Managing Merged Data by Vague Functional Dependencies[⋆]

An Lu and Wilfred Ng

Department of Computer Science
The Hong Kong University of Science and Technology
Hong Kong, China
{anlu, wilfred}@cs.ust.hk

Abstract. In this paper we propose a new similarity measure between vague sets and apply vague logic in a relational database environment with the objective of capturing the vagueness of the data. By introducing a new vague Similar Equality (S_{EQ}) for comparing data values, we first generalize the classical Functional Dependencies (FDs) into Vague Functional Dependencies (VFDs). We then present a set of sound and complete inference rules. Finally, we study the validation process of VFDs by examining the satisfaction degree of VFDs, and the merge-union and merge-intersection on vague relations.

1 Introduction

The relational data model [8] has been extensively studied for over three decades. This data model basically handle precise and exact data in an information source. However, many real life applications such as merging data from many sources involve imprecise and inexact data. It is well known that Fuzzy database models [11, 2], based on the fuzzy set theory by Zadeh [13], have been introduced to handle inexact and imprecise data. In [5], Gau et al. point out that the drawback of using the single membership value in fuzzy set theory is that the evidence for $u \in U$ and the evidence against $u \in U$ are in fact mixed together. (Here U is a classical set of objects, called the universe of discourse. An element of U is denoted by u.) Therefore, they propose vague sets, which is similar to that of intuitionistic fuzzy sets proposed in [1]. A true membership function $\alpha_V(u)$ and a false membership function $\beta_V(u)$ are used to characterize the lower bound on $\mu_F(u)$. (Here V means a vague set and F means a fuzzy set.) The lower bounds are used to create a subinterval $[\alpha_V(u), 1-\beta_V(u)]$ of the unit interval [0,1], where $0 \leq \alpha_V(u) \leq \mu_F(u) \leq 1 - \beta_V(u) \leq 1$, in order to generalize the membership function of fuzzy sets.

There have been many studies which discuss the topic concerning how to measure the degree of similarity or distance between vague sets or intuitionistic fuzzy sets [3, 4, 7, 9, 12, 6]. However, the proposed methods have some limitations.

[⋆] This work is supported in part by grants from the Research Grant Council of Hong Kong, Grant Nos HKUST6185/02E and HKUST6165/03E.

For example, Hong's similarity measure in [7] means that the similarity measure between the vague value with the most imprecise evidence (the precision of the evidence is 0) and the vague value with the most precise evidence (the precision of the evidence is 1) is equal to 0.5. In this case, the similarity measure should be equal to 0. Our view is that the similarity measure should include two factors of vague values. One is the difference between the evidences contained by the vague values; another is the difference between the precisions of the evidences. However, the proposed measures or distances consider only one factor (e.g. in [3, 4]) or do not combine both the factors appropriately (e.g. in [7, 9, 12, 6]). Our new similarity measure is able to return a more reasonable answer.

In this paper, we extend the classical relational data model to deal with vague information. Our first objective is to extend relational databases to include vague domains by suitably defining the *Vague Functional Dependency* (VFD) based on our notion of similarity measure. A set of sound and complete inference rules for VFD is then established. We discuss the satisfaction degree of VFD and apply VFD in merged vague relations as the second objective. The main contributions of the paper are as follows: (1) A new similarity measure between vague sets is proposed to remedy some problems for similar definitions in literature. We argue that our measure gives a more reasonable estimation; (2) VFD is proposed in order to capture more semantics in vague relations; (3) The satisfaction degree of VFD in merged vague relations are studied.

The rest of the paper is organized as follows. Section 2 presents some basic concepts related to databases and the vague set theory. In Section 3, we propose a new similarity measure between vague sets. In Section 4, we introduce the concept of a *Vague Functional Dependency* (VFD) and the associated inference rules. We then explain the validation process which determines the satisfaction degree of $VFDs$ in vague relations. In Section 5, we give the definitions of merge operators of vague relations and discuss the satisfaction degree of $VFDs$ after merging. Section 6 concludes the paper.

2 Preliminaries

In this section, some basic concepts related to the classical relational data model and the vague set theory are given.

2.1 Relational Data Model

We assume the readers are familiar with the basic concepts of the relation data model [8]. There are two operations on relations that are particularly relevant in subsequent discussion: *projection* and *natural join*. The projection of a relation r of *R(XYZ)* over the set of attributes X is obtained by taking the restriction of the tuples of r to the attributes in X and eliminating duplicate tuples in what remains. This operation is denoted by $\pi_X(r) = \{t[X] \mid t \in r\}$. Let r_1 and r_2 be two relations of *R(XY)* and *R(XZ)*, respectively. The natural join $r_1 \bowtie r_2$ is a relation over *R(XYZ)* defined by $r = r_1 \bowtie r_2 = \{t \mid t[XY] \in r_1 \text{ and } t[YZ] \in r_2\}$. *Functional Dependencies* (FDs) are an important integrity constraint in relational databases. An FD is a statement, $X \to Y$, where X and Y are sets of attributes. A relation r satisfies the FD, if for all t_p and t_q in r, $t_p[X] = t_q[X]$ implies $t_p[Y] = t_q[Y]$.

2.2 Vague Data Model

Let U be a classical set of objects, called the universe of discourse, where an element of U is denoted by u.

