A Fuzzy Probabilistic Relational Database Model and Algebra

Li Yan and Z. M. Ma

Abstract

This paper describes an extended relational database model based on probability theory and possibility theory. Fuzzy information and probabilistic information are incorporated into the relational databases simultaneously to represent fuzzy probability of events in the real-world applications. The tuples in such a relation are associated with a possibility distribution, and their attribute values may be uncertain and represented by probabilistic distributions. Such relational databases are called fuzzy probabilistic relational databases. The structure of the fuzzy probabilistic relational model is proposed in the paper. On the basis, the primitive and natural join operations of the algebra are developed.

Keywords: Possibility distribution, probabilistic distribution, fuzzy probabilistic relational databases, relational algebra.

1. Introduction

One of the major areas of database research has been the continuous effort to enrich existing database models with a more extensive collection of semantic concepts because traditional database models often suffer from some inadequacy of necessary semantics. One of these inadequacies can be generalized as the inability to handle imprecise and uncertain information. In real-world applications, information is often imperfect, and human knowledge and natural language have a big deal of imprecision and vagueness. Traditional database models assume that the models are a correct reflection of the world and further assume that the stored data is known, accurate and complete. It is rarely the case in real life that all or most of these assumptions are met. For this reason, imprecise and uncertain data have been introduced into databases [15].

Intuitively, the imprecision is relevant to the content of a value, and it means that a choice must be made from a given range (interval or set) of values but we do not know exactly which one to choose at present. The uncertainty is relevant to the degree of truth of its value, and it means that we can apportion some, but not all, of our belief to a given value or a group of values. We have two ways to model imprecise and uncertain information: one is based on fuzzy sets and possibility distributions [22, 24]; another one is based on probability distributions. Fuzzy sets are applied for representing one kind of vague concept such as tall, heavy, big, or young, etc. For a stochastic event in the real world, it may appear in different states. All these states are conjunctive and the appearance of each state follows the certainty rule, which is described with probability measure (being less than or equal to one). It is unknown which state will appear in advance. So the complete states of a stochastic event are represented by a set where all states are combined with the corresponding probability measures, namely, probabilistic distribution. In a probabilistic distribution, the sum of all probability measures should be equal to one. In general, it is possible that there exists missing probability measure. So the sum of all probability measures in a probabilistic distribution may be less than one.

Fuzzy set and probability theories have been used to extend various database models and this has resulted in numerous contributions, mainly with respect to the popular relational model or to some related form of it. While the fuzzy and probabilistic relational databases are being developed separately for dealing with subjective uncertainty and objective uncertainty, respectively, each of these two kinds of relational databases actually suffers from the inability to simultaneously handle fuzzy and probabilistic information, which is needed by real-world applications. Let us look at an example of weather forecast. Sometimes it is easy to understand “the probability it will be raining tomorrow is very high” instead of “the probability it will be raining tomorrow is 0.85”. In this example, we face a kind of probabilistic event which probability is fuzzy instead of crisp and their interaction and/or integration becomes a crucial issue. However, although it has received increasing attention, research on combining those two partners together appears to be sporadic [1]. In [9], imprecise probability measures represented by interval probabilities have been introduced into the probabilistic relational databases. As we know, fuzzy descriptions of the real world conform to the law

Corresponding Author: Li Yan is with the School of Software, Northeastern University, 3-11 Wenhua Rd., Shenyang, China, 110819.
E-mail: yanl@swc.neu.edu.cn
Z. M. Ma is with the College of Information Science and Engineering, Northeastern University, 3-11 Wenhua Rd., Shenyang, China, 110819.
E-mail: zongmin_ma@yahoo.com
Manuscript received 14 April 2011; revised 25 Feb. 2012; accepted 27 March 2013.
of human understanding and fuzzy values contain more informative semantics than interval values. It is a natural way to use fuzzy values representing probability measures which are imperfectly known. In [4], a deductive probabilistic and fuzzy object-oriented database language was proposed, where a class property can contain fuzzy set values, and uncertain class membership and property applicability are measured by lower and upper bounds on probability.

This paper focuses on the uncertainty of fuzzy probability in relational databases. Instead of crisp probability measures of tuples in [7] and interval probability measures of tuples in [9], we use possibility distributions to represent imprecise probabilities. Fuzzy probabilistic relational model is hence introduced. Based on the proposed fuzzy probabilistic relational model, we discuss semantic relationship between tuples, and develop the strategies and approaches for eliminating data redundancies. The primitive operations as well as natural join operation in fuzzy probabilistic relational algebra are developed in this paper. It will be shown that the relational operations are consistent and can be reduced into the corresponding conventional operations when it is used in the classical relational databases.

The remainder of this paper is organized as follows. Section 2 provides basic knowledge on relational databases and fuzzy sets/possibility distributions. In Section 3, we introduce the fuzzy probabilistic relational model and discuss how to eliminate the data redundancy. Section 4 defines some relational operations in the extended relational model. The consistencies of the relational operations are illustrated in Section 5. In Section 6, we present related work. Section 7 concludes this paper.

2. Basic Knowledge

A. Relational Databases

Relational scheme is a set of attributes with the form of \( R = (A_1, A_2, \ldots, A_n) \), where \( A_i \) is an attribute. For any attribute \( A_i \in R \) \((i = 1, 2, \ldots, n)\), there exists a domain, denoted by \( D_i \) or Dom (\( A_i \)). Therefore, relational schema can also be represented as \( R = (D_1/A_1, D_2/A_2, \ldots, D_n/A_n) \).

