STUDIA INFORMATICA 2009
Volume 30 Number 2A (83)

Krzysztof MYSZKOROWSKI
Politechnika Łódzka, Instytut Informatyki

INTERWAŁOWE ROZKŁADY MOŻLIWOŚCI I ICH ZASTOSOWANIE W ROZMYTYCH BAZACH DANYCH

Streszczenie. Tematyka artykułu dotyczy modelowania niepełnej informacji w relacyjnych bazach danych za pomocą interwałowych zbiorów rozmytych typu drugiego. Zastosowano podejście posybilistyczne. Wartości atrybutów relacji są reprezentowane za pomocą interwałowych rozkładów możliwości. Przedstawiono analizę oceny stopnia spełnienia warunków stawianych zapytań. Za pomocą interwałowego rozmytego implikatora zdefiniowano pojęcie rozmytej zależności funkcyjnej między atrybutami.

Słowa kluczowe: rozmyte bazy danych, zbiory rozmyte, interwałowe zbiory rozmyte, rozkłady możliwości, implikatory rozmyte, rozmyte zależności funkcyjne

INTERVAL POSSIBLITY DISTRIBUTIONS AND THEIR USE IN FUZZY DATABASES

Summary. The paper deals with modeling of imperfect information in relational databases by means of interval-valued fuzzy sets. The possibility-based approach has been applied. Attribute values are represented by means of interval possibility distributions. The analysis of the extent to which attribute values satisfy query conditions has been presented. An interval-valued fuzzy implicator has been applied in the definition of the fuzzy functional dependency between attributes.

Keywords: fuzzy databases, fuzzy sets, interval-valued fuzzy sets, possibility distributions, fuzzy implicators, fuzzy functional dependencies

1. Wstęp

Teoria zbiorów rozmytych została wprowadzona do klasycznych modeli danych w celu umożliwienia przetwarzania nieprecyzyjnych informacji. Przy tworzeniu rozszerzonych w ten sposób modeli były stosowane różne podejścia. Zostały opracowane różne wersje roz-

mytych modeli danych [7, 13]. Definicje pojęć istniejących w teorii konwencjonalnych baz danych musiały zostać zmodyfikowane. W pracach [14, 18] zaproponowano wykorzystanie do opisu nieprecyzyjnych danych rozkładów możliwości. Jest to podstawowy termin wprowadzonej przez Zadeha w 1978 roku teorii możliwości [17]. Wartość atrybutu podawana jest w postaci zbioru możliwych jego wartości wraz z oceną możliwości ich wystąpienia. Ocena ta jest liczbą należącą do przedziału [0, 1]. Dla wartości całkowicie możliwych jest równa 1. Teoria możliwości jest ściśle związana z teorią zbiorów rozmytych. Występujące w rozkładzie możliwości miary są odpowiednikiem funkcji przynależności zbioru rozmytego.

W większości prac poświęconych rozmytym bazom danych stosowano klasyczne zbiory rozmyte, w których każdemu elementowi była przyporządkowana liczba z przedziału [0, 1] określająca jego przynależność do zbioru. Nie zawsze jednak taki sposób opisywania nie w pełni znanego wycinka rzeczywistości jest wystarczający [11, 12]. Wymaganie precyzyjnego określenia stopnia przynależności może być trudne do spełnienia. Bardziej właściwym rozwiązaniem byłoby podanie zakresu wartości, które może on przyjmować. Wymaga to rozszerzenia definicji klasycznego zbioru rozmytego. Przez zastąpienie liczby przedziałem zawartym w [0, 1] otrzymuje się zbiory rozmyte z interwałową funkcją przynależności, znane w literaturze jako interwałowe zbiory rozmyte typu drugiego [3, 6]. Odpowiadają im interwałowe rozkłady możliwości, w których możliwość wystąpienia pewnej wartości została określona za pomocą przedziału. Ich zastosowanie jest przedmiotem niniejszego opracowania.

Definicja interwałowego rozkładu możliwości sformułowana w punkcie trzecim została poprzedzona omówieniem w punkcie drugim interwałowych zbiorów rozmytych typu drugiego. Przy nieprecyzyjnych wartościach atrybutów nie można podać jednoznacznej odpowiedzi na stawiane do bazy danych zapytanie, które również może być nieprecyzyjne. Analiza, w jakim stopniu nieprecyzyjne wartości odpowiadają warunkom zapytania, została umieszczona w punkcie czwartym. W punkcie piątym zdefiniowano pojęcie zależności funkcyjnej między atrybutami, których wartości są opisane za pomocą interwałowych rozkładów możliwości. Zastosowano przy tym rozmyte implikatory interwałowe [1, 4].

2. Interwałowe zbiory rozmyte typu drugiego

Opisy obiektów świata rzeczywistego często zawierają nieprecyzyjne pojęcia. Stosowane terminy lingwistyczne są nazwami zbiorów o niejednoznacznie zdefiniowanych granicach przynależności. Granice te powinny być precyzyjnie określone przy kwalifikacji obiektów przeprowadzanej według klasycznej teorii zbiorów. Rzeczywistość nie jest jednak „zero-jedynkowa", co sprawia, że jednoznaczna kwalifikacja nie zawsze jest oczywista. Obiekty mogą różnić się między sobą stopniem dopasowania do używanego w języku naturalnym niepre-

cyzyjnego terminu. Zbiory rozmyte zostały pomyślane jako narzędzie modelowania takich niejednoznacznych pojęć [16].

