MJSAT

Malaysian Journal of Science and Advanced Technology

journal homepage: https://mjsat.com.my/

A Relational Database Model with Probability Intervals for Uncertain Set-Valued Attributes

Hoa Nguyen*[1], and Thi Nhi Tran[1]

[1] *Information Technology Faculty, Saigon University, Vietnam.*

KEYWORDS

Probability Interval
Uncertain Set-Valued Attributes
Probabilistic Values
Probabilistic Relations
UIRDB model

ARTICLE HISTORY

Received 26 May 2024
Received in revised form 23 August 2024
Accepted 15 September 2024
Available online 29 September 2024

ABSTRACT

Although the conventional relational database model is very useful for modeling, designing and implementing large-scale systems, it is limited for expressing and dealing with uncertain and imprecise information. 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 in practice. To build the new database model, we use three key methods: (1) Extended probabilistic values of set data types are proposed for representing uncertain set-valued attributes; (2) The probabilistic interpretations of binary relations on sets are defined for computing the uncertain degree of relations on set values of relational attributes; and (3) The combination strategies of probability intervals are employed for manipulating uncertain data relations. Then, fundamental concepts as schemas, probabilistic relations, probabilistic relational database, the selection operation and uncertain and imprecise information queries are defined coherently and consistently for the new model.

1. INTRODUCTION

As known, the conventional relational database model (CRDB), as in [1] and [2], is very useful for modeling, designing and implementing large-scale systems, 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. For example, in [3] authors proposed a PRDB model to compute the uncertain membership degree of each tuple in a relation, and in [4] authors introduced another PRDB model that can compute the uncertain degree of attribute values of each tuple in a relation. Probabilistic database models also have been used in many real applications, such as the works in [5] and [6]. More particularly, in [5] probabilistic databases were applied for detecting faulty sensors, and in [6] queries over the relational cross model were processed by using uncertain databases.

Probabilistic relational database models are developed and built as extensions of CRDB based on the probability theory. There are two main types of PRDB models extended from the CRDB model. The first one defines a probabilistic relation as a set of tuples such that each tuple is associated with a probability to express the uncertainty degree of it in the relation. The second one defines a probabilistic relation as a set of tuples such that each tuple attribute is associated with a probability to represent the uncertainty degree of the values that it may take.

The first PRDB model type is the extension of CRDB at the relation level, as the works in [7], [8] and [9], 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 [10], [11], [12] and [13],

*Corresponding author:
E-mail address: Hoa Nguyen <nguyenhoa@sgu.edu.vn>.
https://doi.org/10.56532/mjsat.v4i4.329
2785-8901/

were extended with probability intervals associated with each tuple to overcome the shortcoming.

The second PRDB model type is the extension of CRDB at the attribute level, as the works in [14] and [15], 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 [16], each attribute was associated with a probability distribution on a set of values to express the possibility that the attribute might take one of values of the set with a distributed probability. However, in many real cases, we cannot define precisely the probability distribution function for each value in the set but only can estimate it as an approximate number in a subinterval of [0, 1]. The model in [17] overcame the restriction by using a pair of lower and upper bound probability distribution functions to represent the possibility that an attribute might take a value in a set with a computed probability interval from the distribution function 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 [16] represented the attribute DISEASE of the patient Mary by DISEASE: {⟨{d_1, d_3}, 0.6⟩, ⟨{d_2}, 0.4⟩} 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 [16] 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 models in [18] and [19] have been proposed to overcome the shortcomings of the models in [16] and [17] 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.

Although there are many PRDB models proposed and built as mentioned above, but no model would be so universal that could include all measures and tackle all aspects of uncertainty of information in the real world.

In this paper, we propose a new probabilistic relational database model for uncertain and imprecise information, named UIRDB, as an extension of CRDB with probability intervals for uncertain set-valued attributes to overcome the limitations of models in [16], [18] and [19]. The UIRDB model is consistent with CRDB model by allowing multivalued attributes and more flexibly than the models in [16], [18] and [19] by using probability intervals instead of probability single values and distribution functions.

