The Probabilistic Resource Space Model —Theory and Experiment

Hai Zhuge, Senior Member, IEEE, and **Yunpeng Xing**

Abstract—The development of World Wide Web requires a semantic data model to effectively manage the contents of its heterogeneous resources. This challenges traditional data models, which are mainly for managing structured resources or relatively simple data objects. Classification is one of the most basic methods to organize and manage resources in the real world. The Resource Space Model RSM is a semantic data model for managing the contents of various resources by normalizing the classification semantics. However, uncertainty in our daily life makes it difficult to correctly classify and use resources. Previous probabilistic data models are mainly on the existence of resources or the values of resources' attributes. This paper firstly introduces the basic concepts of RSM by comparing with the relational data model, and then proposes the Probabilistic Resource Space Model P-RSM to manage uncertainty in classification semantics. The uncertain classification semantics of different granularities is proposed to specify and manage resources. Relevant normal forms, operations and integrity constraints are also proposed for dealing with uncertain classification semantics. Experiments show the effectiveness of the proposed P-RSM.

Index Term—Data Model, Classification, Web resource management, Probability.

I. INTRODUCTION

More and more applications require the ability to process uncertain information [11]. To obtain this ability, one way is to incorporate probabilistic techniques into information retrieval approaches [3], the other way is to create a data model that can specify and manage uncertain information.

Traditional data models like classical relational model can only specify and manage deterministic information. Research work has been done on managing uncertain information by extending traditional data models like relational model [1, 8] and XML (eXtensible Markup Language) [5].

Research on modeling relational data falls into two categories depending on whether the model satisfies the first normal form (1NF) of the classical relational model or not. Models satisfying 1NF usually assume that the existence of an object is uncertain and associate probabilities with a whole tuple to indicate this type of uncertainty [7, 12]. Models using non-1NF usually assume that the existence of an object is certain, but the attribute values of an object are uncertain [4, 13]. They associate probabilities with attributes of a tuple. Above two types of probabilistic relational models have limitations. It is difficult for the probabilistic relational models satisfying 1NF to represent the probabilities of attribute values of an object. It could lead to information loss or combinatorial explosion of tuples to specify attribute value probabilities using tuple probabilities. The non-1NF probabilistic relational models are often accompanied by complicated algebras and querying mechanisms. ProbView is an attempt to overcome the two types of limitations [16]. It firstly transforms non-1NF data to its corresponding annotated 1NF patterns, and then all manipulating and querying operations are applied to these corresponding 1NF data. But its limitation is that the transformation from non-1NF to 1NF

is not completely equivalent so that some useful information may be lost during the transformation [18].

Uncertain classification [24] and uncertain relational databases have been studied [6] [21] [22]. Previous probabilistic relational data models are mainly on the existence of a certain entity or the attribute values of a certain entity. Little work has been done on establishing data model based on uncertain classification.

On the other hand, the development of the Web requires a powerful semantic data model to effectively manage the contents of its heterogeneous resources. This challenges traditional data models, which are mainly for managing structured or relatively simple data objects.

Classification is the most basic way to effectively retrieve and efficiently organize information. The Resource Space Model (RSM) is proposed to manage resources based on the normalization of the classification [26, 27]. A resource space is a multi-dimensional classification semantic space. Each dimension of a resource space specifies a type of classification method. The normal forms and operations based on orthogonal classification semantics are also proposed to manage deterministic classification semantics. To establish a powerful semantic data model for the Web, integration and mappings between RSM, OWL (Web Ontology Language, www.w3.org/TR/owl-features/), and database have been studied [28].

Inaccurate information usually leads to uncertain resource classification. How to represent and manage uncertain classification semantics challenges the original RSM. This paper proposes a probabilistic Resource Space Model to effectively and efficiently organize and manage uncertain classification semantics by extending the original RSM.

Relevant works also concern querying from the view of probability and query evaluation on probabilistic database [9,

10, 19]. A system for managing data, accuracy and lineage in an integral manner is introduced [25]. Much work has been done to manage probabilistic data in XML [2, 14, 15, 18]. A framework is proposed to acquire, maintain and query XML documents with incomplete information, in which order in documents and DTDs is ignored [2]. A probabilistic XML approach is proposed to resolve conflicts during data integration, where the order in documents and DTDs plays an important role [17]. A complexity analysis for managing probabilistic XML data is discussed [20]. Modelling uncertainty of ontology has attracted research interest [23].

II. THE RESOURCE SPACE MODEL RSM

A. Basic Concepts

A resource space is an *n*-dimensional space, denoted as $RS(X_1, X_2, \dots, X_n)$ or just by name *RS* in simple. Every point in the space uniquely determines a set of related resources. Axis is defined by a set of coordinates denoted as $X_i = \{C_{i1}, C_{i2}, \dots, C_{im}\}$. A point $p(C_{1,j1}, C_{2,j2}, \dots, C_{n,jn})$ is determined by the coordinate values at all axes. A point can uniquely determine a resource set, where each element is called a resource entry. Point and resource entry are two basic operation units of RSM.
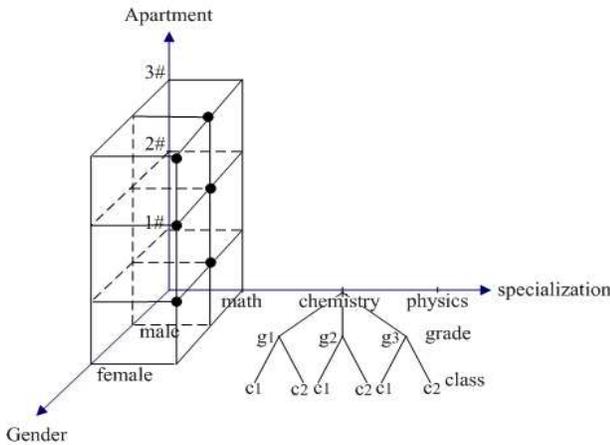

Fig. 1. A 3-dimentional Resource Space *Spec-Apart-Gen*.

Fig. 1 is an example of a 3-dimentional resource space *Spec-Apart-Gen*(*Specialization*, *Apartment*, *Gender*) that specifies student information in a college. Three axes are *Specialization* = {*math*, *chemistry*, *physics*}, *Apartment* = {1#, 2#, 3#} and *Gender* = {*male*, *female*}. Each point denotes a class of students. For example, the point (*math*, 1#, *male*) represents all the male students belonging to the department of math and living in apartment 1 in this college. And each resource entry in this point corresponds to a student of the college.

Each coordinate directly residing at an axis is called the top-level coordinate. For each top-level coordinate, the resource partition hierarchy can be defined top-down. Take Fig. 1 for example, the coordinate *chemistry* on axis *Specialization* is partitioned into g_1, g_2 and g_3 in terms of *grade*, and then they can be further divided according to *class*. Note that in this hierarchy tree, the label of each node is determined by the full path from the root. Thus, the leaf node

'*chemistry.g_1.c_1*' can be distinguished from '*chemistry.g_2.c_1*'. We have indicated the equivalence of the 'flat' resource space and the hierarchical resource space by projecting each leaf node of the hierarchy tree onto the axis where the root resides [26, 27]. So here only discusses the flat resource space.

RSM mainly includes operations on resource spaces and their completeness, normal forms for ensuring the correctness of storing and retrieving resources, relations between the operations and the normal forms, algebra and calculus, expressiveness of query language, search complexity, storage mechanism, and the P2P mechanisms of RSM [26] [27].

B. Resource Space Model and Relational Database Model

The following example shows the characteristics of the RSM. Multi-layer tables provide integrated information of multiple abstraction levels. The higher layers provide more abstract information. The lower layers constitute a fine classification of the higher layer. Fig. 2 is such a multi-layer table on university human resources, which naturally constitutes a classification hierarchy [27].

Fig. 2. Multi-layer table of university human resources.

Since the first normal form of the relational data model requires flat table and atomic values, it is inappropriate to use relational tables to represent such a multi-layer table.

However, it can be naturally converted to the 3-dimensional resource space shown in Fig. 3. The more layers the table has, the more advantages of the RSM exhibit.
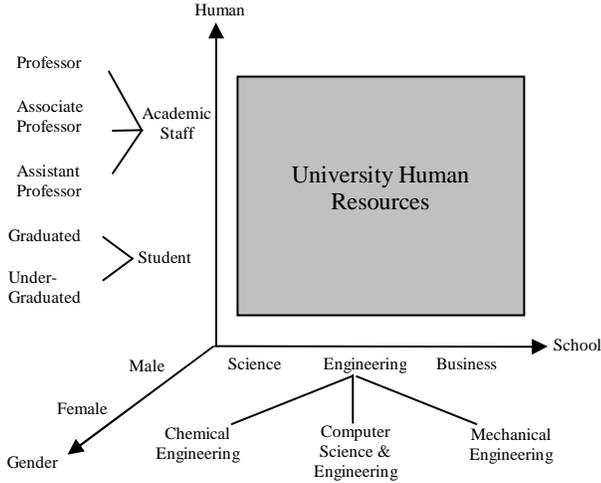

Fig. 3. A resource space for specifying university human resources.

