STUDIA INFORMATICA 2002
Volume 23 Number 4 (51)

Bożena MAŁYSIAK
Politechnika Śląska, Instytut Informatyki

MECHANIZMY WNIOSKOWANIA PRZYBLIŻONEGO W BAZACH DANYCH

Streszczenie. W artykule przedstawiono zastosowanie teorii zbiorów rozmytych oraz metod wnioskowania przybliżonego (na przykładzie modelu Mamdaniego) w procesie zadawania rozmytych (nieprecyzyjnych) pytań do bazy danych i generacji odpowiedzi na tak sformułowane pytania.

APPROXIMATE REASONING IN DATABASES

Summary. This article presents fuzzy sets theory and approximate reasoning and their applications in database. Database can include precise or imprecise data, and queries can be precise or imprecise too.

1. Wprowadzenie

W pracy przedstawiono elementy teorii zbiorów rozmytych oraz metody wnioskowania przybliżonego, oparte na logice rozmytej, stosowane w bazach danych. Informacja rozmyta w niektórych przypadkach jest pełniejsza i trafniej oddaje specyfikę rzeczywistości. Potrzeba stosowania logiki rozmytej rośnie wraz ze złożonością i kompleksowością baz danych.

W bazach danych istnieje wiele możliwości korzystania z idei nieostrości (rozmycia). Daje się wśród nich wyróżnić dwa podejścia [4]:

- zapamiętywanie rozmytych informacji w bazie danych,
- korzystanie z rozmytych języków wyszukiwania informacji w bazach danych.

Biorąc pod uwagę dokładność informacji przechowywanej w bazie danych oraz sposoby zadawania pytań do bazy danych można wyodrębnić następujące sytuacje:

1. Baza danych zawiera dokładne dane, tzn. wartości poszczególnych atrybutów opisujących obiekty bazy danych są określone zgodnie z ich dziedzinami, a także do bazy danych zadawane są dokładne pytania, jednoznacznie specyfikujące własności poszukiwanych obiektów.

   SZBD (system zarządzania bazą danych) generuje wtedy odpowiedzi zwracające listę obiektów (może to być także lista pusta), spełniających warunki wyspecyfikowane w pytaniu. Jest to klasyczny przypadek wykorzystywania baz danych, wyczerpująco zbadany, wykorzystujący różne, efektywne modele reprezentacji danych (np. relacyjny model danych) oraz języki manipulowania danymi (SQL, QBE, QUEL, ...).
2. Baza danych zawiera dokładne dane (w rozumieniu jak w przypadku 1), ale przy ich wyszukiwaniu zadawane są niedokładne pytania, tzn. nie specyfikujące dokładnie własności obiektów reprezentowanych w bazie danych.

   Sytuacja taka może wystąpić wtedy, gdy użytkownik, zainteresowany dostępem do informacji zawartej w bazie danych, nie zna jej szczegółowej reprezentacji (np. schematu relacji) lub gdy nie potrafi sformułować dokładnego pytania. Do SZBD zostaje więc skierowane niedokładne pytanie i system ten musi je „zrozumieć", aby wyszukać w bazie danych określone odpowiedzi.

   W tym przypadku mamy więc do czynienia z dwoma problemami.

   Pierwszy z nich to, jak przekształcić niedokładne pytanie na postać akceptowalną przez SZBD (inicjujący proces wyszukiwania odpowiedzi).

   Drugi to, jak wyselekcjonować i reprezentować odpowiedzi na niedokładne pytanie. Jest to więc problem uzyskania dokładnej odpowiedzi na niedokładne pytanie.

   Przykład:

   Select imię, nazwisko
   From personalia
   Where wiek około 50 and staż_pracy około 20;

   gdzie wartościami atrybutów wiek i staż_pracy są określone wartości numeryczne.
3. Baza danych zawiera niedokładne dane, np. wartości niektórych atrybutów specyfikujących jej poszczególne obiekty nie są w pełni określone, są wartościami podzbiorów rozmytych lub przyjmują wartości nieprecyzyjne (rozmyte) wyrażone w języku naturalnym. Do bazy zawierającej takie dane zadawane są dokładne pytania. Pojawia się więc problem otrzymania zadowalających użytkownika odpowiedzi na tak postawione pytania.

   Wartości atrybutów mogą być podane w postaci liczb rozmytych, przedziałów wartości lub podzbiorów rozmytych [5].

Na przykład wiek – młody; średni; około 20; między 30 a 40; 26-28; 12,14; 2 albo 3.

Do takiej bazy danych może być zadawane dokładne pytanie, np. postaci:

Select imię, nazwisko
From personalia
Where wiek = 50 and staż_pracy = 20;

4. W tym przypadku zarówno baza danych zawiera niedokładne dane (w rozumieniu jak w przypadku 3), jak i zadawane pytania są niedokładne (w rozumieniu jak w przypadku 2).

Wartości atrybutów mogą być podane w postaci liczb rozmytych, przedziałów wartości lub podzbiorów rozmytych (podobnie jak w przypadku 3).

Do takiej bazy danych zadawane jest nieprecyzyjne (rozmyte) pytanie (takie jak w przypadku 2):

Select imię, nazwisko
From personalia
Where wiek około 50 and staż_pracy około 20;

W przeprowadzonych badaniach przeanalizowano różne rodzaje przechowywanych danych w bazie. W celu znalezienia nieprecyzyjnie podanej informacji w bazie danych zastosowano algorytm wnioskowania przybliżonego wg Mamdaniego [1]. Algorytm ten został nieco zmodyfikowany w celu dostosowania go do potrzeb realizacji procesu wyszukiwania.

2. Teoria zbiorów rozmytych

2.1. Pojęcia podstawowe

Teoria zbiorów rozmytych została wprowadzona przez L.A. Zadeha. Podstawowym pojęciem tej teorii jest zbiór rozmyty, zdefiniowany jako [5]:

zbiór par, w pewnej numerycznej przestrzeni rozważań X

$A = \{(\mu_A(x), x)\}$, dla każdego $x \in X$,

gdzie:

μ_A – **funkcja przynależności zbioru rozmytego A**, która każdemu elementowi zbioru $x \in X$ przypisuje **stopień jego przynależności** $\mu_A(x)$ do zbioru A, przy czym: $\mu_A(x) \in [0,1]$.