Definicja 1. Niech $\mathcal{R}$ oznacza obszar odniesienia (uniwersum). Zbiór rozmyty F elementów obszaru $\mathcal{R}$ definiuje się następująco:

$$F = \{< x, \mu_F(x) >: x \in \mathcal{R}, \mu_F(x): \mathcal{R} \rightarrow [0, 1] \}, \quad (1)$$

gdzie μF(x) jest funkcją przynależności elementów obszaru $\Re$ do zbioru F.
Jeżeli sup x∈$\Re$ μF (x) = 1, to F jest zbiorem normalnym.

Nie zawsze przynależność może być określona precyzyjnie. Zaproponowano więc rozszerzenie powyższej definicji przez zastąpienie przedziału [0, 1] rodziną zbiorów rozmytych $\mathcal{I}$ ([0, 1]) zdefiniowanych na tym przedziale. Prowadzi to do bardziej ogólnego pojęcia: zbioru rozmytego typu drugiego [6]. Jego funkcja przynależności jest odwzorowaniem $\mathcal{R} \rightarrow \mathcal{I}$([0, 1]). Przykładem takiego odwzorowania jest $\mathcal{R} \rightarrow$ Int ([0, 1]), gdzie Int ([0, 1]) oznacza zbiór zamkniętych podprzedziałów przedziału [0, 1], czyli Int ([0, 1]) = {[a, b]: a, $b \in$ [0, 1]}. Każdemu elementowi x zostaje przyporządkowany interwałowy zbiór rozmyty typu pierwszego. Odpowiada to określeniu przedziału zawartego w [0, 1], wyznaczającego granice dla wartości stopnia przynależności elementu x. Każda należąca do tego przedziału wartość jest w pełni możliwa (w stopniu 1). Wystąpienie pozostałych wartości jest całkowicie niemożliwe. Tak zdefiniowany zbiór jest nazywany interwałowym zbiorem rozmytym typu drugiego [15].

Definicja 2. Niech $\mathcal{R}$ oznacza obszar odniesienia (uniwersum). Interwałowy zbiór rozmyty typu drugiego F elementów obszaru $\mathcal{R}$ definiuje się następująco:

$$F = \{< x, \mu_F(x) >: x \in \mathcal{R}, \mu_F(x) = [\mu_{F_L}(x), \mu_{F_U}(x)],$$
$$\mu_{F_L}(x), \mu_{F_U}(x): \mathcal{R} \rightarrow [0, 1] \} \quad (2)$$

gdzie μF(x) jest przedziałem przynależności elementów obszaru $\Re$ do zbioru F.

Jeżeli sup x∈$\Re$ μFL(x) = sup x∈$\Re$ μFU(x) = 1, to F jest zbiorem normalnym.

Funkcje przynależności – dolna μFL(x) i górna μFU(x) – spełniają warunek:

$$0 \leq \mu_{F_L}(x) \leq \mu_{F_U}(x) \leq 1 \quad (3)$$

Interwałowemu zbiorowi rozmytemu typu drugiego odpowiadają dwa „graniczne" zbiory rozmyte typu pierwszego:

$$F_L = \{< x, \mu_{F_L}(x) >: x \in \mathcal{R}, \mu_{F_L}(x): \mathcal{R} \rightarrow [0, 1] \} \quad (4)$$

$$F_U = \{< x, \mu_{F_U}(x) >: x \in \mathcal{R}, \mu_{F_U}(x): \mathcal{R} \rightarrow [0, 1] \} \quad (5)$$

Jeśli $\mu_{F_L}(x) = \mu_{F_U}(x)$ dla każdego x, to zbiór F jest klasycznym zbiorem rozmytym. Podstawowe charakterystyki interwałowych zbiorów rozmytych typu drugiego określa się na podstawie odpowiednich charakterystyk zbiorów granicznych. Zostały one omówione w pracach Niewiadomskiego [11, 12]. Przykładowo, liczność zbioru interwałowego $card(F)$ jest przedziałem $card(F) = [card(F)_L, card(F)_U]$, gdzie

$$card(F)_L = card(F_L) = \sum_{x \in \mathfrak{R}} \mu_{F_L}(x) \tag{6}$$

$$card(F)_U = card(F_U) = \sum_{x \in \mathfrak{R}} \mu_{F_U}(x) \tag{7}$$

W podobny sposób można opisać inne charakterystyki. Dla zbioru interwałowego definiuje się dwa nośniki $supp(F)_L$ i $supp(F)_U$. Są to zbiory klasyczne elementów należących do odpowiednio do zbiorów F_L i F_U w stopniu niezerowym [11]:

$$supp(F)_L = supp(F_L) = \{ x: \mu_{F_L}(x) > 0 \} \tag{8}$$

$$supp(F)_U = supp(F_U) = \{ x: \mu_{F_U}(x) > 0 \} \tag{9}$$

Z warunku (3) oraz wzorów (8) i (9) wynika, że $supp(F)_L \subseteq supp(F)_U$.

Podobnie, za pomocą dwóch zbiorów $core(F)_L$ i $core(F)_U$ można opisać rdzeń zbioru interwałowego:

$$core(F)_L = core(F_L) = \{ x: \mu_{F_L}(x) = 1 \} \tag{10}$$

$$core(F)_U = core(F_U) = \{ x: \mu_{F_U}(x) = 1 \} \tag{11}$$

Są to zbiory klasyczne elementów w pełni należących odpowiednio do F_L i F_U. Z warunku (3) oraz wzorów (10) i (11) wynika, że $core(F)_L \subseteq core(F)_U$.

