WYŻSZA SZKOŁA
INFORMATYKI STOSOWANEJ
I ZARZĄDZANIA

ANALIZA SYSTEMOWA W FINANSACH I ZARZĄDZANIU

Wybrane problemy

Tom 2

Pod redakcją

Macieja KRAWCZAKA i Jerzego HOŁUBCA

Warszawa 2000

WYŻSZA SZKOŁA
INFORMATYKI STOSOWANEJ
I ZARZĄDZANIA

ANALIZA SYSTEMOWA W FINANSACH I ZARZĄDZANIU

Wybrane problemy
Tom 2

Pod redakcją
Macieja KRAWCZAKA i Jerzego HOŁUBCA

Warszawa 2000

Wykaz opiniodawców artykułów zamieszczonych w tomie:

doc dr hab. Dariusz **GĄTAREK**

prof. dr hab. Jakub **GUTENBAUM**

prof. dr hab. Jerzy **HOŁUBIEC**

doc. dr hab. Marek **LIBURA**

prof. dr hab. Stanisław **PIASECKI**

prof. dr hab. Andrzej **STRASZAK**

doc. dr hab. Sławomir **WIERZCHOŃ**

doc dr. hab. Leszek **ZAREMBA**

Warszawa 2000

ISBN 83-85847-54-5

ZASTOSOWANIE TESTÓW DO HIPOTEZ ROZMYTYCH W TWORZENIU ROZMYTYCH ZAPYTAŃ

Edyta Mrówka[*] *i Przemysław Grzegorzewski*[$,*]
[*]*Wyższa Szkoła Informatyki Stosowanej i Zarządzania*
[$]*Instytut Badań Systemowych, Polska Akademia Nauk*

Praca przedstawia metodę konstrukcji zapytania rozmytego przeznaczonego do wyszukiwania informacji w bazie danych zawierającej dane obciążone niepewnością o charakterze losowym. Zapytanie takie łączy w sobie elementy zwykłego zapytania rozmytego z testowaniem hipotez rozmytych.

Słowa kluczowe: wyszukiwanie w bazie danych, zapytania rozmyte, logika rozmyta, testowanie hipotez, hipotezy rozmyte, obliczenia na słowach.

1. Wstęp

Podstawową funkcją systemów zarządzania bazami danych (SZBD), oprócz przechowywania zbiorów danych jest również możliwość ich przetwarzania. Proces przetwarzania danych polega między innymi na konstrukcji zapytań za pomocą właściwych języków zapytań do baz danych (QUEL, QBE i najpopularniejszy SQL). Posługując się klasycznymi narzędziami, użytkownik jest w stanie skonstruować zapytanie oparte na logice dwuwartościowej. Zapytania takie choć dają się w dość łatwy sposób przełożyć na język maszynowy stanowią dość znaczące ograniczenie dla sposobu myślenia użytkownika. Poza tym zbiór elementów stanowiących odpowiedź na tak skonstruowane zapytanie jest dość często zbiorem zawężonym, tzn. nie uwzględniane są elementy, które w dość wysokim stopniu, choć nie stuprocentowo, spełniają warunki zapytania. Język naturalny, którym posługuje się człowiek do przetwarzania informacji oraz opisu otaczających go zjawisk, dopuszcza stosowanie pojęć rozmytych (lingwistycznych) takich jak: wysoki mężczyzna, zanieczyszczone jezioro, niska cena itp. Niestety, systemy zarządzania bazami danych nie dostarczają narzędzi do konstrukcji zapytań z użyciem terminów nieprecyzyjnych. Istnieje jednak możliwość wzbogacenia klasycznych języków zapytań do baz danych o elementy umożliwiające konstruowanie takich zapytań. Przekształcenie zapytania opartego na logice zerojedynkowej na zapytanie oparte na logice rozmytej nazywać będziemy *relaksacją rozmytą*.

W niniejszej pracy przez relaksację rozmytą rozumieć będziemy proces wykorzystujący takie narzędzia matematyczne, jak logika rozmyta i teoria zbiorów rozmytych Zadeha [20].

