APPLICATION OF CONDITIONAL PROBABILITY IN PREDICTING INTERVAL PROBABILITY OF DATA QUERYING

Rolly Intan
Faculty of Industrial Technology, Informatics Engineering Department,
Petra Christian University
Email: rintan@petra.ac.id

ABSTRACT: This paper discusses fuzzification of crisp domain into fuzzy classes providing fuzzy domain. Relationship between two fuzzy domains, X_i and X_j, can be represented by using a matrix, w_{ij}. If X_i has n elements of fuzzy data and X_j has m elements of fuzzy data then w_{ij} is $n \times m$ matrix. Our primary goal in this paper is to generate some formulas for predicting interval probability in the relation to data querying, i.e., given John is 30 years old and he has MS degree, how about his probability to get high salary.

Keywords: Fuzzy Conditional Probability Relation, Data Querying, Interval Probability, Mass Assignment, Point Semantic Unification.

1. INTRODUCTION

In a system, we will find that every component has relation one to each other. For example, system CAREER has some components such as *education*, *age*, and *salary* in which we realize that all of them has interrelationship as in general *higher education means higher salary*, or *older someone will get higher salary*, or *in a certain area, percentage of mid-40's of persons who has doctoral degree is 30 percent*, etc.

In this paper, we process a certain relational database by classifying every domain into several value of data or elements of the component, i.e., domain or component *age* can be classified into {*about_20, about_25, …, about_60*}. With assumption that every classified data is a fuzzy set, we must determine a membership function which represents degree of element belonging to the fuzzy set, i.e., *about_30*={0.2/26, 04/27, 0.6/28, 0.8/29, 1/30, 0.8/31, 0.6/32, 0.4/33, 0.2/34}. Next, we apply the all membership function into the previous relational database to find every membership value for every item data. And then, by using conditional probabilistic theory, we construct a model to describe interrelationship among all components of the system. Relationship between two components, X_1 **and** X_2, of a system is expressed in a matrix w_{12}. If component X_1 has *n* elements, X_2 has *m* elements then matrix w_{12} is $n \times m$ matrix, where $a_{ij}^{12} \in w_{12}$ expresses *weight* or degree dependency of $x_{2j} \in X_2$ from $x_{1i} \in X_1$, for $1 \le i \le n$, $1 \le j \le m$. Through this model, we generate some formulas to predict any value of component related to a given query of input data i.e., given John is 30 years old and he has MS degree, how many his probability to get high salary, and off course we have to define high salary as a fuzzy data value.

Given input of data querying can be precise as well as imprecise data (fuzzy data), so first, before the data can be used to make prediction, we must to find their probabilistic matching related to element of components of system by using *Point Semantic Unification Process* as introduced in paper of Baldwin [1]. In this case, *Point Value Semantic Unification* can be considered as a conditional probability between two fuzzy sets [3]. In calculating prediction, we generate two different formulas to provide upper and lower bound probability of prediction. Hence, result of prediction works into a interval truth value [*a*,*b*] where $a \le b$ as proposed in [2].

2. BASIC CONCEPT

2.1 Conditional Probability

Definition 2.1 $P(H \mid D)$ is defined as a conditional probability for *H* given *D*. Relation between conditional and unconditional probability satisfies the following equation [5].

$$P(H \mid D) = \frac{P(H \cap D)}{P(D)}, \qquad (1)$$

where $P(H \cap D)$ is an unconditional probability of compound events 'H and D happen'. *P*(*D*) is unconditional probability of event *D*.

2.2 Mass Assignment

Definition 2.2 Given *f* is a fuzzy set defined on the discrete space $X = \{x_1, \ldots, x_n\}$, namely

$$f = \sum_{i=1}^{n} \frac{\chi_i}{x_i}$$

Suppose *f* is a normal fuzzy set whose elements are ordered such that: $c_1 = 1$, $c_i \le c_j$, if $i \le j$; The mass assignment corresponding to the fuzzy set *f* is [6]

$$mf = \{\{x_1, x_2, \ldots, x_i\} : \chi_i - \chi_{i+1}\} \text{ with } x_{n+1} = 0 \qquad (2)$$

For example, given a fuzzy set *low_numbers* = {1/1, 1/2, 0.5/3, 0.2/4}, the mass assignment of the fuzzy set *low_numbers* is

$m_{low\_numbers}$ = {1,2}:0.5, {1,2,3}:0.3,
{1,2,3,4}:0.2.

2.3 Point Semantic Unification

Definition 2.3 Let m_f = {L_i:l_i} and m_g={N_i:n_i} be mass assignment associate with the fuzzy set *f* and *g*, respectively. From the matrix,

$$M = \{m_{ij}\} = \left\{ \frac{\text{card}(L_i \cap N_j)}{\text{card}(M_j)} \right\} \cdot l_i \cdot n_j. \qquad (3)$$

The probability Pr(*f* | *g*) is given by [3]:

$$\Pr(f \mid g) = \sum_{ij} m_{ij}. \qquad (4)$$

For example, let *f* = {*1/a,0.7/b,0.2/c*} and *g* = {*0.2/a,1/b,0.7/c,0.1/d*} are defined on *X* = {*a,b,c,d,e* }, so that

m_f = {*a*}:0.3, {*a,b*}:0.5, {*a,b,c*}:0.2,
m_g = {*b*}:0.3, {*b,c*}:0.5, {*a,b,c*}:0.1,
{*a,b,c,d*}:0.1.

From the following matrix,

	0.3 {b}	0.5 {b,c}	0.1 {a,b,c}	0.1 {a,b,c,d}
0.3 {a}	0	0	0.01	0.00075
0.5 {a,b}	0.15	0.125	0.0333	0.025
0.2 {1,b,c}	0.06	0.1	0.02	0.015

the probability Pr(*f* | *g*)= 0.53905. It can be proved that Point Semantic Unification satisfies

$$\Pr(f \mid g) + \Pr(\bar{f} \mid g) = 1. \qquad (5)$$

Thus, Point Semantic Unification is considered as a conditional probability [3].

2.4 Interval Probability

Definition 2.4 An interval probability *IP*(*E*) can be interpreted as a scope of probability of event *E*, *P*(*E*), i.e *IP*(*E*)=[e_1,e_2] means $e_1 \le P(E) \le e_2$, where e_1 and e_2 are minimum and maximum probability of *E* respectively[2].

For example, given two probabilities *P*(*A*)=*a* and *P*(*A*)=*b* for event *A* and *B*, where $a,b \in [0,1]$.

Minimum probability of compound event 'A and B happen', $P(A \cap B)_{\min}$, is the least intersection between A and B, given by the following equation:

$$P(A \cap B)_{\min} = \max(0, a + b - 1).$$

Maximum probability of compound event 'A and B happen', $P(A \cap B)_{\max}$, is the most intersection between A and B, given by;

Thus interval probability of compound event 'A and B happened' is defined as

$$IP(A \cap B) = [\max(0, a + b - 1), \min(a, b)]. \qquad (6)$$

Similarly, minimum and maximum probability of compound event 'A or B happens', are max(a,b) and min(1,a+b) respectively.