The instances of \( R \) expressed as \( t = \{t_1, t_2, \ldots, t_m\} \). A tuple \( t \) can be expressed as \( t = <v_1, v_2, \ldots, v_p> \), where \( v_i \in D_i \) \((1 \leq i \leq n)\), i.e., \( t \in D_1 \times D_2 \times \ldots \times D_n \). Therefore, \( r \) is a subset of Cartesian product of attribute domains, i.e., \( r \subseteq D_1 \times D_2 \times \ldots \times D_n \). Viewed from the content of a relation, a relation is a simple table, where tuples are its rows and attributes are its columns. The value of a tuple \( t \) on attribute(s) \( S \) is generally written \( t[S] \), where \( S \subseteq R \). Note that \( t(R) \) denotes that \( t \) is a tuple of the relational instance over schema \( R \).

If the value(s) of attribute(s) in a relation can solely identify a tuple from other tuples, the attribute(s) will be called the super key of the relation. If any true subset of a super key is not a super key of the relation, this super key is called candidate key or shortly key. A relation may have several candidate keys. We choose one candidate as the primary key, and other candidates are called alternate key. It is clear that the values of primary key of all tuples in a relation are different and are not null. The attributes included in a candidate key are called prime attributes and the others called non-prime attributes. If an attribute or a set of attributes in relation \( r \) is not a key of \( r \) but a key of another relation \( s \), the attribute(s) is called the foreign key of \( r \).

B. Fuzzy Sets and Possibility Distribution

Fuzzy data are applied for representing one kind of vague concept such as tall, heavy and young etc. Fuzzy data is originally described as fuzzy set by Zadeh [22]. Let \( U \) be a universe of discourse, then a fuzzy value on \( U \) is characterized by a fuzzy set \( F \) in \( U \). A membership function \( \mu_F : U \rightarrow [0, 1] \) is defined for the fuzzy set \( F \), where \( \mu_F(u) \), for each \( u \in U \), denotes the degree of membership of \( u \) in the fuzzy set \( F \). Thus the fuzzy set \( F \) is described as \( F = \{ \mu_F(u_1)/u_1, \mu_F(u_2)/u_2, \ldots, \mu_F(u_n)/u_n \} \).

When the \( \mu_F(u) \) above is explained as a measure of the possibility that a variable \( X \) has the value \( u \) in this approach, in which \( X \) takes values in \( U \), a fuzzy value is described by a possibility distribution [24] \( \pi_X = \{ \pi_X(u_1)/u_1, \pi_X(u_2)/u_2, \ldots, \pi_X(u_n)/u_n \} \), where \( \pi_X(u_i) \) \((0 \leq i \leq n)\) denotes the possibility that \( X \) takes value \( u_i \). The set of the elements in \( \pi_X \) which possibilities are non-zero is called the support of \( \pi_X \), denoted by \( \text{supp}(\pi_X) = \{ u | u \in U \text{ and } \pi_X(u) > 0 \} \).

The extension principle introduced by Zadeh [23] is regarded as one of the most basic ideas of fuzzy set theory. By providing a general method, the extension principle has been extensively employed to extend nonfuzzy mathematical concepts. The idea is to induce a fuzzy set from a number of given fuzzy sets using a mapping. Zadeh’s extension principle can be sometimes referred as maximum-minimum principle. Let \( \pi_A \) and \( \pi_B \) be two fuzzy data based on possibility distribution on the universe of discourse \( U = \{ u_1, u_2, \ldots, u_n \} \), and \( \pi_A \) and \( \pi_B \) are respectively represented by \( \pi_A = \{ \pi_A(p_i)/p_i | p_i \in U \land 1 \leq i \leq n \} \) and \( \pi_B = \{ \pi_B(q_i)/q_i | q_i \in U \land 1 \leq j \leq n \} \). Following the Zadeh’s extension principle, the operation with an infix operator \( \Theta \) on \( \pi_A \) and \( \pi_B \) can be defined as \( \pi_A \Theta \pi_B = \{ \pi_A(p_i)/p_i | p_i \in U \land 1 \leq i \leq n \} \) \( \Theta \{ \pi_B(q_i)/q_i | q_i \in U \land 1 \leq j \leq n \} = \{ \max (\min (\pi_A(p_i), \pi_B(q_i))/p_i, \theta q_i)) | p_i, q_i \in U \land 1 \leq i, j \leq n \} \).
3. Fuzzy Probabilistic Relational Model

A. Model Structure

Probabilistic data has been introduced into relational databases, in which probability distributions can be attribute values [2] or tuples are associated with probability measures [7, 9, 16, 26]. For the former, there exists a problem of implementing NF² model. Since the former can be converted into the latter [7], here we concentrate on such a probabilistic relational database that its attribute values are crisp atomic ones and its tuples are associated with probability measures. In [7], a probabilistic relational model is introduced, in which each tuple is associated with a crisp probability measure, indicating the probability that a stochastic event in the real world has the state described by the tuple. In other words, it is the joint probability of the given realizations of all the attributes taken together. For this purpose, an additional attribute, written pS, is led to the probabilistic relational scheme. It is clear that Dom (pS) = [0, 1], which is a set of crisp atomic values. In order to represent imprecise probability measures of tuples, i.e., interval probability measures, two additional attributes, written LB and UB, are led to the probabilistic relational scheme [9]. These two attributes are used to represent the lower boundary and upper boundary of probability measures of tuples. It is clear that Dom (LB) = [0, 1] and Dom (UB) = [0, 1].

