STUDIA INFORMATICA 2010
Volume 31 Number 2A (89)

Krzysztof MYSZKOROWSKI
Politechnika Łódzka, Instytut Informatyki

ZALEŻNOŚCI FUNKCYJNE MIĘDZY ATRYBUTAMI ROZMYTYMI POZIOMU DRUGIEGO Z INTERWAŁOWĄ FUNKCJĄ PRZYNALEŻNOŚCI

Streszczenie. Przedmiotem artykułu są zależności funkcyjne w rozmytych bazach danych. Do opisu atrybutów zostały wykorzystane interwałowe zbiory rozmyte poziomu drugiego. Zastosowano podejście posybilistyczne. Stopień rozmytej zależności funkcyjnej został określony na podstawie miar bliskości wartości atrybutów. Zależności takie podlegają rozszerzonym regułom Armstronga. Sformułowano definicje rozmytych postaci normalnych.

Słowa kluczowe: interwałowe zbiory rozmyte, zbiory rozmyte poziomu drugiego, rozkłady możliwości, rozmyte zależności funkcyjne

FUNCTIONAL DEPENDENCIES IN DATABASES WITH ATTRIBUTES REPRESENTED BY INTERVAL-VALUED LEVEL-2 FUZZY SETS

Summary. The paper deals with functional dependencies in fuzzy databases. Attribute values are represented by means of interval-valued level-2 fuzzy sets. A possibility-based approach has been applied. The degree of a fuzzy functional dependency has been determined by means of closeness measures of attribute values. Such dependencies are submitted to extended Armstrong's axioms. Definitions of fuzzy normal forms have been formulated.

Keywords: interval-valued fuzzy sets, level-2 fuzzy sets, possibility distributions, fuzzy functional dependencies

1. Wstęp

Jednym z zadań systemów zarządzania bazą danych jest zapewnienie zachowania zgodności gromadzonych informacji. Oznacza to spełnienie przez dane narzuconych warunków nazywanych więzami integralności. Baza danych pozostaje wtedy w stanie zgodnym. Szczególnie istotnymi warunkami integralności są zależności między atrybutami, a wśród nich zależności funkcyjne. Muszą być one uwzględniane w procesie projektowania. Zależność funkcyjna $X \rightarrow Y$ między atrybutami X i Y relacji R oznacza, że każdej wartości atrybutu X odpowiada dokładnie jedna wartość atrybutu Y. Jest to podstawowe pojęcie procesu normalizacji schematów relacji. Identyfikacja zależności funkcyjnych jest możliwa przy założeniu, że wartości atrybutów są precyzyjne. Przy braku precyzji można jedynie mówić o pewnym stopniu występowania takiej zależności.

Jednym z rozwiązań umożliwiających przedstawienie niekompletnej informacji jest zastosowanie teorii zbiorów rozmytych [17]. Zostały opracowane różne wersje rozmytych modeli danych [5, 9]. Uwzględniono w nich również zależności między atrybutami. Ze względu na znaczenie pojęcia zależności funkcyjnej jego rozszerzenie było przedmiotem wielu prac [3, 10, 11, 12, 14, 16]. Zaproponowano różne definicje rozmytej zależności funkcyjnej. Każda z nich musiała być zgodna z zastosowanym przez jej autora podejściem do modelowania rozmytych baz danych. W większości prac stosowano klasyczne zbiory rozmyte, w których każdemu elementowi była przyporządkowana liczba określająca jego przynależność do zbioru. W pracy [13] zaproponowano zastosowanie zbiorów rozmytych poziomu drugiego, w których liczba z przedziału [0,1] jest przyporządkowana klasycznemu zbiorowi rozmytemu. Dalszym rozszerzeniem jest wyrażenie przynależności za pomocą przedziału zawartego w [0,1]. Zbiory takie będą nazywane interwałowymi zbiorami rozmytymi poziomu drugiego.

Zależności funkcyjne między tak określonymi atrybutami są przedmiotem niniejszego artykułu. W punkcie drugim zdefiniowano interwałowe zbiory rozmyte poziomu drugiego oraz podano miary pozwalające ocenić stopień ich równości. Definicja rozmytej zależności funkcyjnej została sformułowana w punkcie trzecim. Stanowi ona rozszerzenie definicji podanej w pracy [3] dla atrybutów opisanych za pomocą klasycznych zbiorów rozmytych. Na jej podstawie zdefiniowano również pojęcie (α, β)-klucza relacji. Tematem punktu czwartego jest normalizacja schematów relacji. Zostały sformułowane rozszerzone definicje postaci normalnych, w których wykorzystano opisane wcześniej pojęcia.

2. Interwałowe zbiory rozmyte poziomu drugiego

W klasycznej teorii zbiorów można zdefiniować funkcję charakterystyczną określającą przynależność elementu e do zbioru A. Jest ona odwzorowaniem $\mathfrak{R} \rightarrow \{0,1\}$, gdzie $\mathfrak{R}$ jest obszarem odniesienia (uniwersum). Przyjmuje wartość 1, jeżeli $e \in A$ oraz 0 w przypadku przeciwnym. Zbiór rozmyty jest uogólnieniem zbioru klasycznego. Zamiast funkcji charakterystycznej w definicji występuje funkcja przynależności, stanowiąca odwzorowanie $\mathfrak{R} \rightarrow [0,1]$.