Thus, interval probability of compound event 'A or B happened' is defined as:

$$IP(A \cup B) = [\max(a, b), \min(1, a + b)]. \qquad (7)$$

3. CONSTRUCTION MODEL OF SYSTEM

Definition 3.1 System is defined as $S(Er,X,Nm)$, where

Er: Number of entry data or number of record or respondent of system.

X: Domain or components of system, if there are n components then $X=(X_1,\ldots,X_n)$.

Nm: Name of system.

For example, given CAREER DATABASE in Table 1. By assuming that CAREER is a system which has 24 entries and three components, *education*, *age*, and *salary*, therefore Er=24, $X=(X_1{:}education, X_2{:}age, X_3{:}\ salary)$, Nm=CAREER. Now, we try to find relation among *education*, *age*, and *salary*.

Table 1. CAREER DATABASE

Name	Education	Age	Sallary
Nm-1	*MS*	*35*	*400,000*
Nm-2	*SHS*	*24*	*150,000*
Nm-3	*PhD*	*44*	*470,000*
Nm-4	*JHS*	*45*	*200,000*
Nm-5	*ES*	*35*	*125,000*
Nm-6	*SHS*	*37*	*250,000*
Nm-7	*MS*	*39*	*420,000*
Nm-8	*SHS*	*27*	*175,000*
Nm-9	*MS*	*45*	*415,000*
Nm-10	*SHS*	*56*	*275,000*
Nm-11	*N*	*60*	*100,000*
Nm-12	*JHS*	*33*	*300,000*
Nm-13	*BA*	*54*	*350,000*
Nm-14	*SHS*	*47*	*315,000*
Nm-15	*BA*	*41*	*355,000*
Nm-16	*SHS*	*21*	*150,000*
Nm-17	*BA*	*52*	*374,000*
Nm-18	*PhD*	*49*	*500,000*
Nm-19	*ES*	*58*	*125,000*
Nm-20	*JHS*	*59*	*200,000*
Nm-21	*BA*	*35*	*360,000*
Nm-22	*SHS*	*37*	*255,000*
Nm-23	*BA*	*31*	*340,000*
Nm-24	*SHS*	*29*	*250,000*

First, we classify all three domains or components as follows.

education = (*low_edu*,*mid_edu*,*hi_edu*),
age = (*about_20*,…,*about_60*),
salary = (*low_slr*,*mid_slr*,*hi_slr*).

where we assume that membership functions of *low_edu*, *mid_edu*, and *high_edu*:

$$m(low\_edu) = \{1/N, 0.8/ES, 0.5/JHS\},$$
$$m(mid\_edu) = \{0.2/ES, 0.5/JHS, 0.9/SHS, 0.2/BA\},$$
$$m(hi\_edu) = \{0.1/SHS, 0.8/BA, 1/MS, 1/PhD\}.$$

Membership function of *age*,

$$m(about\_n) = \{0.2/(n-4), 0.4/(n-3), 0.6/(n-2), 0.8/(n-1), 1/n, 0.8/(n+1), 0.6/(n+2), 0.4/(n+3), 0.2/(n+4)\}.$$

Membership function of *low_salary*, *mid_salary*, and *high_salary* :

$$m(low\_slr) = [1/0, 1/100000, 0/150000],$$
$$m(mid\_slr) = [0/100000, 1/150000, 1/250000, 0/300000],$$
$$m(hi\_slr) = [0/250000, 1/300000].$$

By using all membership functions above, we calculate and transform table 1 into table 2.

Table 2. CAREER FUZZY VALUE

Nama	*Education*			*Age*				*Salary*		
	LE	*ME*	*HE*	*20*	*25*	…	*60*	*LS*	*MS*	*HS*
Nm-1	0	0	1	0	0	...	0	0	0	1
Nm-2	0	0.9	0.1	0.2	0.8	...	0	0	1	0
Nm-3	0	0	1	0	0	...	0	0	0	1
Nm-4	0.5	0.5	0	0	0	...	0	0	1	0
Nm-5	0.8	0.2	0	0	0	...	0	0.5	0.5	0
Nm-6	0	0.9	0.1	0	0	...	0	0	1	0
Nm-7	0	0	1	0	0	...	0	0	0	1
Nm-8	0	0.9	0.1	0	0.6	...	0	0	1	0
Nm-9	0	0	1	0	0	...	0	0	0	1
Nm-10	0	0.9	0.1	0	0	...	0.2	0	0.5	0.5
Nm-11	1	0	0	0	0	...	1	1	0	0
Nm-12	0	0.9	0.1	0	0	...	0	0	0	1
Nm-13	0	0.2	0.8	0	0	...	0	0	0	1

Nm-14	0	0.9	0.1	0	0	...	0	0	0	1
Nm-15	0	0.2	0.8	0	0	...	0	0	0	1
Nm-16	0	0.9	0.1	0.8	0.2	...	0	0	1	0
Nm-17	0	0.2	0.8	0	0	...	0	0	0	1
Nm-18	0	0	1	0	0	...	0	0	0	1
Nm-19	0.8	0.2	0	0	0	...	0.6	0	0.5	0.5
Nm-20	0	0.9	0.1	0	0	...	0.8	0	1	0
Nm-21	0	0.2	0.8	0	0	...	0	0	0	1
Nm-22	0	0.9	0.1	0	0	...	0	0	0.9	0.1
Nm-23	0	0.2	0.8	0	0	...	0	0	0	1
Nm-24	0	0.9	0.1	0	0.2	...	0	0	1	0
$\sum$	3.1	10.9	10	1	1.8	...	2.6	1.5	9.4	13.1

Note: *LE:low_edu, ME:mid_edu, HE = hi_edu, 20:about_20, 25:about_25, …, LS=low_slr, MS:mid_slr* and *HS:hi_slr*.

Definition 3.2 X_n is defined as compound attribute to express component of the system . X_n is a vector. If there are k elements of X_n then $X_n=(x_{n1},\ldots,x_{nk})$, where x_{ni} is element i of compound attribute X_n and for further, x_{ni} is called attribute.
For example, if system CAREER has three compound attributes and their attributes as follows,

X_1 : *education* = (*low_edu,mid_edu,hi_edu*),
X_2 : *age* = (*about_20,...,about_60*),
X_3 : *salary* = (*low_slr,mid_slr,hi_slr*).

then x_{11} =*low_edu*, x_{25} =*about_40*, etc.

Definition 3.3 e_j^{ni} is defined as membership's value of entry j for attribute x_{ni}. If compound attribute X_n has k attributes then,

$$\forall j \sum_{1\le i\le k} e_j^{ni} = 1 \tag{8}$$

Example, as shown in Table 2., $e_4^{11} = 0.5,\ e_2^{12} = 0.9,$ etc.