To represent fuzzy probability measures of tuples based on possibility distributions, it is necessary to introduce an additional attribute into the relational schema. The domain of this attribute, called fuzzy probabilistic attribute, is a fuzzy set or a set of fuzzy subset.

Definition 1: Let R (A₁, A₂, ..., Aₙ) be a fuzzy probabilistic relational schema, in which Aₙ₊₁ ∈ R is the fuzzy probabilistic attribute. The fuzzy probabilistic attribute is written as fpD in the paper. A fuzzy probabilistic relation instance r over R is a subset of the Cartesian product of D₁ × D₂ × ... × Dₙ × Dₙ₊₁.

Example 1: Table 1 and Table 1 show two fuzzy probabilistic relational instances.

<table>
<thead>
<tr>
<th>Card</th>
<th>Dept</th>
<th>Age</th>
<th>fpD</th>
</tr>
</thead>
<tbody>
<tr>
<td>9106</td>
<td>CS</td>
<td>20</td>
<td>(1.0/0.3)</td>
</tr>
<tr>
<td>9106</td>
<td>CS</td>
<td>30</td>
<td>(0.5/0.2, 0.4/0.3)</td>
</tr>
<tr>
<td>9711</td>
<td>ME</td>
<td>33</td>
<td>(0.4/0.2, 0.5/0.3)</td>
</tr>
<tr>
<td>9711</td>
<td>ME</td>
<td>38</td>
<td>(0.3/0.1, 0.4/0.4)</td>
</tr>
</tbody>
</table>

Table 2. Fuzzy probabilistic relation s.

<table>
<thead>
<tr>
<th>Card</th>
<th>Dept</th>
<th>Age</th>
<th>fpD</th>
</tr>
</thead>
<tbody>
<tr>
<td>9106</td>
<td>CS</td>
<td>20</td>
<td>(0.3/0.2, 0.5/0.3)</td>
</tr>
<tr>
<td>9106</td>
<td>CS</td>
<td>40</td>
<td>(0.4/0.1, 0.5/0.2)</td>
</tr>
<tr>
<td>9711</td>
<td>ME</td>
<td>32</td>
<td>(0.3/0.1, 0.4/0.3)</td>
</tr>
<tr>
<td>9711</td>
<td>ME</td>
<td>38</td>
<td>(0.4/0.2, 0.5/0.3)</td>
</tr>
</tbody>
</table>

Definition 2: A fuzzy probability measure assigned to the tuple t, denoted by t [fpD] = {π₁ (p₁)/p₁, π₂ (p₂)/p₂, ..., πₙ (pₙ)/pₙ}, means that the joint probability of all the attribute values in t is fuzzy, i.e. t’s probability measure may be p₁ (1 ≤ i ≤ n) with possibility degree πᵢ (pᵢ), where πᵢ (pᵢ) ∈ [0, 1] and pᵢ ∈ [0, 1].

Example 2: Let t be a tuple of the fuzzy probabilistic relation r (R) and t [fpD] = {0.3/0.3, 0.6/0.4, 0.2/0.1}. Then the joint probability of the values of non-fpD attribute in t is fuzzy, represented by possibility distribution {0.3/0.3, 0.6/0.4, 0.2/0.1}, where its probability measure 0.3, 0.4, or 0.1 may be a true value with possibility 0.3, 0.6, or 0.2, respectively.

As is the situation in [7], each tuple in the fuzzy probabilistic relational databases cannot stand for a unique object in the real world. Associated with a stochastic event, there may be several tuples which represent the complete joint distribution of its attributes. Therefore, the notion of the primary key should only be the unique identifier of the event, not being a unique identifier of tuples. In addition, since the probability measures of tuples in the fuzzy probabilistic relation are possibility distributions, the probabilistic constraint that the sum of probability measures of all tuples with the same primary key values is less than or equal to one [7] should be adjusted. The new constraint for the fuzzy probabilistic relational databases, called fuzzy probabilistic constraint, is that for all tuples with the same primary key values in the relation, the sum of their maximum joint probabilities must be less than or equal to one. Here, the maximum probability of tuple tᵢ is represented by the maximum of the support of tᵢ [fpD].

Let t₁, t₂, ..., tₖ be the tuples with the same primary key values in a fuzzy probabilistic relation. Let tᵢ [fpD] = {π₁ (p₁)/p₁, π₂ (p₂)/p₂, ..., πₙ (pₙ)/pₙ} (1 ≤ i ≤ n). Then we have max (supp (t₁ [fpD])) + max (supp (t₂ [fpD])) + ... + max (supp (tₖ [fpD])) ≤ 1.

B. Data Redundancies and Removal

In [7], two operations Plus-operation and Max-operation are defined to merge two equivalent tuples of the probabilistic relational databases according to different relational operation requirements. Let u and v be two equivalent tuples of the probabilistic relational databases r (R) in [7]. Their probability measures are crisp and denoted by u [pS] and v [pS], respectively. For any attribute A of R and A ≠ pS, u and v are considered equivalent, written u ≈ v. Assume merging u and v results in a new tuple t. Then we have t ≈ u ≈ v and

- t [pS] = min {tₙ + v = u [pS] + v [pS]} when Plus-operation is utilized, and
- t [pS] = max {u [pS], v [pS]} when Max-operation is utilized.
Two tuples \( u \) and \( v \) in the fuzzy probabilistic relation \( r (R) \) are considered equivalent if and only if \((\forall A) (A \in R \land A \not\equiv fpD \Rightarrow u \uparrow [A] = v \uparrow [A])\), written \( u \equiv v \).