Interwałowe zbiory rozmyte F i G są równe, jeżeli każdemu elementowi rozważanego obszaru $x \in \mathfrak{R}$ odpowiadają takie same przedziały przynależności, czyli $\mu_{F_L}(x) = \mu_{G_L}(x)$ oraz $\mu_{F_U}(x) = \mu_{G_U}(x)$. Bardziej złożonym zagadnieniem jest zdefiniowanie inkluzji zbiorów interwałowych. Należy bowiem porównywać przynależność wyrażoną nie jedną, jak w przypadku zbiorów rozmytych pierwszego typu, lecz dwiema liczbami określającymi granice przedziałów. Powstaje więc konieczność określenia sposobu porównywania przedziałów. W dalszych rozważaniach zostanie przyjęte uporządkowanie według następującej reguły [1]:

$$[a_L, a_U] \leq [b_L, b_U] \Leftrightarrow a_L \leq b_L \text{ i } a_U \leq b_U \tag{12}$$

Spełnienie przez przedziały przynależności interwałowych zbiorów rozmytych typu drugiego F i G warunków

$$\mu_{F_L}(x) \leq \mu_{G_L}(x) \text{ oraz } \mu_{F_U}(x) \leq \mu_{G_U}(x) \tag{13}$$

dla każdego elementu $x \in \mathfrak{R}$ oznacza, że zbiór F jest zawarty w zbiorze G: $F \subseteq G$ [10].

Stopień równości zbiorów interwałowych F i G można ocenić porównując stopnie wzajemnego ich zawierania, czyli tego, że: $F \subseteq G$ oraz $G \subseteq F$. Dla zbiorów klasycznych stopień zawierania wyraża się wzorem [2]:

$$\subseteq (F, G) = \inf_x I(\mu_F(x), \mu_G(x)) \tag{14}$$

gdzie I oznacza operator rozmytej implikacji: $[0, 1] \times [0, 1] \rightarrow [0, 1]$, nazywany rozmytym implikatorem. Zaproponowano różne wersje takich operatorów [2]. Jednym z nich jest implikator Gödela:

$$I(a, b) = 1 \text{ dla } a \le b \text{ oraz } I(a, b) = b \text{ dla } a > b \tag{15}$$

Zbiory F i G są równe, jeśli $F \subseteq G$ i $G \subseteq F$, czyli [2]

$$\approx (F, G) = \min (\inf_x I(\mu_F(x), \mu_G(x)), \inf_x I(\mu_G(x), \mu_F(x))) \tag{16}$$

Przy zbiorach interwałowych muszą być zastosowane rozszerzone operatory implikacji, których argumentami są przedziały zawarte w [0, 1], czyli implikatory interwałowe. Implikator interwałowy jest odwzorowaniem $\text{Int}([0, 1]) \times \text{Int}([0, 1]) \rightarrow \text{Int}([0, 1])$. Wynikiem działania takiej operacji jest więc para liczb I_L oraz I_U, które określają granice przedziału. Rozszerzeniem implikacji Gödela jest następujący implikator [1]:

$$I([a,b],[c,d]) = \begin{cases} [c,d] \text{ dla } a > c \text{ i } b > d \\ [c,1] \text{ dla } a > c \text{ i } b \le d \\ [1,1] \text{ dla } a \le c \text{ i } b \le d \\ [d,d] \text{ dla } a \le c \text{ i } b > d \end{cases} \tag{17}$$

Stopień równości zbiorów F i G jest przedziałem $\approx (F, G) = [\approx (F, G)_L, \approx (F, G)_U]$. Jego granice można wyznaczyć przez dwukrotne zastosowanie wzoru (16):

$$\approx (F, G)_L = \min (\inf_x I(\mu_F(x), \mu_G(x))_L, \inf_x I(\mu_G(x), \mu_F(x))_L) \tag{18}$$

$$\approx (F, G)_U = \min (\inf_x I(\mu_F(x), \mu_G(x))_U, \inf_x I(\mu_G(x), \mu_F(x))_U) \tag{19}$$

gdzie $\mu_F(x)$ i $\mu_G(x)$ oznaczają przedziały przynależności zbiorów F i G, a I jest rozmytym implikatorem interwałowym.

3. Interwałowe rozkłady możliwości

Definicja zbioru rozmytego pozwala na określenie stopnia przynależności jego elementów. Warunkiem jest znajomość precyzyjnej wartości odpowiedniego atrybutu. Jeśli wartość ta nie jest znana, to można jedynie ocenić możliwość jej wystąpienia, czyli rozkład możliwości. Pojęcie to zostało po raz pierwszy użyte przez Zadeha w roku 1978 [17].

Definicja 3. Niech $\mathcal{R}$ oznacza obszar odniesienia(uniwersum), X – zmienną określoną na $\mathcal{R}$, F – zbiór rozmyty w $\mathcal{R}$ z funkcją przynależności $\mu_F(x)$. Rozkładem możliwości zmiennej X w obszarze $\mathcal{R}$ względem F jest zbiór

$$\prod_X = \{\pi_X(x) / x: x \in \mathcal{R} \text{ i } \pi_X(x) = \mu_F(x)\} \tag{20}$$

gdzie πX (x) oznacza miarę możliwości przyjęcia przez zmienną X wartości x.

Wyrażanie takich ocen za pomocą przedziału zawartego w [0, 1] prowadzi do interwałowego rozkładu możliwości.

Definicja 4. Niech $\mathcal{R}$ oznacza obszar odniesienia (uniwersum), X – zmienną określoną na $\mathcal{R}$, F – interwałowy zbiór rozmyty typu drugiego w $\mathcal{R}$ z przedziałem przynależności $\mu_F(x) = [\mu_{F_L}(x), \mu_{F_U}(x)]$. Interwałowy rozkład możliwości zmiennej X w obszarze $\mathcal{R}$ względem F definiuje się następująco:

$$\prod_X = \{\pi_X(x) / x: x \in \mathcal{R} \text{ i } \pi_X(x) = \mu_F(x)\} \tag{21}$$

gdzie πX (x) jest przedziałem zawartym w [0, 1].