Our proposed UIRDB model is a second type PRDB model. To build UIRDB, we extend the definition of the probabilistic value on a set in [20] to the new definition of the probabilistic value on a set of sets (i.e., the definition of the extended probabilistic value) for representing uncertain set-valued attributes of relations and employ probabilistic interpretations of binary relations on sets in [18] to define the selection expressions and conditions for computing uncertain and imprecise data. The combination operators of probability intervals in [18] are also used to build the new selection operation for manipulating and querying uncertain and imprecise information on UIRDB relations.

The UIRDB has the capability of expressing uncertain information better than the first type PRDB models, as in [10], [11], [12] and [13], since using probabilistic values instead of certain, single values. Moreover, the UIRDB also has the ability of querying uncertain data more effectively than the second type PRDB models, as in [17], [18], and [19], since computing on probability intervals instead of on probability distribution function pairs.

The new built UIRDB model is able to represent and manipulate effectively uncertain and imprecise information and can be applied to solve problems in real databases.

The mathematical base for UIRDB is presented in Section 2. Schemas and relations of UIRDB are defined in Section 3. The methodology for building the data model, defining the selection operation and query on UIRDB is introduced in Section 3. Section 4 shows out the achieved results and discussion of UIRDB model. Finally, Section 5 concludes the paper and outlines further research directions in the future.

2. PROBABILITY DEFINITIONS

The mathematical base for UIRDB model includes some probability definitions and notions for representing and handling uncertain and imprecise information.

2.1 Extended Probabilistic Values

For expressing uncertain set-valued attributes in UIRDB, probabilistic values over a set in [20] are extended to probabilistic values over a set of sets as follows.

Definition 1. Let τ be a data type and D be the domain of τ, an *extended probabilistic value* on the domain of τ is a finite set of pairs {(v_1, [l_1, u_1]), …, (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 $i, j = 1, 2, \ldots, m$.

Informally, an extended probabilistic value pv = {(v_1, [l_1, u_1]), …, (v_m, [l_m, u_m])} says that pv's value is exactly one member (set) v_i of the set V = {v_1,…, v_m} and the probability that pv's value is v_i lies in the interval [l_i, u_i]. Thus, an extended probabilistic value represents both the uncertainty of its value and the imprecision of the probability for that value. An extended probabilistic value pv = {(v_1, [l_1, u_1]), …, (v_m, [l_m, u_m])} corresponds with a probability distribution function p over V = {v_1,…, v_m} such that $p(v_i) \in [l_i, u_i]$, $i = 1,\ldots, m$ and $\Sigma_{v_i \in V} p(v_i) \le 1$.

Example 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 extended 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 an extended 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. Also, "an extended probabilistic value" is called "a probabilistic value".

2.2 Probabilistic Interpretation of Binary Relations on Sets

For computing the uncertain degree of relations on attribute values in UIRDB, we use the probabilistic interpretation of binary relations on sets in [18] as below.

Definition 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. $Pr(A\ \theta\ B) = \begin{cases} p(u \in B |\ u\in A), \theta \text{ is } \subseteq \\ p(u \in A |\ u\in B), \theta \text{ is } \supseteq \end{cases}$
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 [12] with relations "$\subseteq$" and "$\supseteq$", meanwhile no probabilistic interpretation of binary relations on sets was introduced in [17].

Example 2. Let $A = \{4, 5\}$ and $B = \{5, 6\}$ be two sets on the domain $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Then

$$\begin{aligned} Pr(A = B) &= p(u = v |\ u\in A, v \in B) \\ &= p(u=v | u\in\{4, 5\}, v\in\{5, 6\}) \\ &= 0.25. \\ Pr(A < B) &= p(u < v |\ u\in A, v \in B) \\ &= p(u< v | u\in\{4, 5\}, v\in\{5, 6\}) \\ &= 0.75. \\ Pr(A \subseteq B) &= p(u \in B |\ u\in A) \\ &= p(u\in\{5, 6\} \,|\, u\in\{4, 5\}) \\ &= 0.5. \end{aligned}$$

2.3 Combination Strategies of Probability Intervals

In many real situations, the probability of an event may not be defined or computed exactly [21] and [22], a probability interval can be used instead of a precise single probability value. 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 [20], where $\otimes$, $\oplus$, and $\ominus$ denote the conjunction, disjunction, and difference operators, respectively and in turn defined as follows.