Definition 3.4 $N(x_{ni})$ is defined as sum of entries value for attribute x_{ni}. If there are Er number of entries, then

$$\mathrm{N}(x_{ni}) = \sum_{1\le i\le Er} e_j^{ni} \tag{9}$$

If compound attribute X_n has k attributes then,

$$Er = \sum_{1\le i\le k} \mathrm{N}(x_{ni}) \tag{10}$$

For example, as shown in Table 2., $N(x_{ni})=N(low\_edu) = 3.1$.

Definition 3.5 $P(x_{ni})$ is defined as probability of attribute x_{ni} as follows.

$$\mathrm{P}(x_{ni}) = \frac{\mathrm{N}(x_{ni})}{Er} \tag{11}$$

If compound attribute X_n has k attributes then,

$$\sum_{1\le i\le k} \mathrm{P}(x_{ni}) = 1 \tag{12}$$

3.1 Relation Among Compound Attributes

Given three compound attributes, X_1, X_2 and X_3. Relation among them can be illustrated in Fig. 1. as follows.

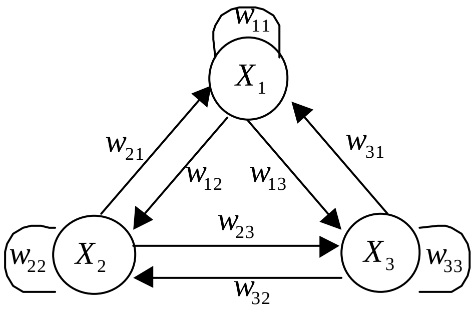


Figure 1. Relation Among Compound Attributes, X_1, X_2 and X_3.

Definition 3.6 w_{nm} is defined as *weight matrix*, to express degree of dependency of X_m from X_n. For a k-compound attribute X_n and a j-compound attribute X_m, w_{nm} and w_{mn} present two different matrices, as follows.

$$w_{nm} = \begin{bmatrix} a_{11}^{nm} & a_{12}^{nm} & \cdots & a_{1j}^{nm} \\ a_{21}^{nm} & a_{22}^{nm} & \cdots & a_{2j}^{nm} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1}^{nm} & a_{k2}^{nm} & \cdots & a_{kj}^{nm} \end{bmatrix},$$

$$w_{mn} = \begin{bmatrix} a_{11}^{mn} & a_{12}^{mn} & \cdots & a_{1k}^{mn} \\ a_{21}^{mn} & a_{22}^{mn} & \cdots & a_{2k}^{mn} \\ \vdots & \vdots & \ddots & \vdots \\ a_{j1}^{mn} & a_{j2}^{mn} & \cdots & a_{jk}^{mn} \end{bmatrix}.$$

Definition 3.7 Each element of matrix w_{nm}, entry a_{ih}^{nm} expresses numerical probabilistic value of relation from $x_{ni} \in X_n$

to $x_{mh} \in X_n$. a_{ih}^{nm} can also be interpreted as conditional probability as follows.

$$a_{ih}^{nm} = P(x_{ni} | x_{mh}) = \frac{P(x_{ni} \cap x_{mh})}{P(x_{mh})} \qquad (13)$$

If there are *Er* number of entries, then

$$a_{ih}^{nm} = \frac{\sum_{1 \le j \le Er} \min(e_j^{ni}, e_j^{mh})}{\sum_{1 \le j \le Er} e_j^{mh}} \qquad (14)$$

where $P(x_{ni} \cap x_{mh})$ express probability of entries which be inside x_{ni} and x_{mh}.

On the other hand, a_{hi}^{mn} expresses numerical probabilistic value of relation from $x_{mh} \in X_m$ to $x_{ni} \in X_n$. a_{hi}^{mn} can also be interpreted as conditional probability as follows.

$$a_{hi}^{mn} = P(x_{mh} | x_{ni}) = \frac{P(x_{ni} \cap x_{mh})}{P(x_{ni})} \qquad (15)$$

If there are *Er* number of entries, then

$$a_{hi}^{mn} = \frac{\sum_{1 \le j \le Er} \min(e_j^{ni}, e_j^{mh})}{\sum_{1 \le j \le Er} e_j^{ni}} \qquad (16)$$

From equations (13), (14) and (15), (16), we conclude that a_{hi}^{mn} and a_{ih}^{nm} are in general different.

The above definition leads to the conclusion that every attribute can be used to determine itself perfectly.

$$\forall X_n, x_{ni} \in X_n, \quad \frac{P(x_{ni} \cap x_{ni})}{P(x_{ni})}. \qquad (17)$$

If compound attribute X_n has *k* attributes, then,

$$w_{nn} = \begin{bmatrix} 1 & a_{12}^{nn} & \cdots & a_{1k}^{nn} \\ a_{21}^{nn} & 1 & \cdots & a_{2k}^{nn} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1}^{nn} & a_{k2}^{nn} & \cdots & 1 \end{bmatrix}.$$

3.2 Relation Among Attributes In System

Given three attributes, $x_{1u} \in X_1$, $x_{2v} \in X_2$ and $x_{3r} \in X_3$. Relation among these three attributes can be seen in Fig. 2.

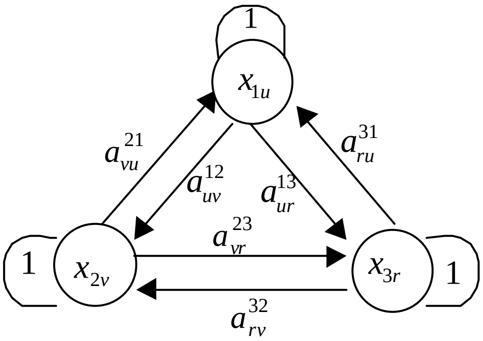
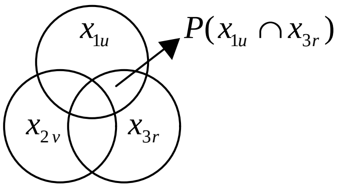


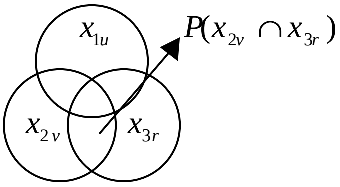
Figure 2. Relation Among Attributes, *x₁ᵤ*, *x₂ᵥ*, and *x₃ᵣ*.

In order to understand the meaning of this connection, we use relation of set.