Definition 1. (Vague Set) *A vague set V in a universe of discourse U is characterized by a true membership function, α_V, and a false membership function, β_V, as follows: $\alpha_V : U \rightarrow [0,1], \beta_V : U \rightarrow [0,1]$, and $\alpha_V(u) + \beta_V(u) \leq 1$, where $\alpha_V(u)$ is a lower bound on the grade of membership of u derived from the evidence for u, and $\beta_V(u)$ is a lower bound on the grade of membership of the negation of u derived from the evidence against u.*

Suppose $U = \{u_1, u_2, \ldots, u_n\}$. A vague set V of the universe of discourse U can be represented by $V = \sum_{i=1}^{n} [\alpha(u_i), 1 - \beta(u_i)]/u_i$, where $0 \leq \alpha(u_i) \leq 1 - \beta(u_i) \leq 1$ and $1 \leq i \leq n$.

This approach bounds the grade of membership of u to a subinterval $[\alpha_V(u), 1-\beta_V(u)]$ of $[0,1]$. In other words, the exact grade of membership $\mu_V(u)$ of u may be unknown, but is bounded by $\alpha_V(u) \leq \mu_V(u) \leq 1 - \beta_V(u)$, where $\alpha_V(u) + \beta_V(u) \leq 1$. We depict these ideas in Fig. 1. Throughout this paper, we simply use α and β for u if no ambiguity of V arising.

Fig. 1. The true (α) and false (β) membership functions of a vague set

For a vague set $[\alpha(u), 1 - \beta(u)]/u$, we say that the interval $[\alpha(u), 1 - \beta(u)]$ is the vague value to the object u. For example, if $[\alpha(u), 1 - \beta(u)] = [0.6, 0.9]$, then we can see that $\alpha(u) = 0.6$, $1 - \beta(u) = 0.9$ and $\beta(u) = 0.1$. It is interpreted as "the degree that object u belongs to the vague set V is 0.6, the degree that object u does not belong to the vague set V is 0.1." In a voting process, the vague value [0.6,0.9] can be interpreted as " the vote for resolution is 6 in favor, 1 against, and 3 neutral (abstentious)."

The precision of the knowledge about u is characterized by the difference $(1 - \beta(u) - \alpha(u))$. If this is small, the knowledge about u is relatively precise; if it is large, we know correspondingly little. If $(1 - \beta(u))$ is equal to $\alpha(u)$, the knowledge about u is exact, and the vague set theory reverts back to fuzzy set theory. If $(1-\beta(u))$ and $\alpha(u)$ are both equal to 1 or 0, depending on whether u does or does not belong to V, the knowledge about u is very exact and the theory reverts back to ordinary sets. Thus, any crisp or fuzzy value can be regarded as a special case of a vague value. For example, the ordinary set $\{u\}$ can be presented as the vague set $[1,1]/u$, while the fuzzy set $0.8/u$ (the membership of u is 0.8) can be presented as the vague set $[0.8,0.8]/u$.

Definition 2. (Empty Vague Set) *A vague set V is an empty vague set, if and only if, its true membership function $\alpha = 0$ and false membership function $\beta = 1$ for all u. We use $\emptyset$ to denote it.*

Definition 3. (Complement) *The complement of a vague set V is denoted by V' and is defined by $\alpha_{V'}(u) = \beta_V(u)$, and $1-\beta_{V'}(u) = 1-\alpha_V(u)$.*

Definition 4. (Containment) *A vague set A is contained in another vague set B, $A \subseteq B$, if and only if, $\alpha_A(u) \leq \alpha_B(u)$, and $1-\beta_A(u) \leq 1-\beta_B(u)$.*

Definition 5. (Equality) *Two vague sets A and B are equal, written as $A = B$, if and only if, $A \subseteq B$ and $B \subseteq A$; that is $\alpha_A(u) = \alpha_B(u)$, and $1-\beta_A(u) = 1-\beta_B(u)$.*

Definition 6. (Union) *The union of two vague sets A and B is a vague set C, written as $C = A \cup B$, whose true membership and false membership functions are related to those of A and B by $\alpha_C(u) = max(\alpha_A(u), \alpha_B(u))$, and $1-\beta_C(u) = max(1-\beta_A(u), 1-\beta_B(u)) = 1 - min(\beta_A(u), \beta_B(u))$.*

Definition 7. (Intersection) *The intersection of two vague sets A and B is a vague set C, written as $C = A \cap B$, whose true membership and false membership functions are related to those of A and B by $\alpha_C(u) = min(\alpha_A(u), \alpha_B(u))$, and $1-\beta_C(u) = min(1-\beta_A(u), 1-\beta_B(u)) = 1 - max(\beta_A(u), \beta_B(u))$.*

Definition 8. (Cartesian Product) *Let $U = U_1 \times U_2 \times \cdots \times U_m$, be the Cartesian product of m universes, and $A_1, A_2, \ldots, A_m$ be the vague sets in their corresponding universe of discourse $U_1, U_2, \cdots, U_m$, respectively, $u_i \in U_i, i = 1, \ldots, m$. The Cartesian product $A = A_1 \times A_2 \times \cdots \times A_m$ is defined to be a vague set of $U = U_1 \times U_2 \times \cdots \times U_m$, where the memberships are defined as follows: $\alpha_A(u_1 \cdots u_m) = min\{\alpha_{A_1}(u_1), \ldots, \alpha_{A_m}(u_m)\}$, and $1-\beta_A(u_1 \cdots u_m) = min\{(1-\beta_{A_1}(u_1)), \ldots, (1-\beta_{A_m}(u_m))\} = 1 - max\{\beta_{A_1}(u_1), \ldots, \beta_{A_m}(u_m)\}$.*