Istnieją dwa sposoby zastosowania zbiorów rozmytych w bazach danych. W pierwszym z nich struktura bazy danych nie ulega zmianie i opiera się na modelu relacyjnym. Logika rozmyta wykorzystywana jest na poziomie wyciągania informacji z bazy danych (wyżej opisany proces relaksacji rozmytej). Istnieje już kilka „systemów nakładek" umożliwiających tworzenie i wykorzystanie zapytań rozmytych do relacyjnych baz danych, między innymi system FQUERY for MS ACCES, opracowany przez Kacprzyka i Zadrożnego [10-15], jako nakładka na system zarządzania relacyjnymi bazami danych Microsoft ACCESS.

Drugie podejście polega na takiej ingerencji w strukturę bazy danych aby możliwa była opcja manipulowania informacją niepewną. Podejście to prowadzi więc do stworzenia rozmytej bazy danych. Niestety takie bazy danych pozostają, jak dotąd, jedynie w fazie projektów.

Niniejsza praca zawiera rozważania odwołujące się do pierwszego podejścia. Rozpatrujemy relacyjny model bazy danych. Przechowywane w bazie informacje są zadane precyzyjnie natomiast zastosowanie logiki rozmytej następuje na poziome przetwarzania informacji.

Sytuacja zaczyna wyglądać szczególnie interesująco w momencie gdy jeden atrybut jest opisywany przez wiele wartości liczbowych obarczonych niepewnością o charakterze losowym. Zilustrujmy to przykładem: załóżmy, że w naszej bazie danych chcemy przechowywać informacje dotyczące głębokości jezior znajdujących się na terenie Polski. Określając głębokość konkretnego jeziora korzystamy z pomiarów dokonanych w kilku losowo wybranych miejscach. Nietrudno zauważyć, że nasze dane mają charakter losowy i stanowią jedynie pewną realizację próby losowej. W najprostszym przypadku do oszacowania nieznanej wielkości może posłużyć nam średnia z próby. Jednakże wnioskowanie o "przeciętnej" głębokości jeziora wyłącznie na podstawie średniej zawierałoby poważny błąd metodologiczny, mający swe źródło w zignorowaniu losowego charakteru przetwarzanych danych.

W niniejszej pracy proponujemy metodę konstrukcji zapytania rozmytego odnoszącego się do danych obarczonych niepewnością o charakterze losowym, wykorzystującą odpowiedni test statystyczny. Nie jest to jednak testowanie hipotez w ujęciu klasycznym, ale weryfikacja hipotez rozmytych. Oprócz opisu budowy zapytań (rozdz. 2 i 3) i hipotez rozmytych (rozdz. 4) pokażemy w jaki sposób połączyć te dwa elementy, by następnie efektywnie wykorzystać je w procesie przetwarzania informacji (rozdz. 5). Schemat wyszukiwania informacji z bazy danych za pomocą logiki rozmytej

zaprezentujemy na przykładzie wyżej wymienionej nakładki FQUERY for MS ACCES. Rozważane zagadnienie można potraktować jako próbę praktycznej realizacji wysuniętej przez Zadeha [21, 23] koncepcji „obliczeń na słowach" (ang. computing with words).

2. Konstrukcja zapytania rozmytego

Budując zapytanie do bazy danych za pomocą języka SQL posługujemy się instrukcją SELECT z odpowiednimi klauzulami (obowiązkowo FROM, opcjonalnie: WHERE, GROUP BY, HEAVING, ORDER BY, itp.). W niniejszej pracy ograniczymy się do polecenia SELECT ... FROM ...WHERE z pominięciem pozostałych klauzul. Całą uwagę skoncentrujemy na klauzuli WHERE, ponieważ jej postać w sposób decydujący wpływa na liczbę rekordów stanowiących odpowiedź. Proponujemy rozszerzenie tradycyjnej składni o pewne elementy bazujące na logice rozmytej i teorii zbiorów rozmytych. Oto przykład prostego zapytania wykorzystującego logikę rozmytą:

SELECT jezioro FROM baza danych o jeziorach WHERE większość głębokość = „głębokie jezioro" I zanieczyszczenie = "bardzo małe".

