TRƯỜNG ĐẠI HỌC SÀI GÒN
TẠP CHÍ KHOA HỌC
ĐẠI HỌC SÀI GÒN
ISSN 1859 - 3208
Số 89, Tập 02/2023

SAIGON UNIVERSITY
SCIENTIFIC JOURNAL
OF SAIGON UNIVERSITY
No. 89, Vol 02/2023

Email: tcdhsg@sgu.edu.vn ; Website: http://sj.sgu.edu.vn/

EXTENDING RELATIONAL DATABASE MODEL WITH PROBABILITY INTERVALS FOR UNCERTAIN SET-VALUED ATTRIBUTES

Mở rộng mô hình cơ sở dữ liệu quan hệ với các khoảng xác suất cho các thuộc tính có giá trị tập hợp không chắc chắn

Nguyễn Hòa
Trường ĐH Sài Gòn

ABSTRACT

In this paper, we introduce a new relational database model whose relational attributes may take a set of values associated with a probability interval for representing and handling uncertain and imprecise information. Probabilistic values of set types are proposed to represent uncertain set-valued attributes. The probabilistic interpretations of binary relations on sets are used to compute the uncertain degree of relations on attribute values. Then, fundamental concepts as schemas, probabilistic relations, probabilistic relational database, selection operation and queries are defined coherently and consistently for the new model.

Keywords: *probability interval, uncertain set-valued attributes, probabilistic interpretation, probabilistic values, probabilistic relation, selection operation.*

TÓM TẮT

Trong bài báo này, chúng tôi giới thiệu một mô hình cơ sở dữ liệu quan hệ mới mà các thuộc tính quan hệ của nó có thể nhận một tập giá trị được kết hợp với một khoảng xác suất để biểu diễn và xử lý thông tin không chắc chắn và không chính xác. Các giá trị xác suất của các kiểu dữ liệu tập hợp được đề nghị để biểu diễn các thuộc tính có giá trị tập hợp không chắc chắn. Các diễn dịch xác suất của các quan hệ hai ngôi trên các tập hợp được sử dụng để tính mức độ không chắc chắn của các quan hệ trên các giá trị thuộc tính. Sau đó, các khái niệm cơ bản như lược đồ, quan hệ, cơ sở dữ liệu xác suất, phép chọn và các truy vấn được định nghĩa một cách chặt chẽ và nhất quán cho mô hình mới.

Từ khóa*: khoảng xác suất, thuộc tính có giá trị tập hợp không chắc chắn, diễn dịch xác suất, giá trị xác suất, quan hệ xác suất, phép chọn.*

1. Introduction

As known, the conventional relational database model (CRDB) is very useful for modeling, designing and implementing large-scale systems [1-2], but it is limited for representing and handling uncertain and imprecise information in practice. Currently, there have been many non-conventional database models, including probabilistic relational database models (PRDB), studied and built to overcome the limitation of CRDB (e.g. [3-16]). Probabilistic database models also have been used in many real applications (e.g. [17-21]). However, no model would be so universal that could include all

Corresponding author: nguyenhoa@sgu.edu.vn

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 to extend CRDB to a probabilistic relational database model.

The first approach is the extension of CRDB at the relation level, as the works in [3-7], thereby 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 [8-11] were extended with probability intervals associated with each tuple to overcome the shortcoming.

The second approach is the extension of CRDB at the attribute level, as in [12-13], thereby 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 and generally, in [14], each attribute was associated with a probability distribution on a set of subsets of the attribute domain to express the possibility that the attribute might take a value as one of elements of some subset with a distributed probability. The model in [15] used a pair of lower and upper bound probability distributions on a set instead of a probability distribution as in [14] to represent the possibility that the attribute might take a value in the set with a computed probability interval from the distribution pair.