Granice przedziału πX (x) = [πXL (x), πXU (x)] oznaczają miary możliwości, odpowiednio dolną i górną, przyjęcia przez zmienną X wartości x. Odpowiadają im dwa „graniczne" rozkłady możliwości:

$$\prod_{X_L} = \{\pi_{X_L}(x) / x: x \in \mathcal{R} \text{ i } \pi_{X_L}(x) = \mu_{F_L}(x)\} \tag{22}$$

$$\prod_{X_U} = \{\pi_{X_U}(x) / x: x \in \mathcal{R} \text{ i } \pi_{X_L}(x) = \mu_{F_U}(x)\} \tag{23}$$

$\prod_X$ jest rozkładem normalnym, jeżeli istnieje element x, dla którego $\pi_{X_L}(x) = \pi_{X_U}(x) = 1$. Jeżeli $\pi_{X_L}(x) = \pi_{X_U}(x)$ dla każdego x, to $\prod_X$ jest klasycznym rozkładem możliwości. Dwa rozkłady możliwości są równe, jeżeli zawarte w nich oceny możliwości wystąpienia dowolnej wartości $x \in \mathcal{R}$ są takie same. Oceny stopnia równości klasycznych rozkładów możliwości można dokonać za pomocą miary możliwości *Poss* i konieczności *Nec*. Są one względem siebie dualne [8]. Wartość *Nec* (*Z*) wyraża stopień przekonania o konieczności zajścia pewnego zdarzenia *Z*. Konieczność oznacza brak możliwości wystąpienia zdarzenia przeciwnego $\neg Z$. Analogicznie można zdefiniować możliwość jako brak konieczności zajścia $\neg Z$. Ocena stopnia równości rozkładów możliwości wyraża się wzorami [14]:

$$Poss(\prod_X = \prod_Y) = \sup_x \min(\pi_X(x), \pi_Y(x)) \tag{24}$$

$$Nec(\prod_X = \prod_Y) = 1 - \sup_{x \neq y} \min(\pi_X(x), \pi_Y(y)) \tag{25}$$

Przy interwałowych rozkładach możliwości powyższe miary ulegają rozszerzeniu. Pojedyncze wartości zastępuje się przedziałami: $Poss = [Poss_L, Poss_U]$ i $Nec = [Nec_L, Nec_U]$. Wyznaczenie granic wymaga modyfikacji wzorów (24) i (25). Miary możliwości $Poss_L$

i $Poss_U$ przedstawiają ocenę stopnia równości rozkładów granicznych, odpowiednio $\prod_{X_L}$ i $\prod_{Y_L}$ oraz $\prod_{X_U}$ i $\prod_{Y_U}$. Miary konieczności Nec_L i Nec_U wyrażają ocenę braku możliwości tego, że rozpatrywane rozkłady są różne. Stopień nierówności można przedstawić za pomocą przedziału:

$$[\sup_{x \neq y} \min(\pi_{X_L}(x), \pi_{Y_L}(y)), \sup_{x \neq y} \min(\pi_{X_U}(x), \pi_{Y_U}(y))] \tag{26}$$

Dla rozkładów możliwości $\prod_X$ i $\prod_Y$ otrzymuje się więc następujące wyrażenia:

$$Poss\ (\prod_X = \prod_Y)_L = \sup_x \min(\pi_{X_L}(x), \pi_{Y_L}(x)), \tag{27}$$

$$Poss\ (\prod_X = \prod_Y)_U = \sup_x \min(\pi_{X_U}(x), \pi_{Y_U}(x)) \tag{28}$$

$$Nec((\prod_X = \prod_Y)_L = 1 - \sup_{x \neq y} \min(\pi_{X_U}(x), \pi_{Y_U}(y)) \tag{29}$$

$$Nec((\prod_X = \prod_Y)_U = 1 - \sup_{x \neq y} \min(\pi_{X_L}(x), \pi_{Y_L}(y)) \tag{30}$$

Przykład 1. Dane są dwa rozkłady możliwości: $\prod_X$ = {[1, 1]/*a*, [0.2, 0.4]/*b*, [0.3, 0.6]/*c*, [0.8, 0.9]/*d*} oraz $\prod_Y$ = {[0.2, 0.3]/*a*, [0.3, 0.4]/*b*, [1, 1]/*c*}. Miary możliwości i konieczności tego, że $\prod_X = \prod_Y$) są równe:

$Poss\ (\prod_X = \prod_Y)_L$ = max(min(1, 0.2), min(0.2, 0.3), min(0.3, 1)) = max(0.2, 0.2, 0.3) = 0.3,

$Poss\ (\prod_X = \prod_Y)_U$ = max(min(1, 0.3), min(0.4, 0.4), min(0.6, 1)) = max(0.3, 0.4, 0.6) = 0.6,

$Nec(\prod_X = \prod_Y)_L$ = 1 – max(min(1, 0.4), min(1, 1), min(0.4, 0.3), min(0.4, 1), min(0.6, 0.3), min(0.6, 0.4), min(0.9, 0.3), min(0.9, 0.4), min(0.9, 1)) = 1 – max(0.4, 1, 0.3, 0.4, 0.3, 0.4, 0.3, 0.4, 0.9) = 0.

$Nec((\prod_X = \prod_Y)_U$ = 1 – max(min(1, 0.3), min(1, 1), min(0.2, 0.2), min(0.2, 1), min(0.3, 0.2), min(0.3, 0.3), min(0.8, 0.2), min(0.8, 0.3), min(0.8, 1)) = 1 – max(0.3, 1, 0.2, 0.2, 0.2, 0.3, 0.2, 0.3, 0.8) = 0.