Klauzula WHERE zawiera specyficzne elementy, które przyczyniają się do interpretacji tego zapytania jako rozmytego. Omówimy teraz poszczególne elementy występujące w naszym przykładzie, którymi są:

- wartości rozmyte,
- spójniki rozmyte,
- kwantyfikatory lingwistyczne,
- modyfikatory

WARTOŚCI ROZMYTE

SELECT jezioro FROM baza danych o jeziorach WHERE większość głębokość = „**głębokie jezioro**" I zanieczyszczenie = "bardzo **małe**".

Jednym z najważniejszych elementów zapytania rozmytego są **wartości rozmyte**. Mogą one być definiowane i przechowywane w bazie jako zbiory rozmyte. Przykładowe wartości rozmyte to: wysoki, niski, głębokie, płytkie, itp. Termin rozmyty jak widać jest pojęciem subiektywnym i będzie przybierał rożne znaczenia w zależności od interpretacji danego użytkownika. W systemie FQUERY for MS ACCESS, udostępnione są

mechanizmy pozwalające użytkownikowi na konstruowanie elementów rozmytych m.in. takich jak: terminy, relacje i kwantyfikatory rozmyte.

SPÓJNIKI ROZMTYE

SELECT jezioro FROM baza danych o jeziorach WHERE większość głębokość = „głębokie jezioro" **I** zanieczyszczenie = "bardzo małe".

Zastosowanie **rozmytych spójników** ma uzasadnienie w przypadku warunków złożonych, istnieje wówczas potrzeba agregacji wyników cząstkowych. W ujęciu klasycznym do łączenia sekwencji prostych warunków używane są m.in. spójniki I oraz LUB. Jak nie trudno się domyślić, mają one swoje odpowiedniki rozmyte.

Rozmyty spójnik I modeluje się w naturalny sposób przy pomocy przecięcia dwóch zbiorów rozmytych. Korzystając z definicji podanej przez Zadeha [20] zdaniu *A* **I** *B* odpowiada zbiór rozmyty o funkcji przynależności

$$\mu_{A\cap B}(x) = \mu_A(x) \wedge \mu_B(x) = \min\{\mu_A(x), \mu_B(x)\}, \ \forall x \in X,$$

gdzie μ_A i μ_B są funkcjami przynależności zbiorów rozmytych odpowiadających terminom *A* i *B*.

Podobnie rozmyty spójnik LUB modelujemy przy pomocy sumy dwóch zbiorów rozmytych. Zdaniu *A* **LUB** *B* odpowiada zbiór rozmyty o funkcji przynależności

$$\mu_{A\cup B}(x) = \mu_A(x) \vee \mu_B(x) = \max\{\mu_A(x), \mu_B(x)\}, \ \forall x \in X.$$

Watro tu wspomnieć, że do modelowania spójników I oraz LUB można użyć, w zależności od konkretnych potrzeb, inne t-normy bądź t-konormy.

KWANTYFIKATORY ROZMYTE

SELECT jezioro FROM baza danych o jeziorach WHERE **większość** głębokość = „głębokie jezioro" I zanieczyszczenie = "bardzo małe".

Logika dwuwartościowa dopuszcza stosowanie dwóch kwantyfikatorów: $\forall$ - dla każdego, $\exists$ - istnieje. Posługując się logiką rozmytą jesteśmy w stanie skonstruować kwantyfikatory lingwistyczne takie jak: prawie wszystkie, około *x*, powyżej *x*, itp. Według Zadeha [22] wyróżniamy następujące typy kwantyfikatorów:

- absolutne (około *x*% , prawie *x*, itp.). Kwantyfikator ten możemy opisać jako podzbiór rozmyty Q w zbiór nieujemnych liczb rzeczywistych.

Wartość funkcji przynależności dla dowolnego argumentu należącego do zbioru nieujemnych liczb rzeczywistych oznacza stopień zgodności tego argumentu z pojęciem rozmytym Q.

- proporcjonalne (mniejszość, prawie wszystkie). Możemy go opisać jako podzbiór rozmyty Q pewnego przedziału liczbowego. Przeważnie przyjmuje się przedział [0,1].

MODYFIKATORY

SELECT jezioro FROM baza danych o jeziorach WHERE większość głębokość = „głębokie jezioro" I zanieczyszczenie = "**bardzo** małe".