As we know, in the CRDB model, the relational attribute can take a set of values [1]. In other words, the CRDB model can allow multivalued attributes. However, in above presented PRDB models, 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 authors in [14] represented the attribute DISEASE of the patient Mary by DISEASE: $\{\langle\{d_1, d_3\}, 0.6\rangle, \langle\{d_2\}, 0.4\rangle\}$ to say that Mary's disease was either d_2 with a probability 0.4 or one of $\{d_1, d_3\}$ with a probability 0.6. According to the meaning of this presentation, the model in [14] did not allow the attributes to take multivalues or set values. In practice, Mary may have both d_1 and d_3 (not one of $\{d_1, d_3\}$) with the probability 0.6 or d_2 with the probability 0.4. In addition, in many real situations, we cannot know exactly the probability for $\{d_1, d_3\}$ and $\{d_2\}$ being 0.6 and 0.4, respectively but only can estimate these probabilities as approximate or imprecise numbers in subintervals of [0, 1]. Recently, the model in [16] has been proposed to overcome the shortcomings of the models in [14-15] by representing the value of each relational attribute as a set of sets associated with two probability distribution functions. However, when the relations have many attributes, the number of generated probability distribution functions is too large to lead the low performance in manipulating data of the model.

In this paper, we propose a new probabilistic relational database model for uncertain and imprecise information, named EPRDB, as an extension of CRDB

with probability intervals for uncertain set-valued attributes to overcome the limitations of models in [14] and [16]. The EPRDB model is consistent with CRDB model by allowing multivalued attributes and more flexibly than the models in [14] and [16] by using probability intervals instead of probability single values and distribution functions. To build EPRDB, we extend probabilistic values on a set in [22] to probabilistic values on a set of sets for representing uncertain set-valued attributes of relations and employ probabilistic interpretations of binary relations on sets in [16] to define the selection operation for computing and querying uncertain and imprecise information on EPRDB relations. The new EPRDB model is built to allow expressing and executing the soft queries on multivalued attributes associated with probability intervals to handle uncertain and imprecise information in real databases.

The mathematical base for EPRDB is presented in Section 2. Schemas and relations of EPRDB are defined in Section 3. Section 4 introduces and defines formally the syntax and semantics of the selection operation as a fundamental one to query EPRDB databases. Finally, Section 5 concludes the paper and outlines further research directions in the future.

2. Probability definitions and notions

The mathematical base for EPRDB model includes some probability definitions and notions for representing and handling uncertain and imprecise information in turn introduced as follows.

Definition 2.1. Let τ be a data type and D be the domain of τ, a *probabilistic value* on the domain of τ is a finite set of pairs $\{(v_1, [l_1, u_1]), \ldots, (v_m, [l_m, u_m])\}$, where v_i belongs to 2^D, v_i and v_j are disjointed and $0 \le l_i \le u_i \le 1$, for every $i, j = 1, 2, \ldots, m$.

We note that this definition is an extension of that in [22] with set data types. Informally, a probabilistic value $pv = \{(v_1, [l_1, u_1]), \ldots, (v_m, [l_m, u_m])\}$ says that pv's value is exactly one member (set) v_i of the set $V = \{v_1,\ldots, v_m\}$ and the probability that pv's value is v_i lies in the interval $[l_i, u_i]$. A probabilistic value $pv = \{(v_1, [l_1, u_1]), \ldots, (v_m, [l_m, u_m])\}$ corresponds with a probability distribution function p over $V = \{v_1,\ldots, v_m\}$ such that $p(v_i) \in [l_i, u_i]$ and $\Sigma_{v_i \in V} p(v_i) \le 1$.

Example 2.1. While 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 with a probability between 0.5 and 0.7 or cholecystitis with a probability between 0.3 and 0.5, then this knowledge may be encoded by the probabilistic value {({hepatitis, cirrhosis}, [0.5, 0.7]), ({cholecystitis}, [0.3, 0.5])}.

We note that an element e in D is also considered as a special set $\{e\}$ on D, thus a probabilistic value $\{(\{e_1\}, [l_1, u_1]), (\{e_2\}, [l_2, u_2]),\ldots, (\{e_k\}, [l_k, u_k])\}$ can be written as $\{(e_1, [l_1, u_1]), (e_2, [l_2, u_2]),\ldots, (e_k, [l_k, u_k])\}$ for simplicity.

For computing uncertain degree of relations on attribute values in EPRDB, we use a probabilistic interpretation of binary relations on sets in [16] as below.