Zbiory rozmyte dopuszczają częściową przynależność obiektów do zbioru (w klasycznych zbiorach – element zbioru należy do zdefiniowanego zbioru lub nie, funkcja charakterystyczna przyjmuje wtedy odpowiednio wartość 1 lub 0).

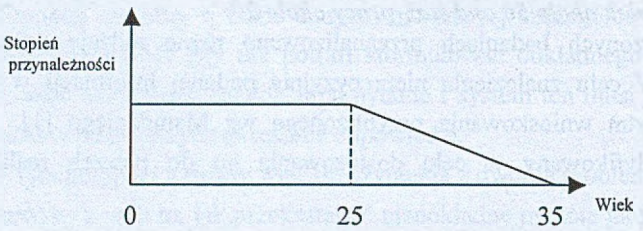
W teorii zbiorów rozmytych funkcja charakterystyczna została zastąpiona funkcją przynależności, która określa, czy dany element zbioru na pewno do niego należy ($\mu_A(x) = 1$), czy być może należy w pewnym stopniu ($\mu_A(x) > 0$) lub z całą pewnością nie należy do zdefiniowanego zbioru ($\mu_A(x) = 0$).

Funkcja przynależności może być wyrażona w postaci: diagramu (ciągłego lub dyskretnego), wzoru matematycznego, tabeli, wektora przynależności.

W przypadku gdy przestrzeń X jest prostą rzeczywistą, można funkcję przynależności przedstawić w postaci funkcyjnej, np.: $\mu_A(x) = 1/(1+x^2)$. Podzbiory rozmyte są wykorzystywane do reprezentacji pojęć o niesprecyzowanych granicach.

Przykład podzbioru rozmytego

Należy utworzyć podzbiór ludzi młodych oparty na dziedzinie *wiek* ludzi. Dziedziną zbioru *wiek* ludzi jest przedział [0, 100]. Funkcja przynależności do podzbioru rozmytego reprezentującego wyraz rozmyty *„młody"* może wyglądać np. jak na rys.1.



Rys. 1. Funkcja przynależności podzbioru rozmytego *'młody'*
Fig. 1. The membership function of fuzzy set '*young*'

W teorii zbiorów rozmytych wprowadza się pojęcie wyraz (element, wartość, liczba) rozmyty, nieprecyzyjny. Wartość rozmyta jest nieprecyzyjną wartością pewnego atrybutu, np. na zbiorze *wiek* może być określona liczba rozmyta *około 30-tki*. Każdy z wyrazów rozmytych jest scharakteryzowany precyzyjnie przez podzbiór rozmyty, który posiada określoną funkcję przynależności.

Liczbę rozmytą charakteryzuje zbiór rozmyty określony na uniwersum rzeczywistym. Jest to zbiór normalny, wypukły o funkcji przynależności przedziałami ciągłej [2].

Zbiór rozmyty normalny to taki zbiór, którego funkcja przynależności przyjmuje wartości od 0 do 1 (łącznie z 1). Zbiór rozmyty, który nie jest normalny, nazywany jest **zbiorem subnormalnym**.

Zbiór rozmyty wypukły cechuje się tym, że wszystkie jego α-przekroje są zwartymi, jednoczęściowymi przedziałami przestrzeni rozważań. W **zbiorach niewypukłych** istnieją niezwarte wieloczęściowe α-przekroje. **Zbiory niewypukłe** powstają w wyniku wykonywania operacji arytmetycznych na zbiorach pierwotnych, które zwykle są wypukłe.

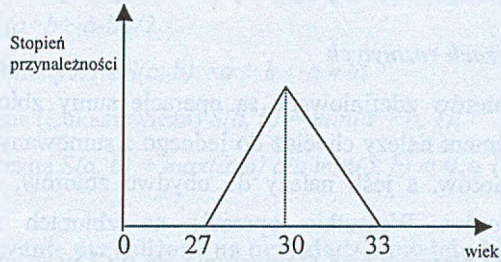
α-przekroj zbioru A (α-cut) jest to ostry zbiór $A_\alpha = \{x: \mu_A(x) \geq \alpha\}$, tzn. zbiór, dla którego funkcja charakterystyczna przyjmuje wartości 1 dla $\mu_A(x) \geq \alpha$ oraz 0 dla $\mu_A(x) \leq \alpha$.

Przykłady liczb rozmytych to

– około zera, mniej więcej 5, trochę więcej niż 9, mniej więcej pomiędzy 10 i 12, np.: liczba rozmyta *około 30-tki* może mieć dziedzinę [27,33], np. jak na rys. 2.

Zmienna lingwistyczna jest to wielkość wejściowa, wyjściowa bądź zmienna stanu, którą zamierza się oceniać stosując oceny lingwistyczne, zwane wartościami lingwistycznymi.

Przykłady zmiennych lingwistycznych to np.: wysokość, temperatura.



Rys. 2. Funkcja przynależności podzbioru rozmytego *'około 30'*
Fig. 2. The membership function of fuzzy set '*about 30*'

Wartość lingwistyczna jest to słowna ocena zmiennej lingwistycznej, podzbiór używany jako nośnik znaczenia pojęcia, np.: można określić słowo *wysoki* jako podzbiór rozmyty zbioru wysokości; pojęcie *zimny* jako podzbiór rozmyty *temperatury.*

Przykłady wartości lingwistycznych to np.: bardzo duży ujemny, średni ujemny, bardzo duży, przyjemny, nieprzyjemny, młody, ładny, stary itd.

Gdy mówi się o wartościach lingwistycznych, należy rozumieć to jako próbę skojarzenia podzbioru rozmytego z wartością zmiennej. Przypisanie podzbioru rozmytego wartości zmiennej wprowadza niepewność do wiedzy o rzeczywistej wartości tej zmiennej. Na przykład stwierdzając, że Katarzyna jest młoda, wie się na pewno, że nie ma ona 65 lat, ale, że ma 17 lub 20 lat jest już możliwe, są to zatem możliwe wartości zmiennej *wiek Katarzyny.* Potrzebny jest w tym momencie nowy rodzaj niepewności, nazwany przez Zadeha **niepewnością możliwą** [1].