Przez modyfikator w rozmyty rozumiemy funkcję rzeczywistą przekształcającą pierwotną funkcję przynależności zbioru. Przykłady modyfikatorów to terminy takie jak: bardzo, mniej więcej, itp. W literaturze przedmiotu ([22], [8]) wyróżniamy modyfikatory wzmacniające dany termin (np. bardzo) oraz osłabiające (np. mniej niż). Najpopularniejszymi modyfikatorami są funkcje zaproponowane przez Zadeha [22] oraz Boucheon-Meuniera i Yao [4].

3. Wyszukiwanie informacji z bazy danych na podstawie zapytania rozmytego

Po przedstawieniu podstawowych elementów wchodzących w skład zapytania rozmytego, przejdziemy teraz do omówienia sposobu wyciągania informacji z bazy danych.

Wyszukiwanie informacji z bazy danych polega na przypisaniu każdemu rekordowi z bazy danych stopnia spełnienia przez ten rekord zapytania. W ujęciu klasycznym stopień spełnienia może przybierać tylko dwie wartości: 0 lub 1, natomiast w przypadku rozmytym jest to liczba rzeczywista z przedziału [0,1]. Więcej informacji na temat obliczania stopnia spełnienia przez dany rekord zapytania znaleźć można w pracach, Dobrzyńskiego [6] oraz Kacprzyka i Zadrożnego [10-15].

Algorytm wyszukiwania informacji z bazy danych, przy założeniu, że rekordy są przetwarzane sekwencyjne, wygląda następująco:

1) Pobierz rekord z bazy danych.
2) Wylicz stopnie spełnienia prostych warunków zapytania dla wartości pochodzących z danego rekordu.
3) Wylicz całościowy stopień spełnienia: agregacja częściowych stopni spełnienia obliczonych w kroku 2.

4) Uwzględnij rekord w odpowiedzi gdy stopień spełnienia posiada odpowiednio wysoką wartość, odrzuć w przeciwnym wypadku.
5) Zakończ proces gdy nie ma więcej rekordów, w przeciwnym przypadku przejdź do kroku 1.

Kluczowym elementem powyższego algorytmu jest sposób obliczania stopnia spełnienia zapytania. W przypadku *zapytania prostego*, to znaczy takiego, w którym klauzula WHERE zawiera tylko jeden warunek z użyciem terminu rozmytego, stopień S_i spełnienia zapytania przez i-ty rekord w bazie danych, określony jest wartością funkcji przynależności terminu rozmytego T w określonym punkcie, jakim jest wartość A_i atrybutu A dla tego rekordu, tzn.

$$S_i = \mu_T(A_i).$$

W *zapytaniu złożonym*, w którym klauzula WHERE składa się z kilku warunków połączonych ze sobą rozmytym spójnikiem I / LUB (patrz rozdz. 2), stopień spełnienia zapytania wyznacza się ze wzoru:

$$S_i = \mu_{T_1}(A_{i1}) \wedge \ldots \wedge \mu_{T_k}(A_{ik})$$

albo

$$S_i = \mu_{T_1}(A_{i1}) \vee \ldots \vee \mu_{T_k}(A_{ik}),$$

gdzie μ_{T_j} oznacza funkcję przynależności terminu rozmytego T_j ($j = 1,\ldots,k$), zaś A_{ij} jest wartością j-tego atrybutu ($j = 1,\ldots,k$) dla i-tego rekordu.

W przypadku zapytania złożonego uwzględniającego w sekwencji warunków zarówno modyfikatory, jak i kwantyfikatory lingwistyczne, posługujemy się wzorem:

$$S_i = \mu_Q(\eta_1(\mu_{T1}(A_{i1}) \otimes \ldots \otimes \eta_k(\mu_{Tk}(A_{ik})),$$

gdzie μ_Q jest funkcją przynależności kwantyfikatora rozmytego, η_k - funkcją przynależności modyfikatora, natomiast $\otimes$ oznacza spójnik lub operator agregacji zastosowany w zapytaniu.