Zbiór dwuelementowy został zastąpiony przedziałem [0,1]. Określone w taki sposób zbiory są nazywane klasycznymi zbiorami rozmytymi lub zbiorami rozmytymi pierwszego typu.

Definicja 1. Niech $\Re$ oznacza obszar odniesienia (uniwersum). Zbiorem rozmytym A elementów obszaru $\Re$ o funkcji przynależności μA (x) nazywamy zbiór uporządkowanych par:

$$A = \{< x, \mu A\,(x) >: x \in \boldsymbol{\Re},\ \mu A\,(x): \boldsymbol{\Re} \rightarrow [0,1]\ \} \tag{1}$$

Dalszą generalizację można uzyskać przez odpowiednią modyfikację pary $<x,\ \mu_A\,(x)>$. Modyfikacja $\mu_A\,(x)$ prowadzi do zbiorów rozmytych typu drugiego [4]. Oznaczmy przez Int ([0,1]) zbiór zamkniętych podprzedziałów przedziału [0,1], czyli Int ([0,1]) = $\{[a, b]: a, b \in [0,1]\}$. Przez zastosowanie odwzorowania $\Re \rightarrow$ Int ([0,1]) zamiast $\Re \rightarrow [0,1]$ otrzymuje się możliwość bardziej elastycznego określenia przynależności do zbioru. Każdemu elementowi x zostaje przyporządkowany przedział $[\mu_{A_L}(x),\ \mu_{A_U}(x)]$ zawarty w [0,1], wyznaczający granice dla wartości stopnia przynależności. W dalszych rozważaniach przyjmiemy, że przedziały są uporządkowane według następującej reguły [1]:

$$[a_L, a_U] \le [b_L, b_U] \Leftrightarrow a_L \le b_L \text{ i } a_U \le b_U \tag{2}$$

Każda wartość należąca do przedziału przynależności jest w pełni możliwa (w stopniu 1). Wystąpienie pozostałych wartości jest całkowicie niemożliwe. Tak zdefiniowany zbiór jest nazywany interwałowym zbiorem rozmytym typu drugiego [6, 8, 15].

Definicja 2. Niech $\Re$ oznacza obszar odniesienia (uniwersum). Interwałowy zbiór rozmyty typu drugiego Ā elementów obszaru $\Re$ definiuje się następująco:

$$\bar{A} = \{< x, \mu\bar{A}\,(x) >: x \in \Re,\ \mu\bar{A}\,(x) = [\mu\bar{A}L(x),\ \mu\bar{A}U(x)],$$
$$\mu\bar{A}L(x),\ \mu\bar{A}U(x) : \Re \rightarrow [0,1]\ \},$$

gdzie μĀ (x) jest przedziałem przynależności elementów obszaru $\Re$ do zbioru Ā.

Inna generalizacja polega na zastąpieniu w definicji 1 elementu x zbiorem rozmytym określonym na $\Re$. Otrzymuje się wtedy definicję zbioru rozmytego poziomu drugiego [13]. Jego elementami są więc klasyczne zbiory rozmyte.

Definicja 3. Niech $\Re$ oznacza obszar odniesienia (uniwersum). Zbiorem rozmytym poziomu drugiego $\breve{A}$ elementów obszaru $\Re$ o funkcji przynależności $\mu_{\breve{A}}$ nazywamy zbiór uporządkowanych par:

$$\breve{A} = \{< A, \mu_{\breve{A}}(A) >:\ \ \mu_{\breve{A}} : \Psi(\Re) \rightarrow [0,1], A \in \Psi(\Re)\}, \tag{4}$$

gdzie $\Psi(\Re)$oznacza rodzinę zbiorów rozmytych pierwszego typu określonych na $\boldsymbol{\Re}$.

Tematyka artykułu dotyczy zbiorów, w których zostały połączone opisane wyżej modyfikacje. Są to interwałowe zbiory rozmyte poziomu drugiego. Ich elementami są zbiory rozmyte typu pierwszego, których przynależność jest określona za pomocą podprzedziału zawartego w [0,1].

Definicja 4. Niech $\mathfrak{R}$ oznacza obszar odniesienia (uniwersum). Interwałowym zbiorem rozmytym poziomu drugiego Â elementów obszaru $\mathfrak{R}$ o funkcji przynależności μÂ nazywamy zbiór uporządkowanych par:

$$\hat{A} = \{< A, \mu\hat{A}\,(A) >:\ \ \mu\hat{A}\,(x) = [\mu\hat{A}L(A), \mu\hat{A}U(A)],$$

$$\mu\hat{A}L(A), \mu\hat{A}U(A) : \Psi\,(\mathfrak{R}) \rightarrow [0,1],\ \ A \in \Psi\,(\mathfrak{R})\} \qquad (5),$$

gdzie $\Psi\,(\mathfrak{R})$ oznacza rodzinę zbiorów rozmytych pierwszego typu określonych na $\mathfrak{R}$.