Definition 2.2. Let A and B be sets, U and V be value domains, *and* θ be a binary relation from $\{=, \neq, \le, \ge, <, >, \subseteq, \supseteq\}$. 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 [11] with relations "$\subseteq$" and "$\supseteq$", meanwhile no probabilistic interpretation of binary relations on sets was introduced in [15].

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 $Pr(A \subseteq B) = p(u \in B|\ u\in A) = p(u\in\{4, 5\}\ |\ u\in\{3, 4\}) = 0.5$.

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 $e_1 \vee e_2$, and difference event $e_1 \wedge \neg e_2$ can be computed by alternative strategies. In this work, we use the conjunction, disjunction, and difference strategies given in [22] as presented in Table 1, where $\otimes$, $\oplus$, and $\ominus$ denote the conjunction, disjunction, and difference operators, respectively.

Table 1: Definitions of probabilistic combination strategies

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	$([l_1, u_1] \otimes_{in} [l_2, u_2]) \equiv [l_1 . l_2, u_1 . u_2]$ $([l_1, u_1] \oplus_{in} [l_2, u_2]) \equiv [l_1 + l_2 - (l_1 . l_2), u_1 + u_2 - (u_1 . u_2)]$ $([l_1, u_1] \ominus_{in} [l_2, u_2]) \equiv [l_1 . (1 - u_2), u_1 . (1 - l_2)]$
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)	$([l_1, u_1] \otimes_{me} [l_2, u_2]) \equiv [0, 0]$ $([l_1, u_1] \oplus_{me} [l_2, u_2]) \equiv [min(1, l_1 + l_2), min(1, u_1 + u_2)]$ $([l_1, u_1] \ominus_{me} [l_2, u_2]) \equiv [l_1, min(u_1, 1 - l_2)]$

In the following sections, the notation $[l_1, u_1] \subseteq [l_2, u_2]$ is used to denote $l_2 \leq l_1$ and $u_1 \leq u_2$. Also, a single probability value p can be treated as the probability interval $[p, p]$ and the operation $p.[l, u]$ is computed as $[p.l, p.u]$.

3. EPRDB schema and relation

As in CRDB, fundamental concepts in EPRDB are the schema, relation and database. An EPRDB schema consists of a set of relational attributes respectively associated with domains that define probabilistic values of those attributes. The EPRDB schema is extended from that of CRDB with uncertain set-valued attributes as follows.

Definition 3.1. *An EPRDB schema* is a pair $R = (\boldsymbol{U}, \wp)$, where

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

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

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

An EPRDB relation is an instance of an EPRDB schema, where each relational attribute is associated with a probabilistic value to represent its uncertain set-valued information. The EPRDB relation is extended from that of CRDB in [2] with uncertain multivalued relational attributes as the following definition.

Definition 3.2. Let $\boldsymbol{U} = \{A_1, A_2, \ldots, A_k\}$ be a set of k pairwise different attributes. *An EPRDB relation* r over the schema $R(\boldsymbol{U}, \wp)$ is a finite set of elements $\{t_1, t_2, \ldots, t_n\}$, where each $t_i = (pv_{i1}, pv_{i2}, \ldots, pv_{ik})$ is a list of k probabilistic values such that pv_{ij} belongs to the set $\wp(A_j)$ for every $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k$.

Each element t in the relation r over $R(\boldsymbol{U}, \wp)$ is called a tuple on $\boldsymbol{U}$. For each tuple t_i, the probabilistic value pv_{ij} represents the uncertain valued set of the attribute A_j of the tuple t_i. We write $t_i.A_j$ or $t_i[A_j]$ to denote pv_{ij}.

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 EPRDB relation, named PATIENT, over the EPRDB schema **PATIENT**({NAME, AGE, DISEASE, D_COST}, $\wp$) can be given as Table 2.