Jeżeli V jest zmienną o wartościach ze zbioru X, a A jest podzbiorem rozmytym X, to wg Zadeh'a zdanie o postaci: *V jest A* indukuje rozkład możliwości Π określony na zbiorze X. W szczególności dla każdego $x \in X$ możliwość, że x jest wartością V, jest określana jako Π(x), przy czym, Π(x) = A(x), gdzie A(x) jest stopniem przynależności x w A. W tej sytuacji Π(x) wskazuje, w jakim stopniu x jest możliwą wartością V.

Ilustruje to następujący przykład:

Dana jest wiadomość, że Katarzyna jest trzydziestolatką. Na zadane pytanie: czy Katarzyna ma 45 lat?, odpowiedź jest prosta – z całą pewnością nie. Gdy zadane zostanie pytanie, czy Katarzyna ma więcej niż 20 lat? odpowiedź również jest prosta – na pewno tak. Ale gdy sformułowane zostanie pytanie, czy Katarzyna ma 33 lata? Najdokładniejszą odpowiedzią jest – możliwe, ale niepewne.

W celu rozwiązania tego problemu Zadeh [7] wprowadził dwie miary.

Niech A i B są podzbiorami rozmytymi zbioru X. Dane jest zdanie „V jest A" oraz interesuje nas potwierdzenie zdania: „V jest B".

Miarą potwierdzenia jest możliwość B przy danym A: Poss[B/A] = max[A(x) ∧B(x)], drugą miarą potwierdzenia jest pewność B przy danym A: Cert[B/A] = 1 - Poss[B/A].

2.1.1. Operacje na zbiorach rozmytych

W klasycznej teorii zbiorów zdefiniowane są operacje sumy zbiorów, iloczynu czy dopełnienia. Jeżeli dany element należy chociaż do jednego z sumowanych zbiorów, należy również do sumy tych zbiorów, a jeśli należy do obydwu zbiorów, należy również do przecięcia (iloczynu) zbiorów. Wszystkie operacje na zbiorach rozmytych będące uogólnieniem pojęć zbiorów nierozmytych sprowadzają się do klasycznych, gdy podzbiory rozmyte mają stopień przynależności dwuwartościowy {0,1}.

W teorii zbiorów rozmytych należy zadać pytanie „w jakim stopniu" badany element należy do sumy lub przecięcia zbiorów, jeśli wiadomo „w jakim stopniu" należy do każdego z sumowanych lub przecinanych zbiorów.

Operatory sumy, iloczynu muszą spełniać pewne własności, by można było je stosować w odniesieniu do zbiorów rozmytych. Wymagania te znane są pod nazwą norm trójkątnych (T-normy dla operacji iloczynu i S-normy dla operacji sumy).

Cztery warunki: przemienność, łączność, monotoniczność i odpowiednie elementy neutralne są stosowane do scharakteryzowania operatorów T-normy i S-normy [1].

T-norma: to funkcja T: $[0,1] \times [0,1] \rightarrow [0,1]$ taka, że $\forall a, b, c \in [0,1]$ o własnościach:

- $T(a, b) = T(b, a)$ – *przemienność,*
- $T(T(a, b), c) = T(a, T(b, c))$ – *łączność,*
- $T(a, b) \leq T(a, c)$ *dla* $b \leq c$ – *monotoniczność,*
- $T(a, 1) = a$ – *warunek brzegowy.*

S-norma: to funkcja S: $[0,1] \times [0,1] \rightarrow [0,1]$ taka, że $\forall a, b, c \in [0,1]$ o własnościach:

- $S(a, b) = S(b, a)$ – *przemienność,*
- $S(S(a, b), c) = S(a, S(b, c))$ – *łączność,*
- $S(a, b) \leq S(a, c)$ *dla* $b \leq c$ – *monotoniczność,*

– *S(a, 0) = a – warunek brzegowy.*

W przypadku gdy zbiór rozmyty degeneruje się do zbioru ostrego, operator sumy rozmytej (iloczynu rozmytego) zachowuje się jak klasyczny operator sumy (iloczynu).

Najczęściej stosowane *T-normy* to:

– Zadeha – minimum *T(a, b) = min(a, b)*, postać algebraiczna *min(a,b) = (a+b-|a-b|)/2*

– iloczyn algebraiczny *T(a, b) = a • b*,

– iloczyn logiczny (Łukasiewicza) *T(a, b) = max(a + b – 1, 0)*,

– iloczyn drastyczny *T(a, b) = min(a, b) dla max(a, b) = 1*, a *0* dla pozostałych.

Najczęściej stosowane *S-normy* to:

– Zadeha – maksimum *S(a, b) = max(a, b)*, postać algebraiczna *max(a,b) = (a+b+|a-b|)/2*

– suma probabilistyczna *S(a, b) =a + b - a • b*,

– suma logiczna (Łukasiewicza) *S(a, b) = min(a + b, 1)*,

– suma drastyczna *S(a, b) = max(a, b) dla min(a, b) = 0*, a *1* dla pozostałych.

2.2. Wnioskowanie przybliżone na przykładzie modelu Mamdaniego

Elementy pierwotne systemu wnioskowania przybliżonego to [1]:

– zbiór zmiennych wejściowych $U_1, ..., U_n$ wraz z odpowiadającą im rodziną zbiorów $X_1, ..., X_n$, przy czym X_k nazywany jest zbiorem bazowym (przestrzenią lub dziedziną) zmiennej U_k i zawiera zbiór wartości dopuszczalnych zmiennej U_k,

– zbiór zmiennych wyjściowych $V_1, ..., V_m$ wraz z odpowiadającą im rodziną zbiorów $Y_1, ..., Y_m$, przy czym Y_k nazywany jest zbiorem bazowym (przestrzenią lub dziedziną) zmiennej Y_k i zawiera zbiór wartości dopuszczalnych zmiennej Y_k,

– system reguł rozmytych $\{R_i; i=1, ..., r\}$.