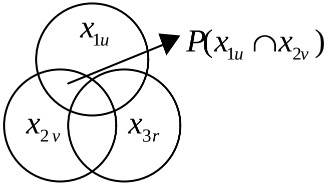


$$P(x_{1u} \cap x_{3r}) = \frac{P(x_{1u} \cap x_{3r})}{P(x_{1u})} \cdot P(x_{1u}) = a_{ru}^{31} \cdot P(x_{1u})$$
$$= \frac{P(x_{1u} \cap x_{3r})}{P(x_{3r})} \cdot P(x_{3r}) = a_{ur}^{13} \cdot P(x_{3r}).$$

Both $a_{ru}^{31} \cdot P(x_{1u})$ and $a_{ur}^{13} \cdot P(x_{3r})$, point to the same area or quantity, are intersection between x_{1u} and x_{3r}. In the same way, we can find two other relations, $a_{vr}^{23} \cdot P(x_{3r}) = a_{rv}^{32} \cdot P(x_{2v})$ and $a_{uv}^{12} \cdot P(x_{2v}) = a_{vu}^{21} \cdot P(x_{1u})$, which are proved as follows.



$$P(x_{2v} \cap x_{3r}) = \frac{P(x_{2v} \cap x_{3r})}{P(x_{2v})} \cdot P(x_{2v}) = a_{rv}^{32} \cdot P(x_{2v})$$
$$= \frac{P(x_{2v} \cap x_{3r})}{P(x_{3r})} \cdot P(x_{3r}) = a_{vr}^{23} \cdot P(x_{3r}).$$



$$P(x_{1u} \cap x_{2v}) = \frac{P(x_{1u} \cap x_{2v})}{P(x_{1u})} \cdot P(x_{1u}) = a_{vu}^{21} \cdot P(x_{1u})$$
$$= \frac{P(x_{1u} \cap x_{2v})}{P(x_{2v})} \cdot P(x_{2v}) = a_{uv}^{12} \cdot P(x_{2v}).$$

From the relations above, we find the following equation.

$$\frac{a^{12}_{uv}.a^{23}_{vr}}{a^{32}_{rv}} = \frac{a^{21}_{vu}.a^{13}_{ur}}{a^{31}_{ru}} \qquad (19)$$

Proof:

$$a^{12}_{uv} \cdot P(x_{2v}) = a^{21}_{vu} \cdot P(x_{1u}),$$

$$a^{12}_{uv} \cdot (a^{23}_{vr} \cdot \frac{P(x_{3r})}{a^{32}_{rv}}) = a^{21}_{vu} \cdot (a^{13}_{ur} \cdot \frac{P(x_{3r})}{a^{31}_{ru}})$$

$$a^{12}_{uv} \cdot \frac{a^{23}_{vr}}{a^{32}_{rv}} = a^{21}_{vu} \cdot \frac{a^{13}_{ur}}{a^{31}_{ru}}.$$

Important characteristic of relation among attributes is *transitive relation*, i.e. given $a^{12}_{uv}, a^{21}_{vu}, a^{23}_{vr}, a^{32}_{rv}$ and we would like to find interval value of a^{13}_{ur}, which satisfy the two following equations.

Lower bound of a^{13}_{ur},

$$a^{13}_{ur} \geq \max\{0,(a^{12}_{uv} + a^{32}_{rv} - 1)\}.\frac{a^{23}_{vr}}{a^{32}_{rv}}. \qquad (20)$$

Upper bound of a^{13}_{ur},

$$a^{13}_{ur} \leq \min\{a^{32}_{rv}, a^{12}_{uv}\}.\frac{a^{23}_{vr}}{a^{32}_{rv}} + \min\{(1-a^{21}_{vu}).\frac{a^{12}_{uv}}{a^{21}_{vu}},(1-a^{23}_{vr}).\frac{a^{32}_{rv}}{a^{23}_{vr}}\}.\frac{a^{23}_{vr}}{a^{32}_{rv}}. \qquad (21)$$

Proof :

To find the upper bound of a^{13}_{ur}, first we take the maximum area inside x_{2v}, result of intersection between two intersection areas which are intersection between x_{1u} and x_{2v}, expressed in a^{12}_{uv} and intersection between x_{3r} and x_{2v}, expressed in a^{32}_{rv}. The maximum area that is result of overlapping between the two intersection areas, shown in Fig. 3., can be expressed in min function applied to a^{12}_{uv} and a^{32}_{rv}. The next, we plus with maximum intersection between remain x_{1u} and x_{2v} which be outside of x_{2v}. Again, this area can be expressed in min function applied to$(1-a^{21}_{vu})$ and $(1-a^{23}_{vr})$. Value of these two area point to two different area, x_{1u} and x_{3r}. However, in order to be able to be compared, they must be point to the same area, in this case we use x_{2v} as base for their comparison. Therefore, we must convert them into x_{2v} by multiplying with $\frac{a^{12}_{uv}}{a^{21}_{vu}}$ and $\frac{a^{32}_{rv}}{a^{23}_{vr}}$, respectively. Finally, again we must convert all from x_{2v} into x_{3r} by multiplying with $\frac{a^{32}_{rv}}{a^{23}_{vr}}$.

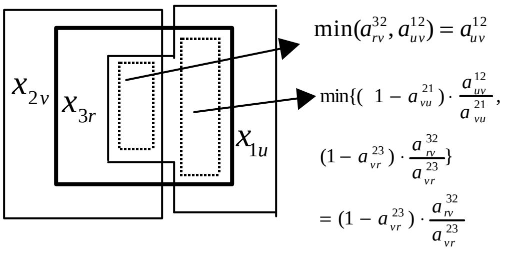

x_{2v}
x_{3r}
x_{1u}
$\min(a^{32}_{rv}, a^{12}_{uv}) = a^{12}_{uv}$
$\min\{(1 - a^{21}_{vu}) \cdot \frac{a^{12}_{uv}}{a^{21}_{vu}},$
$(1 - a^{23}_{vr}) \cdot \frac{a^{32}_{rv}}{a^{23}_{vr}}\}$
$= (1 - a^{23}_{vr}) \cdot \frac{a^{32}_{rv}}{a^{23}_{vr}}$

Figure 3. Maximum Area of Intersection between x_{1u} and x_{3r} inside x_{2v}.

To find the lower bound of a^{13}_{ur}, we take the minimum area inside x_{2v}, result of intersection between two intersection areas which are intersection between x_{1u} and x_{2v}, expressed in a^{12}_{uv} and intersection between x_{3r} and x_{2v}, expressed in a^{32}_{rv}. The minimum area which is result of as much as possible avoid overlapping between the two intersection areas, shown in Fig. 4., can be expressed in max function applied to a^{12}_{uv} and a^{32}_{rv} as shown in (13). The next, we convert quantity of the maximum area from x_{2v} into x_{3r} by multiplying with $\frac{a^{32}_{rv}}{a^{23}_{vr}}$.