The essential differences between RSM and the relational data model are as follows:

(1) RSM can manage structured or semi-structured resources, while the traditional relational data model only manages the atomic values.
(2) The data model of RSM is a uniform coordinate system, while the relational data model is a flat relational table.
(3) The normalization basis of the RSM is an independent and orthogonal coordinate system, while the normalization basis of the relational data model is the function dependence relationship. Above three differences determine that the RSM concerns the content of resources and supports content-based operation, but the relational model concerns the attribute values of entities being managed and supports attribute-based operation.
(4) The RSM enables a uniform and universal resource view when operating resources, while the relational model essentially supports views of tables. This feature enables the RSM to uniformly share and manage Web resources.

The difference between the existing techniques like the data cube and the RSM has been discussed in chapter 1 of [27]. In the following, we will develop the RSM into a probabilistic RSM.

III. THE PROBABILISTIC RESOURCE SPACE MODEL P-RSM

A. Probabilistic Resource Space

The probabilistic event in the Probabilistic Resource Space Model P-RSM is that a resource belongs to a certain class. $Prob(r \in T)$ denotes the membership probability of resource r belonging to class T. T represents a class of resources of a resource space, an axis, a coordinate, a point, or any of their combination by set operations.

The following are two strategies on how to specify the probabilistic distribution of a given resource r belonging to a resource space RS.

(1) For any resource r, specify its membership probability distribution on every point of RS.
(2) For each axis X of RS, specify the membership probability distribution of any resource r on every coordinate of X.

The second strategy is more feasible and more efficient because of the following reasons:

(1) The number of points in resource space $RS(X_1, X_2, \ldots, X_n)$ is $|X_1| \times |X_2| \times \ldots \times |X_n|$ and the number of coordinates is $|X_1| + |X_2| + \ldots + |X_n|$, where $|X|$ is the number of coordinates on X. The large number of points makes it difficult to specify and manage the membership probability of every resource to every point.
(2) Each axis in a resource space represents a resource classification method. For each point $p(C_{i1}, C_{i2}, \ldots, C_{in})$, C_{ij} is from axis X_j $(1 \leq j \leq n)$, and $R(p) = R(C_{i1}) \cap R(C_{i2}) \cap \ldots \cap R(C_{in})$. Each point is a combination of all axes of RS and involves all classification methods used in the resource space. Thus, specifying the membership probability distribution of a resource belonging to every point requires multiple classification methods simultaneously. It is more feasible for users and automatic classification algorithms to specify the membership probability distribution of a resource belonging to coordinates axis by axis.

Definition 1. The resource space $RS(X_1, X_2, \ldots, X_n)$ is a *probabilistic resource space* if for any resource r and any axis X_i of RS, there exists a membership probability function β_{ri}: $X_i \rightarrow [0, 1]$, such that $\beta_{ri}(C_{ij})$ represents the membership probability of resource r belonging to class $R(C_{ij})$ for any top-level coordinate C_{ij} at X_i; for any coordinate C' and its parent coordinate C on axis X_i, $\beta_{ri}(C')$ represents the membership probability of r belonging to C' under the condition that resource r belonging to coordinate C.
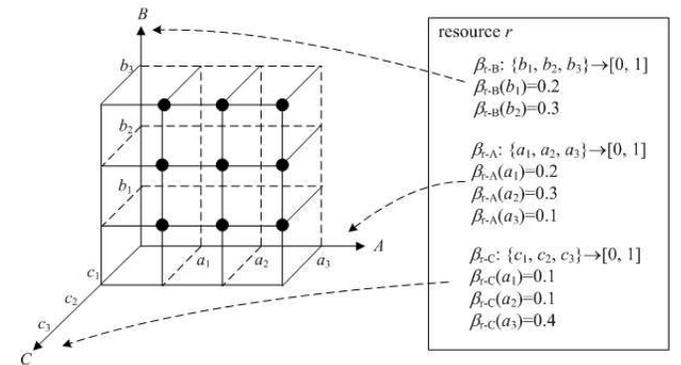

Fig. 4. An example of probabilistic resource space. $\beta_{r\text{-}B}(b_1) = 0.2$ means that the probability of resource r belonging to coordinate b_1 is 0.2 at axis B.

According to above definition, any resource r in a probabilistic resource space $RS(X_1, X_2, \ldots, X_n)$ has n membership probabilistic functions, each of which corresponds to an axis of RS. Take the probabilistic resource space $RS(A, B, C)$ in Fig. 4 for example, resource r has the following three membership probabilistic functions: β_{r-A}: $A \rightarrow [0, 1]$; β_{r-B}: $B \rightarrow [0, 1]$; and β_{r-C}: $C \rightarrow [0, 1]$. Resource r belongs to resource space $RS(X_1, X_2, \ldots, X_n)$ if and only if there exists at least one axis X_i such that the membership probabilistic function of r on X_i has been explicitly specified.

From the definition of the probabilistic resource space, we can specify the membership probability of a resource belonging to each coordinate. Axis and point are the other two resource sets in a resource space, both of which consist of a series of set operations on coordinates. Thus, the membership probability of a resource belonging to an axis or to a point is a complex probabilistic events. Without knowledge about relationships between coordinates, it is very difficult to calculate the membership probability of a resource belonging to an axis or a point from the membership probability of the resource belonging to each coordinate. In the P-RSM, we use an real number interval to specify the possible membership probability of a resource belonging to an axis or a point. From the membership probability on each coordinate, the membership probability on each axis/point in a probabilistic resource space can be evaluated as follows:

(1) For axis $X_i = \{C_{i1}, C_{i2}, \ldots, C_{im}\}$, since $R(X_i) = R(C_{i1}) \cup R(C_{i2}) \cup \ldots \cup R(C_{im})$, the probability of r belonging to X_i falls into the interval $[max\{\beta_{ri}(C_{i1}), \ldots, \beta_{ri}(C_{im})\}, min\{1, \beta_{ri}(C_{i1}) + \beta_{ri}(C_{i2}) + \ldots + \beta_{ri}(C_{im})\}]$.

(2) For point $p(C_{1,j1}, C_{2,j2}, \ldots, C_{n,jn})$, the probability of resource r belonging to p is equal to the probability of resource r simultaneously belonging to $C_{1,j1}, C_{2,j2}, \ldots,$ and $C_{n,jn}$, that is, $Prob(r \in R(p)) = Prob(r \in (R(C_{1,j1}) \cap R(C_{2,j2}) \cap \ldots \cap R(C_{n,jn})))$ holds. For event A and event B, the event that A and B occur simultaneously satisfies $max\{0, Prob(A) + Prob(B) - 1\} \leq Prob(A \wedge B) \leq min\{Prob(A), Prob(B)\}$. Thus, the membership probability of resource r belonging to p falls into the interval $[max\{0, \ldots max\{0, max\{0, Prob(r \in R(C_{1,j1})) + Prob(r \in R(C_{2,j2})) - 1\} + Prob(r \in R(C_{3,j3})) - 1\} \ldots + Prob(r \in R(C_{n,jn})) - 1\}, min\{Prob(r \in R(C_{1,j1})), \ldots, Prob(r \in R(C_{n,jn}))\}]$.

(3) For any coordinate C' and its parent coordinate C on axis X_i, $\beta_{ri}(C')$ is defined as the membership probability of r belonging to C' under the condition that resource r belonging to coordinate C, i.e., $\beta_{ri}(C') = Prob(r \in R(C') \mid r \in R(C))$. Since C' is a child of coordinate C, $Prob(r \in R(C')) = Prob(r \in R(C') \wedge r \in R(C))$ holds. Since $Prob(r \in R(C') \wedge r \in R(C)) = Prob(r \in R(C)) \times Prob(r \in R(C') \mid r \in R(C))$, we have $Prob(r \in R(C')) = \beta_{ri}(C) \times \beta_{ri}(C')$. So the probability of r belonging to $R(C')$ is $\beta_{ri}(C) \times \beta_{ri}(C')$.

The membership probabilities of resource r belonging to axes or to points specified above are trivial and non-trivial membership probabilities will be proposed after the introduction of the normal forms of the P-RSM. Take Fig. 4

for example, the probability of resource r belonging to axis A is $Prob(r \in R(A)) \in [max\{\beta_{r-A}(a_1), \beta_{r-A}(a_2), \beta_{r-A}(a_3)\}, min\{1, \beta_{r-A}(a_1) + \beta_{r-A}(a_2) + \beta_{r-A}(a_3)\}] = [0.3, 0.6]$. The probability of resource r belonging to point $p(a_2, b_2)$ is $Prob(r \in R(p(a_2, b_2))) \in [max\{0, \beta_{r-A}(a_2) + \beta_{r-B}(b_2) - 1\}, min\{\beta_{r-A}(a_2), \beta_{r-B}(b_2)\}] = [0, 0.3]$.