4. Relacje z atrybutami o wartościach określonych za pomocą interwałowych rozkładów możliwości

Rozkłady możliwości są jedną z form reprezentacji danych przy niepełnej ich znajomości. Zastosowanie tego podejścia w relacyjnych bazach danych wymaga rozszerzenia pojęcia relacji. W konwencjonalnych bazach danych relacja jest podzbiorem iloczynu kartezjańskiego dziedzin jej atrybutów. Niech A_1 , A_2 , ... , A_n oraz D_1, D_2, ... , D_n oznaczają atrybuty relacji i ich dziedziny. Relacja R (A_1 , A_2 , ... , A_n) jest zbiorem krotek $r = < a_1 , a_2 , \ldots , a_n >$, gdzie a_i oznacza wartość *i*-tego atrybutu. Jest to więc podzbiór zbioru $D_1 \times D_2 \times \ldots \times D_n$. Jeżeli wartości atrybutów A_i są podane za pomocą normalnych rozkładów możliwości $\prod_{A_i}$, to krotki

relacji R przyjmą postać: $r = < \prod_{A_1}, \prod_{A_2}, \ldots, \prod_{A_n} >$. Tak rozumiana relacja jest określona na iloczynie kartezjańskim zbiorów rozkładów możliwości na dziedzinach jej atrybutów. Jest to więc podzbiór następującego zbioru:

$$\Gamma(D_1) \times \Gamma(D_2) \times \ldots \times \Gamma(D_n) \tag{31}$$

gdzie $\Gamma(D_1) = \{\prod_{A_i} :$ jest normalnym rozkładem możliwości atrybutu A_i o dziedzinie $D_i\}$.

Przy wyszukiwaniu danych dla wartości atrybutów formułowane są warunki, które muszą one spełniać, by należeć do zbioru stanowiącego odpowiedź na zapytanie. W konwencjonalnych bazach danych zarówno warunki, jak i wartości atrybutów są precyzyjne. Odpowiedź jest więc określona jednoznacznie. Przy nieprecyzyjnych atrybutach lub warunkach powstaje problem określenia stopnia dopasowania danych do warunków zapytania.

Ocenę stopnia spełnienia zapytania przez atrybut A można otrzymać przez wyznaczenie miary możliwości $Poss(W, A)$ i konieczności $Nec(W, A)$ spełnienia warunku o postaci $A = W$, gdzie W określa wartość odpowiadającą kryterium kwalifikacji. Wartość ta może być podana precyzyjnie, za pomocą przedziału lub zbioru rozmytego. Jako przykład można podać warunek WIEK = NOWY, gdzie WIEK jest atrybutem relacji SAMOCHODY, a NOWY jest nazwą zbioru rozmytego o tej samej dziedzinie co atrybut WIEK. Należy określić możliwość i konieczność zdarzenia polegającego na tym, że rozpatrywany samochód jest nowy. Miary te wskazują, w jakim stopniu wartość atrybutu określona za pomocą rozkładu możliwości odpowiada wartości określonej za pomocą zbioru rozmytego. Przy założeniu że warunek W jest określony za pomocą zbioru rozmytego pierwszego typu, a wartości atrybutów są reprezentowane za pomocą klasycznych rozkładów możliwości $Poss(W, A)$ i $Nec(W, A)$, wyrażają się wzorami [5]:

$$Poss(W, A) = \sup_{x \in \mathfrak{R}} \min(\mu_W(x), \pi_A(x)) \tag{32}$$

$$Nec(W, A) = \inf_{x \in \mathfrak{R}} \max(\mu_W(x), 1 - \pi_A(x)) \tag{33}$$

gdzie $\pi_A(x)$ jest miarą możliwości przyjęcia przez zmienną X wartości x, a $\mu_W(x)$ jest funkcją przynależności występującego w zapytaniu zbioru rozmytego W.

Jeżeli wartości atrybutów są określone za pomocą interwałowych rozkładów możliwości i/lub warunki zapytania są sformułowane za pomocą interwałowych zbiorów rozmytych typu drugiego, to sposób oceny stopnia spełnienia zapytania powinien zostać zmodyfikowany. Liczby zostają zastąpione przedziałami $[Poss_L, Poss_U]$ i $[Nec_L, Nec_U]$ zawartymi w [0, 1]. Ich granice określają dolne i górne wartości miar możliwości i konieczności spełnienia warunku zapytania. Sposób wyznaczania granic przedziałów otrzymuje się przez odpowiednią modyfikację wzorów (32) i (33):

$$Poss(W, A)_L = \sup_{x \in \mathfrak{R}} \min(\mu_{W_L}(x), \pi_{A_L}(x)) \tag{34}$$

$$Poss(W, A)_U = \sup_{x \in \mathfrak{R}} \min(\mu_{W_U}(x), \pi_{A_U}(x)) \tag{35}$$

$$Nec\ (W, A)_L = \inf_{x \in \mathfrak{R}} \max(\mu_{W_L}(x), 1 - \pi_{A_U}(x)) \tag{36}$$

$$Nec\ (W, A)_U = \inf_{x \in \mathfrak{R}} \max(\mu_{W_U}(x), 1 - \pi_{A_L}(x)) \tag{37}$$

Można wyróżnić charakterystyczne przypadki położenia względem siebie interwałowego rozkładu możliwości $\prod_A$ i funkcji przynależności $\mu_W(x) = [\mu_{W_L}(x), \mu_{W_U}(x)]$ interwałowego zbioru rozmytego W:

1. Nie istnieje wartość x, dla której $\pi_{A_U}(x) > 0$ i $\mu_{W_U}(x) > 0$.
   $Poss\ (W, A)_L = Poss\ (W, A)_U = Nec\ (W, A)_L = Nec\ (W, A)_U = 0$
2. Nie istnieje wartość x, dla której $\pi_{A_L}(x) > 0$ i $\mu_{W_L}(x) > 0$.
   $Poss\ (W, A)_L = Nec(W, A)_L = 0$
3. Istnieje wartość x, dla której $\pi_{A_L}(x) = 1$ i $\mu_{W_L}(x) = 1$.
   $Poss\ (W, A)_L = \mathrm{Poss}(W, A)_U = 1$
4. Istnieje wartość x, dla której $\pi_{A_U}(x) = 1$ i $\mu_{W_U}(x) = 1$.
   $Poss\ (W, A)_U = 1$
5. Wszystkie wartości atrybutu A, dla których górna miara możliwości jest dodatnia, w pełni należą do W, czyli $supp\ (A)_U \subseteq core\ (W)_L$.
   $Nec\ (W, A)_L = Nec\ (W, A)_U = 1$
6. Wszystkie możliwe w stopniu niezerowym wartości atrybutu A w pełni należą do zbioru W_U, czyli $supp\ (A)_L \subseteq core\ (W)_U$.
   $Nec\ (W, A)_U = 1$
7. Wszystkie wartości atrybutu A, dla których górna miara możliwości jest równa 1 należą do zbioru W w stopniu niezerowym, czyli $core\ (A)_U \subset supp\ (W)_L$.
   $Nec\ (W, A)_L > 0$, $Nec\ (W, A)_U > 0$
8. Wszystkie wartości atrybutu A, dla których górna miara możliwości jest równa 1, należą do zbioru W_U w stopniu niezerowym, czyli $core\ (A)_L \subset supp\ (W)_U$.
   $Nec\ (W, A)_U > 0$
9. Wartość atrybutu jest określona precyzyjnie, czyli $\exists\ a\ \pi_{A_L}(a) = \pi_{A_U}(a) = 1$ oraz $\forall\ b \neq a\ \pi_{A_L}(b) = \pi_{A_U}(b) = 0$. Warunek jest podany za pomocą interwałowego zbioru rozmytego typu drugiego W. Miary możliwości i konieczności są sobie równe.
   $Poss\ (W, A)_L = Nec\ (W, A)_L = \mu_{W_L}(a)$, $Poss\ (W, A)_U = Nec\ (W, A)_U = \mu_{W_U}(a)$
10. Wartość atrybutu jest określona za pomocą interwałowego rozkładu możliwości, a warunek jest precyzyjny, czyli $\exists\ w\ \mu_{W_L}(w) = \mu_{W_U}(w) = 1$ oraz $\forall\ u \neq w\ \mu_{W_L}(w) = \mu_{W_U}(w) = 0$.
    $\mathrm{Poss}\ (W, A)_L = \pi_{A_L}(w)$, $\mathrm{Poss}\ (W, A)_U = \pi_{A_U}(w)$
    W tym przypadku wartości miar konieczności są mniejsze niż 1. Ponadto, jeżeli istnieje w pełni możliwa wartość atrybutu różna od wartości w występującej w warunku, to miary te są równe 0.

Przykład 2. Wiek pewnego samochodu został podany za pomocą rozkładu możliwości: $\prod_{\mathrm{WIEK}} = \{[1, 1]/1, [0.5, 0.7]/2, [0.4, 0.5]/3, [0.2, 0.3]/4\}$. Miary możliwości i konieczności

spełnienia warunku WIEK = „Około 3 lata", gdzie „Około 3 lata" = {[0.2, 0.4]/1, [0.7, 0.8]/2, [1, 1]/3, [0.7, 0.8]/4. [0.2, 0.4]/5}, wynoszą:

$Poss_L$ = max(min(0.2, 1),min(0.7,0.5),min(1, 0.4), min(0.7,0.2)) = max(0.2,0.5,0.4,0.2) = 0.5

$Poss_U$ = max(min(0.4, 1),min(0.8,0.7), min(1, 0.5), min(0.8,0.3)) = max(0.4,0.7,0.5,0.3) = 0.7

Nec_L = min(max(0.2, 0), max(0.7, 0.3), max(1, 0.5), max (0.7, 0.7), max(0.2, 1)) = min (0.2, 0.7, 1, 0.7, 1) = 0.2

Nec_U = min(max(0.4, 0), max(0.8, 0.5), max(1, 0.6), max (0.8, 0.8), max(0.4, 1)) = min (0.4, 0.8, 1, 0.8, 1) = 0.4

Miara możliwości jest większa od miary konieczności, co jest zgodne z intuicją. Naruszenie normalności rozkładu możliwości może prowadzić do wyników przeciwnych (*Nec* > *Poss*). Po usunięciu z $\prod_{WIEK}$ pierwszego elementu otrzymuje się:

$Poss_L = 0.5$, $Poss_U = 0.7$, $Nec_L = 0.7$ i $Nec_U = 0.8$.