4. Testowanie hipotez rozmytych

Problem decyzyjny, którego celem jest potwierdzenie bądź falsyfikacja dowolnego stwierdzenia o badanej populacji, na podstawie

danych obarczonych niepewnością o charakterze losowym, jest zadaniem weryfikacji hipotez statystycznych. Zadanie takie można opisać przez trójkę uporządkowaną $(\mathscr{X},\mathscr{H},\mathscr{W})$, gdzie $\mathscr{X}$ jest przestrzenią możliwych obserwacji, $\mathscr{H}$ oznacza rozważane hipotezy (zerową i alternatywną), natomiast $\mathscr{W}$ opisuje wymagania nakładane na procedurę decyzyjną, jaką w tym wypadku jest test statystyczny. W klasycznej teorii weryfikacji hipotez $\mathscr{X}$, $\mathscr{H}$ i $\mathscr{W}$ muszą być precyzyjnie opisane, co okazuje się być zbyt restrykcyjnym wymaganiem w wielu sytuacjach spotykanych w praktyce. Ma to miejsce szczególnie często wtedy, gdy w rozważanym zagadnieniu istotną rolę odgrywa tzw. „czynnik ludzki", a więc na przykład wówczas, kiedy dostępne dane lub stawiane hipotezy, czy też wymagania, wyrażane są w sposób nieprecyzyjny – w szczególności – przy pomocy języka naturalnego. Tego typu sytuacje wymagają nowych, nieklasycznych narzędzi statystycznych. Ogólną metodę konstrukcji testu dla nieprecyzyjnych danych zaproponował Grzegorzewski [7]. Zagadnienie testowania hipotez przy nieprecyzyjnie określonych wymaganiach rozważał Arnold [1]. Z kolei problem weryfikacji nieprecyzyjnie sformułowanych hipotez podejmowało niezależnie wielu autorów, w szczególności Delgado, Verdegay i Vila [5], Saade i Schwarzlander [17], Saade [16], Watanabe i Imaizumi [19], Arnold [2, 3], oraz Taheri i Behboodian [18]. W niniejszej pracy przedstawimy metodę weryfikacji nieprecyzyjnych hipotez zbliżoną do podejścia Watanabe i Imaizumi. Jednocześnie przyjmiemy, że zarówno dostępne dane, jak i wymagania dotyczące charakterystyk statystycznych testu będą wyrażone precyzyjnie.

Niech $(\mathscr{X},\mathscr{A},\{P_\theta : \theta \in \Theta\})$ oznacza przestrzeń statystyczną, gdzie $\mathscr{A}$ jest σ-ciałem podzbiorów zbioru $\mathscr{X}$, a $\{P_\theta : \theta \in \Theta\}$ jest rodziną rozkładów prawdopodobieństwa na $\mathscr{A}$. Rozpatrujemy hipotezę $H : \theta \in \Theta_H$ dotyczącą parametru θ, wobec alternatywy $K : \theta \in \Theta_K$, gdzie $\Theta_H, \Theta_K \in \mathscr{P}(\Theta)$, $\Theta_H \cap \Theta_K = \varnothing$, tzn. Θ_H i Θ_K są pewnymi rozłącznymi zbiorami należącymi do rodziny wszystkich podzbiorów przestrzeni parametrów Θ. W szczególności, mogą to być typowe hipotezy jednostronne lub dwustronne postaci $H : \theta = \theta_0$, $K : \theta \neq \theta_0$; $H : \theta \leq \theta_0$, $K : \theta > \theta_0$; $H : \theta \geq \theta_0$, $K : \theta < \theta_0$, przy czym owym parametrem θ może być, w zależności od rozważanego problemu, średnia, mediana, wariancja, wskaźnik struktury itd. W tak postawionym zadaniu regułą decyzyjną, zwaną testem statystycznym, nazywa się odpowiednio skonstruowaną funkcję $\phi : \mathscr{X} \to \{0,1\}$, przypisującą możliwym obserwacjom decyzję o odrzuceniu ($\phi = 1$) albo przyjęciu ($\phi = 0$) weryfikowanej hipotezy H.

W rozważanym przez nas zagadnieniu konstrukcji rozmytych zapytań, odwołujących się do danych obarczonych niepewnością o charakterze losowym, interesuje nas weryfikacja hipotez wyrażonych w sposób mniej "sztywny", niż wspomniane powyżej, np. "θ wynosi około 5", "θ jest mniej więcej równe 10", "θ jest dużo większe niż 100", czy wręcz hipotez ujętych językiem potocznym, jak "θ jest małe", "θ jest duże" itp.