1. Independence conjunction, disjunction, and difference strategies, denoted $\otimes_{in}$, $\oplus_{in}$, and $\ominus_{in}$ respectively, are determined by:
   - $[l_1, u_1] \otimes_{in} [l_2, u_2] = [l_1 \,.\, l_2, u_1 \,.\, u_2]$
   - $[l_1, u_1] \oplus_{in} [l_2, u_2] = [l_1 + l_2 - (l_1 \,.\, l_2), u_1 + u_2 - (u_1 \,.\, u_2)]$
   - $[l_1, u_1] \ominus_{in} [l_2, u_2] = [l_1 \,.\, (1 - u_2), u_1 \,.\, (1- l_2)]$
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:
   - $[l_1, u_1] \otimes_{me} [l_2, u_2] = [0, 0]$
   - $[l_1, u_1] \oplus_{me} [l_2, u_2] = [min(1, l_1 + l_2), min(1, u_1 + u_2)]$
   - $[l_1, u_1] \ominus_{me} [l_2, u_2] = [l_1, min(u_1, 1 - l_2)]$
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:
   - $[l_1, u_1] \otimes_{pc} [l_2, u_2] = [min(l_1, l_2), min(u_1, u_2)]$
   - $[l_1, u_1] \oplus_{pc} [l_2, u_2] = [max(l_1, l_2), max(u_1, u_2)]$
   - $[l_1, u_1] \ominus_{pc} [l_2, u_2] = [max(0, l_1 - u_2), max(0, u_1 - l_2)]$
4. Ignorance conjunction, disjunction, and difference strategies, denoted $\otimes_{ig}$, $\oplus_{ig}$, and $\ominus_{ig}$ respectively, are determined by:
   - $[l_1, u_1] \otimes_{ig} [l_2, u_2] = [max(0, l_1 + l_2 - 1), min(u_1, u_2)]$
   - $[l_1, u_1] \oplus_{ig} [l_2, u_2] = [max(l_1, l_2), min(1, u_1 + u_2)]$
   - $[l_1, u_1] \ominus_{ig} [l_2, u_2] = [max(0, l_1 - u_2), min(u_1, 1 - l_2)]$

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. PROPOSED METHODOLOGY

The proposed UIRDB including the data model and query operations is defined and built by extending the conventional relational database model [2] using the probability definitions and notions presented above.

3.1 UIRDB Data Model

As CRDB data model, UIRDB data model is a structure including fundamental components as the schema, relation and database.

A UIRDB schema consists of a set of relational attributes respectively associated with domains that define (extended) probabilistic values of those attributes. The UIRDB schema is extended from that of CRDB with uncertain set-valued attributes as follows.

Definition 3. *A UIRDB 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 (extended) probabilistic values on the domain of A.

For simplicity, the notation $R(\boldsymbol{U}, \wp)$ and R can be used to denote $R = (\boldsymbol{U}, \wp)$, the domain of A is denoted by $dom(A)$.

A UIRDB relation is an instance of a UIRDB schema, where each relational attribute is associated with a probabilistic value to represent an uncertain value set that the attribute may take. The UIRDB relation is extended from that of CRDB in [2] with uncertain multivalued relational attributes as the following definition.

Definition 4. Let $\boldsymbol{U} = \{A_1, A_2, \ldots, A_k\}$ be a set of k pairwise different attributes. *A UIRDB 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. In the database about patients at the clinic of a hospital, a simple UIRDB relation, named PATIENT, over the UIRDB schema **PATIENT**({NAME, AGE, DISEASE, D_COST}, $\wp$) can be given as Table 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 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.

Table 1. 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])}	{(43, [1, 1])}	{(cholecystitis, [1, 1])}	{(\$8, [1, 1])}
{(**Selena**, [1, 1])}	{(15, [1, 1])}	{({bronchitis, angina}, [1, 1])}	{(\$12, [0.5, 0.5]), (\$13, [0.5, 0.5]}
{(Alice, [1, 1])}	{(36, [1, 1])}	{(duodenitis, [0.4, 0.5]), (gastritis, [0.5, 0.6])}	{(\$8, [0.3, 0.5]), (\$9, [0.5, 0.7])}