Za pomocą tak zdefiniowanych zbiorów można opisać różnego rodzaju niekompletne informacje.

Przykład 1. Określenie preferencji dotyczących wieku kandydata na określone stanowisko w przedsiębiorstwie:

$\hat{A} = \{ [0.4, 0.6]/A_{\text{około\_30}}, [1, 1]/A_{\text{około\_35}}, [0.7, 0.8]/A_{\text{około\_40}}\}$,

gdzie $A_{\text{około\_30}}$, $A_{\text{około\_35}}$, oraz $A_{\text{około\_40}}$ są odpowiednio zdefiniowanymi zbiorami rozmytymi pierwszego typu. Należą one do liczb rozmytych – zbiorów rozmytych określonych na zbiorze liczb rzeczywistych [18].

Przykład 2. Określenie preferencji dotyczących znajomości języków obcych kandydata przyjmowanego do przedsiębiorstwa:

Â = { [1, 1]/AEF, [0.7, 0.9]/AED, [0.5, 0.7]/AES},

gdzie AEF, AED oraz AES są następującymi zbiorami rozmytymi:

AEF = {1/E, 0.8/F}, AED = {1/E, 0.8/D}, AES = {1/E, 0.8/S}.

Liczby występujące w definicjach zbiorów określają żądany poziom znajomości danego języka (E – angielski, D – niemiecki, S – hiszpański). Zauważmy, że konstrukcja tych zbiorów zawiera semantykę typu *and*. Oznacza to, że wymagana jest znajomość każdego z wymienionych języków. Inna sytuacja wystąpiła w poprzednim przykładzie, gdzie możliwa była tylko jedna wartość określająca wiek kandydata.

Zawarte w definicji zbioru liczby mogą również oznaczać ocenę możliwości wystąpienia określonego zdarzenia. Tak więc zbiór $\hat{A}$ z przykładu 1 odnoszący się do pewnej osoby oznacza, że jest całkowicie możliwe, iż jej wiek wynosi około 35 lat. Pozostałe warianty (około 30 i około 40) uzyskały niższą ocenę.

W dalszych rozważaniach będziemy przyjmować tę drugą interpretację. Oznacza to założenie, że wartość atrybutu X w krotce t ma postać interwałowego rozkładu możliwości poziomu drugiego $\prod_X$:

$$t\,[X] = \prod_X = \{\pi_X(A)\,/\,A\colon\ \pi_X(A) = [\pi_{X_L}(A), \pi_{X_U}(A)],$$
$$\pi_{X_L}(A), \pi_{X_U}(A)\colon\ \Psi(\mathfrak{R}) \rightarrow [0, 1],\ A \in \Psi(\mathfrak{R})\ \}, \qquad (6)$$

gdzie $\pi_X(A)$ jest przedziałem zawartym w [0, 1]. Jego granice $\pi_{X_L}(A)$ i $\pi_{X_U}(A)$ oznaczają miary możliwości, odpowiednio dolną i górną, przyjęcia przez X wartości zdefiniowanej za pomocą zbioru rozmytego A.

Porównanie wartości określonych za pomocą rozkładów możliwości $\prod_1$ i $\prod_2$ wymaga wyznaczenia dwóch miar: możliwości *Poss* ($\prod_1 = \prod_2$) i konieczności *Nec* ($\prod_1 = \prod_2$). Stanowią one ocenę stopnia równości porównywanych rozkładów. Przy interwałowych rozkładach możliwości miary te mają postać przedziałów: $Poss = [Poss_L, Poss_U]$ i $Nec = [Nec_L, Nec_U]$. Ich granice wyrażają się wzorami:

$$Poss\,(\prod_X = \prod_Y)_L = \sup_x \min(\pi_{X_L}(x), \pi_{Y_L}(x)), \tag{7}$$

$$Poss\,(\prod_X = \prod_Y)_U = \sup_x \min(\pi_{X_U}(x), \pi_{Y_U}(x)), \tag{8}$$

$$Nec((\prod_X = \prod_Y)_L = 1 - \sup_{x\neq y} \min(\pi_{X_U}(x), \pi_{Y_U}(y)), \tag{9}$$

$$Nec((\prod_X = \prod_Y)_U = 1 - \sup_{x\neq y} \min(\pi_{X_L}(x), \pi_{Y_L}(y)). \tag{10}$$

Powyższe wzory, podane w pracy [7], stanowią modyfikację ocen stopnia równości klasycznych rozkładów możliwości. Dla zbiorów rozmytych poziomu drugiego wyrażenia te powinny zostać zmodyfikowane w taki sposób, by można było uwzględnić bliskość zbiorów rozmytych występujących w ich definicji. W przeciwnym razie dokonana ocena nie będzie właściwa.