W systemach z jednym wejściem i jednym wyjściem zakodowaną wiedzę można wyrazić za pomocą reguły JEŻELI-TO (jeśli przesłanka - to konkluzja):

JEŻELI U jest B_1 **TO** V jest D_1
TAKŻE
...
TAKŻE
JEŻELI U jest B_n **TO** V jest D_m,
gdzie:
U – zmienna wejściowa,

V – zmienna wyjściowa,

B_i, D_i – wartości lingwistyczne (etykiety) reprezentowane jako podzbiory rozmyte odpowiednich przestrzeni X i Y dla zmiennych U i V. Funkcje przynależności tych wartości lingwistycznych oznaczone są odpowiednio: $B_i(x)$, $D_i(y)$. Lewa strona reguły – poprzednik jest związana z wejściem systemu, prawa – następnik z wyjściem.

TAKŻE – spójnik alternatywny, który przekształca się w sumę logiczną.

Przykład

JEŻELI ciśnienie **JEST** duże **TO** zmiana_objętości **JEST** bardzo_duża

Procedura otrzymywania wyjścia rozmytego z bazy reguł jest następująca:

- znalezienie poziomu zapłonu każdej z reguł,
- znalezienie wyjścia każdej z reguł,
- agregacja poszczególnych wyjść reguły w celu otrzymania całkowitego wyjścia systemu.

Poszczególne kroki zostaną szczegółowo omówione na podstawie metody wnioskowania Mamdaniego.

Założenie:

Pojedyncza reguła ma postać:

JEŻELI U_1 jest B_{i1} I U_2 jest B_{i2} **TO** V jest D_i,

gdzie spójnik „I" jest interpretowany jako koniunkcja rozmyta.

Metoda wnioskowania Mamdaniego (oparta na regule max-min) [1] polega na wykonaniu następujących kroków:

1. Dla każdej reguły należy obliczyć stopień jej zapłonu (dopasowania do danych) τ_i, w następujący sposób:
   - $\tau_i = \forall_{x1} [B_{i1}(x_1) \wedge A_1(x_1)] \wedge \forall_{x2} [B_{i2}(x_2) \wedge A_2(x_2)]$, jeśli zmienne wejściowe przyjmują wartości rozmyte, reprezentowane odpowiednio przez podzbiory rozmyte A_1, A_2 ($U_1 = A_1$, $U_2 = A_2$), to poziom dopasowania między wejściową wartością rozmytą A_1 i etykietą lingwistyczną B_{i1} otrzymuje się z możliwości warunkowej:

     $\text{Poss}(B_{i1} | A_1) = \text{MAX}_{x1} [B_{i1}(x_1) \wedge A_1(x_1)]$, a nie ze stopnia przynależności, podobnie dla wartości rozmytej A_2 i etykiety rozmytej B_{i2}.

     Wtedy stopień zapłonu reguły obliczany jest:

     $\tau_i = \text{Poss}(B_{i1} | A_1) \wedge \text{Poss}(B_{i2} | A_2)$

     Przykład

     JEŻELI wiek jest około 50 I staż_pracy jest około 25 **TO** premia jest duża
   - $\tau_i = \forall_{x1} [B_{i1}(x_1^*)] \wedge \forall_{x2} [B_{i2}(x_2^*)]$, jeśli wejścia są liczbami nierozmytymi x_1^* i x_2^* np. wiek jest podany jako wartość liczbowa np. 48 i staż pracy również np. 24.
2. Znalezienie wyjścia i-tej reguły ($F_i(y)$) jako zbioru rozmytego zgodnie ze wzorem:

   $F_i(y) = \tau_i \wedge D_i(y)$.

3. Agregacja poszczególnych wyjść reguły, by otrzymać całkowite wyjście z systemu (F), które także jest zbiorem rozmytym nad przestrzenią Y.
   Przeprowadzenie agregacji wyprowadzonych zbiorów rozmytych F_i, za pomocą spójnika alternatywnego (**TAKŻE**), który przekształca się w agregację typu sumy logicznej wyjść Fi(y).

Podsumowując - w celu znalezienia wartości zmiennych wyjściowych należy wykonać cztery kroki:

1. Pobrać wartości zmiennych wejściowych (rozmyte lub nierozmyte).
2. Dla każdej reguły R_i ustalić stopień zapłonu (dopasowania). Stopień dopasowania τ_i obliczany jest na podstawie zbiorów rozmytych w przesłance reguły:
   Jeśli przesłanka ma postać **JEŻELI** U_1 jest B_1 I U_2 jest B_2, a pobrano dane nierozmyte $U_1 = x_1$, $U_2 = x_2$, wtedy stopień dopasowania danych do R_i można obliczyć stosując np. operator MIN.
   Stopień zapłonu i-tej reguły: $\tau_i = \text{MIN}(\mu_{Bi1}(x_1), (\mu_{Bi2}(x_2))$.
3. Na podstawie zbiorów rozmytych w przesłankach reguł oraz stopni dopasowania danych do reguł należy wyznaczyć rozmyty wynik wnioskowania (wyjście). Wynik ten jest zbiorem rozmytym w przestrzeni możliwych wartości zmiennej wejściowej i pokazuje, na ile poszczególne wyniki są zgodne z danymi.
4. Ostatni etap to defuzzyfikacja wyniku rozmytego, czyli uzyskanie ze zbioru możliwych wyników konkretnej wartości ostrej.

Sposób otrzymywania jednej wartości ostrej z rozmytego wyniku wnioskowania stanowi część mechanizmu wnioskowania. Na przykład, jeśli zastosowano k-reguł rozmytych, w konkluzjach których jest ta sama zmienna wyjściowa V, to efektem wnioskowania prowadzonego dla każdej reguły z osobna jest k zbiorów rozmytych, tzw. rozmytych wyników wnioskowania. Na podstawie kształtów zbiorów rozmytych należy znaleźć dokładną wartość wyjścia F(y).