Table 2. Relation PATIENT

NAME	AGE	DISEASE	D_COST
{(John, [1, 1])}	{(65, [1, 1])}	{(lung cancer, [0.5, 0.5]), (tuberculosis, [0.5, 0.5])}	{($30, [0.3, 0.6]), ($35, [0.4, 0.7])}
{(Paul, [1, 1])}	{(43,[0.5,0.5]), (44, [0.5,0.5])}	{({hepatitis, cirrhosis}, [0.5, 0.7]), ({cholecystitis}, [0.3, 0.5])}	{($6, [0.4, 0.6]), ($7, [0.4, 0.6])}
{(Helen, [1, 1])}	{(36, [1, 1])}	{(cholecystitis, [1, 1])}	{($8, [1, 1])}
{(Anne, [1, 1])}	{(15, [1, 1])}	{({bronchitis, angina}, [1, 1])}	{($7, [1, 1)}

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. For instance, the information of the patient John says that John's age is 65, the patient's disease may be lung cancer or tuberculosis with the probability 0.5 and John has to pay the daily treatment cost \$30 with the probability between 0.3 and 0.6 or \$35 with the probability between 0.4 and 0.7. Note that, for each attribute A in the schema **PATIENT**, $\wp(A)$ includes all probabilistic values on the domain of A (Definition 3.1).

The EPRDB relational database is defined as an extension of CRDB with uncertain set-valued attributes as follows.

Definition 3.3. *An EPRDB relational database* over a set of uncertain set-valued attributes is a set of EPRDB relations corresponding to the set of their EPRDB schemas.

4. Selection operation and queries on EPRDB

As in CRDB model, the selection is a basic algebraic operation in EPRDB model for querying data on relations of databases. The selection operation in EPRDB is extended from that of CRDB taking into account uncertain set-valued relational attributes. Before defining the selection operation, we present the formal syntax and semantics of selection expressions and conditions as below.

Definition 4.1. Let R be an EPRDB 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.

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

Definition 4.2. Let R be an EPRDB 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 4.1. 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 $(x.\text{AGE} > 40)[0.8, 1] \vee (x.\text{DISEASE} = \text{tuberculosis} \otimes x.\text{D_COST} \geq 30)[0.5, 0.6]$.

The probabilistic interpretation (i.e., semantics) of selection expressions in EPRDB is defined using the probabilistic interpretation of binary relations on sets as below.

Definition 4.3. Let R be an EPRDB 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 $Prob_{R,r,t}$, 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) $Prob_{R,r,t}(x.A\ \theta\ c) = \bigoplus_{i=1}^{k} [l_i, u_i].Pr(v_i\ \theta\ c)$, where $t.A = \{(v_1, [l_1, u_1]), \ldots, (v_k, [l_k, u_k])\}$ and $\oplus$ is the mutual exclusion probabilistic disjunction operator.

2) $Prob_{R,r,t}(x.A_1 =_{\otimes} x.A_2) = \bigoplus_{i=1}^{m}\bigoplus_{j=1}^{n}(([l_{1i}, u_{1i}] \otimes [l_{2j}, u_{2j}]).Pr(v_{1i} = v_{2j}))$, where $t.A_1 = \{(v_{11}, [l_{11}, u_{11}]), \ldots, (v_{1m}, [l_{1m}, u_{1m}])\}$, $t.A_2 = \{(v_{21}, [l_{21}, u_{21}]), \ldots, (v_{2n}, [l_{2n}, u_{2n}])\}$ and $\oplus$ is the mutual exclusion probabilistic disjunction operator.

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

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

We note that the probabilistic disjunction operator $\oplus_{me}$ is used in the item 1 and 2 of Definition 4.3 because the intervals $[l_1, u_1], \ldots, [l_k, u_k]$ represent a probability distribution function over $\{v_1, \ldots, v_k\}$, likewise for $[l_{11}, u_{11}], \ldots, [l_{1m}, u_{1m}]$ and $[l_{21}, u_{21}], \ldots, [l_{2n}, u_{2n}]$. Intuitively, $Prob_{R,r,t}(x.A\ \theta\ c)$ is the probability interval for the attribute A of the tuple t having a value v_i such that $v_i\ \theta\ c$, while $Prob_{R,r,t}(x.A_1 =_{\otimes} x.A_2)$ is the probability interval for the attributes A_1 and A_2 of the tuple t having values v_{1i} and v_{2j}, respectively, such that $v_{1i} = v_{2j}$.