In Fig. 5, the axis *Area* is used to classify scientific publications according to their areas. In the classification hierarchy of coordinate *CS* (Computer Science) on axis *Area*, *DB* (DataBase) is a subclass of *CS* and *RDB* (Relational DataBase) is a subclass of *DB*. For resource r and its membership probability function β_r, $\beta_r(RDB)$ represents the conditional probability of r belonging to *RDB* given r belonging to *DB* has occurred, i.e. $\beta_r(RDB) = Prob(r \in R(RDB) \mid r \in R(DB))$. Similarly, $\beta_r(DB) = Prob(r \in R(DB) \mid r \in R(CS))$. Since *DB* is a subclass of *CS*, the probability of r belonging to *DB* is $Prob(r \in R(DB)) = Prob(r \in R(DB) \wedge r \in R(CS)) = Prob(r \in R(CS)) \times Prob(r \in R(DB) \mid r \in R(CS)) = \beta_r(CS) \times \beta_r(DB)$. In fact, the probability of r belonging to a sub-coordinate is the multiplication of all the conditional probabilities along the path from the top-level coordinate to this sub-coordinate. So the probability of r belonging to *RDB* is $\beta_r(CS) \times \beta_r(DB) \times \beta_r(RDB)$.
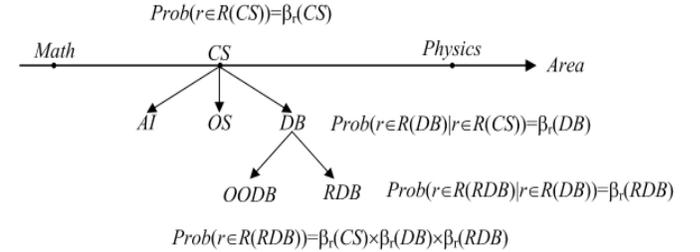

Fig. 5. Conditional probabilities in coordinate hierarchy.

B. Normal Forms of the Probabilistic Resource Space Model

The dependence between categories often makes it difficult to correctly classify resources, and it also affects the precision of evaluation on the membership probabilities of resources belonging to points or axes. Normalization of probabilistic resource spaces can help eliminate this dependence.

1) The First Normal Form and the Second Normal Form

The first normal form of RSM is used to eliminate the redundancy caused by name duplication between coordinates. It also applies to the P-RSM. The second normal form of RSM is to eliminate the redundancy caused by coordinate dependency.

Definition 2. A probabilistic resource space $RS(X_1, X_2, \ldots, X_n)$ is a 2NF resource space if for any resource r and any two coordinates C and C' on X_i ($1 \leq i \leq n$), $Prob(r \in R(C) \wedge r \in R(C')) = 0$ holds.

According to above definition, any second normal form resource space is also a first normal form resource space. For coordinates C and C', if for any resource r, $Prob(r{\in}R(C) \wedge r{\in}R(C'))=0$ holds, then we say that C and C' are independent of each other.

Theorem 1. For axis X, if any two coordinates on X are independent of each other, then for any resource r, $Prob(r{\in}R(X)) = \sum_{C\in X} Prob(r \in R(C)) \le 1$ holds.

Proof. According to the definition of resource space, $R(X) = R(C_1){\cup}R(C_2){\cup}{\ldots}{\cup}R(C_m)$ for axis $X = \{C_1, C_2, \ldots, C_m\}$. Because any two coordinates C_i and C_j ($1{\le}i{\ne}j{\le}m$) on X are independent of each other, $Prob(r{\in}R(C_i) \wedge r{\in}R(C_j))=0$ holds. So the probability of resource r belonging to $R(C_i){\cup}R(C_j)$ is the sum of the probability of r belonging to $R(C_i)$ and the probability of r belonging to $R(C_j)$, i.e. $Prob(r{\in}R(C_i) \vee r{\in}R(C_j)) = Prob(r{\in}R(C_i)) + Prob(r{\in}R(C_j))$. So the probability of r belonging to X is the sum of the probability of r belonging to each coordinate on X, i.e. $Prob(r{\in}R(X)) = \sum_{C\in X} Prob(r \in R(C))$ holds.

In theorem 1, $\sum_{C\in X} Prob(r \in R(C)) < 1$ represents the case that there is the probability that r belongs to a certain coordinate not at axis X. For a resource space RS satisfying 2NF, the probability of r belonging to any axis can be evaluated from the membership probability function of r on this axis.

2) The Third Normal Form

Fine classification and orthogonal classification semantics are to normalize a resource space. A more general definition of fine classification in the P-RSM is given as follows:

Definition 3. Let $X=\{C_1, C_2, \ldots, C_n\}$ be an axis and C' be a coordinate at another axis X', we say that *X finely classifies C*' (denoted as C'/X) if and only if for any resource r:

a) $Prob((r{\in}R(C'){\cap}R(C_i)) \wedge (r{\in}R(C'){\cap}R(C_j))) = 0$, for $1{\le}i{\ne}j{\le}n$; and,
b) $Prob(r{\in}R(C'))= \sum_{C\in X}Prob(r{\in}R(C')|r{\in}R(C)){\times}Prob(r{\in}R(C))$

hold.

According to definition 3 and the total probability theorem, coordinate C' can be finely classified by axis X if and only if the probability of resource r belonging to $R(C')$ can be partitioned into the probabilities of r belonging to $R(C'){\cap}R(C_1)$, $R(C'){\cap}R(C_2)$, ..., and $R(C'){\cap}R(C_n)$ respectively.

For two axes $X=\{C_1, C_2, \ldots, C_n\}$ and $X'=\{C_1', C_2', \ldots, C_m'\}$, *X finely classifies X*' (denoted as X'/X) if and only if X finely classifies C_1', C_2', ..., and C_m' respectively. Two axes X and X' are called orthogonal with each other (denoted as $X{\perp}X$') if X finely classifies X' and vice versa, i.e., both X'/X and X/X' hold.

Definition 4. A probabilistic resource space $RS(X_1, X_2, \ldots , X_n)$ satisfies the third normal form of RSM if for any two axes X_i and X_j ($1{\le}i{\ne}j{\le}n$) in RS, $X_i{\perp}X_j$ holds.

In the P-RSM, any given 3NF resource space $RS(X_1, X_2, \ldots , X_n)$ satisfies the following theorems.

Theorem 2. Let $RS(X_1, X_2, \ldots , X_n)$ be a probabilistic resource space satisfying 3NF. For any two axes X_i and X_j ($1{\le}i, j{\le}n$) and resource r in RS, $\sum_{C\in Xi} Prob(r \in R(C)) = \sum_{C\in Xj} Prob(r \in R(C'))$ holds.

Proof. Since RS satisfies 3NF, coordinate C at axis X_i can be finely classified by axis X_j. So $Prob(r{\in}R(C)) = \sum_{C'\in Xj} (Prob(r \in R(C) \wedge r \in R(C')))$ holds. Thus, we can get that $\sum_{C\in Xi} Prob(r \in R(C)) = \sum_{C\in Xi}\sum_{C'\in Xj} (Prob(r \in R(C) \wedge r \in R(C')))$ holds. On the other hand, coordinate C' at axis X_j can be finely classified by axis X_i. So, $Prob(r{\in}R(C')) = \sum_{C\in Xi} (Prob(r \in R(C') \wedge r \in R(C)))$ holds. Thus, we can get that $\sum_{C\in Xj} Prob(r \in R(C')) = \sum_{C'\in Xj}\sum_{C\in Xi} (Prob(r \in R(C') \wedge r \in R(C)))$ holds. Therefore $\sum_{C\in Xi} Prob(r \in R(C)) = \sum_{C'\in Xj} Prob(r \in R(C'))$ holds.

Theorem 2 indicates that for any two axes X_i and X_j of a probabilistic resource space satisfying 3NF, the probability of resource r belonging to X_i is equal to the probability of r belonging to X_j.

Theorem 3. Let $RS(X_1, X_2, \ldots , X_n)$ be a 2NF probabilistic resource space. For any axis X_i ($1{\le}i{\le}n$) and any coordinate C at X_i, $Prob(r{\in}R(C)) \ge \sum_{p[Xi]=C} Prob(r \in R(p))$ holds, where p represents point in RS and $p[X_i]$ is the projection of p at axis X_i. And if RS is in 3NF, we have $Prob(r{\in}R(C)) = \sum_{p[Xi]=C} Prob(r \in R(p)) \cdot$

Proof. Let T be the union of all points whose projections on X_i are C. So $R(T) = R(C) \cap \bigcap_{1\le j\ne i\le n} \bigcup_{Cj\in Xj} R(C_j)$. Since resource space RS satisfies 2NF, any two points in RS are independent of each other. Thus we have $Prob(r{\in}R(T)) = \sum_{p[Xi]=C} Prob(r \in R(p))$. So $\sum_{p[Xi]=C} Prob(r \in R(p)) = Prob(r{\in}(R(C)\cap \bigcap_{1\le j\ne i\le n} \bigcup_{Cj\in Xj} R(C_j)))$ holds. $Prob(r{\in}R(C)) \ge \sum_{p[Xi]=C} Prob(r \in R(p))$ holds. On the other hand, $Prob(r{\in}R(T)) = Prob(r{\in}(R(C)\cap \bigcap_{1\le j\ne i\le n} \bigcup_{Cj\in Xj} R(C_j))) = Prob(r{\in}(R(C) \cap \bigcap_{1\le j\ne i\le n} R(X_j)))$ holds. If resource space RS satisfies 3NF, coordinate C can be finely classified by axis X_j ($1{\le}j{\ne}i{\le}n$). We can get that $R(C)$ is a subclass of $R(X_j)$. So $Prob(r{\in}R(T)) =$

$Prob(r \in (R(C) \cap \bigcap\limits_{1 \leq j \neq i \leq n} R(X_j))) = Prob(r \in R(C))$ holds.