Jest wiele strategii uzyskiwania najbardziej reprezentatywnej wartości ostrej (defuzzyfikacji). W modelu Mamdaniego zastosowano operator MAX. Są jednak przypadki, gdzie operator ten może generować kilka lub nieskończenie wiele rozwiązań o tej samej wartości funkcji przynależności. Aby tego uniknąć stosuje się inne metody defuzzyfikacji.

3. Logika rozmyta i metoda wnioskowania przybliżonego w aproksymacyjnym wyszukiwaniu informacji w bazach danych

3.1. Wykorzystanie teorii wnioskowania przybliżonego w procesie generacji odpowiedzi na niedokładne pytania zadawane do bazy danych zawierającej dokładne dane

Założenia:

W systemie istnieje tabela z atrybutami A_1, ..., A_n (wartości atrybutów są liczbami nierozmytymi), zdefiniowane są zbiory rozmyte B_1, ..., B_n oraz odpowiednie funkcje przynależności.

Użytkownik zadaje pytanie rozmyte względem atrybutów A_i i A_j o postaci np.:

Select A_k, A_l from nazwa_tabeli

Where A_i jest B_i and A_j jest B_j

Korzystając z najprostszego algorytmu wnioskowania MAMDANIEGO opartego na regule wnioskowania max-min proces wyszukiwania informacji najbardziej zbliżonej do podanej w zapytaniu powinien w tym przypadku przebiegać następująco:

- Dla każdego wiersza w tabeli należy określić stopień przynależności wartości atrybutu A_i do podanego zbioru rozmytego B_i, następnie stopień przynależności wartości atrybutu A_j do podanego zbioru B_j.
- Stopień przynależności wiersza do zbioru wyjściowego (stopień dopasowania) odpowiedzi obliczany jest jako:
  $\tau = \text{MIN}(\mu_{Bi}(A_i), (\mu_{Bj}(A_j))$.
- Po tych krokach uzyskano zbiór wynikowy złożony z wierszy z określonym stopniem dopasowania do danych specyfikowanych w pytaniu.
- W zależności od zastosowań należy na wyjściu określić najbardziej reprezentatywne wiersze tabeli, np. jako kryterium wybrać operator MAX (wybierający wiersze o maksymalnym stopniu dopasowania) lub zastosować inne metody defuzzyfikacji.

Przykład 1

W systemie istnieje tabela *Personalia*, zawierająca informacje o pracownikach, kolumny tej tabeli zawierają dokładne dane (tab. 1). Użytkownik, nie znając dokładnych danych o tych pracownikach, chce uzyskać informacje o pracownikach około pięćdziesiątki, mających staż pracy około 20 lat.

Select imię, nazwisko
From personalia

Where wiek około 50 and staż_pracy około 20;

Tabela 1

Personalia

Nr	Imię	Nazwisko	Wiek	Staż_pracy	Płeć	Adres
1	Jan	Kowalski	48	19	M	Zabrze
2	Kasia	Nowak	38	10	K	Chorzów
3	Marcin	Sowa	21	1	M	Gliwice
4	Jakub	Sroka	53	22	M	Kraków
5	Anna	Maj	47	8	K	Katowice

Wyraz rozmyty **około** zdefiniowany jest (dla porównania) za pomocą dwóch różnych funkcji przynależności:

1. **Funkcji Gaussa, opisanej wzorem** [3]:

$$\mu_G(x) = e^{-\left(\frac{x-b}{a}\right)^2}. \tag{1}$$

Przebieg funkcji określony jest przez dwa parametry *a, b*, gdzie odpowiednio:

b – wartość modalna funkcji (najbardziej typowa wartość x dla zbioru rozmytego),

a – szerokość tej funkcji (funkcja Gaussa ma szerokość 2a na poziomie $\mu(x)=e^{-1}$).

W celu wyznaczenia parametru *a* można posłużyć się pojęciem **punktu krytycznego k** funkcji przynależności. Jest to taki punkt, dla którego wartość funkcji przynależności wynosi 0.5. Krzywa Gaussa posiada dwa takie punkty.

$$\mu_G(x_k) = e^{-\left(\frac{x-b}{a}\right)^2} = 0.5. \tag{2}$$

Po przekształceniach, wartość *a* określana jest następująco:

$$a = \frac{|x_k - b|}{\sqrt{\ln 2}}. \tag{3}$$

2. **Funkcji trójkątnej, opisanej wzorem** [3]:

$$\mu_T(x) = w\left(\frac{a-|x-b|}{a}\right), \tag{4}$$

gdzie zmienna logiczna *w* ma wartość:

$$w = \begin{cases} 1, & \text{gdy } (b-a) \le x < (b+a) \\ 0, & \text{w każdym pozostałym przypadku} \end{cases} \tag{5}$$

Funkcje przynależności określane dla poszczególnych wyrazów rozmytych rozpatrywanego przykładu mają postać:

Wiek około 50

– $\mu_G(x) = e^{-\left(\frac{x-50}{6}\right)^2}$, gdzie x – wiek, a parametry b *i* a z równania (1) wynoszą odpowiednio 50 i 6.

– $\mu_T(x) = w\left(\frac{10-|x-50|}{10}\right)$, gdzie x – wiek, parametry b *i* a z równania (4) wynoszą odpowiednio 50 i 10, a zmienna logiczna w z równania (5) wynosi:

$$w = \begin{cases} 1, \text{ gdy } 40 \le x < 60 \\ 0, \text{ w każdym pozostałym przypadku} \end{cases}$$

Staż pracy około 20

– $\mu_G(x) = e^{-\left(\frac{x-20}{6}\right)^2}$, gdzie x – staż pracy, a parametry b *i* a z równania (1) wynoszą odpowiednio 20 i 6.