2.3 Vague Relations

Definition 9. (Vague Relation) *A vague relation r on a relation scheme $R = \{A_1, A_2, \ldots, A_m\}$ is a vague subset of $Dom(A_1) \times Dom(A_2) \times \cdots \times Dom(A_m)$. A tuple $t = (a_1, a_2, \ldots, a_m)$ in $Dom(A_1) \times Dom(A_2) \times \cdots \times Dom(A_m)$ is a vague subset of $U = U_1 \times U_2 \times \cdots \times U_m$.*

A relation scheme R is denoted by $R\{A_1, A_2, \dots, A_m\}$ or simply by R if the attributes are understood. Corresponding to each attribute name $A_i, 1 \leq i \leq m$, the domain of A_i is written as $Dom(A_i)$. However, unlike classical and fuzzy relations, in vague relations, we define $Dom(A_i)$ as a set of vague sets. Vague relations may be considered as an extension of classical relations and fuzzy relations, which can capture more information about imprecision.

Example 1. Consider the vague relation r over Product(ID, Weight, Price) given in Table 1. In r, $Weight$ and $Price$ are vague attributes. To make the attribute ID simple, we express it as the ordinary value. The first tuple in r means the product with $ID = 1$ has the weight of $[1, 1]/10$ and the price of $[0.4, 0.6]/50 + [1, 1]/80$, which are vague sets. In the vague set $[1, 1]/10$, $[1, 1]$ means the evidence in favor "the weight is 10" is 1 and the evidence against it is 0.

Table 1. A Product Relation r

ID	Weight	Price
1	[1,1]/10	[0.4,0.6]/50+[1,1]/80
2	[1,1]/20	[1,1]/100+[0.6,0.8]/150
3	[1,1]/20	[1,1]/100+[0.6,0.8]/150
4	[1,1]/10+[0.6,0.8]/15	[1,1]/80+[0.6,0.8]/100
5	[0.6,0.8]/10+[1,1]/15+[0.6,0.8]/20	[0.6,0.8]/60+[1,1]/90

3 Similarity Measure Between Vague Sets

In this section, we review the notions of similarity measures between vague sets proposed by Chen [3, 4], Hong [7] and Li [9], together with distances between intuitionistic fuzzy sets proposed by Szmidt [12] and Grzegorzewski [6]. We show by some examples that these measures are not able to reflect our intuitions. A new similarity measure between vague sets is proposed to remedy the limitations.

3.1 Similarity Measure Between Two Vague Values

Let x and y be two vague values to a certain object such that $x = [\alpha_x, 1 - \beta_x]$, $y = [\alpha_y, 1 - \beta_y]$. In general, there are two factors should be considered in measuring the similarity between two vague values. One is the difference between the difference of the true and false membership values, which is given by $D_d = |(\alpha_x - \beta_x) - (\alpha_y - \beta_y)|/2 = |(\alpha_x - \alpha_y) - (\beta_x - \beta_y)|/2$, such that $0 \leq D_d \leq 1$; another is the difference between the sum of the true and false membership values, which is given by $D_s = |(\alpha_x + \beta_x) - (\alpha_y + \beta_y)| = |(\alpha_x - \alpha_y) + (\beta_x - \beta_y)|$, such that $0 \leq D_s \leq 1$. The first factor implies the difference between the evidences contained by the vague values, and the second factor implies the difference between the precisions of the evidences.

In [3, 4], Chen defines a similarity measure between two vague values x and y as follows:

$$M_C(x, y) = 1 - \frac{|(\alpha_x - \alpha_y) - (\beta_x - \beta_y)|}{2}, \tag{1}$$

which is equal to $(1 - D_d)$. This similarity measure ignores the difference between the precisions of the evidences (D_s). For example, consider $x = [0,1], y = [a, 1 - a], 0 < a \leq 0.5$,

$$M_C(x,y) = 1 - \frac{|(0-a)-(0-a)|}{2} = 1. \tag{2}$$

This means that x and y are equal. On the one hand, $x = [0,1]$ means $\alpha_x = 0$ and $\beta_x = 0$, that is to say, we have no information about the evidence, and the precision of the evidence is zero. On the other hand, $y = [a, 1-a]$ means $\alpha_x = a$ and $\beta_x = a$, that is to say, we have some information about the evidence, and the precision of the evidence is not zero. So it is not intuitive to have the similarity measure of x and y being equal to 1.

In order to solve this problem, Hong et al. [7] propose another similarity measure between vague values as follows:

$$M_H(x,y) = 1 - \frac{|\alpha_x - \alpha_y| + |\beta_x - \beta_y|}{2}. \tag{3}$$

However, this definition also has some problems. Here is an example.

Example 2. The similarity measure between $[0,1]$ and $[a,a], 0 \leq a \leq 1$ is equal to 0.5. This means that the similarity measure between the vague value with the most imprecise evidence (the precision of the evidence is equal to zero) and the vague value with the most precise evidence (the precision of the evidence is equal to one) is equal to 0.5. However, our intuition shows that the similarity measure in this case should be equal to 0.

Li et al. in [9] also give a similarity measure in order to remedy the problems in Chen's definition as follows:

$$M_L(x,y) = 1 - \frac{|(\alpha_x - \alpha_y) - (\beta_x - \beta_y)| + |\alpha_x - \alpha_y| + |\beta_x - \beta_y|}{4}. \tag{4}$$

It can be checked that $M_L(x,y) = (M_C(x,y) + M_H(x,y))/2$. This means Li's similarity measure is just the arithmetic mean of Chen's and Hong's. So Li's similarity measure still contains the same problems.