Do modelowania tak formułowanych hipotez posłużymy się aparatem teorii zbiorów rozmytych. Nawiązując do podanych powyżej oznaczeń, rozważać będziemy problem testowania hipotezy zerowej $H:\theta\in\Theta_H$ przy alternatywie $K:\theta\in\Theta_K$ z tym, że obecnie Θ_H i Θ_K będą rozmytymi podzbiorami przestrzeni parametrów Θ (tzn. $\Theta_H,\Theta_K\in\mathscr{F}(\Theta)$) o funkcjach przynależności, odpowiednio, μ_H i μ_K, gdzie $\mu_H,\mu_K:\Theta\to[0,1]$. Dokładniej, przyjmiemy założenie, że Θ_H będzie zwykłą liczbą rozmytą, bądź liczbą rozmytą jednostronną (tzn. że będzie to zbiór rozmyty normalny, wypukły, mający co najmniej górno półciągłą funkcję przynależności i nośnik ograniczony przynajmniej z jednej strony), natomiast Θ_K będzie rozmytym dopełnieniem Θ_H, a więc zbiorem rozmytym o funkcji przynależności $\mu_K(\theta)=1-\mu_H(\theta)$ dla każdego $\theta\in\Theta$.

Przy takich założeniach okazuje się, że zadanie weryfikacji rozmytej hipotezy zerowej $H:\theta\in\Theta_H$, wobec alternatywy $K:\theta\in\Theta_K=\neg\Theta_H$, można sprowadzić - drogą dekompozycji - do problemu testowania rodziny standardowych hipotez nierozmytych, względem odpowiednich nierozmytych hipotez alternatywnych. Na szczególną uwagę zasługują tu następujące trzy przypadki hipotez rozmytych $H:\theta\in\Theta_H$:

a) Θ_H ma nośnik ograniczony. Z przypadkiem tym mamy do czynienia przy modelowaniu hipotez zawierających wyrażenia typu "około ...", "mniej więcej ...", "w przybliżeniu ...", "mniej więcej pomiędzy ... i ...", etc. Wówczas zadanie testowania hipotezy rozmytej $H:\theta\in\Theta_H$ sprowadza się do rodziny zadań testowych dla hipotez nierozmytych
$$\{H_{\theta_0}:\theta=\theta_0 \quad \text{vs} \quad K_{\theta_0}:\theta\neq\theta_0 \mid \theta_0\in\operatorname{supp}\Theta_H\}. \qquad (1)$$

b) Θ_H ma nośnik ograniczony tylko z góry. Sytuacja ta odpowiada modelowaniu hipotez zawierających wyrażenia typu "raczej mniejsze niż ...", "dużo mniejsze niż ...", "małe", "bardzo małe", etc. W tym przypadku problem weryfikacji hipotezy $H:\theta\in\Theta_H$ możemy sprowadzić do następującej rodziny zadań testowych
$$\{H_{\theta_0}:\theta=\theta_0 \quad \text{vs} \quad K_{\theta_0}:\theta>\theta_0 \mid \theta_0\in\operatorname{supp}\Theta_H\}. \qquad (2)$$

c) Θ_H ma nośnik ograniczony tylko z dołu. Jest to przypadek modelowania hipotez rozmytych zawierających wyrażenia typu "raczej większe niż ...", "dużo większe niż ...", "duże", "bardzo duże", etc. Tym razem problem testowania hipotezy rozmytej $H:\theta\in\Theta_H$ sprowadzamy do rodziny zadań testowych dla hipotez nierozmytych

$$\{H_{\theta_0}:\theta=\theta_0 \quad \text{vs} \quad K_{\theta_0}:\theta<\theta_0 \mid \theta_0\in\text{supp}\Theta_H\}. \tag{3}$$

W tym miejscu pozostaje jeszcze do wyjaśnienia, jak explicite wygląda test statystyczny do weryfikacji hipotez rozmytych. Niech $\{\phi_{\theta_0}:\mathscr{X}\to\{0,1\} \mid \theta_0\in\text{supp}\Theta_H\}$ oznacza rodzinę klasycznych testów statystycznych na poziomie istotności α do weryfikacji hipotez nierozmytych postaci (1), (2) lub (3). Wówczas