– $\mu_T(x) = w\left(\frac{10-|x-20|}{10}\right)$, gdzie x – staż pracy, parametry b *i* a z równania (4) wynoszą odpoeidnio 20 i 10, a zmienna logiczna w z równania (5) wynosi:

$$w = \begin{cases} 1, \text{ gdy } 10 \le x < 30 \\ 0, \text{ w każdym pozostałym przypadku} \end{cases}$$

Przykład 1a

Dla wyrazów rozmytych ***około***, których funkcje przynależności zdefiniowano za pomocą funkcji Gaussa:

– wyznaczono stopnie przynależności szukanych atrybutów w każdym wierszu tabeli **Personalia**, zgodnie z punktami przedstawionymi powyżej,

– po wykonaniu niezbędnych obliczeń dla wyrazów rozmytych ***około*** określonych za pomocą funkcji Gaussa, otrzymano wyniki, przedstawione w tab. 2.

Tabela 2

Personalia z wyliczonymi stopniami przynależności

Nr	Imię	Nazwisko	Wiek	μ_{50}(Wiek)	Staż_pracy	μ_{20}(Staż_pracy)	Płeć	Adres
1	Jan	Kowalski	48	0,895	19	0,973	M	Zabrze

cd. tab. 2

2	Kasia	Nowak	38	0,018	10	0,062	K	Chorzów
3	Marcin	Sowa	21	0,001	1	0,001	M	Gliwice
4	Jakub	Sroka	53	0,779	22	0,895	M	Kraków
5	Anna	Maj	47	0,779	8	0,018	K	Katowice

– Dla każdego wiersza wyznaczono stopień dopasowania do zadawanego pytania, zgodnie ze wzorem:

$$\tau = \text{MIN}\ (\mu_{50}(\text{Wiek}), (\mu_{20}(\text{Staż\_pracy}))).$$

Wyniki po wykonaniu tej operacji przedstawiono w tab. 3.

Tabela 3

Personalia z obliczonym stopniem zapłonu (dopasowania) dla każdego wiersza

Nr	Imię	Nazwisko	Wiek	τ	Staż_pracy	Płeć	Adres
1	Jan	Kowalski	**48**	**0,895**	**19**	M	Zabrze
2	Kasia	Nowak	38	0,018	10	K	Chorzów
3	Marcin	Sowa	21	0,001	1	M	Gliwice
4	Jakub	Sroka	**53**	**0,779**	**22**	M	Kraków
5	Anna	Maj	47	0,018	8	K	Katowice

Otrzymano zbiór wyjściowy zawierający wiersze z odpowiadającymi im stopniami dopasowania do zadanego pytania. Na tym etapie, w zależności od wymagań użytkownika, można:

– znaleźć elementy najbardziej odpowiadające pytaniu (wykorzystując operator MAX),

– znaleźć zbiór wyjściowy tych wierszy, których stopień dopasowania jest większy od zadanego (np. zwykle 0,5),

– wykorzystać inną metodę defuzzyfikacyjną.

Przykład 1b

Odpowiednio dla liczb rozmytych ***około*** zdefiniowanych za pomocą funkcji trójkątnej, po wykonaniu opisanych wcześniej operacji, otrzymano wyniki przedstawione w tab. 4:

Tabela 4

Personalia z wyliczonymi stopniami przynależności

Nr	Imię	Nazwisko	Wiek	μ_{50}(Wiek)	Staż_pracy	μ_{20}(Staż_pracy)	Płeć	Adres
1	Jan	Kowalski	48	0,8	19	0,9	M	Zabrze
2	Kasia	Nowak	38	0,0	10	0,0	K	Chorzów
3	Marcin	Sowa	21	0,0	1	0,0	M	Gliwice
4	Jakub	Sroka	53	0,7	22	0,8	M	Kraków
5	Anna	Maj	47	0,7	8	0,0	K	Katowice

– Dla każdego wiersza wyznaczono stopień dopasowania:

$\tau = MIN(\mu_{50}(Wiek), (\mu_{20}(Staż\_pracy))$

Wyniki przedstawiono w tab. 5.

Tabela 5

Personalia z obliczonym stopniem zapłonu (dopasowania) dla każdego wiersza

Nr	Imię	Nazwisko	Wiek	τ	Staż_pracy	Płeć	Adres
1	Jan	Kowalski	48	0,8	19	M	Zabrze
2	Kasia	Nowak	38	0,0	10	K	Chorzów
3	Marcin	Sowa	21	0,0	1	M	Gliwice
4	Jakub	Sroka	53	0,7	22	M	Kraków
5	Anna	Maj	47	0,0	8	K	Katowice

Otrzymano zbiór wyjściowy – wiersze z odpowiadającymi im stopniami dopasowania do zadanego pytania. Podobnie jak w przykładzie poprzednim należy ograniczyć zbiór wynikowy zgodnie z wymaganiami użytkownika, wykorzystując odpowiednie metody defuzzyfikacyjne.

3.2. Wykorzystanie teorii wnioskowania przybliżonego w procesie generacji odpowiedzi na niedokładne pytania zadawane do bazy danych zawierającej rozmyte dane

Rozpatrywany przypadek jest uogólnieniem wszystkich opisanych w rozdziale 1 przypadków. Zastosowano tu również najprostszy algorytm wnioskowania MAMDANIEGO oparty na regule wnioskowania max-min. W tym przypadku zakłada się, że:

– w systemie istnieje tabela z atrybutami A_1, ..., A_n (wartości atrybutów mogą być liczbami lub zbiorami rozmytymi), zdefiniowane są zbiory rozmyte B_1, ..., B_n oraz odpowiednie funkcje przynależności,

– wartościami atrybutów mogą być liczby rozmyte, przedziały wartości lub podzbiory rozmyte, np. wiek – młody; średni; około 20; między 30 lub 40; 26-28; 12,14,

– użytkownik zadaje pytanie rozmyte względem atrybutów A_i i A_j:

Select A_k, A_l *from nazwa_tabeli*

Where A_i *jest* B_l *and* A_j *jest* B_j

Zgodnie z algorytmem Mamdaniego proces wyszukiwania przebiega następująco:

– Dla każdego wiersza w tabeli należy określić stopień przynależności wartości atrybutu A_i do podanego zbioru rozmytego B_i, następnie stopień przynależności wartości atrybutu A_j do podanego zbioru B_j. W tym przypadku dla każdego atrybutu, o który pytano w zapytaniu, należy wyznaczyć stopnie przynależności dla każdej jego

wartości lub wykonać iloczyn funkcji przynależności wartości atrybutu z funkcją przynależności zbioru wejściowego.