[12, 6] adopt Hamming distance and Euclidean distance to measure the distances between intuitionistic fuzzy sets as follows:

1. Hamming distance is given by

$$D_H(x,y) = \frac{|\alpha_x - \alpha_y| + |\beta_x - \beta_y| + |(\alpha_x - \alpha_y) + (\beta_x - \beta_y)|}{2}; \tag{5}$$

2. Euclidean distance is given by

$$D_E(x,y) = \sqrt{\frac{(\alpha_x - \alpha_y)^2 + (\beta_x - \beta_y)^2 + ((\alpha_x - \alpha_y) + (\beta_x - \beta_y))^2}{2}}. \tag{6}$$

These methods also have some problems. Here is an example.

Example 3. We still consider the vague values x, y_1 and y_2 in Example 2. For the Hamming distance, it can be calculated that $D_H(x, y_1) = D_H(x, y_2) = 0.6$. This means that the Hamming distance between x and y_1 are equal to that between x and y_2. In a voting process, as mentioned in Example 2, since both x and y_2 have identical votes in favor and against, the Hamming distance between x and y_2 should be less than that between x and y_1. For the Euclidean distance, consider the Euclidean distance between $[0, 1]$ and $[a, a], 0 \leq a < 1$, which is equal to $(\sqrt{a^2 - a + 1})$. This means that the distance between the vague value with the most imprecise evidence and the vague value with the most precise evidence is not equal to 1. (Actually, the Euclidean distance in this case is in the interval $[\frac{\sqrt{3}}{2}, 1)$.) However, our intuition shows that the distance in this case should always be equal to 1.

In order to solve all the problems mentioned above, we define a new similarity measure between the vague values x and y as follows:

Definition 10. (Similarity Measure Between Two Vague Values)

$$\begin{aligned} M(x,y) &= \sqrt{(1 - D_d)(1 - D_s)} \\ &= \sqrt{(1 - \frac{|(\alpha_x - \alpha_y) - (\beta_x - \beta_y)|}{2})(1 - |(\alpha_x - \alpha_y) + (\beta_x - \beta_y)|)}. \end{aligned} \quad (7)$$

Furthermore, we define a distance between the vague values x and y as $D(x, y) = 1 - M(x, y)$.

The similarity measure given in Definition 10 takes into account of both the difference between the evidences contained by the vague values and the difference between the precisions of the evidences. Here is an example.

Example 4. We still consider the vague values x, y_1 and y_2 in Example 2. It can be calculated that $M(x, y_1) = 0.53$, $M(x, y_2) = 0.63$. So $M(x, y_1) < M(x, y_2)$. This means that the similarity measures between x and y_1 are less than that between x and y_2. As mentioned in Example 2, this result is accordant to our intuition. Another example is the similarity measure between $[0, 1]$ and $[a, a], 0 \leq a \leq 1$, which is equal to 0. This means that the similarity measure between the vague value with the most imprecise evidence and the vague value with the most precise evidence is equal to 0. This result is also accordant to our intuition.

From Definition 10, we can obtain the following theorem.

Theorem 1. *The following statements are true:*

1. *The similarity measure is bounded, i.e.,* $0 \leq M(x, y) \leq 1$*;*
2. $M(x, y) = 1$*, if and only if, the vague values x and y are equal (i.e., $x = y$);*
3. $M(x, y) = 0$*, if and only if, the vague values x and y are* $[0, 0]$ *and* $[1, 1]$ *or* $[0, 1]$ *and* $[a, a], 0 \leq a \leq 1$*;*
4. *The similarity measure is commutative, i.e.,* $M(x, y) = M(y, x)$*.*

3.2 Similarity Measure Between Two Vague Sets

We generalize the similarity measure to two given vague sets.

Definition 11. (Similarity Measure Between Two Vague Sets) *Let X and Y be two vague sets, where $X = \sum_{i=1}^{n}[\alpha_X(u_i), 1-\beta_X(u_i)]/u_i$, and $Y = \sum_{i=1}^{n}[\alpha_Y(u_i), 1-\beta_Y(u_i)]/u_i$. The similarity measure between the vague sets X and Y can be evaluated as follows:*

$$\begin{aligned} M(X,Y) &= \frac{1}{n}\sum_{k=1}^{n} M([\alpha_X(u_k), 1-\beta_X(u_k)], [\alpha_Y(u_k), 1-\beta_Y(u_k)]) \\ &= \frac{1}{n}\sum_{k=1}^{n}\sqrt{(1-\frac{|(\alpha_X(u_k)-\alpha_Y(u_k))-(\beta_X(u_k)-\beta_Y(u_k))|}{2})}\cdot \\ &\qquad \sqrt{(1-|(\alpha_X(u_k)-\alpha_Y(u_k))+(\beta_X(u_k)-\beta_Y(u_k))|)} \end{aligned} \tag{8}$$

Similarly, we give the definition of distance between two vague sets as $D(X,Y) = 1 - M(X,Y)$.

From Definition 11, we obtain the following theorem for vague sets, which is similar to Theorem 1.