5. Rozmyte zależności funkcyjne

Występowanie zależności funkcyjnej $X \rightarrow Y$ między atrybutami X i Y oznacza jednoznaczną identyfikację wartości atrybutu Y przez atrybut X. Taka interpretacja tego pojęcia przy nieprecyzyjnych danych nie jest wystarczająca. Do jej formalnego zapisu nie można zastosować logiki dwuwartościowej. Przy braku precyzji można jedynie mówić o pewnym stopniu identyfikacji. Oznaczałoby to, że bliskim wartościom atrybutu X odpowiadają bliskie wartości atrybutu Y. Sformułowanie definicji rozmytej zależności funkcyjnej wymaga ustalenia miary bliskości oraz rozmytego implikatora. Na ich podstawie można wyznaczyć stopień zależności. W pracy [2] sformułowano definicję zależności funkcyjnej między atrybutami, których wartości są opisane za pomocą klasycznych rozkładów możliwości – wzór (20). Uwzględnienie rozszerzonej formy opisu – rozkładów interwałowych – wymaga rozszerzenia definicji. Miara bliskości interwałowych rozkładów możliwości jest określona za pomocą przedziału $[\theta_L, \theta_U]$, gdzie θ_L i θ_U wyrażają się wzorami (27) i (28). Drugim elementem rozszerzonej definicji jest rozmyty implikator interwałowy, za pomocą którego zostanie wyznaczony stopień zależności.

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 $\theta = [\theta_L, \theta_U]$, $X \rightarrow_{\theta} Y$, wtedy i tylko wtedy, gdy dla każdej relacji o schemacie R jest spełniony następujący warunek:

$$\min_{t,\,t'} I\,(Poss(X), Poss(Y))_L \geq \theta_L, \quad \min_{t,\,t'} I\,(Poss(X), Poss(Y))_U \geq \theta_U \tag{38}$$

gdzie I jest interwałowym implikatorem rozmytym.

Przy klasycznych rozkładach możliwości – określonych wzorem (20) – wartości $Poss_L(X)$ i $Poss_U(X)$ są równe. Otrzymuje się wtedy definicję podaną przez Chena [2]. Przy wyborze operatora implikacji konieczna jest analiza własności poszczególnych implikatorów. Dla danego schematu można sformułować wiele zależności. Niektóre z nich mogą być konsekwencją logiczną innych lub zawierać nadmiarowe atrybuty. Redundancja powinna zostać wyeliminowana. Dlatego istotne znaczenie miałaby możliwość wnioskowania prowadzącego do utworzenia zbioru minimalnego. Ponadto, rozmyta zależność funkcyjna powinna posiadać odpowiednie własności. Zależą one od zastosowanego implikatora. Przykładowo, z wysokiego stopnia bliskości wartości atrybutu X powinny wynikać odpowiednio bliskie wartości atrybutu Y. Jeżeli przy tym jest spełniony warunek $Poss(X) \leq Poss(Y)$, to zależność funkcyjna istnieje w stopniu całkowitym, czyli zastosowany implikator powinien dawać wynik [1, 1]. Znaczne różnice między wartościami atrybutu X w poszczególnych krotkach powinny implikować wysoki stopień zależności. Zastosowanie implikatora (17) daje pożądany efekt, jeżeli $Poss(X) \leq Poss(Y)$. Jeżeli natomiast stopień bliskości dla atrybutu Y jest również mały i ponadto mniejszy niż dla atrybutu X, to otrzymuje się przy zastosowaniu tego implikatora niską ocenę poziomu zależności $X \rightarrow_\theta Y$.

Definicja zależności funkcyjnej jest bardzo istotna przy definiowaniu innych pojęć w dziedzinie baz danych. Jej rozszerzenie implikuje więc odpowiednią modyfikację pozostałych definicji. Dotyczy to między innymi pojęcia klucza relacji. Jest nim atrybut K, dla którego zachodzi zależność $K \rightarrow_\theta S$, przy czym jest to zależność pełna (nie istnieje podzbiór $K' \subset K$, taki że $K' \rightarrow_\theta S$). Poziom tej zależności jest określony za pomocą przedziału zawartego w [0, 1]. Jego granice określają dolną i górną wartość oceny stopnia identyfikacji krotek przez klucz K. Podobny sposób powinien być zastosowany przy ocenie stopnia spełnienia warunków poszczególnych postaci normalnych [2], stanowiących kolejne etapy projektowania za pomocą procedury normalizacji.

Przykład 3. Relacja ETATY (tabela 1) przedstawia związek między zajmowanym stanowiskiem (S), wykształceniem (W) oraz zarobkami (Z). Funkcje przynależności do interwałowych zbiorów rozmytych typu drugiego: „niskie” oraz „wysokie” wyrażają się wzorami:

$\mu_{niskie_L}(x) = 1$ dla $x \leq 1$, $\mu_{niskie_L}(x) = -x + 2$ dla $x \in [1, 2]$, $\mu_{niskie_L}(x) = 0$ dla $x > 2$

$\mu_{niskie_U}(x) = 1$ dla $x \leq 1.5$, $\mu_{niskie_U}(x) = -x + 2.5$ dla $x \in [1.5, 2.5]$, $\mu_{niskie_U}(x) = 0$ dla $x > 2.5$

$\mu_{wysokie_L}(x) = 0$ dla $x < 6$, $\mu_{wysokie_L}(x) = 0.2\,x - 1.2$ dla $x \in [6, 11]$, $\mu_{wysokie_L}(x) = 1$ dla $x > 11$

$\mu_{wysokie_U}(x) = 0$ dla $x < 5$, $\mu_{wysokie_U}(x) = 0.2\,x - 1$ dla $x \in [5, 10]$, $\mu_{wysokie_U}(x) = 1$ dla $x > 10$

Tabela 1

ETATY

	Stanowisko	Wykształcenie	Zarobki (w tys PLN)
t_1	S1	W1	niskie
t_2	S2	W2	9
t_3	S3	[0.8, 0.9]/W3, [1,1]/W4	8
t_4	1/S2, 1/S3	[1,1]/W2, [0.6, 0.8]/W3	wysokie
t_5	S1	W2	3
t_6	S4	W1	1