Therefore $Prob(r \in R(C)) = \sum\limits_{p[Xi]=C} Prob(r \in R(p))$ holds.

C. Membership Probability on Points

Besides the membership probability of resource *r* belonging to each coordinate, another important issue is how to specify the membership probabilities of *r* belonging to points.

For point $p(C_{1,j1}, C_{2,j2}, …, C_{n,jn})$ in resource space $RS(X_1, X_2, … , X_n)$, $R(p)$ is defined as $R(C_{1,j1}) \cap R(C_{2,j2}) \cap … \cap R(C_{n,jn})$, where $C_{i,ji}$ is the projection of point *p* on axis X_i ($1 \leq i \leq n$). So the probability of resource *r* belonging to point *p* is the probability of the complexity event that *r* belongs to $R(C_{1,j1})$, $R(C_{2,j2})$, …, and $R(C_{n,jn})$ simultaneously. Since no sufficient information is available, we can only get an interval for the membership probability of *r* belonging to *p*.

For a resource space only satisfying 1NF, the probability of *r* belonging to *p* falls into the interval [$max\{0, … max\{0, max\{0, Prob(r \in R(C_{1,j1})) + Prob(r \in R(C_{2,j2})) − 1\} + Prob(r \in R(C_{3,j3})) − 1\} … + Prob(r \in R(C_{n,jn})) − 1\}$, $min\{Prob(r \in R(C_{1,j1})), …, Prob(r \in R(C_{n,jn}))\}$].

For a resource space $RS(X_1, X_2, … , X_n)$ satisfying 2NF, according to theorem 1 and theorem 3, the interval for the membership probability of *r* belonging to *p* can be obtained by resolving the following linear programming problem:

Object function: $Prob(r \in R(p))$;

Subject to:

1. $\sum\limits_{C \in Xi} Prob(r \in R(C)) \leq 1$, $1 \leq i \leq n$;

2. $\sum\limits_{p[Xi]=C} Prob(r \in R(p)) \leq Prob(r \in R(C))$, for any coordinate

C at axis X_i ($1 \leq i \leq n$);

3. $\sum\limits_{p' \in RS} Prob(r \in R(p')) \geq max \{0, … max\{0, max\{0,$

$Prob(r \in R(X_1)) + Prob(r \in R(X_2)) − 1\} + Prob(r \in R(X_3)) − 1\} … + Prob(r \in R(X_n)) − 1\}$; and,
4. $L_i \leq Prob(r \in R(p_i) \leq U_i$, for any point p_i in *RS*.

L_i and U_i in item 4 of above constraint are respectively the lower bound and the upper bound of the membership probability of *r* belonging to p_i set by users. If they are not set explicitly, the default value of L_i is 0 and the default value of U_i is 1.

If *RS* satisfies 3NF, then the linear programming problem will be:

Object function: $Prob(r \in R(p))$;

Subject to:

1. $\sum\limits_{C \in Xi} Prob(r \in R(C)) \leq 1$, $1 \leq i \leq n$;

2. $\sum\limits_{p[Xi]=C} Prob(r \in R(p)) = Prob(r \in R(C))$, for any coordinate

C at axis X_i ($1 \leq i \leq n$);
3. $\sum\limits_{C \in Xi} Prob(r \in R(C)) = \sum\limits_{C' \in Xj} Prob(r \in R(C'))$, for $1 \leq i \neq j \leq n$
4. $L_i \leq Prob(r \in R(p_i)) \leq U_i$, for any point p_i in *RS*.

When *RS* is in 3NF, it is obvious that item 3 of the constraint is redundant. Item 3 will be satisfied if both item 1 and item 2 are satisfied.

Theorem 4. For any probabilistic resource space *RS*, the membership probability interval of any resource *r* belonging to point *p* can be obtained in polynomial time of the number of points in *RS*.

Proof. For resource space $RS(X_1, X_2, … , X_n)$, if *RS* is in 1NF, then the membership probability interval of resource *r* belonging to point *p* is [$max\{0, … max\{0, max\{0, Prob(r \in R(C_1)) + Prob(r \in R(C_2)) − 1\} + Prob(r \in R(C_3)) − 1\} … + Prob(r \in R(C_n)) − 1\}$, $min\{Prob(r \in R(C_{1,j1})), …, Prob(r \in R(C_n))\}$], where C_i is the projection of *p* on axis X_i ($1 \leq i \leq n$). It is obvious that both the lower bound and the upper bound can be computed within $n-1$ steps.

If *RS* is in 2NF, the issue on identifying the membership probability intervals of resource *r* belonging to points can be converted to the following linear program problem *LP*:

Object function:
    $Prob(r \in R(p_j))$, for any point p_j in *RS* ;

Subject to:

1. $\sum\limits_{C \in Xi} Prob(r \in R(C)) \leq 1$, $1 \leq i \leq n$;

2. $\sum\limits_{p[Xi]=C} Prob(r \in R(p)) \leq Prob(r \in R(C))$, for any coordinate

C at axis X_i ($1 \leq i \leq n$);
3. $\sum\limits_{p' \in RS} Prob(r \in R(p')) \geq max \{0, … max\{0, max\{0,$

$Prob(r \in R(X_1)) + Prob(r \in R(X_2)) − 1\} + Prob(r \in R(X_3)) − 1\} … + Prob(r \in R(X_n)) − 1\}$;
4. $L_i \leq Prob(r \in R(p_i)) \leq U_i$, for any point p_i in *RS*.

In *LP*, both $Prob(r \in R(C))$ and $Prob(r \in R(X_i))$ ($1 \leq i \leq n$) are constants and $Prob(r \in R(p_j))$ are variables. It is obvious that both the number of variables and the number of inequalities in *LP* are polynomial in the number of points in *RS*. Since linear programming problem is tractable in polynomial time, the membership probability interval of any resource *r* belonging to point *p* can be obtained in polynomial time of the number of points in *RS*.

In the similar way, we can prove that when *RS* is in 3NF, the membership probability interval of any resource *r* belonging to point *p* can also be calculated in polynomial time of the number of points in *RS*.

IV. OPERATIONS OF PROBABILISTIC RESOURCE SPACES

A. Point Query

The first query approach of the P-RSM is point query. The result of a point query is a set of points, each of which contains a set of resources with membership probability.

For a resource space *RS*, the point query operation is used to select the desirable points according to given restrictions. This type of query can be denoted as $\sigma_p(RS)=\{p \mid p \in RS \wedge F_p(p)\}$, where F_p is a logic expression. The basic form of F_p is: $p_m[X_i] \theta Y$, where Y may be $p_n[X_j]$ or just a noun and noun phase in domain ontology, p_m and p_n are points and θ represents $=, \neq, <, \leq, \geq$ or $>$. F_p is usually a logical combination of basic forms by using $\wedge$, $\vee$ and $\neg$.

The P-RSM uses the following statement to support point queries. The *conditional expression* in this statement is the logical combination of restrictions on the projections on axes of points.

SELECT POINT *p* FROM *RS*($X_1, \ldots X_n$)
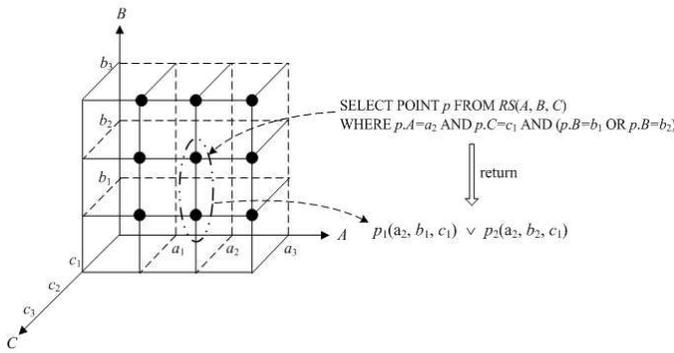[WHERE <*conditional expression*>]

Fig. 6. An example of point query.

Take Fig. 6 for example, if the user wants to query all resources in points $p_1(a_2, b_1, c_1)$ and $p_2(a_2, b_2, c_1)$, the logical expression should be $\sigma_p(RS)=\{p \mid p \in RS \wedge p[A]=a_2 \wedge p[C]=c_1 \wedge (p[B]=b_1 \vee p[B]=b_2)\}$ and the issued point query statement should be:

SELECT POINT *p* FROM *RS*(*A, B, C*)
WHERE $p[A]=a_2$ AND $p[C]=c_1$ AND ($p[B]=b_1$ OR $p[B]=b_2$)

Thus, points $p_1(a_2, b_1, c_1)$ and $p_2(a_2, b_2, c_1)$ will be returned with resources and their membership probabilities belonging to these two points.

B. Resource Query

The query result of a resource query is a set of resources, each of which satisfies the specified restrictions on membership probabilities.

This type of query can be represented as $\sigma_r(RS) = \{r \mid r \in RS \wedge F_r(r, T)\}$, where F_r is a logic expression. The basic form of F_r is $Prob(r \in T) \theta Y$, where Y represents a real number and θ

represents $=, \neq, <, \leq, \geq$, or $>$, T represents a resource set of points in *RS*. Complex logic expressions can be created by combining basic ones with $\forall, \exists, \wedge, \vee$ and $\neg$.