Theorem 2. *The following statements related to $M(X,Y)$ are true:*

1. *The similarity measure is bounded, i.e., $0 \leq M(X,Y) \leq 1$;*
2. *$M(X,Y) = 1$, if and only if, the vague sets X and Y are equal (i.e., $X = Y$);*
3. *$M(X,Y) = 0$, if and only if, all the vague values $[\alpha_X(u_k), 1-\beta_X(u_k)]$ and $[\alpha_Y(u_k), 1-\beta_Y(u_k)]$ are $[0,0]$ and $[1,1]$ or $[0,1]$ and $[a,a], 0 \leq a \leq 1$;*
4. *The similarity measure is commutative, i.e., $M(X,Y) = M(Y,X)$.*

4 Vague Functional Dependency and Inference Rules

In this section, we first give the definition of *Similar Equality* (S_{EQ}) of vague relations, which can be used to compare vague relations. Then we present the definition of a *Vague Functional Dependency* (VFD). Next, we present a set of sound and complete inference rules for *VFDs*, which is an analogy to Armstrong's Axiom for classical *FDs*.

4.1 Similar Equality of Vague Relations

Similar Equality (S_{EQ}) of vague relations defined below can be used as a vague similarity measure to compare elements of a given domain. Suppose t_p and t_q are any two tuples in a relation r over the scheme R.

Definition 12. (Similar Equality of Tuples) *The Similar Equality of two vague tuples t_p and t_q on the attribute A_i in a vague relation is given by:*

$$\begin{aligned} & S_{EQ}(t_p[A_i], t_q[A_i]) \\ &= \frac{1}{n}\sum_{k=1}^{n}\sqrt{(1-\frac{|(\alpha_{t_p[A_i]}(u_k)-\alpha_{t_q[A_i]}(u_k))-(\beta_{t_p[A_i]}(u_k)-\beta_{t_q[A_i]}(u_k))|}{2})}\cdot \\ &\qquad \sqrt{(1-|(\alpha_{t_p[A_i]}(u_k)-\alpha_{t_q[A_i]}(u_k))+(\beta_{t_p[A_i]}(u_k)-\beta_{t_q[A_i]}(u_k))|)} \end{aligned} \tag{9}$$

The Similar Equality of two vague tuples t_p and t_q on attributes $X = \{A_1, \ldots, A_n\}$ ($X \subseteq R$) in a vague relation is given by:

$$\begin{aligned} S_{EQ}(t_p[X], t_q[X]) &= S_{EQ}(t_p[A_1 \cdots A_n], t_q[A_1 \cdots A_n]) \\ &= min\{S_{EQ}(t_p[A_1], t_q[A_1]), \ldots, S_{EQ}(t_p[A_n], t_q[A_n])\} \quad (10) \end{aligned}$$

From Definition 12 and Theorem 2, we have the following theorem.

Theorem 3. *The following statements of the properties of $S_{EQ}(t_p[X], t_q[X])$ are true:*

1. *The similar equality is bounded: $0 \leq S_{EQ}(t_p[X], t_q[X]) \leq 1$;*
2. *$S_{EQ}(t_p[X], t_q[X]) = 1$, if and only if, all vague sets $t_p[A_s]$ and $t_q[A_s]$ ($i \leq s \leq j$) are equal (i.e., $t_p[A_s] = t_q[A_s], i \leq s \leq j$);*
3. *$S_{EQ}(t_p[X], t_q[X]) = 0$, if and only if, $\exists A_i \in X$, $S_{EQ}(t_p[A_i], t_q[A_i]) = 0$, if and only if, $\exists A_i \in X$, all the vague values $[\alpha_{t_p[A_i]}(u_k), 1 - \beta_{t_p[A_i]}(u_k)]$ and $[\alpha_{t_q[A_i]}(u_k), 1 - \beta_{t_q[A_i]}(u_k)]$ are $[0, 0]$ and $[1, 1]$, or $[0, 1]$ and $[a, a]$, where $0 \leq a \leq 1$;*
4. *The similar equality is commutative: $S_{EQ}(t_p[X], t_q[X]) = S_{EQ}(t_q[X], t_p[X])$.*

4.2 Vague Functional Dependencies

Informally, a VFD captures the semantics of the fact that, for given two tuples, Y values should not be less similar than X values. We now give the following definition of a VFD.

Definition 13. (Vague Functional Dependency) *Given a relation r over a relation schema $R(A_1, A_2, \ldots, A_m)$, where $Dom(A_i), i = 1 \cdots m$, are sets of vague sets, a Vague Functional Dependency (VFD) $X \hookrightarrow Y$ where $X, Y \subseteq R$ holds over r, if for all tuples t_p and t_q in r, we have $S_{EQ}(t_p[X], t_q[X]) \leq S_{EQ}(t_p[Y], t_q[Y])$.*

In the database literature [8], a set of inference rules is generally used to derive new data dependencies from the given set of dependencies. We now present a set of sound and complete inference rules for $VFDs$, which is similar to Armstrong's Axiom for FDs.

Definition 14. (Inference Rules) *Let us consider a relation scheme $R(A_1, A_2, \ldots, A_m)$ and a set of $VFDs$ F. Let X, Y, and Z be subsets of the relation scheme R. We define a set of inference rules as follows:*

1. *Reflexivity: If $Y \subseteq X$, then $X \hookrightarrow Y$;*
2. *Augmentation: If $X \hookrightarrow Y$ holds, then $XZ \hookrightarrow YZ$ also holds;*
3. *Transitivity: If $X \hookrightarrow Y$ and $Y \hookrightarrow Z$ hold, then $X \hookrightarrow Z$ holds.*

The following theorem follows by assuming that there are at least two elements a and b in each data domain such that $S_{EQ}(a, b) = 0$.