Na podstawie (27-30) oraz (17) otrzymuje się:
dla krotek t_2 i t_4: $Poss(\mathrm{SW})_L = Poss(\mathrm{SW})_U = 1$, $Poss(\mathrm{Z})_L = 0.6$, $Poss(\mathrm{Z})_U = 0.8$,
dla krotek t_3 i t_4: $Poss(\mathrm{SW})_L = 0.6$, $Poss(\mathrm{SW})_U = 0.8$, $Poss(\mathrm{Z})_L = 0.4$, $Poss(\mathrm{Z})_U = 0.6$.
$I([1, 1], [0.6, 0.8]) = [0.6, 0.8]$ oraz $I([0.6, 0.8], [0.4, 0.6]) = [0.4, 0.6]$.
Atrybuty relacji ETATY spełniają więc zależność $\mathrm{SW} \rightarrow_{[0.4, 0.6]} \mathrm{Z}$.

6. Podsumowanie

W artykule przedstawiono wybrane elementy rozmytego modelu relacyjnego. Zastosowano przy tym zbiory rozmyte o interwałowej funkcji przynależności. Oznaczało to zastosowanie bardziej złożonego narzędzia przy modelowaniu niepełnej informacji. Prezentowane zagadnienia wymagają dalszych badań. W szczególności dotyczy to zależności funkcyjnych i zbudowania dla nich adekwatnego oraz zupełnego zbioru reguł wnioskowania.

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. Cornelis C., Deschrijver G., Kerre E.E.: Advances and challenges in interval-valued fuzzy logic, Fuzzy Sets and Systems, 2006, 157, s. 622÷627.
4. Deschrijver G., Kerre E.E.: Implicators based on binary aggregation operators in interval valued fuzzy set theory, Fuzzy Sets and Systems, 2005, 153, s. 229÷248.

5. Dubois D., Prade H., Testemale C.: Weighted Fuzzy Pattern Matching, Fuzzy Sets and Systems, 1988, 28, s. 313÷331.
6. Karnik N.N., Mendel J.M.: An Introduction to Type-2 Fuzzy Logic Systems. University of Southern California, Los Angeles 1998.
7. 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.
8. Łachwa A.: Rozmyty świat zbiorów, liczb, relacji, faktów, reguł i decyzji, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2001.
9. Mańko J., Niewiadomski A.: Cardinality and Probability under intuitionistic and interval-valued fuzzy sets. Journal of Applied Computer Science, 2006, 14, s. 31÷41.
10. Mondal T.K., Samanta S.K.: Topology of interval-valued intuitionistic fuzzy sets, Fuzzy Sets and Systems, 2001, 119, s. 483÷494.
11. Niewiadomski A.: Interval-Valued and Interval Type-2 Fuzzy Sets: A Subjective Comparison. Proceedings of FUZZ-IEEE'07, Londyn 2007, s. 1198÷1203.
12. Niewiadomski A.: Methods for the linguistic summarization of data: applications of fuzzy sets and their extensions. EXIT, Warszawa 2008.
13. Petry F.: Fuzzy Databases: Principles and Applications, Kluwer Academic Publishers, 1996.
14. Prade H., Testemale C.: Generalizing database relational algebra for the treatment of incomplete or uncertain information and vague queries. Information Sciences, 1984, 34, s. 115÷143.
15. Turksen I.B.: Interval-valued fuzzy sets based on normal forms. Fuzzy Sets and Systems, 1986, 20, s. 191÷210.
16. Zadeh L.A.: Fuzzy sets. Information and Control. 1965, 8, s. 338÷353.
17. Zadeh L.A.: Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets and Systems, 1978, 1, s. 3÷28.
18. Zemankova M., Kandel A.: Implementing imprecision in information systems. Information Sciences, 1985, 37, s. 107÷141.

Recenzent: Dr inż. Alina Momot

Wpłynęło do Redakcji 20 stycznia 2009 r.

Abstract

Conventional database systems are designed with the assumption of the precision of information collected in them. Otherwise one has to apply tools for describing uncertain or imprecise information. One of them is the fuzzy set theory. However in some circumstances traditional fuzzy sets may appear insufficient. Hence in the paper applying of their extension, known as interval-valued fuzzy sets, is proposed. For data representation a possibility-based approach has been used.

In Section 2 interval-valued fuzzy sets (1) are introduced and some of their characteristics are also given. The concept of equality of sets has been based on the notion of set inclusion (14), with the use of interval-valued fuzzy implicator (17). Attribute values are represented by interval possibility distributions (20) – Section 3. Each possible value is associated with a closed interval in [0, 1], which expresses its degree of possibility. Measures of the equality of interval possibility distributions are expressed by (27 – 30).

Because of imperfect data querying databases raises additional problems. Values of attributes may satisfy conditions of the query in different grade. Moreover conditions in queries can also be imprecise. Very often propositions in queries can be written in the following form: "*A* is *W*", where *A* denotes an attribute and *W* is a linguistic term represented by a fuzzy set, in particular by an interval-valued fuzzy set. One has to determine to what extent a given value belongs to the answer. The problem is presented in Section 4. The degrees of possibility *Poss*(*W*, *A*) and necessity *Nec*(*W*, *A*) of matching are expressed by (34 – 37). The analysis of several particular cases is also given.

In fuzzy databases it is possible to evaluate the degree of the closeness of compared values. If close values of attribute *X* correspond to close values of attribute *Y* there exists a fuzzy functional dependency $X \to Y$. Its definition (38) has been formulated in Section 5. It requires establishing an interval-valued fuzzy implicator.

Adres

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