Equivalent tuples must be coalesced in the fuzzy probabilistic relational databases. Note that in the fuzzy probabilistic relational model, the probability measures of tuples may be fuzzy. While merging equivalent tuples in the fuzzy probabilistic relational databases, the fuzzy probability measures of the result tuples must be calculated by using the extension principle [23] according to the fuzzy probabilistic constraint.

Let \( u \) and \( v \) be two equivalent tuples in the fuzzy probabilistic relation \( r (R) \) (i.e., \( u \equiv v \)), and \( u \equiv fpD = \{u_0 (p_1)/p_1, u_0 (p_2)/p_2, \ldots, u_0 (p_m)/p_m\} \) and \( v \equiv fpD = \{v_0 (q_1)/q_1, v_0 (q_2)/q_2, \ldots, v_0 (q_p)/q_p\} \). We define four types of coalescence operations, which will be applied in the definitions of projection, difference, join, and union in the probabilistic relational algebra, respectively.

(a) The coalescence-Plus operation. Coalescence-Plus on two equivalent tuples \( u \) and \( v \), denoted by \( \Theta \), is defined as \( t = u \Theta v \), in which \( t \equiv u \approx v \) and
\[
t \equiv fpD = \{\max (\min_{i,j} (u_0 (p_i), v_0 (q_j))) | 1 \leq i \leq m, 1 \leq j \leq n\} \quad (1)
\]
(b) The coalescence-Minus operation. The coalescence-Minus on two equivalent tuples \( u \) and \( v \), denoted by \( \Theta \), is defined as \( t = u \Theta v \), in which \( t \equiv u \approx v \) and
\[
t \equiv fpD = \{\max (\min_{i,j} (u_0 (p_i), v_0 (q_j))) | 1 \leq i \leq m, 1 \leq j \leq n\} \quad (2)
\]
(c) The coalescence-Times operation. The coalescence-Times on two equivalent tuples \( u \) and \( v \), denoted by \( \Theta \), is defined as \( t = u \Theta v \), in which \( t \equiv u \approx v \) and
\[
t \equiv fpD = \{\max (\min_{i,j} (u_0 (p_i), v_0 (q_j))) | 1 \leq i \leq m, 1 \leq j \leq n\} \quad (3)
\]
(d) The coalescence-Max operation. Coalescence-Max on two equivalent tuples \( u \) and \( v \), denoted by \( \Theta \), is defined as \( t = u \Theta v \), in which \( t \equiv u \approx v \) and
\[
t \equiv fpD = \{\max (\min_{i,j} (u_0 (p_i), v_0 (q_j))) | 1 \leq i \leq m, 1 \leq j \leq n\} \quad (4)
\]

Example 3: Let \( u \) and \( v \) be two tuples of a fuzzy probabilistic relation \( r (R) \). Let \( u \equiv fpD = \{0.3/0.2, 0.5/0.3\} \) and \( v \equiv fpD = \{0.4/0.1, 0.5/0.2\} \). Let \( u \equiv v \). Then applying the operations defined above, we have

(a) \( t \equiv fpD = \{\max (\min (0.3, 0.4) \min (1, 0.2 + 0.1), \min (0.3, 0.5) \min (1, 0.2 + 0.2), \min (0.5, 0.4) \min (1, 0.3 + 0.1), \min (0.5, 0.5) \min (1, 0.3 + 0.2)) = \{0.3/0.3, 0.4/0.4, 0.5/0.5\}\}
\]
(b) \( t \equiv fpD = \{\max (0.3/0.3, 0.4/0.4, 0.5/0.5)\} \quad (1)
\]

4. Fuzzy Probabilistic Relational Operations

Relational operations are classified into two basic types: conventional set operations and special relational operations. The former includes union, intersection, difference and Cartesian product, and the later includes projection, join, selection and division. These operations can also be classified as primitive operations (including union, difference, selection, projection, and Cartesian product) and additional operations (including intersection, join, and division), in which the additional operations can be defined with the primitive ones [19]. Therefore, in the following, we only concentrate on the primitive operations of the fuzzy probabilistic relational databases. Moreover, the definition of the natural join operation for the fuzzy probabilistic relations is given because it is useful for the retrieval of relational data.

Just like the traditional relational operations, the set operations in the fuzzy probabilistic relations require the argument relations are union-compatible. Two fuzzy probabilistic relations, say \( r (A_1, A_2, \ldots, A_n) \) and \( s (B_1, B_2, \ldots, B_n) \), are considered union compatible if they have the same number of attributes in their relational schemas and the domain of \( A_i \) is the same as the domain of \( B_i (1 \leq i \leq n) \). Here \( A_i \) and \( B_i \) can be the attribute \( fpD \).

A. Union

Let \( r (R) \) and \( s (R) \) be two union-compatible fuzzy probabilistic relations. The union of \( r \) and \( s \) produces a new fuzzy probabilistic relation, which contains three types of tuples as follows.