If *RS* does not satisfy 2NF, we do not have any knowledge about the relationship between any two points. So it is difficult to evaluate the membership probability of resource *r* belonging to the set combination of multiple points in *RS*. Thus in $\sigma_r(RS) = \{r \mid r \in RS \wedge F_r(r, T)\}$, if *RS* does not satisfy 2NF, the probabilistic resource space only supports the case that *T* represents a single point in *RS*. For example, $\sigma_r(RS) = \{r \mid r \in RS \wedge \forall p_1 \forall p_2 (Prob(r \in R(p_1) \cup R(p_2))>0.5)\}$ is not a valid resource query if *RS* does not satisfy 2NF since $Prob(r \in R(p_1) \cup R(p_2))$ involves two different points in *RS*. And, $\sigma_r(RS) = \{r \mid r \in RS \wedge \forall p_1(Prob(r \in R(p_1))>0.5) \wedge \forall p_2(Prob(r \in R(p_2))<0.7)\}$ is valid resource query since both $Prob(r \in R(p_1))$ and $Prob(r \in R(p_2))$ only concern a point in *RS*.

If *RS* satisfies 2NF, the event resource *r* belonging to point p_i and the event *r* belonging to another point p_j are mutually exclusive. The membership probability interval of *r* belonging to the set combination of multiple points in *RS* can be obtained by solving a linear programming problem. So in $\sigma_r(RS) = \{r \mid r \in RS \wedge F_r(r, T)\}$, if *RS* satisfies 2NF, the probabilistic resource space supports the case that *T* represent a set combination of multiple points in *RS*.
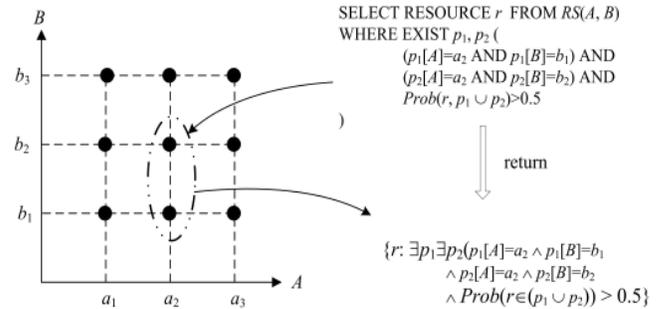

Fig. 7. An example of resource query.

Take Fig. 7 for example, the query is to return all resources in *RS* such that the membership probability of each resource belonging to $p_1(a_2, b_1) \cup p_2(a_2, b_2)$ is greater than 0.5, i.e. $Prob(r \in (R(p_1) \cup R(p_2)) > 0.5$. If *RS* is in 2NF, the event resource *r* belonging to point p_1 and the event *r* belonging to another point p_2 are mutually exclusive. So $Prob(r \in (R(p_1) \cup R(p_2)) = Prob(r \in R(p_1)) + Prob(r \in R(p_2))$ holds. We can use the following linear program to calculate the membership probability interval of *r* belonging to $p_1(a_2, b_1) \cup p_2(a_2, b_2)$.

Object function:
$Prob(r \in R(p_1)) + Prob(r \in R(p_2))$;

Subject to:
1. $\sum_{C \in X_i} Prob(r \in R(C)) \leq 1$, for any axis X_i in *RS*, $1 \leq i \leq n$;

7

2. $\sum_{p[Xi]=C} Prob(r \in R(p)) \leq Prob(r \in R(C))$, for any coordinate

C at axis X_i ($1 \leq i \leq n$);

3. $\sum_{p \in RS} Prob(r \in R(p)) \geq max$ {0, ... max\{0, max\{0,

$Prob(r \in R(X_1)) + Prob(r \in R(X_2)) - 1\} + Prob(r \in R(X_3)) - 1\} \ldots + Prob(r \in R(X_n)) - 1\}$;

4. $L_i \leq Prob(r \in R(p_i)) \leq U_i$, for any point p_i in RS;

5. $Prob(r \in R(p_1)) + Prob(r \in R(p_2)) > 0.5$.

In above linear program, item 5 $Prob(r \in R(p_1)) + Prob(r \in R(p_2)) > 0.5$ is the constraint taken from the query statement. If above linear program is solvable, then the resource r will be returned.

In Fig. 7, the corresponding logical expression of the resource query is $\sigma_F(RS) = \{r: \exists p_1 \exists p_2 (p_1[A] = a_2 \land p_1[B] = b_1 \land p_2[A] = a_2 \land p_2[B] = b_2 \land Prob(r \in (p_1 \cup p_2)) > 0.5\}$ and the SQL-like resource query statement should be in the following form:

SELECT RESOURCE r FROM $RS(A, B)$
WHERE EXIST p_1, p_2 (
          ($p_1[A]=a_2$ AND $p_1[B]=b_1$) AND
          ($p_2[A]=a_2$ AND $p_2[B]=b_2$) AND
               $Prob(r, p_1 \cup p_2) > 0.5$
)

C. Resource Modification

In the original RSM, before a resource r can be inserted into a resource space RS, we have to identify the coordinates which r belongs to at each axis in RS.
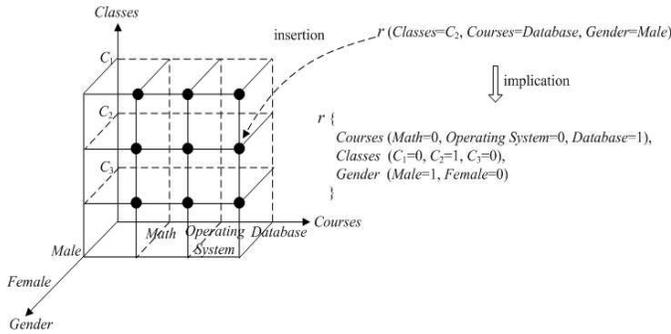

Fig. 8. Insert a resource into a resource space.

Take Fig. 8 for example, the resource space $RS(Classes, Courses, Gender)$ is used to manage student information according to their *classes*, *courses* and *gender*. Once the resource r has been identified that it belongs to coordinate *Database* on axis *Courses*, belongs to coordinate C_2 on axis *Classes* and belongs to coordinate *Male* on axis *Gender*, it can be inserted into the point ($Database$, C_2, $Male$).

From the perspective of probability, $r(Courses=Database, Classes=C_2, Gender=Male)$ implies the fact that the membership probability functions of resource r on axes

Courses, *Classes* and *Gender* are $\beta_{r\text{-}Courses}$, $\beta_{r\text{-}Classes}$ and $\beta_{r\text{-}Gender}$ respectively such that:

(1) $\beta_{r\text{-}Courses}(Math)=0$, $\beta_{r\text{-}Courses}(Operating\ System)=0$, and $\beta_{r\text{-}Courses}(Database)=1$.

(2) $\beta_{r\text{-}Classes}(C_1)=0$, $\beta_{r\text{-}Classes}(C_2)=1$, and $\beta_{r\text{-}Classes}(C_3)=0$,

(3) $\beta_{r\text{-}Gender}(Male)=1$ and $\beta_{r\text{-}Gender}(Female)=0$.

The process of inserting a resource into a probabilistic resource space is the same as the original resource space except that the membership probability functions in the P-RSM can take value within the range [0, 1].

The following is the insertion statement used to insert a resource r into a resource space RS. β_1, β_2, …, and β_n are the membership probability functions of r on axes X_1, X_2, …, and X_n respectively.

INSERT $r<\beta_1, \beta_2, …, \beta_n>$ INTO $RS<X_1, X_2, …, X_n>$

The P-RSM also supports the following delete operation and update operation:

DELETE r FROM RS
[WHERE <*conditional expression*>]
UPDATE $r<\beta_i,…, \beta_j>$ INTO $RS<X_i,…, X_j>$
[WHERE <*conditional expression*>]

D. Operations on Probabilistic Resource Spaces

Join, Disjoin, Merge and Split are four major operations of the original RSM. The following defines the corresponding operations in the P-RSM.

(1) **Join**. If two resource spaces $RS_1(X_1, …, X_m, Y_1, …, Y_n)$ and $RS_2(Y_1, …, Y_n, Z_1, …, Z_k)$ store the same type of resources and have n common axes, then they can be *joined* together as one resource space $RS(X_1, …, X_m, Y_1, …, Y_n, Z_1, …, Z_k)$ such that RS_1 and RS_2 share these n common axes and $|RS|=|RS_1| + |RS_2| - n$. For any resource r in RS, the membership probability functions of r on axes X_i ($1 \leq i \leq m$), Y_j ($1 \leq j \leq n$) and Z_h ($1 \leq h \leq k$) are the same as those functions in RS_1 and RS_2.

Let $p(x_1, …, x_m, y_1, …, y_n, z_1, …, z_k)$, $p_1(x_1, …, x_m, y_1, …, y_n)$ and $p_2(y_1, …, y_n, z_1, …, z_k)$ be the points in RS, RS_1 and RS_2 respectively. The event of resource r belonging to point p corresponds to the event that both r belonging to p_1 and r belonging to p_2 occur simultaneously. If the membership probability interval of r belonging to p_1 is $[L_1, U_1]$ and the membership probability interval of r belonging to p_2 is $[L_2, U_2]$, then we can obtain the following restriction: $max\{0, L_1+L_2-1\} \leq Prob(r \in R(p)) \leq min\{U_1, U_2\}$. The membership probability interval of r belonging to p can be calculated according to the approaches introduced in section 3.2.3.