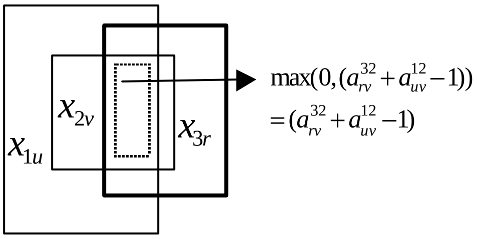

x_{1u}
x_{2v}
x_{3r}
$\max(0,(a^{32}_{rv}+a^{12}_{uv}-1))$
$=(a^{32}_{rv}+a^{12}_{uv}-1)$

Figure 4. Minimum Area of Intersection between x_{1u} and x_{3r} inside x_{2v}.

4. CALCULATING PREDICTION

After constructed the model of system, it can be used to predict interval probability (find lower and upper bound) of any query data. In this section, we generate formulas to calculate interval probability of the query data. First, user must give input related to data type of compound attributes.

Definition 4.1 Q is define as set of input data that be given by user to do query for a certain data. If there are *n* compound attributes then $Q=\{q_1,\ldots,q_n\}$ where q_i is data input related to compound attribute X_i.

For example, suppose CAREER system has been constructed, given John is *old* man and has *MS* degree as input for *age* and *education*, respectively, then q_1 = old and q_2=MS.

Definition 4.2 $P(X_i,q_i)$ is defined as probabilistic matching of compound attribute X_i toward given input data q_i. If there are *k* elements or attributes of compound attribute X_i, then,

$$P(X_i,q_i)=(p_{i1},\ldots,p_{ik}), \qquad (22)$$

where

$$p_{ij}= P(x_{ij} \mid q_i), \qquad (23)$$

expresses conditional probability for x_{ij} given q_i. In this case Point Semantic Unification Process [1,3] can be used to calculate p_{ij}.

For example, given q_i =old which is a *fuzzy set* defined as q_i ={0/55,1/60}.

X_i = age has 9 attributes as defined in section 2, as follows.

X_i = {*about_20,about_25,…,about_60*}.

By using *point semantic unification process* applying to membership function of *age* which has been defined in section 2 and membership function of q_i, we calculate $P(X_i,q_i)$ as follows. First, we calculate *the mass assignment* for q_i. It is equivalent to the basic probability assignment of Dempster Shafer Theory which we can write as

$$m_{q_i} = \{56,57,\ldots,60\}:0.2,\{57,\ldots 60\}:0.2,$$
$$\{58,59,60\}:0.2,\{59,60\}:0.2,\{60\}:0.2.$$

Next, i.e. *mass assignment* for x_{i8}*=55* as one attribute of X_i is given by

$$m_{x_{i8}} = \{51,\ldots,59\}:0.2,\{52,\ldots 58\}:0.2,$$
$$\{53,\ldots,57\ \}:0.2,\ \{54,55,56\}\ :0.2,\ \{55\}\ :0.2.$$

Process to calculate *Point Value Semantic Unification* of relation between two fuzzy set, *old* and *about_55* or P(*about_55,old*) is shown in the following table.

	0.2 {56,...,60}	0.2 {57,...,60}	0.2 {58,59,60}	0.2 {59,60}	0.2 {60}
0.2 {51,...,59}	0.032	0.03	0.026	0.02	0
0.2 {52,...,58}	0.024	0.02	0.013	0	0
0.2 {53,...,57}	0.016	0.01	0	0	0
0.2 {54,55,56}	0.008	0	0	0	0
0.2 {55}	0	0	0	0	0

From the table, we calculate

$$\begin{aligned} P(about\_55,old) &= 0.032+0.03+0.026+0.02+ \\ &\quad 0.024+0.02+0.013+0.016+ \\ &\quad 0.01+0.008 \\ &= 0.199. \end{aligned}$$

In the same way, we find *P(about_60,old) = 0.799,* where *P(about_20,old) = P(about_25,old) = ... = P(about_50,old) = 0,* because there is no intersection between their members. Finally, we find,

$$P(X_{i,}q_i) = P(age,old) = (0,0,0,0,0,0,0,0.199,0.799).$$

Definition 4.3 $P(X_{i,}q_j)$ is defined as probability of attribute $X_{i,}$ influenced by given input data q_j. X_i and q_j have different type of data, therefore to find their probabilistic matching, first, we must find $P(X_{j,}q_j)$ and then apply max-multiply (*) operation between $P(X_{j,}q_j)$ and w_{ji} as follows. If X_i has k attributes and X_j has s attributes then,

$$P(X_{i,}q_j) = P(X_{j,}q_j) * w_{ji} \quad (24)$$

$$= (p_{j1},\dots,p_{js}) * \begin{matrix} a^{ji}_{11} & a^{ji}_{12} & \dots a^{ji}_{1k} \\ a^{ji}_{21} & a^{ji}_{22} & \dots a^{ji}_{2k} \\ . & . & . \quad . \\ . & . & . \quad . \\ . & . & . \quad . \\ a^{ji}_{s1} & a^{ji}_{s2} & \dots a^{ji}_{sk} \end{matrix} \quad (25)$$

$$= (\max\{p_{j1}.a^{ji}_{11},\dots,p_{js}.a^{ji}_{s1}\},\dots, \quad (26)$$

$$\max\{\mathrm{p}_{j1}.a^{ji}_{1k},\dots,p_{js}.a^{ji}_{sk}\})$$

$$= (P(x_{i1},q_j),\dots,P(x_{ik},q_j)), \quad (27)$$

where $P(x_{ir,}q_j) = \max\{p_{j1}.a^{ji}_{1r},\dots,p_{js}.a^{ji}_{sr}\}$.

Definition 4.4 $P(x_{ir},Q)$, which is defined as probability of attribute x_{ir} influenced by given set input data Q, is $\vee$ operation for all probabilities of relation between x_{ir} and all members of Q. $\vee$ operation will be explained in the latter. If there are n members of Q, $\{(q_1,\dots,q_n)\}$, then,

$$P(x_{ir,}Q) = \underset{1\le j\le n}{\vee} P(x_{ir},q_j). \quad (28)$$

Definition 4.5 $P(X_{i,}Q)$ is defined as probability of compound attribute X_i influenced by given set input data Q. If there are n members of Q and k attributes of X_i, then

$$P(X_{i,}Q) = (P(x_{i1},Q),\dots,P(x_{ik},Q)), \quad (29)$$

$$P(X_{i,}Q) = (\underset{1\le j\le n}{} P(x_{i1},q_j),\dots, \underset{1\le j\le n}{} P(x_{ik},q_j)). \quad (30)$$

4.1 Calculating minimum probability truth of $P(x_{ir,}Q)$

Now, we generate formula for calculating minimum probability of attribute x_{ir} given $Q=\{q_1,\dots,q_n\}$, as input data. Related to (27), we defined minimum probability truth of $P(x_{ir},Q)$ as follows.

$$P_{\min}(x_{ir},Q) = \overset{\min}{\underset{1\le j\le n}{\bigvee}} P(x_{ir},q_j). \quad (31)$$