- Jeśli w wyniku poprzedniej operacji uzyska się kilka stopni przynależności, dla każdej wartości atrybutu w tym wierszu lub w wyniku wykonania operacji iloczynu funkcji przynależności, należy spośród wszystkich obliczonych stopni jako stopień przynależności (dopasowania) atrybutu wybrać stopień o maksymalnej wartości.
- Operacje te należy wykonać dla wszystkich rozmytych atrybutów, o które pytano w zapytaniu.
- Stopień przynależności wiersza do zbioru wyjściowego (stopień dopasowania) odpowiedzi obliczany jest jako:
  $\tau = \text{MIN}\,(\mu_{Bi}(A_i), (\mu_{Bj}(A_j))$.
- Po tych krokach uzyskuje się zbiór wynikowy złożony z wierszy z określonym stopniem dopasowania (zapłonu) do danych specyfikowanych w pytaniu.
- W zależności od zastosowań należy na wyjściu określić najbardziej reprezentatywne wiersze tabeli, np. jako kryterium wybrać operator MAX (wybierający wiersze o maksymalnym stopniu dopasowania) lub zastosować inne metody defuzzyfikacji.

Przykład 2

W systemie istnieje tabela *Personalia* (tab. 6), zawierająca informacje o pracownikach; kolumny tej tabeli nie zawierają dokładnych danych. Atrybuty *wiek* oraz *staż pracy* nie są dokładnie sprecyzowane i ich wartości są zbiorami liczb (np. wiek 24 lub 25, lub 26 albo między 23 a 26). Użytkownik ponownie, nie znając dokładnych danych o tych pracownikach, chce uzyskać informacje o pracownikach koło pięćdziesiątki, mających staż pracy około 20 lat.

Select imię, nazwisko
From personalia
Where wiek około 50 and staż_pracy około 20;

Tabela 6

Personalia

Nr	Imię	Nazwisko	Wiek	Staż_pracy	Płeć	Adres
1	Jan	Kowalski	47 – 49	19 – 21	M	Zabrze
2	Kasia	Nowak	37 - 39	10 – 12	K	Chorzów
3	Marcin	Sowa	20 – 22	1 – 3	M	Gliwice
4	Jakub	Sroka	52 – 54	21 – 23	M	Kraków
5	Anna	Maj	46 – 48	7 – 9	K	Katowice

W przykładzie tym wyrazy rozmyte ***około 50*** i ***około 20*** mają funkcje przynależności określone za pomocą funkcji Gaussa lub trójkątnej, podobnie jak w przykładzie 1.

Przykład 2a

Wyrazy rozmyte ***około 50*** i ***około 20*** określają funkcje przynależności zdefiniowane za pomocą funkcji Gaussa. Proces generacji odpowiedzi na zadane przez użytkownika pytanie przebiega następująco:

- Dla każdego atrybutu, na który nałożony jest warunek w pytaniu, wyznacza się stopnie przynależności dla każdej wartości tego atrybutu rozmytego w każdym wierszu tabeli *Personalia*.
- Po wykonaniu niezbędnych obliczeń dla wyrazów rozmytych ***około*** określonych za pomocą funkcji Gaussa otrzymano wyniki przedstawione w tab. 7.

Tabela 7

Personalia z wyliczonymi stopniami przynależności dla każdej wartości rozmytego atrybutu

Nr	Imię	Nazwisko	Wiek	μ_{50}(Wiek)	Staż_pracy	μ_{20}(Staż_pracy)	Płeć	Adres
1	Jan	Kowalski	47	0,779	19	0,973	M	Zabrze
			48	0,895	20	**1**		
			49	**0,973**	21	0,973		
2	Kasia	Nowak	37	0,009	10	0,062	K	Chorzów
			38	0,018	11	**0,105**		
			39	**0,032**	12	**0,169**		
3	Marcin	Sowa	20	≈0,001	1	≈0,001	M	Gliwice
			21	≈0,001	2	≈0,001		
			22	≈0,001	3	≈0,001		
4	Jakub	Sroka	52	**0,895**	21	**0,973**	M	Kraków
			53	0,779	22	0,895		
			54	0,641	23	0,779		
5	Anna	Maj	46	0,641	7	0,009	K	Katowice
			47	0,779	8	0,018		
			48	**0,895**	9	**0,032**		

- Spośród wyznaczonych stopni dopasowania poszczególnych wartości atrybutów rozmytych wybiera się maksymalny stopień dopasowania dla każdego atrybutu w wierszu.
- Po wykonaniu tej operacji, wiersze z wyznaczonymi stopniami dopasowania dla atrybutu przedstawione są w tab. 8.

Tabela 8

Personalia z max wartościami dopasowania atrybutów rozmytych

Nr	Imię	Nazwisko	Wiek	μ_{50}(Wiek)	Staż_pracy	μ_{20}(Staż_pracy)	Płeć	Adres
1	Jan	Kowalski	49	**0,973**	20	**1**	M	Zabrze
2	Kasia	Nowak	39	**0,032**	12	**0,169**	K	Chorzów
3	Marcin	Sowa	22	≈0,001	3	≈0,001	M	Gliwice

cd. tab. 8

4	Jakub	Sroka	52	**0,895**	21	**0,973**	M	Kraków
5	Anna	Maj	48	**0,895**	9	0,032	K	Katowice

– Dla każdego wiersza wyznaczono stopień dopasowania do zadawanego pytania, zgodnie ze wzorem:

$$\tau = \text{MIN}\ (\mu_{50}(\text{Wiek}), (\mu_{20}(\text{Staż\_pracy}))).$$

– Po wykonaniu tej operacji wiersze z wyznaczonymi stopniami dopasowania przedstawione są w tab. 9.