Note that, for each attribute A in the schema **PATIENT**, $\wp(A)$ includes all extended probabilistic values on the domain of A (Definition 3). In other words, each attribute A in the relation PATIENT is associated with an extended probabilistic value $\{(v_1, [l_1, u_1]), \ldots, (v_m, [l_m, u_m])\}$ for A taking some v_i with a probability in the interval $[l_i, u_i]$. For instance, the value of the attribute DISEASE of the patient Paul represented by {({hepatitis, cirrhosis}, [0.5, 0.7]), ({cholecystitis}, [0.3, 0.5])} says that Paul's diseases may be hepatitis and cirrhosis with the probability between 0.5 and 0.7 or cholecystitis with the probability between 0.3 and 0.5. In the patient database, we can query uncertain and imprecise information about patients such as "Find all patients who are not over 45 years old and have cholecystitis with a probability of at least 0.3" or "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 0.4 and 0.7" and so on. The formal query langue for the UIRDB model will be defined in the next section to answer the queries.

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

Definition 5. *A UIRDB relational database* over a set of uncertain set-valued attributes is a set of UIRDB relations corresponding to the set of their UIRDB schemas.

3.2 Selection Operation and Queries on UIRDB

As in CRDB model, the selection is a basic algebraic operation in UIRDB model for querying data on relations of databases. The selection operation as the formal query langue in UIRDB 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 6. *Let R* be a UIRDB 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 $\{=, \neq, \leq, \geq, <, >, \subseteq, \supseteq\}$, $c \in 2^D$, and D is the domain of A.
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. $\alpha \otimes \beta$, where α and β are selection expressions on the same relational tuple variable, and $\otimes$ is a probabilistic conjunction strategy.
4. $\alpha \oplus \beta$, where α and β are selection expressions on the same relational tuple variable, and $\oplus$ is a probabilistic disjunction strategy.

Example 4. Consider the schema **PATIENT** in Example 3, the selection of "all patients who get bronchitis and pay the daily treatment cost over 10 USD" can be represented by the selection expression x.DISEASE = bronchitis $\otimes$ x.D_COST > 10.

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

Definition 7. Let R be a UIRDB schema. Then *selection conditions* are inductively defined as follows:

1. If α is a selection expression and $[l, u]$ is a subinterval of $[0, 1]$, then $(\alpha)[l, u]$ is a selection condition.
2. If φ and ω are selection conditions on the same tuple variable, then $\neg\varphi$, $(\varphi \wedge \omega)$, $(\varphi \vee \omega)$ are selection conditions.

Example 5. Given the schema **PATIENT** in Example 3, the selection of "all patients who are over 50 years old with a probability of at least 0.7 or have lung cancer and pay the daily treatment cost not less than 35 USD with a probability from 0.4 to 0.6" can be done using the selection condition (x.AGE > 50)[0.7, 1.0] $\vee$ (x.DISEASE = lung cancer $\otimes$ x.D_COST $\geq$ 35)[0.4, 0.6].

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

Definition 8. Let R be a UIRDB 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. $Prob_{R,r,t}(\alpha \otimes \beta) = Prob_{R,r,t}(\alpha) \otimes Prob_{R,r,t}(\beta)$.
4. $Prob_{R,r,t}(\alpha \oplus \beta) = Prob_{R,r,t}(\alpha) \oplus Prob_{R,r,t}(\beta)$.

We note that the probabilistic disjunction operator $\oplus_{me}$ is used in the item 1 and 2 of Definition 8 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 (set) 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 6. Let R denote the schema **PATIENT** and r denote the relation PATIENT in Example 3. Consider the second tuple in r, denoted by t_2. We have

$Prob_{R,r,t_2}(x.\text{DISEASE} = \text{cholecystitis})$
$= [0.5, 0.7].Pr(\{\text{hepatitis, cirrhosis}\} = \text{cholecystitis})$
$\oplus_{me} [0.3, 0.5].Pr(\text{cholecystitis} = \text{cholecystitis})$
$= [0.5, 0.7] \times 0.0 \oplus_{me} [0.3, 0.5] \times 1.0$
$= [0, 0] \oplus_{me} [0.3, 0.5]$
$= [0.3, 0.5]$.

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

Definition 9. Let R be a UIRDB 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. $Prob_{R,r,t} \vDash (\alpha)[l, u]$ if and only if (iff) $Prob_{R,r,t}(\alpha) \subseteq [l, u]$.
2. $Prob_{R,r,t} \vDash \neg\varphi$ iff $Prob_{R,r,t} \vDash \varphi$ does not hold.
3. $Prob_{R,r,t} \vDash \varphi \wedge \omega$ iff $Prob_{R,r,t} \vDash \varphi$ and $Prob_{R,r,t} \vDash \omega$.
4. $Prob_{R,r,t} \vDash \varphi \vee \omega$ iff $Prob_{R,r,t} \vDash \varphi$ or $Prob_{R,r,t} \vDash \omega$.