- The first type of tuples is the tuples from \( r \), which are not equivalent to any tuples in \( s \).
- The second type of tuples is the tuples from \( s \), which are not equivalent to any tuples in \( r \).
- The third type of tuples is the result tuples after coalescing the equivalent tuples in \( r \) and \( s \) by using the coalescence-Max operation.

Based on the discussion above, we define the union operation as follows.
\[
r \cup s = \{t (r) \mid (\forall v) (v \in s \land t \in r \land t \not\equiv v) \land (\forall u) (u \in s \land \exists v \in t \not\equiv v) \land (\exists v) (u \in r \land v \in s \land u \equiv v \cup \Theta v)\} \quad (5)
\]
Here the result relation $r \cup s$ is a fuzzy probabilistic relation on the schema $R$, which is the same as the schema of $r$ and $s$. One tuple of $r \cup s$, say $t$, is one of the two types of tuples mentioned above.

**Example 4:** Let us look at the fuzzy probabilistic relation $r \cup s$ in Table 3, which is the union of $r$ in Table 1 and $s$ in Table 2.

In Table 3, the last tuple of $r \cup s$, say $t$, is the result of merging the last tuple of $r$, say $u$, and the last tuple of $s$, say $v$, by using the coalescence-Max operation ($u \approx v$).

$$t [fpD] = \{ \max (0.3, 0.4)/(0.2, 0.2), \min (0.4, 0.4)/(0.4, 0.2), \min (0.4, 0.3), 0.4/0.4 \}$$

Similarly, the first tuple of $r \cup s$ is produced through the merging of the first tuple of $r$ and the first tuple of $s$ with the coalescence-Min operation. The second and third tuples of $r \cup s$ are directly from $r$ because they are not equivalent to any tuples in $s$. The fourth and fifth tuples of $r \cup s$ are directly from $s$ and they are not equivalent to any tuples in $r$.

### B. Difference

Let $r (R)$ and $s (R)$ be two union-compatible fuzzy probabilistic relations. The difference of $r$ and $s$ results in a fuzzy probabilistic relation, which contains two types of tuples as follows.

- The first type of tuples is the tuples from $r$, which is not equivalent to any tuples in $s$.
- The second type of tuples is the result tuples after coalescing the equivalent tuples in $r$ and $s$ by using the coalescence-Min operation.

We thus define the difference operation as follows.

$$r - s = \{ t (R) \mid (\forall v) (v \in s \land t \in r \land t \not= v) \lor (\exists u) (u \in r \land v \in s \land x \cup y = x \Theta y) \}$$

Here the result relation $r - s$ is a fuzzy probabilistic relation on the schema $R$, which is the same as the schema of $r$ and $s$. One tuple of $r - s$, say $t$, is one of the two types of tuples mentioned above.

**Example 5:** The result relation of the difference operation in the fuzzy probabilistic relations in Table 1 and Table 2 is shown in Table 4.

In Table 4, the last tuple of $r - s$, say $t$, is the result of merging the last tuple of $r$, say $u$, and the last tuple of $s$, say $v$, by using the coalescence-Max operation ($u \approx v$).

### C. Projection

Projection operation provides a view relation on the specified attribute(s). Let $r (R)$ be a fuzzy probabilistic relation on the schema $R$ and attribute subset $S \subset R (fpD$ attribute can be included in $S)$. The projection of $r (R)$ on $S$ results in a fuzzy probabilistic relation on the schema $S$. For each tuple in $r$, say $u$, we first get $u [S]$ and then we use the coalescence-Plus operation to merge all tuples like $u [S]$. The final result tuples become the tuples of the projection of $r (R)$ on $S$.

We define the projection operation as follows.

$$\Pi_S (r) = \{ t (S) \mid (\forall u) (u \in r \land u [S] \approx t (S) \land t (S) = \Theta u [S]) \}$$

Here the result relation $\Pi_S (r)$ is a fuzzy probabilistic relation on the schema $S$, where $S \subset R$.

**Example 6:** Consider the fuzzy probabilistic relation $r$ shown in Table 5. The projection operation of $r$ on attribute subset $S = \{ \text{Card}, \text{First Name}, \text{Dept}, \text{Nationality}, \text{Office}, \text{fpD} \}$ is shown in Table 6.
In Table 6, the last tuple of \( \Pi_S (r) \), say \( t \), comes from the last two tuples of \( r \), say \( u \) and \( v \). But \( t \) is the result of merging \( u [S] \) and \( v [S] \) by using the coalescence-Plus operation. Here
\[
t \{fpD\} = \{ \max (\min (0.3, 0.4)/\min (1, 0.2 + 0.3), \min (0.3, 0.5)/\min (1, 0.2 + 0.3), \min (0.5, 0.4)/\min (1, 0.2 + 0.3), \min (0.5, 0.4)/\min (1, 0.2 + 0.3), \min (0.5, 0.4)/\min (1, 0.2 + 0.3), \min (0.5, 0.4)/\min (1, 0.2 + 0.3), \min (0.5, 0.4)/\min (1, 0.2 + 0.3), \min (0.5, 0.4)/\min (1, 0.2 + 0.3) \}
\]
The other tuples of \( \Pi_S (r) \) are the tuples which come from \( r \), but only consist of the values of the attributes in \( S \). No coalescence-Plus operation is used to merge them.