Tabela 9

Personalia z obliczonym stopniem zapłonu (ufności) dla każdego wiersza

Nr	Imię	Nazwisko	Wiek	τ	Staż_pracy	Płeć	Adres
1	Jan	Kowalski	49	**0,973**	20	M	Zabrze
2	Kasia	Nowak	39	0,032	12	K	Chorzów
3	Marcin	Sowa	22	≈0,001	3	M	Gliwice
4	Jakub	Sroka	52	**0,895**	21	M	Kraków
5	Anna	Maj	48	0,032	9	K	Katowice

Otrzymany zbiór wyjściowy zawiera wiersze z odpowiadającymi im stopniami dopasowania do zadanego pytania. Na tym etapie, w zależności od wymagań użytkownika, można:

– znaleźć elementy najbardziej odpowiadające pytaniu (wykorzystując operator MAX),

– znaleźć zbiór wyjściowy tych wierszy, których stopień dopasowania jest większy od zadanego (zwykle 0,5),

– wykorzystać inną metodę defuzzyfikacyjną.

Przykład 2b

Odpowiednio dla liczb rozmytych ***około 50 i około 20***, których funkcje przynależności zdefiniowano za pomocą funkcji trójkątnej, wykonano kolejno takie same operacje jak w przykładzie 2a. Otrzymane wyniki przedstawione są w tab. 10, 11 i 12.

Tabela 10

Personalia z wyliczonymi stopniami przynależności dla każdej wartości rozmytego atrybutu

Nr	Imię	Nazwisko	Wiek	μ_{50}(Wiek)	Staż_pracy	μ_{20}(Staż_pracy)	Płeć	Adres
			47	0,7	19	0,9		
1	Jan	Kowalski	48	0,8	20	1	M	Zabrze
			49	0,9	21	0,9		

cd. tab. 10

2	Kasia	Nowak	37 38 39	0,0 0,0 0,0	10 11 12	0,0 0,1 0,2	K	Chorzów
3	Marcin	Sowa	20 21 22	0,0 0,0 0,0	1 2 3	0,0 0,0 0,0	M	Gliwice
4	Jakub	Sroka	52 53 54	0,8 0,7 0,6	21 22 23	0,9 0,8 0,7	M	Kraków
5	Anna	Maj	46 47 48	0,6 0,7 0,8	7 8 9	0,0 0,0 0,0	K	Katowice

Tabela 11

Personalia z maksymalnymi wartościami dopasowania atrybutów rozmytych

Nr	Imię	Nazwisko	Wiek	μ_{50}(Wiek)	Staż_pracy	μ_{20}(Staż_pracy)	Płeć	Adres
1	Jan	Kowalski	49	0,9	20	1	M	Zabrze
2	Kasia	Nowak	39	0,0	12	0,2	K	Chorzów
3	Marcin	Sowa	22	0,0	3	0,0	M	Gliwice
4	Jakub	Sroka	52	0,8	21	0,9	M	Kraków
5	Anna	Maj	48	0,8	9	0,0	K	Katowice

Tabela 12

Personalia z obliczonym stopniem zapłonu (ufności)

Nr	Imię	Nazwisko	Wiek	τ	Staż_pracy	Płeć	Adres
1	Jan	Kowalski	49	0,9	20	M	Zabrze
2	Kasia	Nowak	39	0,0	12	K	Chorzów
3	Marcin	Sowa	22	0,0	3	M	Gliwice
4	Jakub	Sroka	52	0,8	21	M	Kraków
5	Anna	Maj	48	0,0	9	K	Katowice

4. Podsumowanie

W świecie rzeczywistym w wielu sytuacjach logika boolowska jest niewystarczająca, ludzie myślą w sposób nieostry (nieprecyzyjny, rozmyty) i w ten również sposób wyrażają swoje opinie, zdania i zadają pytania. Znacznie prościej i wygodniej jest móc zadać pytanie nie do końca sprecyzowane i niezbyt ścisłe. Dlatego też teoria zbiorów rozmytych jak i mechanizmy wnioskowania rozmytego znalazły zastosowanie w bazach danych.

W pracy przedstawiono metody oraz zilustrowano je przykładami, które ukazują możliwość wykorzystania teorii zbiorów rozmytych, logiki rozmytej i metod wnioskowania

przybliżonego w celu wyszukiwania informacji zapisanej w bazie danych w sposób nieprecyzyjny czy niepełny, a także uzyskiwania odpowiedzi na niedokładne i nieprecyzyjne pytania.

LITERATURA

1. Yager R, Filev D.: Podstawy modelowania i sterowania rozmytego. Wydawnictwa Naukowo-Techniczne, Wiley. Warszawa 1995.
2. Łachwa A.: Rozmyty świat zbiorów, liczb, relacji, faktów, reguł i decyzji. Akademicka Oficyna Wydawnicza Exit, Warszawa 2001.
3. Piegat A.: Modelowanie i sterowanie rozmyte. Akademicka Oficyna Wydawnicza Exit, Warszawa 1999.
4. Badurek J.: Logika rozmyta w bazach danych. Informatyka. Styczeń 1999.
5. Zadeh L. A.: Fuzzy sets, Information and Control 8, 338-353, 1965.
6. Dubois D., Prade H.: Possibility Theory: An Approach to Computerized Processing of Uncertainty. Plenum Press: New York, 1988.
7. Zadeh L. A.: „Fuzzy sets an information granularity" in Advances in Fuzzy Set Theory and Applications. Amsterdam: North-Holland, 3-18, 1979.

Recenzent: Dr hab. inż. Stanisław Wołek, prof. Pol. Rzeszowskiej

Wpłynęło do Redakcji 27 grudnia 2001 r.

Abstract

This article presents fuzzy sets theory and approximate reasoning and their applications in database. Database can include precise or imprecise data, and queries can be precise or imprecise too.

In standard relational databases, both user queries and data are precise. But the occurrences of imprecise data in databases are natural. Many of people prefer to submit imprecise queries than the exact queries.

In general, three types of imprecision may arise:

- Imprecise Queries – precise data in database;
- Precise Queries – imprecise data in database;

– Imprecise Queries – imprecise data in database.

In the chapter 1 author introduces problems which occur with imprecision in database. In the next chapter the concept of a fuzzy set and approximate reasoning are defined.

In the third chapters the examples of applications approximate reasoning in processing fuzzy relational query are presented.