To simplify the problem, let's say that system just has three compound attributes, $X_1, X_2,$ and X_3 and their relation shown in Fig. 2. We calculate minimum probability truth of $x_{3r} \in X_3$ based on input $Q=\{q_1,q_2,q_3\}$..

$$P(x_{3r},Q)_{\min} = P(x_{3r},q_1) \vee_{\min} P(x_{3r},q_2) \vee_{\min} P(x_{3r},q_3).$$

We separate formula above into two parts. The first, we call *direct predicted probability of* x_{3r} which is $P(x_{3r},q_3) = P(x_{3r} \mid q_3) = p_{3r}$ and the second, we call *indirect predicted probability truth of* x_{3r} which is predicted from other attributes value, $P(x_{3r,}q_1) \vee_{\min} P(x_{3r,}q_2)$. The next, we compare both of them by applying max function as follows.

$$P(x_{3r,}Q)_{\min} = \max\{P(x_{3r,}q_1) \vee_{\min} P(x_{3r,}q_2), p_{3r})\}. \quad (32)$$

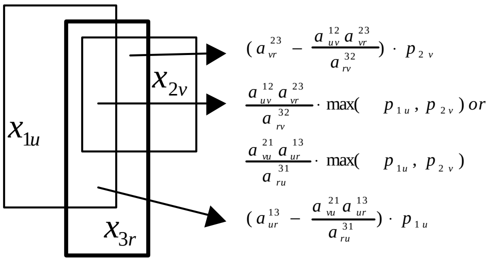
The problem now, is how to calculate $P(x_{3r,}q_i) \vee_{\min} P(x_{3r,}q_2) = d_{\min}$. i.e. X_1 has s attributes, X_2 has t attributes. Let's say that,

$$P(x_{3r,}q_1) = \max\{p_{11}.a^{13}_{1r},\dots,p_{1s}.a^{13}_{sr}\} = p_{1u}.a^{13}_{ur},$$

$$P(x_{3r,}q_2) = \max\{p_{21}.a^{23}_{1r},\dots,p_{2t}.a^{23}_{tr}\} = p_{2v}.a^{23}_{vr}.$$

We solve this problem by imaging interrelationship among x_{1u}, x_{2v}, and x_{3r} as shown in Fig. 2, in the following three conditions.

1. If $|(x_{1u} \cap x_{2v})| \le |(x_{1u} \cap x_{3r})|$ and $|(x_{1u} \cap x_{2v})| \le |(x_{2v} \cap x_{3r})|$, then $(x_{1u} \cap x_{2v})$ will be put in x_{3r} .



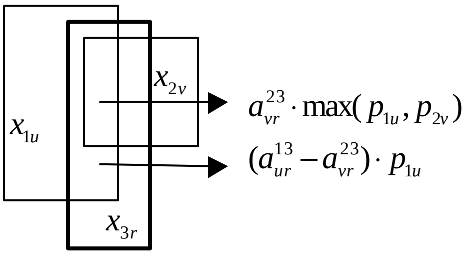
X_{1u}
X_{2v}
X_{3r}
$(a^{23}_{vr} - \frac{a^{12}_{uv} a^{23}_{vr}}{a^{32}_{rv}}) \cdot p_{2v}$
$\frac{a^{12}_{uv} a^{23}_{vr}}{a^{32}_{rv}} \cdot \max(p_{1u}, p_{2v})$ or
$\frac{a^{21}_{vu} a^{13}_{ur}}{a^{31}_{ru}} \cdot \max(p_{1u}, p_{2v})$
$(a^{13}_{ur} - \frac{a^{21}_{vu} a^{13}_{ur}}{a^{31}_{ru}}) \cdot p_{1u}$

$$\delta_{\min} = (a^{23}_{vr} - \frac{a^{12}_{uv} a^{23}_{vr}}{a^{32}_{rv}}) \cdot p_{2v} + (a^{13}_{ur} - \frac{a^{21}_{vu} a^{13}_{ur}}{a^{31}_{ru}}) \cdot p_{1u} + \frac{a^{12}_{uv} a^{23}_{vr}}{a^{32}_{rv}} \cdot \max(p_{1u}, p_{2v}).$$

or

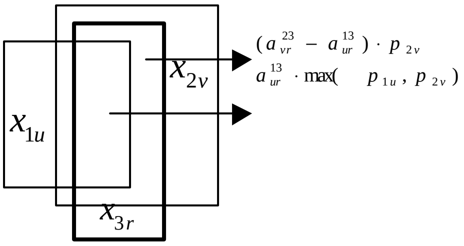
$$\delta_{\min} = (a^{23}_{vr} - \frac{a^{12}_{uv} a^{23}_{vr}}{a^{32}_{rv}}) \cdot p_{2v} + (a^{13}_{ur} - \frac{a^{21}_{vu} a^{13}_{ur}}{a^{31}_{ru}}) \cdot p_{1u} + \frac{a^{21}_{vu} a^{13}_{ur}}{a^{31}_{ru}} \cdot \max(p_{1u}, p_{2v})$$

2. If $|(x_{3r} \cap x_{2v})| \leq |(x_{1u} \cap x_{3r})|$ and $|(x_{3r} \cap x_{2v})| \leq |(x_{2v} \cap x_{1u})|$, then $(x_{2v} \cap x_{3r})$ will be put in $(x_{1u} \cap x_{3r})$.


x_{1u}
x_{2v}
x_{3r}
$a_{vr}^{23} \cdot \max(p_{1u}, p_{2v})$
$(a_{ur}^{13} - a_{vr}^{23}) \cdot p_{1u}$

$$\delta_{\min} = a_{vr}^{23} \cdot \max(p_{1u}, p_{2v}) + (a_{ur}^{13} - a_{vr}^{23}) \cdot p_{1u}$$

3. If $|(x_{1u} \cap x_{3r})| \leq |(x_{1u} \cap x_{2v})|$ and $|(x_{1u} \cap x_{3r})| \leq |(x_{2v} \cap x_{3r})|$, then $(x_{1u} \cap x_{3r})$ will be put in $(x_{2v} \cap x_{3r})$.


$(a_{vr}^{23} - a_{ur}^{13}) \cdot p_{2v}$
$a_{ur}^{13} \cdot \max(p_{1u}, p_{2v})$
x_{1u}
x_{2v}
x_{3r}

$$\ddot{a}_{min} = (a_{vr}^{23} - a_{ur}^{13}) \cdot p_{2v} + a_{ur}^{13} \cdot max(p_{1u}, p_{2v})$$

From the above conditions, we generate a formula that satisfy all conditions as follows.

$$\delta_{\min} = (a_{vr}^{23} - \min(\frac{a_{uv}^{12} a_{vr}^{23}}{a_{rv}^{32}}, a_{vr}^{23}, a_{ur}^{13})) \cdot p_{2v} +$$

$$(a_{ur}^{13} - \min(\frac{a_{uv}^{12} a_{vr}^{23}}{a_{rv}^{32}}, a_{vr}^{23}, a_{ur}^{13})).p_{1u} +$$

$$\min(\frac{a_{uv}^{12} a_{vr}^{23}}{a_{rv}^{32}}, a_{vr}^{23}, a_{ur}^{13}).\max(p_{1u}, p_{2v}). \qquad (33)$$

Finally, we find that $P_{\min}(x_{3r}, Q) = \max\{d_{\min}, P(x_{3r} | q_3)\}$.