Note that, in CRDB, the concepts of the selection expression and selection condition are identical, where probability intervals $[l, u]$ in selection conditions to be 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 UIRDB.

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

Definition 10. Let R be a UIRDB 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 σ_{φ}®, is the relation $r^* = \{t \in r \mid Prob_{R,r,t} \vDash \varphi\}$ over R, including all satisfied tuples of the selection condition φ.

Example 7. Let r denote the relation PATIENT in Example 3 and R denote its schema. The query "Find all patients who are not over 45 years old and have cholecystitis with a probability of at least 0.3" can be done by the selection operation σ_{φ}(PATIENT), where $\varphi = (x.\text{AGE} \leq 45 \otimes_{in} x.\text{DISEASE} = \text{cholecystitis})[0.3, 1.0]$.

There are two patients denoted by the second and third tuples (t_2 and t_3) of the relation PATIENT in Example 3 satisfies φ, because:

For t_2, we have
$Prob_{R,r,t_2}(x.\text{AGE} \leq 45)$
$= [0.5, 0.5] \times Pr(43 \leq 45) \oplus_{me} [0.5, 0.5] \times Pr(44 \leq 45)$
$= [0.5, 0.5] \times 1.0 \oplus_{me} [0.5, 0.5] \times 1.0 = [1.0, 1.0]$.

From the result of the computation in Example 6, we get $Prob_{R,r,t_2}(x.\text{DISEASE} = \text{cholecystitis}) = [0.3, 0.5]$.

Hence
$Prob_{R,r,t_2}(x.\text{AGE} \leq 45 \otimes_{in} x.\text{DISEASE} = \text{cholecystitis})$
$= Prob_{R,r,t_2}(x.\text{AGE} \leq 45)$
$\otimes_{in} Prob_{R,r,t_2}(x.\text{DISEASE} = \text{cholecystitis})$
$= [1.0, 1.0] \otimes_{in} [0.3, 0.5] = [0.3, 0.5] \subseteq [0.3, 1.0]$.

Thus t_2 satisfies φ.

For t_3, we have
$Prob_{R,r,t_3}(x.\text{AGE} \leq 45)$
$= [1.0, 1.0] \times Pr(43 \leq 45) = [1.0, 1.0] \times 1.0 = [1.0, 1.0]$.
$Prob_{R,r,t_3}(x.\text{DISEASE} = \text{cholecystitis})$
$= [1.0, 1.0] \times Pr(\text{cholecystitis} = \text{cholecystitis})$
$= [1.0, 1.0] \times 1.0 = [1.0, 1.0]$.

Hence
$Prob_{R,r,t_3}(x.\text{AGE} \leq 45 \otimes_{in} x.\text{DISEASE} = \text{cholecystitis})$
$= Prob_{R,r,t_3}(x.\text{AGE} \leq 45)$
$\otimes_{in} Prob_{R,r,t_3}(x.\text{DISEASE} = \text{cholecystitis})$
$= [1.0, 1.0] \otimes_{in} [1.0, 1.0] = [1.0, 1.0] \subseteq [0.3, 1.0]$.

Thus t_3 satisfies φ.

For the other tuples, one has $Prob_{R,r,t_i}(x.\text{AGE} \leq 45 \otimes_{in} x.\text{DISEASE} = \text{cholecystitis}) = [0, 0] \not\subset [0.3, 1.0]$, $\forall i \neq 2, 3$. Thus, the result of the query is as Table 2.

Table 2. 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])}
{(Helen, [1, 1])}	{(43, [1, 1])}	{(cholecystitis, [1, 1])}	{($8, [1, 1])}

Example 8. Let r denote the relation PATIENT in Example 3 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 0.4 and 0.7" can be done by the selection operation σ_{ω}(PATIENT), where $\omega = (x.\text{AGE} > 40)[0.9, 1.0] \wedge (x.\text{DISEASE} \supseteq \{\text{hepatitis, cirrhosis}\} \otimes_{in} x.\text{D\_COST} \geq 6)[0.4, 0.7]$.