Przykład 3. Rozważmy dwa interwałowe rozkłady możliwości poziomu drugiego określone na $\mathfrak{R} = \{a, b, c, d\}$:

$\hat{A}1 = \{[1, 1]/A, [0.1, 0.2]/B\}$ oraz $\hat{A}2 = \{[0.2, 0.3]/B, [1, 1]/C\}$, gdzie
$A = \{1/c, 0.7/d\}, B = \{1/a, 0.7/b\}, C = \{0.8/c, 1/d\}$

Na podstawie wzorów (7-10) otrzymujemy:

$Poss\,(\hat{A}1 = \hat{A}2)_L = \sup \min(0.1, 0.2) = 0.1$
$Poss\,(\hat{A}1 = \hat{A}2)_U = \sup \min(0.2, 0.3) = 0.2$
$Nec\,(\hat{A}1 = \hat{A}2)_L = 1 - \sup\,(\min(1, 0.3), \min\,(1, 1), \min\,(0.2, 1)) = 0$
$Nec\,(\hat{A}1 = \hat{A}2)_U = 1 - \sup\,(\min(1, 0.2), \min\,(1, 1), \min\,(0.1, 1)) = 0$

Ocena równości obydwóch rozkładów jest więc dość niska, a przecież widać, że wystąpienie w dużym stopniu sobie równych wartości określonych zbiorami *A* i *C* jest w obydwóch rozkładach całkowicie możliwe.

Przy porównywaniu zbiorów rozmytych stosuje się pojęcie miary równości.

Zaproponowano różnej jej konstrukcje. Jedną z nich jest wysokość iloczynu porównywanych zbiorów *A* i *B*, czyli

$$\approx (A, B) = \sup_x \min\,(\mu_A(x), \mu_B(x)). \tag{11}$$

Dla zbiorów *A* i *C* z przykładu 3 wartość $\approx (A, C)$ wynosi:

$$\approx (A, C) = \max(\min(1, 0.8), \min\,(0.7, 1)) = 0.8.$$

Wartości miar równości zbiorów rozmytych występujących w definicji zbioru rozmytego poziomu drugiego są elementami relacji bliskości R_c. Dla zbiorów z przykładu 3 relacja bliskości została przedstawiona w tabeli 1.

Tabela 1

Relacja bliskości – przykład 1

	A	B	C
A	1	0	0.8
B	0	1	0
C	0.8	0	1

Relacja bliskości stanowi uzupełnienie opisu zbioru rozmytego poziomu drugiego. Jej uwzględnienie przy ocenie równości interwałowych rozkładów możliwości poziomu drugiego prowadzi do następującej modyfikacji wzorów (7-10):

$$Poss\ (\prod_X = \prod_Y)_L = \sup_{x,y} \min(\mu_{R_c}(x, y),\ \pi_{X_L}(x),\ \pi_{Y_L}(y)), \quad (12)$$

$$Poss\ (\prod_X = \prod_Y)_U = \sup_{x,y} \min(\mu_{R_c}(x, y),\ \pi_{X_U}(x),\ \pi_{Y_U}(x)), \quad (13)$$

$$Nec((\prod_X = \prod_Y)_L = 1 - \sup_{x \neq y} \min(1 - \mu_{R_c}(x, y),\ \pi_{X_U}(x),\ \pi_{Y_U}(y)), \quad (14)$$

$$Nec((\prod_X = \prod_Y)_U = 1 - \sup_{x \neq y} \min(1 - \mu_{R_c}(x, y),\ \pi_{X_L}(x),\ \pi_{Y_L}(y)). \quad (15)$$

Dla rozkładów z przykładu 3 otrzymujemy:

$Poss\ (\hat{A}1 = \hat{A}2)_L = \sup \min(0.8, 1, 1), \min\ (1, 0.1, 0.2)) = 0.8$

$Poss\ (\hat{A}1 = \hat{A}2)_U = \sup \min(0.8, 1, 1), \min\ (1, 0.2, 0.3)) = 0.8$

$Nec\ (\hat{A}1 = \hat{A}2)_L = 1 - \sup\ (\min(1, 1, 0.3), \min\ (0.2, 1, 1), \min\ (1, 0.2, 1)) = 0.7$

$Nec\ (\hat{A}1 = \hat{A}2)_U = 1 - \sup\ (\min(1, 1, 0.2), \min\ (0.2, 1, 1), \min\ (1, 0.1, 1)) = 0.8$

Zastosowanie zmodyfikowanych wzorów pozwoliło uzyskać wyniki bardziej zbliżone do intuicyjnej oceny równości obydwóch rozkładów. W rozkładach tych występują podobnie zdefiniowane zbiory rozmyte *A* i *C*, którym odpowiadają wysokie wartości miar możliwości.

3. Rozmyte zależności funkcyjne

Pojęcie rozmytej zależności funkcyjnej było przedmiotem zainteresowania wielu autorów. W najmniej złożonym rozmytym modelu danych uwzględniono tylko możliwość niepełnej przynależności krotki *t* do relacji *R* [14]. Wartości atrybutów są precyzyjne. Stopień przynależności $\mu_R(t)$ krotki wyraża się liczbą z przedziału [0,1]. Przy takich założeniach zależność $X \to Y$ między atrybutami relacji *R* dotyczy nośnika *R*, czyli zbioru krotek, które należą do niej w stopniu większym od 0: supp(R) = $\{t\colon\ \mu_R(t) > 0\}$. W modelu wykorzystującym podobieństwo elementów dziedzin atrybutów dokonuje się ich α-podziału na klasy równoważności [11]. Wartości atrybutu *X* w krotkach *t* i *t'* są α-redundantne, jeżeli należą do tej samej klasy. Im wyższa wartość α, tym podział jest bardziej dokładny. Tak więc wartości α_1-redundantne nie muszą być α_2-redundantne, jeżeli $\alpha_1 < \alpha_2$. Zależność $X \to_{(\alpha_X,\ \alpha_Y)} Y$ istnieje, jeżeli α_X-redundantne wartości atrybutu *X* w krotkach *t* i *t'* implikują w nich α_Y-redundantne wartości atrybutu *Y*.

