*, vol.35, no.4, pp.237-256, 2003.

[10] A. Dekhtyar, R. Ross and V. S. Subrahmanian, Probabilistic temporal databases, I: Algebra, *ACM Trans. Database Systems*, vol.26, no.1, pp.41-95, 2001.

[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] L. V. S. Lakshmanan, N. Leone, R. Ross and V. S. Subrahmanian, ProbView: A flexible probabilistic database system, *ACM Trans. Database Systems*, vol.22, no.3, pp.419-469, 1997.

[13] R. Ross and V. S. Subrahmanian, Aggregate operators in probabilistic databases, *Journal of the ACM*, vol.52, no.1, pp.54-101, 2005.

[14] D. Dey and S. Sarkar, Generalized normal forms for probabilistic relational data, *IEEE Trans. Knowledge and Data Engineering*, vol.14, no.3, pp.485-497, 1992.

[15] D. Barbara, H. Garcia-Molina and D. Porter, The management of probabilistic data, *IEEE Trans. Knowledge and Data Engineering*, vol.4, no.5, pp.487-502, 1992.

[16] T. Eiter, T. Lukasiewicz and M. Walter, A data model and algebra for probabilistic complex values, *Annals of Mathematics and Artificial Intelligence*, vol.33, pp.205-252, 2001.

[17] T. Eiter, J. J. Lu, T. Lukasiewicz and V. S. Subrahmanian, Probabilistic object bases, *ACM Trans. Database Systems*, vol.26, no.3, pp.264-312, 2001.

[18] Y. Kornatzky and S. E. Shimony, A probabilistic object-oriented data model, *Data and Knowledge Engineering*, vol.12, pp.143-166, 1994.

[19] S. K. Lee, An extended relational database model for uncertain and imprecise information, *Proc. of the 18th Conference on Very Large Data Bases*, Vancouver, British Columbia, Canada, pp.211-220, 1992.

[20] H. Nguyen and D. H. Tran, A probabilistic relational data model for uncertain information, *Proc. of the 3rd IEEE International Conference on Information Science and Technology*, Yangzhou, China, pp.607-613, 2013.

[21] H. Nguyen, A probabilistic relational database model and algebra, *Journal of Computer Science and Cybernetics*, vol.31, no.4, pp.305-321, 2015.

[22] H. Nguyen, Extending relational database model for uncertain information, *Journal of Computer Science and Cybernetics*, vol.35, no.4, pp.355-372, 2019.

[23] A. V. Vitianingsih, I. Wisnubhadra, S. S. K. Baharin, R. Marco and A. L. Maukar, Classification of pertussis vulnerable area with location analytics using multiple attribute decision making, *International Journal of Innovative Computing, Information and Control*, vol.16, no.6, pp.1943-1957, 2020.

[24] T. Friedman and G. Broeck, Symbolic querying of vector spaces: Probabilistic databases meets relational embeddings, *Proc. of the 36th Conference on Uncertainty in Artificial Intelligence*, Toronto, Canada, vol.124, pp.1268-1277, 2020.

[25] J. Bernad, C. Bobed and E. Mena, Uncertain probabilistic range queries on multidimensional data, *Information Sciences*, vol.537, pp.334-367, 2020.