4.2 Calculating Maximum Probability Truth of $P(x_{ir}, Q)$

Next, we generate formula for calculating maximum probability of attribute x_{ir} given $Q=(q_1, ... q_n)$, as input data. Related to (27), we defined maximum probability truth of $P(x_{ir}, Q)$ as follows.

$$P_{max}(x_{ir}, Q) = \overset{max}{\underset{1 \leq j \leq n}{\vee}} P(x_{ir}, q_j) \qquad (34)$$

To simplify the problem, let's say that system just has three compound attributes, X_1, X_2 and X_3 and their relation shown in Fig. 2.2. We calculate maximum probability truth of $x_{3r} \in X_3$ based on input $Q=(q_1, q_2, q_3)$.

$$P_{\max}(x_{3r}, Q) = P(x_{3r}, q_1) \vee_{\max} P(x_{3r}, q_2) \vee_{\max} P(x_{3r}, q_3)$$

We separate formula above into two parts. The first, we call *direct predicted probability of* x_{3r} which is $P(x_{3r} | q_3) = P(x_{3r} | q_3) = p_{3r}$ and the second, we call *indirect predicted probability truth of* x_{3r} which is predicted from other attributes value, $P(x_{3r}, q_1) \vee_{\max} P(x_{3r}, q_2)$. The next, we compare both of them by applying *min* function as follows.

$$P_{\max}(x_{3r}, Q) = \min\{1, (P(x_{3r}, q_1) \vee_{\max} P(x_{3r}, q_2)) + p_{3r})\} \qquad (35)$$

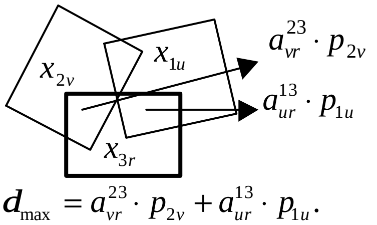
The problem now, is how to calculate $P(x_{3r}, q_2) = \ddot{a}_{max}$ .i.e. X_1 has s attributes, X_2 has t attributes. Let's say that,

$$P(x_{3r}, q_1) = max\{p_{11} \cdot a_{1r}^{13}, ..., p_{1s} \cdot a_{sr}^{13}\} = p_{1u} \cdot a_{ur}^{13}$$

$$P(x_{3r}, q_2) = max\{p_{21} \cdot a_{1r}^{23}, ..., p_{1t} \cdot a_{tr}^{23}\} = p_{2u} \cdot a_{vr}^{23}$$

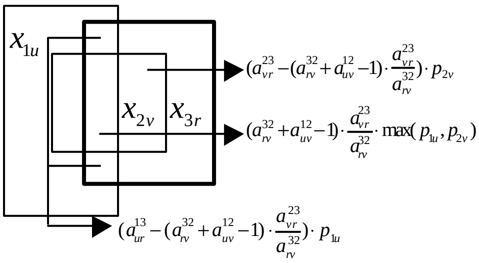
We solve this problem by imaging interrelationship among x_{1u}, x_{2v} and x_{3r} as shown in Fig. 2.2, in the following four conditions.

1. If $(a_{rv}^{32} + a_{uv}^{12} \leq 1)$ and $(a_{r1}^{31} + a_{vu}^{21} \leq 1)$ and $(a_{ur}^{13} + a_{vr}^{23} \leq 1)$ then



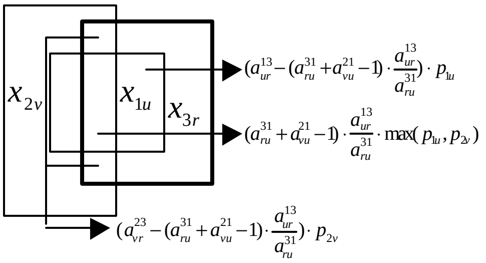
$$d_{\max} = a_{vr}^{23} \cdot p_{2v} + a_{ur}^{13} \cdot p_{1u}.$$

2. If $(a_{rv}^{32} + a_{uv}^{12} > 1)$ and $((a_{rv}^{32} + a_{uv}^{12} - 1).\frac{a_{vr}^{23}}{a_{rv}^{32}} > (a_{ru}^{31} + a_{vu}^{21} - 1).\frac{a_{ur}^{13}}{a_{ru}^{31}})$ and $((a_{rv}^{32} + a_{uv}^{12} - 1).\frac{a_{vr}^{23}}{a_{rv}^{32}} > (a_{ur}^{13} + a_{vr}^{23} - 1))$, then


x_{1u}
x_{2v}
x_{3r}
$(a^{23}_{vr}-(a^{32}_{rv}+a^{12}_{uv}-1)\cdot\frac{a^{23}_{vr}}{a^{32}_{rv}})\cdot p_{2v}$
$(a^{32}_{rv}+a^{12}_{uv}-1)\cdot\frac{a^{23}_{vr}}{a^{32}_{rv}}\cdot\max(p_{1u},p_{2v})$
$(a^{13}_{ur}-(a^{32}_{rv}+a^{12}_{uv}-1)\cdot\frac{a^{23}_{vr}}{a^{32}_{rv}})\cdot p_{1u}$

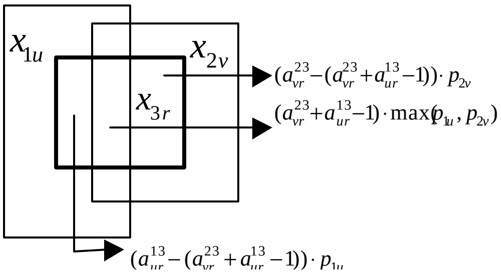
$$\delta_{\max}=(a^{23}_{vr}-(a^{32}_{rv}+a^{12}_{uv}-1)\cdot\frac{a^{23}_{vr}}{a^{32}_{rv}})\cdot p_{2v}+$$
$$(a^{13}_{ur}-(a^{32}_{rv}+a^{12}_{uv}-1)\cdot\frac{a^{23}_{vr}}{a^{32}_{rv}})\cdot p_{1u}+$$
$$(a^{32}_{rv}+a^{12}_{uv}-1)\cdot\frac{a^{23}_{vr}}{a^{32}_{rv}}\cdot\max(p_{1u},p_{2v})$$