### D. Cartesian Product

Let \( r (R) \) and \( s (S) \) be two fuzzy probabilistic relations, \( R' = R - \{fpD\} \), and \( S' = S - \{fpD\} \). The Cartesian product of relation \( r \) and \( s \) is defined as follows.
\[
r \times s = \{ t (R' \cup S' \cup \{fpD\}) | (\exists u) (\exists v) (u \in r \land v \in s \\
\land t [R'] = u [R'] \land t [S'] = v [S'] \land t [fpD] = u [fpD] \otimes v [fpD] \}
\]

### E. Selection

Let \( r (R) \) be a fuzzy probabilistic relation and \( P \) be a predicate denoting selection condition. A predicate is formed through combining the basic clause \( X \theta Y \) as operands with operations \( \neg, \land, \lor \), where \( \theta \in \{>, <, =, \neq, \geq, \leq\} \). Here \( X \) is an attribute, and \( Y \) is a constant or an attribute. Where \( \theta \) is the domain of attribute \( \{fpD\} \) can appear in the conditions. The selection on \( r \) for \( P \) returns a fuzzy probabilistic relation containing a subset of the tuples of \( r \), which satisfies the selection condition represented by \( P \).

The selection is defined as follows,
\[
\sigma_P (r) = \{ t (R) | t (R) \in r \land P (t (R)) \}
\]

### F. Natural Join

Let \( r (R) \) and \( s (S) \) be two fuzzy probabilistic relations. Let \( R' = R - \{fpD\} \), \( S' = S - \{fpD\} \), and \( Q' = R' \cap S' \). It is clear that \( Q' \) is a set of the attributes which are common in \( R' \) and \( S \). The natural join of \( r \) and \( s \) is to take any couple of tuples from \( r \) and \( s \) (say \( u \) and \( v \)), respectively, in which \( u [Q'] = v [Q'] \). And then \( u \) and \( v \) are connected together to form a single tuple, say \( t \). The result fuzzy probabilistic relation is hereby produced after the natural join of \( r \) and \( s \). It is not difficult to see that tuple \( t \) consists of three parts.

- The first part is the values of \( u \) on the attributes \( R' - Q' \), i.e., \( u [R' - Q'] \).
- The second part is the values of \( v \) on the attributes \( S' - Q' \), i.e., \( v [R' - Q'] \).
- The third part is the result after coalescing \( u [Q' \cup \{fpD\}] \) and \( v [Q' \cup \{fpD\}] \) by using the coalescence-Times operation. Here \( u [Q' \cup \{fpD\}] \approx v [Q' \cup \{fpD\}] \).

The Natural join of relation \( r \) and \( s \) is defined as follows.
\[
r \bowtie s = \{ t ((R' - Q') \cup S) | (\exists u) (\exists v) (u \in r \land v \in s \\
\land t [R' - Q'] = u [R' - Q'] \land t [S' - Q'] = v [S' - Q'] \land t [Q' \cup \{fpD\}] = u [Q' \cup \{fpD\}] \otimes v [Q' \cup \{fpD\}] \}
\]

#### Example 7: The result relation of the natural join of fuzzy probabilistic relations in Tables 7 and 8 is shown in Table 9.

### Table 6. Projection operation \( \Pi_{\{\text{Card, First Name, Dept, Nationality, \{fpD\}\}} (r) \).

<table>
<thead>
<tr>
<th>Card</th>
<th>First Name</th>
<th>Dept</th>
<th>Nationality</th>
<th>{fpD}</th>
</tr>
</thead>
<tbody>
<tr>
<td>9106</td>
<td>Mary</td>
<td>CS</td>
<td>America</td>
<td>{0.3/0.2, 0.4/0.3, 0.5/0.5}</td>
</tr>
<tr>
<td>9106</td>
<td>Mary</td>
<td>CS</td>
<td>Canada</td>
<td>{0.4/0.3, 0.5/0.4}</td>
</tr>
<tr>
<td>9609</td>
<td>Tom</td>
<td>ME</td>
<td>Germany</td>
<td>{1.0/0.4}</td>
</tr>
<tr>
<td>9609</td>
<td>Tom</td>
<td>ME</td>
<td>England</td>
<td>{1.0/0.6}</td>
</tr>
<tr>
<td>9705</td>
<td>John</td>
<td>IS</td>
<td>Italy</td>
<td>{0.3/0.1, 0.6/0.3}</td>
</tr>
<tr>
<td>9705</td>
<td>John</td>
<td>IS</td>
<td>France</td>
<td>{0.3/0.4, 0.4/0.5, 0.5/0.6, 0.2/0.7}</td>
</tr>
</tbody>
</table>

### Table 7. Fuzzy probabilistic relation \( r \).

<table>
<thead>
<tr>
<th>First Name</th>
<th>Dept</th>
<th>Age</th>
<th>{fpD}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>CS</td>
<td>20</td>
<td>{0.4/0.2, 0.6/0.3}</td>
</tr>
<tr>
<td>Mary</td>
<td>CS</td>
<td>30</td>
<td>{0.3/0.1, 0.5/0.4}</td>
</tr>
<tr>
<td>John</td>
<td>IS</td>
<td>33</td>
<td>{0.4/0.3, 0.6/0.4}</td>
</tr>
<tr>
<td>John</td>
<td>IS</td>
<td>38</td>
<td>{1.0/0.3}</td>
</tr>
</tbody>
</table>