(2) **Disjoin**. A resource space $RS(X_1, …, X_m, Y_1, …, Y_n, Z_1, …, Z_k)$ can be *disjoined* into two resource spaces $RS_1(X_1, …, X_m, Y_1, …, Y_n)$ and $RS_2(Y_1, …, Y_n, Z_1, …, Z_k)$ that store the same type of resources as that of RS such that they have n common axes and $k + m$ different axes, and $|RS|=|RS_1| + |RS_2| - n$. For any resource r in RS_1, the membership probability functions of r on axes X_i ($1 \leq i \leq m$) and Y_j ($1 \leq j \leq n$) are the same as those functions in RS.

For point $p(x_1, \ldots, x_m, y_1, \ldots, y_n)$ in RS_1, let p_i be the point in RS such that p_i has the same projections on axes $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ as point p ($1 \leq i \leq t$). Suppose that the membership probability interval of resource r belonging to p_i is $[L_i, U_i]$ ($1 \leq i \leq t$). Then we can obtain the following restrictions:

(a) If RS only satisfies 1NF, then $max\{L_1, \ldots, L_t\} \leq Prob(r \in R(p)) \leq min\{1, U_1 + \ldots + U_t\}$ holds;

(b) If RS only satisfies 2NF, then $L_1 + \ldots + L_t \leq Prob(r \in R(p)) \leq 1$ holds; and,

(c) If RS satisfies 3NF, then $L_1 + \ldots + L_t \leq Prob(r \in R(p)) \leq U_1 + \ldots + U_t$ holds.

(3) **Merge**. If two resource spaces $RS_1(X_1, \ldots, X_{n-1}, X')$ and $RS_2(X_1, \ldots, X_{n-1}, X'')$ store the same type of resources and satisfy: a) $|RS_1| = |RS_2| = n$; and, b) they have $n-1$ common axes, and there exist two different axes X' and X'' satisfying the merge condition, then they can be merged into one RS by retaining the $n-1$ common axes and adding a new axis $X^* = X' \cup X''$. RS is called the merge of RS_1 and RS_2, denoted as $RS_1 \cup RS_2 \Rightarrow RS$, and $|RS| = n$. For any resource r in RS, the membership probability functions of r on axes X_i ($1 \leq i \leq n-1$) are the same as those functions in RS_1. Let β' and β'' be the membership probability functions of r at axes X' and X'' respectively. Then the membership probability function β of r on axis X^* is defined as follows: for any coordinate C at axis X^*, $\beta(C) = \beta'(C)$ if C is at axis X', otherwise $\beta(C) = \beta''(C)$.

(4) **Split**. A resource space $RS(X_1, \ldots, X_{n-1}, X)$ can be *split* into two resource spaces RS_1 and RS_2 that store the same type of resources as RS and have $|RS|-1$ common axes by splitting axis X into two: X' and X'', such that $X = X' \cup X''$. For any resource r in RS_1, the membership probability functions of r on axes X_i ($1 \leq i \leq n-1$) are the same as those functions in RS. Let β be the membership probability function of r on axis X. Then, the membership probability function β' of r on axis X' is defined as follows: for any coordinate C at axis X', $\beta'(C) = \beta(C)$.

V. INTEGRITY CONSTRAINTS UNDER PROBABILITY

Integrity constraints play an important role in maintaining consistency of the RSM. In the P-RSM, the meaning of some constraint rules changes and new rules are taken into consideration.

A. Key in Probabilistic Resource Space Model

As a coordinate system, the RSM supports accurate resource location by giving coordinates. However, it is sometimes unnecessary and even arduous to specify all the coordinates to identify a point, especially for high-dimensional resource spaces. The key has been defined in the original RSM to efficiently locate resources.

Definition 5. Let CK be a subset of axis set $\{X_1, X_2, \ldots, X_n\}$, p_1 and p_2 be any two non-null points of resource space $RS(X_1, X_2, \ldots, X_n)$. CK is called a candidate key of resource space RS if we can derive $p_1[X_i] = p_2[X_i]$, $X_i \in \{X_1, X_2, \ldots, X_n\}$ from $p_1[X_j] = p_2[X_j]$, $X_j \in CK$.

Note that there are two concepts in above definition: null point and non-null point. In the original RSM, if a point cannot contain any resources, then this point is denoted as a null point, otherwise it is denoted as non-null point. In the P-RSM, point p is a null point if and only if for any resource r, $Prob(r \in R(p)) = 0$ holds. The key in the probabilistic resource space is defined as follows.

Definition 6. Let CK be a subset of axis set $\{X_1, X_2, \ldots, X_n\}$ and p_1, p_2 be any two points of resource space $RS(X_1, X_2, \ldots, X_n)$ such that $p_1[X_i] = p_2[X_i]$, $X_i \in CK$. CK is called a candidate key of resource space RS if CK satisfies: if there exists an axis X_j such that $X_j \in \{X_1, X_2, \ldots, X_n\} - CK$ and $p_1[X_j] \neq p_2[X_j]$, then $Prob(r_1 \in R(p_1) \wedge r_2 \in R(p_2)) = 0$ holds for any two resources r_1 and r_2.

Above definition implies a kind of resource dependency: if event r_1 belonging to p_1 occurs, the probability of r_2 belonging to p_2 is 0, i.e. $Prob(r_2 \in R(p_2) \mid r_1 \in R(p_1)) = 0$, vice versa.

Most previous probabilistic relational data models manage entities one by one and seldom concern the relationship between entities. They usually assume that the uncertainty of one entity is independent of another entity. The P-RSM should consider some dependency between resources. The following theorem presents a situation where the probabilistic events of two resources should not be supposed to be independent of each other.

Theorem 5. Let CK be a candidate key of 3NF resource space $RS(X_1, X_2, \ldots, X_n)$ and CK' be a subset of $\{X_1, X_2, \ldots, X_n\}$ such that $CK \subset CK'$. Let p_1 and p_2 be two points in RS such that $p_1[X_i] = p_2[X_i]$ ($X_i \in CK$) and $p_1[X_j] \neq p_2[X_j]$ ($X_j \in CK' - CK$). For any two resources r_1 and r_2, the events $r_1 \in \bigcap_{X \in CK' \wedge p1[X]=C} R(C)$ and $r_2 \in \bigcap_{X \in CK' \wedge p2[X]=C} R(C)$ are not independent of each other, and $Prob(r_1 \in \bigcap_{X \in CK' \wedge p1[X]=C} R(C) \wedge r_2 \in \bigcap_{X \in CK' \wedge p2[X]=C} R(C)) = 0$.

Proof. Suppose that both $Prob(r_1 \in \bigcap_{X \in CK' \wedge p1[X]=C} R(C)) \neq 0$ and $Prob(r_2 \in \bigcap_{X \in CK' \wedge p2[X]=C} R(C)) \neq 0$ hold. Since RS satisfies 3NF, both $\bigcap_{X \in CK' \wedge p1[X]=C} R(C) = \bigcup_{X \in CK' \wedge p1[X]=p[X]} R(p)$ and $\bigcap_{X \in CK' \wedge p2[X]=C} R(C) = \bigcup_{X \in CK' \wedge p2[X]=p'[X]} R(p')$ hold. If $Prob(r_1 \in \bigcap_{X \in CK' \wedge p1[X]=C} R(C) \wedge r_2 \in \bigcap_{X \in CK' \wedge p2[X]=C} R(C)) \neq 0$, then there must exist at least two points p_3 and p_4 such that $p_1[X_i] = p_3[X_i]$, $p_2[X_i] = p_4[X_i]$ ($X_i \in CK'$) and $Prob(r_1 \in R(p_3) \wedge r_2 \in R(p_4)) \neq 0$ hold. This contradicts to the fact that CK is a candidate key of RS. So $Prob(r_1 \in \bigcap_{X \in CK' \wedge p1[X]=C} R(C) \wedge r_2 \in \bigcap_{X \in CK' \wedge p2[X]=C} R(C)) = 0$ holds.

B. Integrity Constraints in Probabilistic Resource Space Model

Modification of resources may result in inconsistency in resource spaces. Entity integrity constraint, membership integrity constraint, reference integrity constraint and user-defined integrity constraint have been proposed in the original RSM. More integrity constraint rules should be satisfied in the P-RSM.

Rule 1. For resource space $RS(X_1, X_2, …, X_n)$, let β_{ri} be the membership probabilistic function of resource r at axis X_i, $(1 \le i \le n)$. For any coordinate C at X_i, $0 \le \beta_{ri}(C) \le 1$ must hold. If any two coordinates on X_i are independent of each other, then $\sum_{C \in Xi} \beta_{ri}(C') \le 1$ holds.

Since $\beta_{ri}(C)$ represents the probability of resource r belonging to coordinate C, it is natural to require that $0 \le \beta_{ri}(C) \le 1$ should hold. For axis X_i, $R(X_i) = \bigcup_{C \in X} R(C')$ holds.