Theorem 4. *The inference rules given in Definition 14 are sound and complete.*

The Union, Decomposition, Pseudotransitivity rules follow from these three rules, as in the case of functional dependencies [8]. We skip the proof due to space limitation.

4.3 Validation of VFDs

In this section, we study the validation issues of $VFDs$. We relax the notion that if a VFD does not hold for a pair of tuples in r, then the VFD does not hold. We allow VFD to hold with a certain satisfaction degree over r. The validation process and the calculation of the satisfaction degree of the VFD $X \hookrightarrow A$ are given as follows:

1. For every attribute A_i in $X \cup A$, we calculate $S_{EQ}(t_p[A_i], t_q[A_i])$ between every pair of tuples t_p and t_q in r by constructing two $n \times n$ (n is the cardinality of r) upper triangular matrices X and A. The row and column represent a comparison of different tuples. We ignore the lower part of the matrix and the diagonal, since S_{EQ} is commutative. Thus we get $n(n-1)/2$ entries in the matrix. Each entry is the comparison of a pair of tuples;
2. We check $S_{EQ}(t_p[X], t_q[X]) \leq S_{EQ}(t_p[A], t_q[A])$ for every t_p, t_q in r. If true, then we say that the VFD $X \hookrightarrow A$ holds (with the satisfaction degree of 1). We construct a matrix $W = X - A$ to check this;
3. If the result in Step 2 is not true, in the matrix W, we count the number of entries (denoted by s) which are less than or equal to 0. The satisfaction degree SD of the VFD $X \hookrightarrow A$ in r can be calculated as follows:

$$SD = \frac{s}{(\frac{n(n-1)}{2})}. \tag{11}$$

Obviously, if the inequality given in Definition 13 holds for all tuples in r, the satisfaction degree calculated by (11) is equal to 1.

Suppose there are many $VFDs$ hold over relation r, say $f_1, f_2, \ldots, f_n$, with the satisfaction degrees $SD_1, SD_2, \ldots, SD_n$ respectively. We use a VFD set $F = \{f_1, f_2, \ldots, f_n\}$ to present this. Then the satisfaction degree of VFD set F over relation r can be calculated by the arithmetic mean of the satisfaction degrees of F as follows:

$$SD_F = \frac{SD_1 + SD_2 + \cdots + SD_n}{n}. \tag{12}$$

Here is an example to illustrate the validation process and the calculation of the satisfaction degree of the VFD.

Example 5. Consider the vague relation r presented in Table 1, it can be checked that the VFD $Weight \hookrightarrow Price$ holds to a certain satisfaction degree.

In step 1, we calculate $S_{EQ}(t_p[A_i], t_q[A_i])$ for attributes $X = Weight$ and $A = Price$ and the results are shown by matrix X and A or Tables 2 and 3.

In step 2, we check $S_{EQ}(t_p[X], t_q[X]) \leq S_{EQ}(t_p[A], t_q[A])$ by taking the difference between the two matrices X and A. The result is shown by matrix W or Table 4.

Since $S_{EQ}(t_p[X], t_q[X]) \leq S_{EQ}(t_p[A], t_q[A])$ does not hold for every p, q, we go to step 3.

$$\mathbf{X} = \begin{pmatrix} - & 0 & 0 & 0.74 & 0.41 \\ - & - & 1 & 0.16 & 0.41 \\ - & - & - & 0.16 & 0.41 \\ - & - & - & - & 0.66 \\ - & - & - & - & - \end{pmatrix}$$

Table 2. Weight

Tuples	1	2	3	4	5
1	-	0	0	0.74	0.41
2	-	-	1	0.16	0.41
3	-	-	-	0.16	0.41
4	-	-	-	-	0.66
5	-	-	-	-	-

$$\mathbf{A} = \begin{pmatrix} - & 0.28 & 0.28 & 0.71 & 0.28 \\ - & - & 1 & 0.41 & 0.24 \\ - & - & - & 0.41 & 0.24 \\ - & - & - & - & 0.24 \\ - & - & - & - & - \end{pmatrix}$$

Table 3. Price

Tuples	1	2	3	4	5
1	-	0.28	0.28	0.71	0.28
2	-	-	1	0.41	0.24
3	-	-	-	0.41	0.24
4	-	-	-	-	0.24
5	-	-	-	-	-

In step 3, we get $s = 5$. So the satisfaction degree SD can be calculated as follows:

$$SD = \frac{s}{(\frac{n(n-1)}{2})} = \frac{5}{(\frac{5(5-1)}{2})} = 0.5. \tag{13}$$

Therefore, the *VFD* $Weight \hookrightarrow Price$ over relation r holds with the satisfaction degree 0.5.

Furthermore, for the zero entries in W, we check the corresponding values in the matrix X. If the values are equal to 1, all vague sets ($t_p[A_i]$ and $t_q[A_i]$) (A_i in X) are equal according to Theorem 3. Thus, we can remove some redundancies by decomposing the original relation into two relations.

For instance, there is a value in position (3,2) is 0 in W above. We check the corresponding value in position (3,2) in matrix X, and find the value is 1. So the vague relation in Table 1 can be decomposed into two relations *IW(ID, Weight)*, *WP(Weight,Price)* (Tables 5 and 6), and some redundancies have been removed.