### Table 8. Fuzzy probabilistic relation \( s \).

<table>
<thead>
<tr>
<th>First Name</th>
<th>Dept</th>
<th>Degree</th>
<th>{fpD}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>CS</td>
<td>M. Phil</td>
<td>{0.3/0.25, 0.5/0.3}</td>
</tr>
<tr>
<td>Mary</td>
<td>CS</td>
<td>Ph. D.</td>
<td>{0.4/0.1, 0.4/0.35}</td>
</tr>
<tr>
<td>John</td>
<td>IS</td>
<td>M. Phil</td>
<td>{0.4/0.2, 0.6/0.3}</td>
</tr>
<tr>
<td>John</td>
<td>IS</td>
<td>Ph. D.</td>
<td>{0.3/0.15, 0.5/0.35}</td>
</tr>
</tbody>
</table>

### Table 9. Natural join operation \( r \bowtie s \).

<table>
<thead>
<tr>
<th>Card</th>
<th>First Name</th>
<th>Dept</th>
<th>Degree</th>
<th>Age</th>
<th>{fpD}</th>
</tr>
</thead>
<tbody>
<tr>
<td>9106</td>
<td>Mary</td>
<td>CS</td>
<td>M. Phil</td>
<td>20</td>
<td>{0.3/0.05, 0.4/0.06, 0.3/0.075, 0.5/0.09}</td>
</tr>
<tr>
<td>9106</td>
<td>Mary</td>
<td>CS</td>
<td>Ph. D.</td>
<td>20</td>
<td>{0.4/0.02, 0.4/0.03, 0.4/0.07, 0.4/0.105}</td>
</tr>
<tr>
<td>9106</td>
<td>Mary</td>
<td>CS</td>
<td>M. Phil</td>
<td>30</td>
<td>{0.3/0.025, 0.3/0.03, 0.3/0.1, 0.5/0.12}</td>
</tr>
<tr>
<td>9106</td>
<td>Mary</td>
<td>CS</td>
<td>Ph. D.</td>
<td>30</td>
<td>{0.3/0.01, 0.3/0.035, 0.4/0.04, 0.4/0.14}</td>
</tr>
</tbody>
</table>

In Table 9, the first tuple of \( r \bowtie s \), \( t \), is the result of merging the first tuple of \( r \), say \( u \), and the first tuple of \( s \), say \( v \), by using the coalescence-Times operation. Here
\[
t \{fpD\} = \{ \max (\min (0.4, 0.3)/(0.2 \times 0.25), \min (0.4, 0.5)/(0.2 \times 0.3), \min (0.6, 0.3)/(0.3 \times 0.25), \min (0.6, 0.5)/(0.3 \times 0.3) \}
\]
tuples of $r$ and the second tuple of $s$, both with the coalescence-Times operation.

5. Semantic Consistencies and Properties of Fuzzy Probabilistic Relational Algebra

The fuzzy probabilistic relational operations introduced in the paper satisfies the fuzzy probabilistic constraint, namely, the sum of all maximum joint probabilities of the tuples with the same primary key value must be less than or equal to one. We demonstrate the extended union and other operations can be illustrated in the similar manner. It is reasonable to assume that two union-compatible original relations satisfy the fuzzy probabilistic constraint. Following the definition of the extended union, the result relation contains three types of tuples. The first two types of tuples are directly from the relations participating in union, respectively. The last type of tuples is the result tuples after the equivalent tuples in the relations participating in the union are coalesced with the coalescence-Max operation. The definition of this coalescence operation has ensured the fuzzy probabilistic constraint. Note that the extended union may cause semantic inconsistency. That is the union operation may violate the probabilistic constraints. The reason why such a situation arises is related with gaining the probability distribution of concerned data, not being the problem of the union operation. In fact, the similar problem also exists in the traditional union, in which two tuples belonging to two relations respectively have the same primary key values but different non-primary key values. The strategy resolving this problem is that the extended fuzzy operations in the relational algebra have the same properties as those of classical set operations. Let $r$, $s$, and $u$ be three union-compatible fuzzy relations. Then

- $r \cup s = s \cup r$ and $r \cap s = s \cap r$ (commutativity)
- $r \cup r = r$ and $r \cap r = r$ (idempotence)
- $r \cap (r \cup s) = r \cup (r \cap s)$ (absorption)
- $(r \cup s) \cup u = r \cup (s \cup u)$ and $(r \cap s) \cap u = r \cap (s \cap u)$ (associativity)
- $r \cup (s \cup u) = (r \cap s) \cup (r \cap s)$ and $r \cap (s \cup u) = (r \cup s) \cap (r \cap s)$ (distributivity)
- $r \cup s = r \cup (s - r)$ and $r \cap s = r - (r - s)$ (associativity)

The following properties are also held in the fuzzy probabilistic operations. Let $r$ ($R$), $s$ ($R$), and $u$ ($Q$) be three fuzzy probabilistic relations. Let $P$ be a selection predicate involving attributes of $R$. Then

- $u \bowtie (r \cup s) = (u \bowtie r) \cup (u \bowtie s)$ and $u \bowtie (r \cap s) = (u \bowtie r) - (u \bowtie s)$
- $\sigma_P (r \cup s) = \sigma_P (r) \cup \sigma_P (s)$ and $\sigma_P (r \cap s) = \sigma_P (r) - \sigma_P (s)$
- $\sigma_P (u \bowtie r) = u \bowtie \sigma_P (r)$

These properties can be proven by the definitions of the extended fuzzy operations in the relational algebra.