If any two coordinates at X_i are independent of each other, $Prob(r \in R(X_i)) = \sum_{C \in Xi} \beta_{ri}(C')$. So $\sum_{C \in Xi} \beta_{ri}(C') \le 1$ holds. The insertion and modification of resources and merge operations between resource spaces may violate Rule 1.

Rule 2. For resource space $RS(X_1, X_2, …, X_n)$ and resource r, let β_{ri} and β_{rj} be the membership probabilistic functions of resource r at X_i and X_j $(1 \le i, j \le n)$ respectively. If X_i can be finely classified by X_j and any two coordinates at X_i are independent of each other, then $\sum_{C \in Xi} \beta_{ri}(C) \le \sum_{C \in Xj} \beta_{rj}(C')$ holds.

If X_i is orthogonal with X_j, i.e. $X_i \perp X_j$ holds, then $\sum_{C \in Xi} \beta_{ri}(C) = \sum_{C \in Xj} \beta_{rj}(C')$ must hold.

If X_i/X_j holds, we can conclude that $R(X_i) \subseteq R(X_j)$ holds. So $Prob(r \in R(X_i)) \le Prob(r \in R(X_j))$ holds. Since $R(X_i) = \bigcup_{C \in Xi} R(C')$ and $R(X_j) = \bigcup_{C \in Xj} R(C')$, both $Prob(r \in R(X_i)) = \sum_{C \in Xi} \beta_{ri}(C)$ and $Prob(r \in R(X_j)) = \sum_{C \in Xj} \beta_{rj}(C')$ hold. Thus, $\sum_{C \in Xi} \beta_{ri}(C) \le \sum_{C \in Xj} \beta_{rj}(C')$ must hold.

If X_i is in orthogonal with X_j, both X_i/X_j and X_j/X_i hold, then both $\sum_{C \in Xi} \beta_{ri}(C) \le \sum_{C \in Xj} \beta_{rj}(C')$ and $\sum_{C \in Xi} \beta_{ri}(C) \ge \sum_{C \in Xj} \beta_{rj}(C')$ hold. So $\sum_{C \in Xi} \beta_{ri}(C) = \sum_{C \in Xj} \beta_{rj}(C')$ holds.

Rule 3. For any 3NF resource space $RS(X_1, X_2, …, X_n)$ and resource r, let β_{ri} be the membership probabilistic function of resource r at X_i $(1 \le i \le n)$. For any coordinate C on X_i and point p in RS, $\sum_{p[Xi]=C} Prob(r \in R(p)) = \beta_{ri}(C)$ holds.

According to theorem 3, in any 3NF resource space, the probability of r belonging to coordinate C can be partitioned into all the points having projection C at axis X_i, i.e. $Prob(r \in R(C)) = \sum_{p[Xi]=C} Prob(r \in R(p))$ holds. Rule 3 should be checked to make sure the maintenance of theorem 3 when inserting or updating resources.

So far we have presented a complete P-RSM.

VI. EXPERIMENTS

The following experiments are to compare the original RSM and the proposed P-RSM in managing uncertain classification semantics.

A. Experimental Data and Schemas

Experimental data are the papers collected from the World Wide Web conference from 2001 to 2007. These papers fall into 13 topics such as Browser & Interfaces, Data Mining, E-Applications, Search, and Semantic Web. The following two resource space schemas can be designed to manage these papers.

(1) 1NF resource space RS_1(*Topics*, *Years*, *Locations*), where *Topics*={*Browser&Interfaces*, *DataMining*, *E-Applications*, *Practice&Experience*, *Performance&Scalability*, *Ubiquitous*, *Search*, *Sercurity&Reliability*, *SemanticWeb*, *WebEngineering*, *XML&WebData*, *WebServices*, *Ontologies*, *E-Learning*, *WebMining*, *Multimedia*}, *Years*={2001, 2002, 2003, 2004, 2005, 2006, 2007} and *Locations*={*Hong Kong*, *Hawaii*, *Budapest*, *New York*, *Chiba*, *Edinburgh*, *Banff*}. Since axis *Topics* has several coordinates that are not independent of each other such as *Semantic Web* and *Ontologies*, resource space schema RS_1 does not satisfies 2NF. Two resource space instances ORS_1 and PRS_1 having the same schema as RS_1 are created for the original RSM and the P-RSM respectively.

(2) 2NF resource space RS_2(*Topics*, *Years*, *Locations*), where *Topics*={*Browser & Interfaces*, *DataMining*, *E-Applications*, *Practice & Experience*, *Performance & Scalability*, *Ubiquitous*, *Search*, *Sercurity & Reliability*, *SemanticWeb*, *WebEngineering*, *XML & WebData*, *WebServices*, *Multimedia*}, *Years*={2001, 2002, 2003, 2004, 2005, 2006, 2007} and *Locations*={*Hong Kong*, *Hawaii*, *Budapest*, *New York*, *Chiba*, *Edinburgh*, *Banff* }. Since all coordinates on each axis in resource space schema RS_2 are independent of each other, RS_2 satisfies 2NF. Two resource space instances ORS_2 and PRS_2 having the same schema as RS_2 are created for the original RSM and the P-RSM respectively.

The membership probability of each paper belonging to each topic can be calculated by using the naïve bayes model. Boolean model based keyword vector x is used to represent paper. For topics T_1, …, and T_k, the probability $p(T_i|x)$ is used to represent the possibility of a given paper belonging to topic T_i. $p(T_i|x)$ can be evaluated by $(p(x|T_i) \times P(T_i))/p(x)$, where $P(T_i)$ is prior probability. The required training samples are the papers published in WWW2002 and WWW2005.

B. Queries on Uncertain Resources

To manage uncertain resources in the original RSM, the following strategies will be adopted:

(1) Let *RS* be a resource space satisfying 2NF. For resource *r* and any axis *X* of *RS*, select the coordinate *C* on *X* such that the membership probability of *r* belonging to *C* is the maximum among all the coordinates on *X*. Then, insert resource *r* into coordinate *C*.

(2) If *RS* is just in 1NF, select the coordinate *C* on *X* such that the membership probability of *r* belonging to *C* is the maximum among all the coordinates on *X*. And then insert resource *r* into coordinate *C* and the coordinate *C'* as long as *C'* is not independent of *C* and the membership probability of *r* belonging to *C'* is greater than 0.

Using the original RSM to manage uncertain resources, users have to judge which coordinates a given resource belongs to according to the membership probabilities. Misjudgment will lead to the resources to be classified into improper coordinates.

Different from the original RSM, the P-RSM will maintain all the membership probabilities of each paper belonging to each topic. It supports the resource query according to the membership probability. Using the resource query, a confidence threshold for each paper belonging to a given topic can be set when querying in the P-RSM. For example, in the query "return all papers of which the topic is *search* and the membership probability is equal to or greater than 0.2", the confidence threshold is 0.2.

The following experiments evaluate the recall and precision for querying the original RSM and the proposed P-RSM. The recall is the ratio of the number of the returned relevant papers to the total number of the relevant papers and the precision is the ratio of the number of the returned relevant papers to the total number of the returned papers.

We refer to the maximum among the membership probabilities of a given paper belonging to each topic as the probability upper bound of this paper. According to the probability upper bound, the papers have been classified into 8 categories: the papers of which probability upper bound is equal to or less than 0.3, 0.4, …, or 1.

The first experiment compares the recall and the precision of the original RSM and the proposed model. Fig. 9 and Fig. 10 plot the average recall and precision of the resource spaces ORS_1, ORS_2 and PRS_1. When querying in the probabilistic resource space PRS_1, the confidence threshold is set to 0.2. The following conclusions can be drawn:

(1) The probability upper bound can indicate whether the membership probability distribution of a paper belonging to the topics is even or not. Both the recall and the precision are quite low when the membership probability distribution of a paper belonging to the topics is even. This is mainly because it is easier to misjudge when the probabilities of a paper belonging to several topics are almost equal. Both the recall and the precision are

gradually improved with the increase of the probability upper bound.

(2) To manage the uncertain information by the original RSM, both the recall and precision of 1NF resource space is better than 2NF resource space. It is mainly because a paper can belong to several topics in the 1NF resource space whereas a paper can belong to only one topic in the 2NF resource space.

(3) The probabilistic resource space PRS_1 has better recall and precision than the original resource spaces ORS_1 and ORS_2. It is mainly because the probabilistic resource space can store all the probabilistic information of a paper belonging to the topics regardless of their independence.

The second experiment is to evaluate the impact of confidence threshold on the recall and the precision when querying in the probabilistic resource spaces. Fig. 11 indicates the trend in the recall and the precision of the probabilistic resource space PRS_1 with the increase of the confidence threshold. We can draw the following two conclusions:

(1) The recall of the probabilistic resource space goes down and the precision of probabilistic resource space goes up with the increase of the confidence threshold. In the resource query on the probabilistic resource space, only the papers of which membership probability is equal to or larger than the confidence threshold can be returned. Thus, the number of returned relevant papers goes down and the total number of the returned papers goes down more.