Only one patient denoted by the second tuple t_2 of the relation PATIENT in Example 3 satisfies ω, because:
$Prob_{R,r,t_2}(x.\text{AGE} > 40)$
$= [0.5, 0.5] \times Pr(43 > 40) \oplus_{me} [0.5, 0.5] \times Pr(44 > 40)$
$= [0.5, 0.5] \times 1.0 \oplus_{me} [0.5, 0.5] \times 1.0 = [1.0, 1.0] \subseteq [0.9, 1.0]$.
$Prob_{R,r,t_2}(x.\text{DISEASE} \supseteq \{\text{hepatitis, cirrhosis}\})$
$= [0.5, 0.7].Pr(\{\text{hepatitis, cirrhosis}\} \supseteq \{\text{hepatitis, cirrhosis}\})$
$\oplus_{me} [0.3, 0.5].Pr(\{\text{cholecystitis}\} \supseteq \{\text{hepatitis, cirrhosis}\})$
$= [0.5, 0.7] \times 1.0 \oplus_{me} [0.3, 0.5] \times 0.0$
$= [0.5, 0.7] \oplus_{me} [0, 0] = [0.5, 0.7]$.
$Prob_{R,r,t_2}(x.\text{D\_COST} \geq 6)$
$= [0.4, 0.6] \times Pr(6 \geq 6) \oplus_{me} [0.4, 0.6] \times Pr(7 \geq 6)$
$= [0.4, 0.6] \times 1.0 \oplus_{me} [0.4, 0.6] \times 1.0$
$= [0.4, 0.6] \oplus_{me} [0.4, 0.6] = [0.8, 1.0]$.

$Prob_{R,r,t_2}$(x.DISEASE $\supseteq$ {hepatitis, cirrhosis} $\otimes_{in}$ x.D_COST$\geq$6)
= [0.5, 0.7] $\otimes_{in}$ [0.8, 1.0] = [0.4, 0.7] $\subseteq$ [0.4, 0.7].

Hence, $Prob_{R,r,t_2} \vDash$ (x.AGE > 40)[0.9, 1.0] and $Prob_{R,r,t_2} \vDash$ (x.DISEASE $\supseteq$ {hepatitis, cirrhosis} $\otimes_{in}$ x.D_COST $\geq$ 6)[0.4, 0.7]. Thus t_2 satisfies ω.

For the other tuples, one has $Prob_{R,r,t_i}$(x.DISEASE $\supseteq$ {hepatitis, cirrhosis} $\otimes_{in}$ x.D_COST $\geq$ 6) = [0, 0] $\not\subset$ [0.4, 0.7], $\forall i \neq 2$. Thus, the result of the query is as Table 3.

Table 3. Relation σ_{ω}(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 UIRDB is not dependent on the order of selection conditions as the following theorem.

Theorem 1. Let r be a relation over the schema R in UIRDB, φ and ω be two selection conditions on the same tuple variable, then

$$\sigma_{\varphi}(\sigma_{\omega}(r)) = \sigma_{\omega}(\sigma_{\varphi}(r)) \quad (1)$$

Proof: Let $s = \sigma_{\omega}(r)$, by Definition 9 and 10, we have

$$\begin{aligned}\sigma_{\varphi}(\sigma_{\omega}(r)) &= \{t \in s \mid Prob_{R,s,t} \vDash \varphi\} \\ &= \{t \in r \mid (Prob_{R,r,t} \vDash \omega) \wedge (Prob_{R,s,t} \vDash \varphi)\} \\ &= \{t \in r \mid (Prob_{R,r,t} \vDash \omega) \wedge (Prob_{R,r,t} \vDash \varphi)\} \\ &= \{t \in r \mid Prob_{R,r,t} \vDash \varphi \wedge \omega\} = \sigma_{\varphi \wedge \omega}(r).\end{aligned}$$

Thus, the equation $\sigma_{\varphi}(\sigma_{\omega}(r)) = \sigma_{\varphi \wedge \omega}(r)$ is proven. The equation $\sigma_{\omega}(\sigma_{\varphi}(r)) = \sigma_{\omega \wedge \varphi}(r)$ is similarly proven. Since $\omega \wedge \varphi \Leftrightarrow \varphi \wedge \omega$. So, Theorem 1 is proven.□

4. RESULT AND DISCUSSION

It easy to see that UIRDB is an extension of CRDB and the second type PRDB models as in [14], [15] and [16] with extended probabilistic values (i.e. probabilistic intervals for value sets). Moreover, UIRDB also has the ability of querying data more effectively than the second type PRDB models as in [17], [18] and [19]. A more detailed discussion of the obtained results is as below.