Additionally, the fuzzy probabilistic relational operations can be reduced to the corresponding conventional
operations when they are used in the traditional relational databases. It can be seen from Section 4 that the differences between the fuzzy probabilistic relational operations and traditional relational operations are the processing of equivalent tuples because of the attribute \( fpD \). Under certain information environment, the \( fpD \) attribute values of all tuples in the relation are a crisp 1.0 and there are no multiple tuples with the same primary key value. Four kinds of coalescence operations become only one operation of eliminating data duplicate at this point. The fuzzy probabilistic relational operations are reducible to the traditional relational operations.

6. Related Work

In order to deal with database’s impreciseness and information uncertainty which exist in many real-world applications, fuzzy set and probability theories have been used to extend relational database models, and this has resulted in the fuzzy relational databases and probabilistic relational databases. The fuzzy probabilistic relational database model presented in the paper is closely related to the fuzzy relational databases and probabilistic relational databases.

Various fuzzy relational database models based on similarity relations [3], possibility distributions [17] and fuzzy sets [18] have been proposed. The fuzzy relational databases based on possibility distribution can further be classified into two categories: tuples associated with possibilities and attribute values represented by possibility distributions. Based on the proposed fuzzy relational database models, some major issues such as semantic measures and data redundancies, query and data processing and data dependencies and normalizations have been investigated [18, 20, 21, 10, 6]. More recently, conceptual design of fuzzy relational databases using fuzzy UML data model was developed in [14] and mapping of fuzzy XML into fuzzy relational databases was proposed in [11]. For a comprehensive review of what has been done in the fuzzy relational databases, please refer to [12, 13].

In addition to the fuzzy relational databases, there are also some studies to represent and manipulate probabilistic data in the context of relational databases. In [5], an extended relational model was proposed using the well-known probability calculus, in which a probability measure with every tuple was assigned for indicating the joint probability of all the attribute values in that tuple. The projection and join operations for their relational structure were also discussed. In [2], an extension of the relational model is proposed using probability theory. They adopted a \( \text{NF}^2 \) view of probabilistic relations, namely, probability distribution as attribute value. The projection, selection, and join operations were redefined using semantics of probability theory. For the problem of implementing \( \text{NF}^2 \) model, Dey and Sarkar [7] proposed a probabilistic relational model, in which each tuple was stamped with associated probability measure, and refined the extended relational algebra. Not being the same as the model in [5], the total probability assigned to all the tuples with the same key value may be less than or equal to one in [7]. In addition, the extended probabilistic relational databases in [16, 26] also applied the same probabilistic relational data model as that in [7]. In [9], a comprehensive discussion of the representation and manipulation of probabilistic data was presented in relational databases, in which the attribute values of tuples may be probabilistic distributions. As a result, the probabilistic relations are not in first normal form and they are not directly amenable for algebraic manipulation. So, in [9], the path-annotated relations were introduced and the transformations from the probabilistic relations to the path-annotated relations were achieved. The probabilistic operations based on the path-annotated relations were thus defined also. Note that the imprecise probabilities (i.e., interval probabilities) were used in [9], where two additional attributes are needed to represent the minimum and maximum probabilities of tuples. In addition, an attribute for representing the paths of tuples is also needed. The research and development of probabilistic databases are recently receiving increasing attention, which focus mainly on querying probabilistic relational databases and probabilistic XML. For some more recent results, one can refer to [8].

7. Conclusions

In this paper, a fuzzy probabilistic relational model was introduced, in which possibility distributions arise at the level of tuples as probability measures. Such an extended relational model can be applied for modeling stochastic events which probabilities are represented by possibility distributions. Based on the extended relational model and the fuzzy probabilistic constraint, we discussed semantic relationship between tuples. The strategies and approaches for eliminating data redundancies were thus investigated. The primitive operations as well as natural join operation in the fuzzy probabilistic relational algebra were developed in this paper. Finally the semantic consistencies of fuzzy probabilistic relational operations were discussed.

For modeling real-world problems and constructing intelligent systems, integration of different methodologies and techniques has been the quest and focus of significant interdisciplinary research efforts. The advantages of such a hybrid system are that the strengths of its partners are combined and the weaknesses of its partners are complementary one another. The fuzzy probabilistic
relational databases provide a kind of flexible relational model to represent and process hybrid imprecise and uncertain information. In the near future, we will devote to implement the operations proposed in the paper to manage fuzzy and probabilistic sensor data. Basically we have two strategies on the implementation of the proposed operations: one is based on probabilistic relational database management, which will be enhanced with fuzzy data processing; another is based on fuzzy relational database management, which will be enhanced with probabilistic data processing. In addition, it should be noted that hybrid imprecise and uncertain information generally has multiple and complicated characteristics. In the fuzzy multidatabase systems, for example, the integrated databases are consisted of such tuples that are connected with probability measures and which attribute values are fuzzy [25]. Future work will be on the relational databases with other hybrid imprecise and uncertain information. In particular, a fuzzy and probabilistic relational model, in which the attribute values may be fuzzy ones represented by possibility distributions, will be studied in a forthcoming paper.

Acknowledgment

The authors wish to thank the anonymous referees for their valuable comments and suggestions, which improved the technical content and the presentation of the paper. This work was supported by the National Natural Science Foundation of China (60873010 & 61073139) and the Program for New Century Excellent Talents in University (NCET-05-0288).

References


1983.