5 Merge Operations of Vague Relations

In this section, we first give the definition of merge operators of vague relations and then discuss the evaluation of the satisfaction degree of $VFDs$ over the merged vague relations.

$$\mathbf{W} = \begin{pmatrix} - & -0.28 & -0.28 & 0.03 & 0.13 \\ - & - & 0 & -0.25 & 0.17 \\ - & - & - & -0.25 & 0.17 \\ - & - & - & - & 0.42 \\ - & - & - & - & - \end{pmatrix}$$

Table 4. Weight-Price

Tuples	1	2	3	4	5
1	-	-0.28	-0.28	0.03	0.13
2	-	-	0	-0.25	0.17
3	-	-	-	-0.25	0.17
4	-	-	-	-	0.42
5	-	-	-	-	-

Table 5. IW

ID	Weight
1	[1,1]/10
2	[1,1]/20
3	[1,1]/20
4	[1,1]/10+[0.6,0.8]/15
5	[0.6,0.8]/10+[1,1]/15+[0.6,0.8]/20

Table 6. WP

Weight	Price
[1,1]/10	[0.4,0.6]/50,[1,1]/80
[1,1]/20	[1,1]/100+[0.6,0.8]/150
[1,1]/10+[0.6,0.8]/15	[1,1]/80+[0.6,0.8]/100
[0.6,0.8]/10+[1,1]/15 +[0.6,0.8]/20	[0.6,0.8]/60+[1,1]/90

5.1 Merge Operators

Generally speaking, when multiple data sources merge together, the result may contain objects of three cases [10]: (1) an attribute value is not provided; (2) an attribute value is provided by exactly one source; (3) an attribute value is provided by more than one source. When merging vague data, in the first case, we use an empty vague set to express the unavailable value; in the second case, we keep the original vague set; in the third case, we take the union of the vague sets provided by the source. We now define two new merge operators to serve our purpose.

Definition 15. (Join Merge Operator) *Let t_r be a tuple in the vague relation r over scheme $R = (A_1, A_2, \dots, A_m)$ and t_s be a tuple in the vague relation s over scheme $S = (A_1, A_i, \dots, A_n)$. r and s have a common ID attribute A_1. The attributes $A_i, \dots, A_m$ are common in both vague relations. Then we define the* join merge *of r and s, denoted by $r \wedge s$, as follows: $r \wedge s = \{t | \exists t_r \in r, t_s \in s$ with $t[A_1] = t_r[A_1] = t_s[A_1], t[A_j] = t_r[A_j], j = 2, \dots, i-1; t[A_j] = t_r[A_j] \cup t_s[A_j], j = i, \dots, m; t[A_j] = t_s[A_j], j = m+1, \dots, n\}$, where $t_r[A_j] \cup t_s[A_j]$ means the union of two vague sets as defined in Definition 6.*

Definition 16. (Union Merge Operator) *Let $r' = r - \pi_R(r \wedge s)$, $s' = s - \pi_S(r \wedge s)$. Then we define the union merge of r and s, denoted by $r \vee s$, as follows: $r \vee s = (r \wedge s) \cup \{t | \forall t_{r'} \in r'$ with $t[A_j] = t_{r'}[A_j], j = 1, \dots, m; t[A_j] = \emptyset, j = m+1, \dots, n\} \cup \{t | \forall t_{s'} \in s'$ with $t[A_1] = t_{s'}[A_1], t[A_j] = \emptyset, j = 2, \dots, i-1; t[A_j] = t_{s'}[A_j], j = i, \dots, n\}$, where $\emptyset$ means an empty vague set.*

Since vague sets have the property of associativity given in Theorem **??**, the join merge operator and the union merge operator also have the property of associativity. That is to say, $r \wedge (s \wedge t) = (r \wedge s) \wedge t$ and $r \vee (s \vee t) = (r \vee s) \vee t$ (recall that r, s, t are vague relations). We can also generalize Definitions 15 and 16 to more than two data sources. Definition 16 guarantees that every tuple is contained in the new merged relation. For example, consider the following vague relations r and s given in Tables 7 and 8. We then have $(r \wedge s)$ and $(r \vee s)$ as given in Tables 9 and 10.

5.2 Satisfaction Degree of Merged Relations

Suppose we have m data sources represented by the vague relations $r_1, \dots, r_m$. Each relation r_i $(1 \leq i \leq m)$ has a set of $VFDs$, F_i $(1 \leq i \leq m)$, with

Table 7. Vague Relation r

A_1	A_2	A_3
1	[1,1]/2	∅
2	∅	[0.3,0.7]/a+ [0.6,0.8]/c
3	[0.2,0.3]/6+ [0.5,0.7]/8	[0.7,0.9]/b+ [0.5,0.9]/d

Table 8. Vague Relation s

A_1	A_3	A_4
1	[0.1,0.4]/a	[1,1]/x+[0.6,0.8]/z
3	[0.2,0.8]/a+ [0.6,0.8]/d	∅
5	[0.2,0.3]/b+ [0.5,0.7]/f	[0.7,0.9]/s+ [0.5,0.6]/t

Table 9. Vague Relation $r \wedge s$

A_1	A_2	A_3	A_4
1	[1,1]/2	[0.1,0.4]/a	[1,1]/x+[0.6,0.8]/z
3	[0.2,0.3]/6+[0.5,0.7]/8	[0.2,0.8]/a+[0.7,0.9]/b+[0.6,0.9]/d	∅

the satisfaction degree SD_{F_i} defined in (12). By the union merge operator, we get a new relation $r = r_1 \vee \cdots \vee r_m$. We can also get a new VFD set $F = F_1 \cup F_2 \cup \cdots \cup F_m$ over r. For each VFD in F, we can calculate the new satisfaction degree over r by the validation process proposed in Sect. 4. Then the satisfaction degree SD_F of the new VFD set F over relation r can be calculated by (12).