W modelach posybilistycznych stosuje się miary równości wartości atrybutów wyrażonych za pomocą rozkładów możliwości. W proponowanych definicjach wymaga się, by dla każdych dwóch krotek t i t' był spełniony warunek $c(t[Y], t'[Y]) \geq c(t[X], t'[X])$, gdzie c oznacza stopień bliskości. W pracy [2] zastosowano rozmyty implikator Gödela I, co pozwoliło określić stopień zależności. Zależność $X \to Y$ istnieje w stopniu θ, jeżeli

$$\min_{t,t'} (I (c(t[X], t'[X]), c(t[Y], t'[Y])) \geq \theta. \tag{16}$$

Przy niskich wartościach $c(t[X], t'[X])$ i $c(t[Y], t'[Y])$ takich, że $c(t[X], t'[X]) \geq c(t[Y], t'[Y])$, otrzymuje się niską ocenę poziomu zależności, co jest wadą definicji. W pracy [3] zaproponowano wyrażenie poziomu zależności za pomocą dwóch liczb oznaczających stopień równości wartości atrybutów X i Y. Przy interwałowych rozkładach możliwości liczby te muszą zostać zastąpione przedziałami. Modyfikacja propozycji zawartej w [3] prowadzi do następującej definicji:

Definicja 5. Niech R $(X_1, X_2, \ldots, X_n)$ będzie schematem relacyjnym oraz niech X i Y będą atrybutami złożonymi: $X, Y \subseteq S$, gdzie $S = \{X_1, X_2, \ldots, X_n\}$. Niech $Poss(X) = [Poss_L(X), Poss_U(X)]$ oznacza przedział miary możliwości tego, że wartości atrybutu X w krotkach t i t' są sobie równe. Atrybut Y jest funkcyjnie zależny od atrybutu X w stopniu (α, β), co oznaczamy przez $X \to_{\alpha,\beta} Y$, gdzie α i β są przedziałami zawartymi w [0, 1] wtedy i tylko wtedy, gdy dla każdej relacji o schemacie R jest spełniony następujący warunek:

$$[c_L(X), c_U(X)] \geq \alpha \Rightarrow [c_L(Y), c_U(Y)] \geq \beta, \tag{17}$$

gdzie $c_L(X) = \min_{t,t'} Poss_L(X)$, $c_U(X) = \min_{t,t'} Poss_U(X)$, $c_L(Y) = \min_{t,t'} Poss_L(Y)$, $c_L(Y) = \min_{t,t'} Poss_U(Y)$.

Tabela 2

Relacja PZ – przykład 4

	P	Z
t	$[1,1]/P_{1200}$	$[1,1]/Z_8$
t	$\{[0.7, 0.8]/P_{1200}, [1,1]/P_{1300}\}$	$\{[0.6, 0.7]/Z_8, [1,1]/Z_{8,5}\}$
t	$\{[1,1]/P_{1500}, [0.5, 0.6]/P_{1600}\}$	$\{[0.5, 0.6]/Z_{9,5}, [1,1]/Z_{10}\}$
t	$\{[0.6, 0.7]/P_{1500}, [1,1]/P_{1600}\}$	$\{[0.7, 0.8]/Z_{10}, [1,1]/Z_{10,5}\}$

Przykład 4. Relacja PZ przedstawia związek między pojemnością samochodu i zużyciem paliwa. Wartości atrybutów zostały przedstawione za pomocą interwałowych zbiorów rozmytych poziomu drugiego. Ich elementami są trapezoidalne liczby rozmyte. Załóżmy, że odpowiadające im wyrazy macierzy bliskości przyjmują następujące wartości:

$R_c(P_{1200}, P_{1300}) = 0.8$, $R_c(P_{1500}, P_{1600}) = 0.9$, $R_c(Z_8, Z_{8.5}) = 0.5$, $R_c(Z_{9.5}, Z_{10}) = 0.5$, $R_c(Z_{10}, Z_{10.5}) = 0.5$.

Atrybuty relacji PZ spełniają zależność $P \rightarrow_{[0.7,\ 0.8],[0.6,\ 0.7]} Z$. Na podstawie (12-15) mamy bowiem: Poss $_{t_1,\ t_2}$(P) = [0.7, 0.8], Poss $_{t_3,\ t_4}$(P) = [0.9, 0.9], Poss $_{t_1,\ t_2}$(Z) = [0.6, 0.7], Poss $_{t_3,\ t_4}$(Z) = [0.7, 0.8].