4.1 Extension of UIRDB in representing and handling data

As mentioned above, there are two main types of the PRDB models. The first type one, denoted by T-1PRDB, represents a probabilistic relation as a set of tuples whose membership degree is a probability in [0, 1], such as [8] and [9]. Each attribute of a tuple is associated with a single value to say that the attribute may take the value with a probability computed and inferred from the membership degree of the tuple. The T-1PRDB selection operation and query are defined by extending directly the CRDB selection operation and query based on computing and combining probabilities of tuples in the T-1PRDB relations.

The second type one, denoted by T-2PRDB, represents a probabilistic relation as a set of tuples whose membership degree is a probability in {0, 1}, such as [4], [14] and [15] each relational attribute is associated with a single probability value as (v, p) to say that the attribute may take the value v with the probability p. Some extended models of T-2PRDB such as [16], denoted by ET-2PRDB, where each relational attribute is associated with a probability distribution as {(v_1, p_1),..., (v_m, p_m)} to say that the attribute may take one of values v_i with the probability p_i. The T-2PRDB and ET-2PRDB selection operation and query are defined by extending the CRDB selection operation and query, using operators on single probabilities or probability distributions for computing and combining probabilities of attribute values in the T-2PRDB or ET-2PRDB relations.

As presented in previous sections, the proposed UIRDB model belongs to T-2PRDB. Each relational attribute in UIRDB is associated with an extended probabilistic value pv = {(v_1, [l_1, u_1]), …, (v_m, [l_m, u_m])} (as a distribution of probability intervals on a finite set of value sets) to say that the attribute may take one set of values v_i with a probability in [l_i, u_i]. The UIRDB selection operation and query are defined by extending the CRDB selection operation and query, employing the probabilistic interpretations of binary relations on sets, the combination strategies of probabilistic intervals of attribute values (i.e. extended probabilistic values) in the C-2PRDB relations.

We can see that a special extended probabilistic value in UIRDB as {(v_1, [p_1, p_1]), …, (v_m, [p_m, p_m])} with v_i being a single value also is a probability distribution {(v_1, p_1),..., (v_m, p_m)} in the model [16]. Thus, UIRDB model is an extension of T-2PRDB models, such as [15] and [16] with extended probabilistic values (Definition 1 and 4). Moreover, by associating probabilistic intervals with attribute values (in extended probabilistic values), UIRDB allows representing both the uncertainty of attribute values and the imprecision of the probability for that attribute values, whereas the models as [15] and [16] only allow representing the uncertainty of attribute values but do not allow expressing the imprecision of the probability for that attribute values (because in {(v_1, p_1),..., (v_m, p_m)}, the probability for the value v_i is a precise number p_i). In addition, UIRDB model also allows uncertain multivalued attributes (i.e. uncertain set-valued attributes) whereas the models as [15] and [16] do not permit set-valued attributes. Fig.1 illustrates the extension of UIRDB in comparison with the CRDB, T-2PRDB and ET-2PRDB models.