In the case of non-overlapping sources, we can simplify the calculation as follows. Assume two data sources represented by the vague relations, r_1 and r_2, which have the same VFD $X \hookrightarrow A$ on a common schemas. We let the satisfaction degree be SD_1 and SD_2, and the cardinalities of r_1 and r_2 are c_1 and c_2. (As the sources are non-overlapping, there exists no tuple which has the same value of A_1 (the ID attribute) in both r_1 and r_2.) This implies that the cardinality of $r_1 \vee r_2$ is $(c_1 + c_2)$. In order to calculate the new SD of $X \hookrightarrow A$ over $r_1 \vee r_2$, we need to construct two new $(c_1 \times c_2)$ matrices, X' and A', to calculate the S_{EQ} of every pair of tuples between r_1 and r_2. Then we need to construct a matrix $W' = X' - A'$ and count the number of entries (denoted by s'), which are less than or equal to 0 in W'. According to (11), the satisfaction degree SD of the VFD $X \hookrightarrow A$ over $r_1 \vee r_2$, where $C = (c_1 + c_2)(c_1 + c_2 - 1)$, can be calculated as follows:

$$SD = \frac{c_1(c_1-1)}{C} SD_1 + \frac{c_2(c_2-1)}{C} SD_2 + \frac{2s'}{C} \tag{14}$$

Table 10. Vague Relation $r \vee s$

A_1	A_2	A_3	A_4
1	[1,1]/2	[0.1,0.4]/a	[1,1]/x+[0.6,0.8]/z
2	∅	[0.3,0.7]/a+[0.6,0.8]/c	∅
3	[0.2,0.3]/6+[0.5,0.7]/8	[0.2,0.8]/a+[0.7,0.9]/b+[0.6,0.9]/d	∅
5	∅	[0.2,0.3]/b+[0.5,0.7]/f	[0.7,0.9]/s+[0.5,0.6]/t

6 Conclusions

In this paper, we incorporate the notion of vagueness into the relational data model, with an objective to provide a generalized approach for treating imprecise data. We propose a new similarity measure between vague sets, which gives more reasonable estimation than those proposed in literature. We apply *Similar Equality*(S_{EQ}) in vague relations. The equality measure can be used to compare elements of a given vague data domain. Based on the concept of similar equality of attribute values in vague relations, we develop the notion of *Vague Functional Dependency* (VFD), which is a simple and natural generalization of classical or fuzzy functional dependencies. In spite of this generalization, the inference rules for $VFDs$ share the simplicity of Armstrong's axiom for classical FDs. We also present the validation process of VFD and the formula to determine the satisfaction degree of VFD. Finally, we give the definition of merge operators of vague relations and discuss the satisfaction degree of $VFDs$ over the merged vague data. As a future work, we plan to extend the merge operations over vague data, which provide a flexible means to merge data in modern applications, such as querying internet sources and merging the returned result. We are also studying the notion *Vague Inclusion Dependencies*, which is useful to generalize the foreign keys in vague relations.

References

1. Atanassov, K.: Intuitionistic Fuzzy Sets. Fuzzy Sets and Systems **20(1)** (1986) 87–96
2. Buckles, B.P., Petry F.E.: A Fuzzy Representation of Data for Relational Databases. Fuzzy Sets and Systems **7** (1982) 213–226
3. Chen, S.M.: Similarity Measures Between Vague Sets and Between Elements. IEEE Transactions on System, Man and Cybernetics **27(1)** (1997) 153–159
4. Chen, S.M.: Measures of Similarity Between Vague Sets. Fuzzy Sets and Systems **74(2)** (1995) 217–223
5. Gau, W.L., Danied, J.B.: Vague Sets. IEEE Transactions on Systems, Man, and Cybernetics **23(2)** (1993) 610–614
6. Grzegorzewski, P.: Distances Between Intuitionistic Fuzzy Sets and/or Interval-valued Fuzzy Sets Based on the Hausdorff Metric. Fuzzy Sets and Systems (2003)
7. Hong, D.H., Kim, C.: A Note on Similarity Measures Between Vague Sets and Between elements. Information Sciences **115** (1999) 83–96
8. Levene, M., Loizou, G.: A Guided Tour of Relational Databases and Beyond. Springer Verlag, (1999)
9. Li, F., Xu, Z.: Measures of Similarity Between Vague sets. Journal of Software **12(6)** (2001) 922–927
10. Naumann, F., Freytag, J.C.: Completeness of Information Sources. Ulf Leser Workshop on Data Quality in Cooperative Information Systems 2003 (DQCIS), (2003)
11. Raju, K.V.S.V.N., Majumdar, A.K.: Fuzzy Functional Dependencies and Lossless Join Decomposition of Fuzzy Relational Database Systems. ACM Transactions on Database Systems **13(2)** (1988) 129–166
12. Szmidt, E., Kacprzyk, J.: Distances Between Intuitionistic Fuzzy Sets. Fuzzy Sets and Systems **114** (2000) 505–518
13. Zadeh, L.A.: Fuzzy Sets. Information and Control **8(3)** (1965) 338–353