Na podstawie znajomości pewnych zależności funkcyjnych można uzyskać inne. Oznaczmy przez F zbiór rozmytych zależności funkcyjnych dla schematu R. Jego domknięcie (zbiór wszystkich zależności funkcyjnych wynikających z F) oznaczymy przez F^+. W klasycznych bazach danych podstawą procesu wnioskowania są reguły zwane aksjomatami Armstronga. Dla rozmytych zależności funkcyjnych aksjomaty te muszą być odpowiednio rozszerzone [3]:

A1: $Y \subseteq X \Rightarrow X \rightarrow_{\alpha_X,\, \beta_Y} Y$ dla dowolnych α_X i β_X (zwrotność),

A2: $X \rightarrow_{\alpha_X,\, \beta_Y} Y \Rightarrow XZ \rightarrow_{\alpha_{XZ},\, \beta_{YZ}} YZ$ (rozszerzanie),

A3: $X \rightarrow_{\alpha_X,\, \beta_Y} Y \wedge Y \rightarrow_{\beta_Y,\, \gamma_Z} Z \Rightarrow X \rightarrow_{\alpha_X,\, \gamma_Z} Z$ (przechodniość).

Z aksjomatów Armstronga wynikają następujące reguły:

D1: $X \rightarrow_{\alpha_X,\, \beta_{YZ}} YZ \Rightarrow X \rightarrow_{\alpha_X,\, \beta_Y} Y \wedge X \rightarrow_{\alpha_X,\, \beta_Z} Z$,

D2: $X \rightarrow_{\alpha_X,\, \beta_Y} Y \wedge X \rightarrow_{\alpha_X,\, \beta_Z} Z \Rightarrow X \rightarrow_{\alpha_X,\, \beta_{YZ}} YZ$,

D3: $X \rightarrow_{\alpha_X,\, \beta_Y} Y \wedge YW \rightarrow_{\beta_{YW},\, \gamma_Z} Z \Rightarrow XW \rightarrow_{\alpha_{XW},\, \gamma_Z} Z$,

D4: $X \rightarrow_{\alpha,\, \beta} Y \Rightarrow X \rightarrow_{\alpha',\, \beta'} Y$, gdzie $\alpha' \geq \alpha$ i $\beta' \leq \beta$.

Istnienie pewnych zależności funkcyjnych jest przyczyną niekorzystnych własności schematów relacyjnych w rozmytych bazach danych. Są to zależności częściowe i tranzytywne. Zależność $X \rightarrow_{\alpha,\ \beta} Y$ nazywamy zależnością częściową, jeżeli istnieje podzbiór właściwy X' zbioru X taki, że $X' \rightarrow_{\alpha,\ \beta} Y$. Jeżeli takiego podzbioru nie ma, to zależność nazywamy zależnością pełną. O zależności tranzytywnej mówimy wtedy, gdy atrybut Y nie zależy bezpośrednio od X. Zależność taka jest redundantna. Można ją otrzymać na podstawie aksjomatu przechodniości.

Krotki relacji są identyfikowane przez wartości jej klucza. Ze względu na niedoskonałość danych poziom tej identyfikacji nie jest całkowity. Przy atrybutach reprezentowanych za pomocą interwałowych rozkładów możliwości poziomu drugiego poziom klucza wyraża się za pomocą dwóch przedziałów.

Definicja 6. Podzbiór K zbioru atrybutów S jest (α, β)-kluczem relacji o schemacie $R(S)$, jeżeli zależność $X \rightarrow_{\alpha,\ \beta} S$ należy do F^+ i jest to zależność pełna. Atrybuty należące do K nazywamy (α, β)-kluczowymi.

Przykład 5. Kluczem relacji $R(X, Y, Z)$ z zależnością funkcyjną $XY \rightarrow_{[0.8,\ 0.9],[0.6,\ 0.7]} Z$ jest XY. Jest to ([0.8,0.9], [0.6,0.7])-klucz. Rozszerzenie schematu przez włączenie do niego dodatkowego atrybutu V związanego funkcyjnie z X zależnością $X \rightarrow_{[0.8,\ 0.9],[\gamma_L,\ \gamma_U]} Z$ nie powoduje zmiany klucza. W wyniku tej modyfikacji (α,β)-kluczem jest nadal XY, przy czym $\alpha = [0.8,0.9]$ oraz $\beta = [\beta_L, \beta_U]$, gdzie $\beta_L = \min(0.6, \gamma_L)$ oraz $\beta_U = \min(0.7, \gamma_U)$. Wprowadzony atrybut zależy funkcyjnie od klucza. Nie jest to jednak zależność pełna.

Eliminacja zależności niepełnych i tranzytywnych prowadzi do schematów relacji zdefiniowanych za pomocą odpowiednich postaci normalnych.

4. Normalizacja

Na podstawie pojęcia rozmytej zależności funkcyjnej można rozszerzyć definicje postaci normalnych zdefiniowanych dla konwencjonalnych baz danych. W definicjach tych zostały wykorzystane pojęcia (α, β)-klucza oraz atrybutu (α, β)-kluczowego.

Definicja 7. Schemat R jest w pierwszej rozmytej postaci normalnej (F1NF), jeżeli dla każdej relacji o schemacie R wartości jej atrybutów są wyłącznymi interwałowymi rozkładami możliwości poziomu drugiego.