Example 4.2. 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

$Prob_{R,r,t_2}(x.\text{DISEASE} \supseteq \{\text{hepatitis, cirrhosis}\})$

$= [0.5, 0.7].Pr(\{\text{hepatitis, cirrhosis}\} \supseteq \{\text{hepatitis, cirrhosis}\})$

$\oplus_{me}\ [0.3, 0.5].Pr(\{\text{cholecystitis}\} \supseteq \{\text{hepatitis, cirrhosis}\})$

$= [0.5, 0.7]\times 1.0 \oplus_{me} [0.3, 0.5]\times 0.0 = [0.5, 0.7] \oplus_{me} [0, 0] = [0.5, 0.7]$.

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

Definition 4.4. Let R be an EPRDB 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) $Prob_{R,r,t} \vDash (E)[L,U]$ if and only if (iff) $Prob_{R,r,t}(E) \subseteq [L,U]$.

2) $Prob_{R,r,t} \vDash \neg\phi$ iff $Prob_{R,r,t} \vDash \phi$ does not hold.

3) $Prob_{R,r,t} \vDash \phi \wedge \psi$ iff $Prob_{R,r,t} \vDash \phi$ and $Prob_{R,r,t} \vDash \psi$.

4) $Prob_{R,r,t} \vDash \phi \vee \psi$ iff $Prob_{R,r,t} \vDash \phi$ or $Prob_{R,r,t} \vDash \psi$.

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

Definition 4.5. Let R be an EPRDB 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 $r^* = \{t \in r \mid Prob_{R,r,t} \vDash \phi\}$ over R, including all satisfied tuples of the selection condition ϕ.

Example 4.3. 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 σ_ϕ(PATIENT), where $\phi = (x.\text{AGE} > 40)[0.9, 1] \wedge (x.\text{DISEASE} \supseteq \{\text{hepatitis, cirrhosis}\} \otimes_{in} x.\text{D_COST} \geq 6)[0.3, 0.7]$.

Only the second tuple t_2 of the relation PATIENT in Example 3.1 satisfies ϕ,

because

$Prob_{R,r,t_2}(x.\text{AGE} > 40) = [0.5, 0.5]\times Pr(43 > 40) \oplus_{me} [0.5, 0.5]\times Pr(44 > 40)$

$= [0.5, 0.5]\times 1 \oplus_{me} [0.5, 0.5]\times 1 = [1, 1] \subseteq [0.9, 1]$,

$Prob_{R,r,t_2}(x.\text{D_COST} \geq 6) = [0.4, 0.6]\times Pr(6 \geq 6) \oplus_{me} [0.4, 0.6]\times Pr(7 \geq 6)$

$= [0.4, 0.6] \times 1 \oplus_{me} [0.4, 0.6]\times 1 = [0.4, 0.6] \oplus_{me} [0.4, 0.6] = [0.8, 1]$.

From the result of the computation in Example 4.2, we have

$Prob_{R,r,t_2}(x.\text{DISEASE} \supseteq \{\text{hepatitis, cirrhosis}\} \otimes_{in} x.\text{D_COST} \geq 6)$

$= [0.5, 0.7] \otimes_{in} [0.8, 1] = [0.4, 0.7] \subseteq [0.3, 0.7]$.

For the other tuples, one has $Prob_{R,r,t_i}(x.\text{DISEASE} \supseteq \{\text{hepatitis, cirrhosis}\} \otimes_{in} x.\text{D_COST} \geq 6) = [0, 0] \not\subset [0.3, 0.7]$, $\forall i \neq 2$. Thus, the result of the query is as Table 3.

Table 3. Relation σ_ϕ(PATIENT)

NAME	AGE	DISEASE	D_COST
{(Paul, [1, 1])}	{(43, [0.5, 0.5]), (44, [0.5, 0.5])}	{({hepatitis, cirrhosis}, [0.5, 0.7]), ({cholecystitis}, [0.3, 0.5])}	{($6, [0.4, 0.6]), ($7, [0.4, 0.6])}

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

Theorem 4.1. Let r be a relation over the schema R in EPRDB, ϕ_1 and ϕ_2 be two selection conditions, then $\sigma_{\phi_1}(\sigma_{\phi_2}(r)) = \sigma_{\phi_2}(\sigma_{\phi_1}(r))$. (1)