3. If $(a^{31}_{ru}+a^{21}_{vu}>1)$ and $((a^{31}_{ru}+a^{21}_{vu}-1).\frac{a^{13}_{ur}}{a^{31}_{ru}}>$ $(a^{32}_{rv}+a^{12}_{uv}\text{ -}1).\frac{a^{23}_{vr}}{a^{32}_{rv}})$ and $((a^{31}_{ru}+a^{21}_{vu}\text{ -}1).\frac{a^{13}_{ur}}{a^{31}_{ru}}>$ $(a^{13}_{ur}+a^{23}_{vr}\text{ -}1))$, then


x_{2v}
x_{1u}
x_{3r}
$(a^{13}_{ur}-(a^{31}_{ru}+a^{21}_{vu}-1)\cdot\frac{a^{13}_{ur}}{a^{31}_{ru}})\cdot p_{1u}$
$(a^{31}_{ru}+a^{21}_{vu}-1)\cdot\frac{a^{13}_{ur}}{a^{31}_{ru}}\cdot\max(p_{1u},p_{2v})$
$(a^{23}_{vr}-(a^{31}_{ru}+a^{21}_{vu}-1)\cdot\frac{a^{13}_{ur}}{a^{31}_{ru}})\cdot p_{2v}$

$$\delta_{\max}=(a^{13}_{ur}-(a^{31}_{ru}+a^{21}_{vu}-1)\cdot\frac{a^{13}_{ur}}{a^{31}_{ru}})\cdot p_{1u}+$$
$$(a^{23}_{vr}-(a^{31}_{ru}+a^{21}_{vu}-1)\cdot\frac{a^{13}_{ur}}{a^{31}_{ru}})\cdot p_{2v}+$$
$$(a^{31}_{ru}+a^{21}_{vu}-1)\cdot\frac{a^{13}_{ur}}{a^{31}_{ru}}\cdot\max(p_{1u},p_{2v})$$

4. If $(a^{23}_{vr}+a^{13}_{ur}>1)$ and $((a^{23}_{vr}+a^{13}_{ur}-1).\frac{a^{13}_{ur}}{a^{31}_{ru}}>$ $(a^{31}_{ru}+a^{21}_{vu}\text{ -}1).\frac{a^{13}_{ur}}{a^{31}_{ru}})$ and $((a^{13}_{ur}+a^{23}_{vr}\text{ -}1)>$ $(a^{32}_{rv}+a^{12}_{uv}\text{ -}1).\frac{a^{23}_{vr}}{a^{32}_{rv}})$, then


x_{1u}
x_{2v}
x_{3r}
$(a^{23}_{vr}-(a^{23}_{vr}+a^{13}_{ur}-1))\cdot p_{2v}$
$(a^{23}_{vr}+a^{13}_{ur}-1)\cdot\max(p_{1u},p_{2v})$
$(a^{13}_{ur}-(a^{23}_{vr}+a^{13}_{ur}-1))\cdot p_{1u}$

$$\delta_{\max}=(a^{23}_{vr}-(a^{23}_{vr}+a^{13}_{ur}-1))\cdot p_{2u}+$$
$$(a^{13}_{ur}-(a^{23}_{vr}+a^{13}_{ur}-1))\cdot p_{1u}+$$
$$(a^{23}_{vr}+a^{13}_{ur}-1)\cdot\max(p_{1u},p_{2v}).$$

From the above conditions, we generate a formula that satisfy all condition as follows.

$$\delta_{\max}=(a^{23}_{vr}-\max(0,(a^{32}_{rv}+a^{12}_{uv}-1)\cdot\frac{a^{23}_{vr}}{a^{32}_{rv}},(a^{31}_{ru}+a^{21}_{vu}-1)\cdot$$
$$\frac{a^{13}_{ur}}{a^{31}_{ru}},(a^{23}_{vr}+a^{13}_{ur}-1)))\cdot p_{2v}+(a^{13}_{ur}-\max(0,(a^{32}_{rv}+a^{12}_{uv}-1)\cdot$$
$$\frac{a^{23}_{vr}}{a^{32}_{rv}},(a^{31}_{ru}+a^{21}_{vu}-1)\cdot\frac{a^{13}_{ur}}{a^{31}_{ru}},(a^{23}_{vr}+a^{13}_{ur}-1)))\cdot p_{1u}+$$
$$\max(0,(a^{32}_{rv}+a^{12}_{uv}-1)\cdot\frac{a^{23}_{vr}}{a^{32}_{rv}},(a^{31}_{ru}+a^{21}_{vu}-1)\cdot$$
$$\frac{a^{13}_{ur}}{a^{31}_{ru}},(a^{23}_{vr}+a^{13}_{ur}-1))\cdot\max(p_{1u},p_{2v}).\qquad(36)$$

Finally, we find that
$$P_{\max}(x_{3r},Q)=\max\{1,d_{\max}+P(x_{3r}\mid q_3)\}.$$

5. CONCLUSION

This paper proposed a method based on conditional probability relation to approximately calculate interval probability of dependency of data for data querying. Theoretically the formulation is quite interesting. However, it seems to be too complicated to calculate interaction of three or more components. Practically the formulas should be simplified, even though the accuracy of prediction may be decreased.

REFERENCES

1. Intan, R., Mukaidono, M., 'Application of Conditional Probability in Constructing Fuzzy Functional Dependency (FFD)', *Proceedings of AFSS'00*, 2000, pp.271-276.

2. Intan, R., Mukaidono, M., 'A proposal of Fuzzy Functional Dependency based on Conditional Probability', *Proceeding of FSS'00 (Fuzzy Systems Symposium)*, 2000, pp. 199-202.

3. Intan, R., Mukaidono, M., 'Fuzzy Functional Dependency and Its Application to Approximate Querying', *Proceedings of IDEAS'00*, 2000, pp.47-54.

4. Intan, R., Mukaidono, M., 'Conditional Probability Relations in Fuzzy Relational Database ', *Proceedings of RSCTC'00, LNAI 2005, Springer & Verlag,* 2000, pp.251-260.

5. Baldwin J.F., 'Knowledge from Data using Fril and Fuzzy Methods',*Fuzzy Logic*,John Wiley & Sons Ltd 1996, pp. 33–75.

6. Yukari Yamauchi, Masao Mukaidono , 'Interval and Paired Probabilities for Treating Uncertain Events', *The Institute of Electronics, Information and Communication Engineers*, Vol. E82-D (May 1999), pp. 955–961.

7. Baldwin J.F., Martin T.P., and Pilsworth B.W. *FRIL-Fuzzy and Evidential Reasoning in AI*, Research Studies Press and Willey, 1995.

8. Shafer G. *A Mathematical Theory of Evidence*, Princeton Univ. Press. 1976.

9. Richard Jeffrey, 'Probabilistic Thinking', Priceton University, 1995.

10. Baldwin J.F., Martin T.P., 'A Fuzzy Data Browser in Fril', *Fuzzy Logic*, John Wiley & Sons Ltd. 1996, pp. 101-123.