Atrybuty relacji występujących w F1NF mogą przyjmować dokładnie jedną wartość ze swojej dziedziny. Dla każdej wartości został określony przedział możliwości jej wystąpienia.

Definicja 8. Schemat R (X_1, X_2, …, X_n) jest w (α, β)-rozmytej drugiej postaci normalnej ((α, β)-F2NF), jeżeli każdy atrybut (α, β)-niekluczowy jest w pełni funkcyjnie zależny od (α, β)-klucza.

Schemat relacji R z przykładu 5 jest w postaci ([0.8,0.9], [0.6,0.7])-F2NF. Jego rozszerzenie o atrybut V zależny funkcyjnie tylko od X powoduje naruszenie warunków definicji 8.

Definicja 9. Schemat R (X_1, X_2, …, X_n) jest w (α, β)-rozmytej trzeciej postaci normalnej ((α, β)-F3NF), jeżeli dla każdej zależności funkcyjnej $X \rightarrow_{\alpha,\beta} Y$, gdzie $X \subseteq S$, $S = \{X_1, X_2, \ldots, X_n\}$ oraz $Y \notin X$, X zawiera (α, β)-klucz relacji lub Y jest atrybutem (α, β)-kluczowym.

Przykład 6. Rozważmy relację PWZ z atrybutami P – pracownik, W – wykształcenie i Z – zarobki. Załóżmy, ze między jej atrybutami istnieją zależności $\mathrm{P}\rightarrow_{[0.7,\ 0.8],[0.8,\ 0.9]}\mathrm{W}$ oraz $\mathrm{W}\rightarrow_{[0.8,\ 0.9],[0.6,\ 0.7]}\mathrm{Z}$. Na podstawie aksjomatu A3 otrzymujemy zależność $\mathrm{P}\rightarrow_{[0.7,\ 0.8],[0.6,\ 0.7]}\mathrm{Z}$. Kluczem relacji jest P. Jest to ([0.7,0.8], [0.6,0.7])-klucz. Warunki definicji 9 nie są spełnione, ponieważ W nie jest (α, β)-kluczem oraz Z nie jest atrybutem (α, β)-kluczowym. Postać (α, β)-F3NF można otrzymać w wyniku rozkładu na relacje o schematach: PW(P, W) – postać ([0.7,0.8], [0.8,0.9])-F3NF oraz WZ(W, Z) – postać ([0.8,0.9], [0.6,0.7])-F3NF. Jest to rozkład zachowujący zależności.

Możliwość występowania atrybutu kluczowego po prawej stronie dowolnej zależności powoduje, że mogą istnieć atrybuty zależne funkcyjnie tylko od części klucza. Taka sytuacja nie może wystąpić w relacjach o schematach spełniających warunki postaci normalnej BCNF.

Definicja 10. Schemat R $(X_1, X_2, \ldots, X_n)$ jest w (α, β)-rozmytej postaci normalnej Boyce'a-Codda, ((α, β)-FBCNF), jeżeli dla każdej zależności funkcyjnej $X \rightarrow_{\alpha, \beta} Y$, gdzie $X \subseteq S$, $S = \{X_1, X_2, \ldots, X_n\}$ oraz $Y \notin X$, X zawiera (α, β)-klucz relacji.

Przykład 7. Rozważmy relację KONTRAKTY z atrybutami P – numer pracownika, N – nazwisko pracownika, D – numer projektu, Z – zarobki. Załóżmy, że między jej atrybutami istnieją zależności $P \rightarrow_{[1,\,1],[1,\,1]} N$, $N \rightarrow_{[1,\,1],[1,\,1]} P$, $PD \rightarrow_{[0.8,\,0.9],[0.6,\,0.7]} Z$. Relacja ma dwa klucze kandydujące na poziomie ([1,1], [0.6,0.7]). Są to PD i ND. Jedynym atrybutem niekluczowym jest Z. Relacja ta występuje więc w postaci: ([1,1], [0.6,0.7])-F3NF. Lewe strony zależności $P \rightarrow_{[1,\,1],[1,\,1]} N$, $N \rightarrow_{[1,\,1],[1,\,1]} P$ nie zawierają ([1,1], [0.6,0.7])-klucza, co narusza warunki definicji 10. Rozmytą postać BCNF można uzyskać przez rozkład na relacje o schematach PN(P, N) i PDZ(P, D, Z).

Naruszenie warunków definicji 10 może nastąpić, gdy relacja posiada co najmniej dwa złożone klucze kandydujące, które mają wspólne atrybuty. Zauważmy, że usunięcie atrybutu D ze schematu relacji KONTRAKTY i z zależności $PD \rightarrow_{[0.8,\,0.9],[0.6,\,0.7]} Z$ spowoduje zmianę kluczy kandydujących. Są nimi atrybuty N oraz P. Atrybut niekluczowy Z staje się bezpośrednio zależny od każdego klucza. Warunki definiujące postać BCNF są więc spełnione.