Proof: Let $s = \sigma_{\phi_2}(r)$, by Definition 4.4 and 4.5, we have

$\sigma_{\phi_1}(\sigma_{\phi_2}(r)) = \{t \in s \mid Prob_{R,s,t} \vDash \phi_1\}$

$= \{t \in r \mid (Prob_{R,r,t} \vDash \phi_2) \wedge (Prob_{R,s,t} \vDash \phi_1)\}$

$= \{t \in r \mid (Prob_{R,r,t} \vDash \phi_2) \wedge (Prob_{R,r,t} \vDash \phi_1)\}$

$= \{t \in r \mid Prob_{R,r,t} \vDash \phi_1 \wedge \phi_2\} = \sigma_{\phi_1\wedge\phi_2}(r)$.

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 4.1 is proven.

5. Conclusions

We have presented a new probabilistic relational database model extended with probability intervals for uncertain set-valued attributes. Probabilistic values on the domains of set types have proposed to represent associating probability intervals with uncertain set-valued attributes. The probabilistic interpretation of binary relations on sets has used to define the selection operation for querying uncertain information expressed by relations of this model.

In the next steps, we will extend algebraic operations in the conventional relational database model as the 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 realworld applications.

Acknowledgment. This work is funded by Saigon University in the scientific research project of 2024-2026.

REFERENCES

[1] G. Özsoyoğlu, Z. M. Özsoyoğlu, and V. Matos, "Extending relational algebra and relational calculus with set-valued attributes and aggregate functions", *ACM Transactions on Database Systems*, vol.12, no.4, pp.566-592, 1987.

[2] A. Silberschatz, H.F. Korth and S. Sudarshan, *Database System Concepts*, Seventh Edition, McGraw-Hill, 2019.

[3] 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.

[4] Z. Ma and L. Yan, *Advances in Probabilistic Databases for Uncertain Information Management*, Springer-Verlag Berlin Heidelberg, 2013.

[5] 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.

[6] 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.

[7] H. Debbi, "Explaining Query Answers in Probabilistic Databases", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no.4, pp.140-152, 2023.

[8] 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.

[9] W. Zhao, A. Dekhtyar and J. Goldsmith, "Databases for interval probabilities", *International Journal of Intelligent Systems*, vol.19, no.9, pp.789-815**,** 2004.

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

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

[12] 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, 1992.

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

[14] S.K. Lee, "An extended relational database model for uncertain and imprecise information", *Proceedings of 18th Conference on Very Large Data Bases*, Canada, 1992, pp.211-220.

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

[16] H. Nguyen, T.N, Nguyen and T.T.N. Tran, "A probabilistic relational database model with uncertain multivalued attributes", *International Journal of Innovative Computing, Information and Control (ICIC Express Letters)*, vol. 16, no.3, pp.241-248, 2022.

[17] T. Friedman, G. Broeck, "Symbolic querying of vector spaces: probabilistic databases meets relational embeddings", *Proceedings of 36th Conference on Uncertainty in Artificial Intelligence*, Canada, 2020, vol.124, pp.1268-1277.

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

[19] A. Gilad, A. Imber and B. Kimelfeld, "The consistency of probabilistic databases with independent cells", *Proceedings of 26th International Conference on Database Theory*, Greece, 2023, pp. 22:1–22:19.

[20] V.V. Kheradkar and S. K. Shirgave, "Query processing over relationalcross model in uncertain and probabilistic databases", *Proceedings of 3th International Conference on Artificial Intelligence and Smart Energy*, India, 2023, pp.763 – 769.

[21] T.V. Bremen and K.S. Meel, "Probabilistic query evaluation: The combined FPRAS landscape", *Proceedings of 42th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, USA, 2023, pp 339–347.

[22] V. Biazzo, R. Giugno, T. Lukasiewicz and V. S. Subrahmanian, "Temporal Probabilistic Object Bases", *IEEE Transactions on Knowledge and Data Engineering*, vol.15, no.4, pp. 921–939, 2003.

Ngày nhận bài: 27/02/2024 Ngày chấp nhận đăng: 02/4/2024