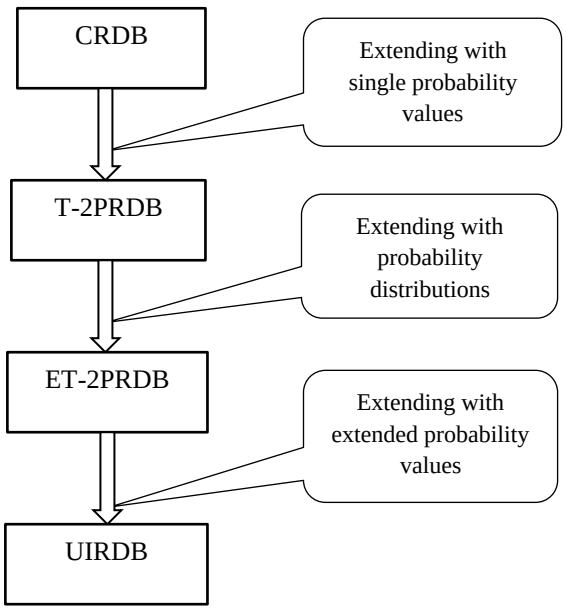

Fig. 1. Extension of UIRDB

4.2 Efficiency of UIRDB in computing and querying data

In CRDB model, as we have known, the computing complexity of a selection query on a CRDB relation having n tuples is O(n). In ET-2PRDB models, such as the model in [16], since each relational attribute is represented by a probability distribution function of a set of values, the computing complexity of a selection query on an ET-2PRDB relation having n tuples is O(kn), where k is the cardinality of the domain of the distribution function.

In UIRDB model, since each relational attribute is represented by a list of some values or data associated with probability intervals (i.e. an extended probabilistic value), the selection queries on a UIRDB relation, defined by the selection operation as in the section 3.2, are more effectively than those on ET-2PRDB models, where each relational attribute is represented by a probability distribution function of a set of values. The computing complexity of a UIRDB selection query is a polynomial under the size of probabilistic relations and it is as effective as the computing complexity of a CRDB selection query. Indeed, because the computation time that a tuple holds or does not hold a selection condition is bounded above by some constant under the constant of some probability intervals of relational attribute values (Definition 8 and 9), then the cost for the selection of each tuple in a UIRDB relation (Definition 10) also is some constant or O(1). From that, the computing time complexity of a selection query on a UIRDB relation having n tuples is O(n).

Because each relational attribute of T-2PRDB models is represented by a single probability value, these models (e.g., [14], [15]), are special cases of UIRDB model. Consequently, the computing complexity of a selection query on a T-2PRDB relation having n tuples also is O(n). However, in the models [17], [18] and [19] that each relational attribute is represented by a probability distribution function pair of a set of values, the computing complexity of a selection query on a relation having n tuples is O(kn), where k is the cardinality of the domain of the distribution function pair.

Table 4 illustrates the efficiency of a selection query on a relation having n tuples in CRDB, T-2PRDB, ET-2PRDB and UIRDB models, where k is the cardinality of the domain of a distribution function that represents a relational attribute value.

Table 4. Efficiency of query on relation of database models

MODEL	RELATIONAL ATTRIBUTE VALUE	EFFICIENCY OF QUERY
CRDB	Single values	O(n)
T-2PRDB	Single probability values	O(n)
ET-2PRDB	Probability distributions	O(kn)
UIRDB	Extended probability values	O(n)

From the discussion above, we can say that the performance of UIRDB model in computing and querying uncertain and imprecise information is good and can apply it in practice.

5. CONCLUSION

We have presented a new probabilistic relational database model extended with probability intervals for uncertain set-valued attributes. Extended 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. The new built model has the ability of representing, querying and dealing with effectively uncertain and imprecise data.

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 real world applications.

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] Z. Ma and L. Yan, *Advances in probabilistic databases for uncertain information management*, Springer-Verlag Berlin Heidelberg, 2013.

[4] 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.

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

[6] 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*, Coimbatore, India, pp.763-769, 2023.

[7] D. Suciu, "Probabilistic databases for all", *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, USA, 2020, pp. 19–31.

[8] 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.

[9] H. Debbi, "Explaining query answers in probabilistic databases", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no.4, pp.140-152, 2023.

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

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

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

[13] C. Zhang, Z. Mei, B. Wu, Z. Zhao, J. Yu, Q. Wang , "Query with assumptions for probabilistic relational databases", *Technical gazette*, vol. 27, no. 3, pp.923-932, 2020.

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

[15] K. Papaioannou, M. Theobald, and M. Böhlen, "Supporting set operations in temporal-probabilistic databases", *Proceedings of the 34th IEEE International Conference on Data Engineering*, France, 2018, pp. 1180-1191.

[16] 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.

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

[18] H. Nguyen, T.N, Nguyen and T.T.N. Tran, "A probabilistic relational database model with uncertain multivalued attributes", *ICIC Express Letters*, vol. 16, no.3, pp.241-248, 2022.

[19] H. Nguyen, "Extending probabilistic relational database model with uncertain multivalued attributes", *International Journal of Innovative Computing, Information and Control*, vol.18, no.5, pp.1477–1492, 2022.

[20] 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.

[21] 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.

[22] 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.