5. Podsumowanie

Interwałowe zbiory rozmyte poziomu drugiego stanowią rozszerzenie klasycznych zbiorów rozmytych. Ich zastosowanie wymaga modyfikacji innych pojęć, a w szczególności omawianego w artykule pojęcia rozmytej zależności funkcyjnej. Dalszą konsekwencją jest uwzględnienie wprowadzonych zmian przy projektowaniu rozmytych baz danych za pomocą procedury normalizacji. Przedstawione w artykule definicje rozmytych postaci normalnych stanowią rozszerzenie klasycznych definicji sformułowanych dla konwencjonalnych baz danych.

BIBLIOGRAFIA

1. Alcade C., Burusco A., Fuentes-Gonzales R.: A constructive method for the definition of interval-valued fuzzy implication operators. Fuzzy Sets and Systems, 2005, 153, s. 211÷227.

2. Chen G.Q.: Fuzzy logic in Data Modeling – semantics, constraints and database design. Kluwer, Boston, 1998.
3. Cubero J.C., Vila M. A.: A new definition of fuzzy functional dependency in fuzzy relational databases. International Journal for Intelligent Systems, 1994, 9, s. 441÷448.
4. Karnik N.N., Mendel J.M.: An Introduction to Type-2 Fuzzy Logic Systems. University of Southern California, Los Angeles, 1998.
5. Kerre E., Chen G.: An overview of fuzzy data models. In Studies in Fuzziness: Fuzziness in Database Management Systems, P. Bosc and J. Kacprzyk (Eds.), Physica Verlag, Heidelberg, 1995, s. 23÷41.
6. Mańko J., Niewiadomski A.: Cardinality and Probability under intuitionistic and interval-valued fuzzy sets. Journal of Applied Computer Science, 2006, 14, s. 31÷41.
7. Myszkorowski K.: Interwałowe rozkłady możliwości i ich zastosowanie w rozmytych bazach danych. Studia Informatica, 2009, 30, (83), s. 107÷120.
8. Niewiadomski A.: Interval-Valued and Interval Type-2 Fuzzy Sets: A Subjective Comparison. Proceedings of FUZZ-IEEE'07, Londyn, 2007, s. 1198÷1203.
9. Petry F.: Fuzzy Databases: Principles and Applications, Kluwer Academic Publishers, 1996.
10. Saxena P.C., Tyagi B.K.: Fuzzy functional dependencies and independencies in extended fuzzy relational database models. Fuzzy Sets and Systems, 1995, 69, s. 65÷89.
11. Shenoi S., Melton A., Fan L.T.: Functional dependencies and normal forms in the fuzzy relational database model. Information Sciences, 60, 1992, s. 153÷170.
12. Sozat M.I., Yazici A.: A complete approximation for fuzzy functional and multi-valued dependencies in fuzzy database relations. Fuzzy Sets and Systems, 2001, 117, s. 161÷181.
13. de Tre G., de Caluwe R.: Level-2 fuzzy sets and their usefulness in object-oriented database modeling. Fuzzy Sets and Systems, 2003, 140, s. 29÷49.
14. Tripathy R.C., Saxena P.C.: Multi-valued dependencies in fuzzy relational databases. Fuzzy Sets & Systems, 1990, 38, s. 267÷279.
15. Turksen I.B.: Interval-valued fuzzy sets based on normal forms. Fuzzy Sets and Systems, 1986, 20, s. 191÷210.
16. Tyagi B.K., Sharfuddin A., Dutta R.N., Tayal D.K.: A complete axiomatization of fuzzy functional dependencies using fuzzy function. Fuzzy Sets and Systems, 2005, 151, s. 363÷379.
17. Zadeh L.A.: Fuzzy sets. Information and Control. 1965, 8, s. 338÷353.
18. Zadrożny S.: Zapytania nieprecyzyjne i lingwistyczne podsumowania baz danych. Akademicka Oficyna Wydawnicza EXIT, Warszawa 2006.

Recenzenci: Dr inż. Jerzy Respondek

Wpłynęło do Redakcji 31 stycznia 2010 r.

Abstract

The paper deals with functional dependencies in fuzzy databases. Imperfect information is modeled by means of fuzzy sets which elements are ordinary fuzzy sets. Their membership is given by means of an interval contained in [0,1]. This construction has been named 'interval-valued level-2 fuzzy set' – definition 4 in section 2. For data representation a possibility-based approach has been used. Attributes vales are represented by means of interval level-2 possibility distributions (6). Each possible value represented by an ordinary fuzzy set is associated with a closed interval contained in [0, 1], which expresses its degree of possibility. Measures of the equality of interval level-2 possibility distributions are expressed by (12 – 15).

The notion of fuzzy functional dependency has been defined in section 3 – definition 5. This is an extension of the definition given by Cubero et al. [3]. The degree of fuzzy functional dependency has been determined by means of closeness measures of attribute values. Such dependencies are submitted to extended Armstrong's axioms. Inclusion of fuzzy functional dependencies requires extension of the notion of relation key – definition 6 and normal forms. The normalization problem is considered in section 4.

Adres

Krzysztof MYSZKOROWSKI: Politechnika Łódzka, Instytut Informatyki, ul. Wólczańska 215, 93-005 Łódź, Polska, kamysz@ics.p.lodz.pl .