(2) Theoretically, the recall of the probabilistic resource space will be 100 percent when the confidence threshold is 0. It is because if the confidence threshold is 0, all the papers having the possibility of belonging to a certain topic will be returned. On the other hand, the 100 percent of recall is due to the expense of somewhat low precision.
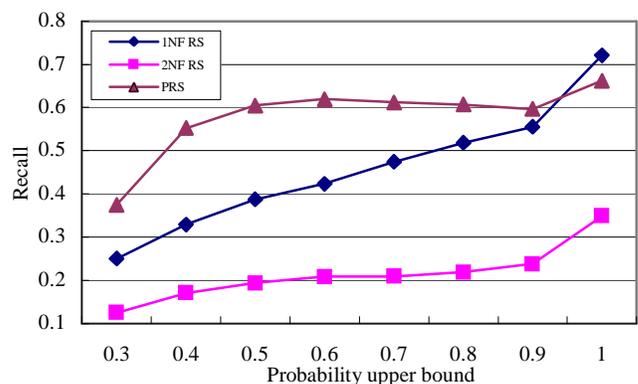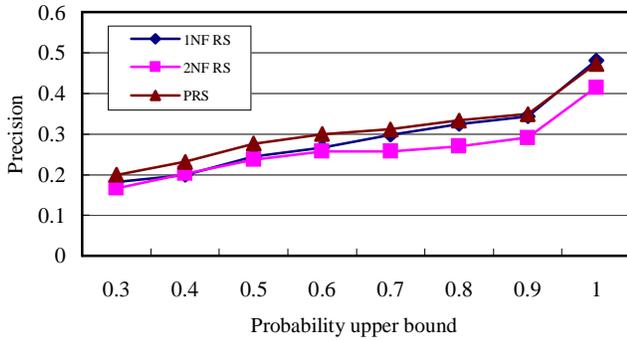

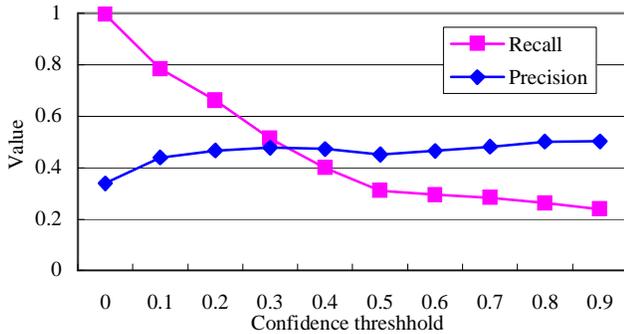Fig. 9. Recall comparison.

Fig. 10. Precision comparison.

Fig. 11. Recall and precision of a probabilistic resource space.

C. Resource Distribution

Here we compare the resource distribution in both the original RSM and the P-RSM. The experiment carries out to evaluate how even the resources are distributed onto all points in a resource space. We use the formula $\sqrt{\sum_{1 \le i \le n} (|p_i| - m/n)^2}$ to represent how even the resources are distributed in a resource space, where m is the total number of resources to be managed, n is the total number of points in a resource space and $|p_i|$ is the number of resources in point p_i.
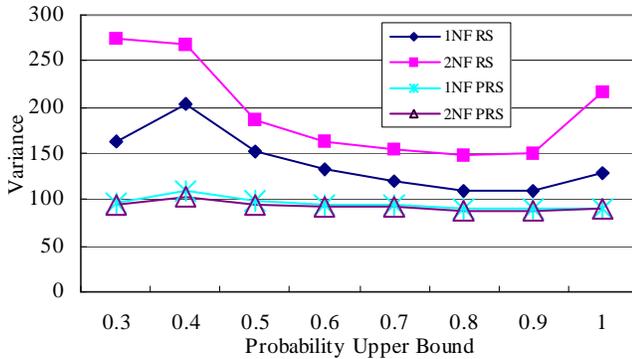

FIG. 12. Resource distribution comparison.

Fig. 12 is the paper distribution comparison in resource spaces ORS_1, ORS_2, PRS_1 and PRS_2. Two conclusions can be drawn from this experiment:

(1) Resources are distributed more even in the probabilistic resource space than the original resource space.
(2) Normal forms have more impact on resource distribution of the original resource spaces than on that of the probabilistic resource spaces. This is mainly because a resource in the probabilistic resource space can be inserted into a point if the membership probability of this resource belonging to this point is larger than 0. But in the original resource space, a resource cannot be inserted into two points independent of each other simultaneously.

VII. CONCLUSION

A powerful semantic data model is the key to efficiently manage various resources on the Web. RSM is a semantic data model for managing the contents of various resources by normalizing classification semantics. By mapping the RSM into probabilistic space, this paper extends the RSM to P-RSM, which can effectively manage resources by uncertain classification semantics. Compared with RSM, the proposed P-RSM can manage richer uncertain classification information, support more flexible resource queries, and have better query performance based on recall/precision. It could be a promising semantic data model for managing Web contents. The trade-off is that P-RSM needs to store the uncertain information to support the flexibility. The cost is the total number of coordinates of a resource space.

REFERENCES

[1] S. Abiteboul, R. Hull and V. Vianu. *Foundations of Databases.* Addison-Wesley, Reading, MA, 1995.
[2] S. Abiteboul et al, "Representing and Querying XML with Incomplete Information". *ACM Transactions on Database Systems*, 31(1) (2006) 208-254.
[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*, Addison Wesley, 1999.
[4] D. Barbara, et al. "The Management of Probabilistic Data", *IEEE Transactions on Knowledge and Data Engineering*, 4(5)(1992)437-502.
[5] T. Bray, J. Paoli and C.M. Sperberg-McQueen. "Extensible Markup Language (XML) 1.0". W3C Recommendation, 1998. http://www.w3.org/TR/REC-xml/.
[6] O. Benjelloun, A.D. Sarma, A.Halevy and J. Widom, "ULDBs: Databases with Uncertainty and Lineage," VLDB2006, pp.953 – 964.
[7] R. Cavallo and M. Pittarelli. "The Theory of Probabilistic Databases". VLDB 1987, pp. 71-81.
[8] E. F. Codd. "A Relational Model of Data for Large Shared Data Banks". *Communications of the ACM*, 13 (6) (1970) 377-387.
[9] N. Dalvi and D. Suciu. "Efficient Query Evaluation on Probabilistic Databases", VLDB 2004, pp. 864-875.
[10] N. Dalvi and D. Suciu. "Answering Queries from Statistics and Probabilistic Views", VLDB 2005, pp. 805-816.

[11] N. Dalvi and D. Suciu. "Management of probabilistic data: foundations and challenges", SIGMOD 2007, pp. 1-12.

[12] D. Dey and S. Sarkar. "A Probabilistic Relational Model and Algebra", *ACM Transactions on Database Systems*, 21(3) (1996) 339-369.

[13] N. Fuhr and T. Rolleke. "A Probabilistic Relational Algebra for the Integration of Information Retrieval and Database Systems", *ACM Transactions on Information Systems*, 15(1) (1997) 32-66.

[14] E. Hung et al. "Probabilistic interval XML", *ACM Transactions on Computational Logic*, 8(4) (2007) No. 24.

[15] B. Kimelfeld and Y. Sagiv. "Matching twigs in probabilistic XML", VLDB 2007, pp. 27-38.

[16] L.V.S. Lakshmanan, et al. "ProbView: A Flexible Probabilistic Database System", *ACM Transactions on Database Systems*, 22(3) (1997) 419-469.

[17] M. Keulen et al. "A Probabilistic XML Approach to Data Integration", ICDE 2005, pp 459-470.

[18] A. Nierman and H. V. Jagadish. "ProTDB: Probabilistic data in XML", VLDB 2002, pp. 646-657.

[19] C. Ré and D. Suciu. "Materialized views in probabilistic databases: for information exchange and query optimization", VLDB 2007, pp. 51-62.

[20] P. Senellart and S. Abiteboul. "On the Complexity of Managing Probabilistic XML Data", PODS 2007.

[21] Y. Takahashi, "Fuzzy Database Query Languages and Their Relational Completeness Theorem," *IEEE Transactions on Knowledge and Data Engineering*, 5(1) (1993)122-125.

[22] Y. Tao, et al., "Indexing Multi-dimensional Uncertain Data with Arbitrary Probability Density Functions," VLDB2005, pp.922 – 933.

[23] Q. T. Tho, S. C. Hui, A.C.M. Fong, and T. H. Cao, "Automatic Fuzzy Ontology Generation for Semantic Web," *IEEE Transactions on Knowledge and Data Engineering*, 18(6) (2006) 842-856.

[24] F. Wang, "Fuzzy Supervised Classification of Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, 28(2) (1990) 194-201.

[25] J. Widom. "Trio: A System for Integrated Management of Data, Accuracy, and Lineage," 2nd Biennial Conference on Innovative Data Systems Research, CIDR 2005, pp. 262-276.

[26] H. Zhuge. *The Knowledge Grid*, World Scientific, 2004.

[27] H. Zhuge. *The Web Resource Space Model*, Springer, 2007.

[28] H.Zhuge, Y.Xing and P.Shi, Resource Space Model, OWL and Database: Mapping and Integration, ACM Transactions on Internet Technology, 8/4, 2008. www.knowledgegrid.net/~h.zhuge/data/ACM-TOIT-Zhuge-final.pdf.