Definicja 1

Testem statystycznym do weryfikacji rozmytej hipotezy zerowej $H:\theta\in\Theta_H$, wobec alternatywy $K:\theta\in\Theta_K=\neg\Theta_H$, na poziomie istotności α, nazywamy przekształcenie rozmyte $\psi:\mathscr{X}\to\mathscr{F}\{0,1\}$ o następującej funkcji przynależności

$$\mu_\psi(0)=\begin{cases}\sup\limits_{\{\theta_0\in\text{supp}\Theta_0:\phi_{\theta_0}(x)=0\}}\mu_H(\theta_0) & \text{gdy} \quad \{\theta_0\in\text{supp}\theta_0:\phi_{\theta_0}(x)=0\}\neq\varnothing\\ 0 & \text{gdy} \quad \{\theta_0\in\text{supp}\theta_0:\phi_{\theta_0}(x)=0\}=\varnothing\end{cases}$$

$$\mu_\psi(1)=1-\mu_\psi(0)$$

A zatem test służący weryfikacji hipotez rozmytych nie prowadzi zawsze do jednoznacznego odrzucenia, bądź przyjęcia testowanej hipotezy, jak ma to miejsce w klasycznym teście dla hipotez nierozmytych. Podaje on raczej stopień przekonania o słuszności odrzucenia $(\mu_\psi(1))$ lub też przyjęcia $(\mu_\psi(0))$ rozważanej hipotezy. Z jednoznacznym wskazaniem na którąś z przeciwstawnych decyzji będziemy mieć do czynienia wtedy, gdy w wyniku przeprowadzonego testowania otrzymamy $\mu_\psi(0)=0$, $\mu_\psi(1)=1$, co oznacza odrzucenie H, albo $\mu_\psi(0)=1$, $\mu_\psi(1)=0$, odpowiadające przyjęciu H.

Jak więc widać test rozmyty do weryfikacji hipotez rozmytych może być z powodzeniem użyty do wspomagania decyzji odnośnie spełnienia pewnych warunków czy też zachodzenia pewnych relacji sformułowanych nieprecyzyjnie, a których to decyzji podstawą są dane obarczone niepewnością o charakterze losowym. W szczególności, przedstawiony test rozmyty może być efektywnie wykorzystany przy tworzeniu rozmytych zapytań w bazach danych.

5. Zapytania wykorzystujące testy dla hipotez rozmytych

Dotychczas zakładaliśmy, ze w bazie danych posiadamy atrybuty jednowartościowe, opisujące w pełni interesujące nas obiekty. W rzeczywistości jednak, często nie jesteśmy w stanie określić wartości danego atrybutu z wykorzystaniem li tylko jednej liczby. Rozważmy prosty przykład dotyczący określenia *wyników w nauce* pewnej grupy studentów. Nie trudno zauważyć, że interesującą nas wielkość musimy opisać za pomocą atrybutu wielowymiarowego, jakim jest zestaw ocen uzyskanych z poszczególnych przedmiotów dla konkretnego studenta. Załóżmy, że chcemy wyszukać w bazie danych "dobrych studentów". Rodzi się więc pytanie jak konstruować zapytania dla baz danych zawierających takie wielowymiarowe atrybuty. Naturalnym wydaje się przejście od atrybutu wielowymiarowego do jednowymiarowego za pomocą stosownej agregacji. W rozważanym przez nas przypadku narzucającym się operatorem jest średnia arytmetyczna z ocen. Zaproponowany operator agregacji w dobry sposób syntetyzuje *wyniki w nauce*. Dzieje się tak dlatego, iż wartość średnia obliczona jest na podstawie ocen uzyskanych dla całej badanej populacji (oceny wszystkich badanych studentów z wszystkich przedmiotów).

Jednakże dosyć często liczba elementów stanowiących populację badanego zjawiska jest tak duża, że dokonanie dokładnych pomiarów czyni badanie zbyt kosztownym bądź też, w przypadku gdy liczba elementów jest nieprzeliczalna, niemożliwym do zrealizowania. Zmuszeni wówczas jesteśmy ograniczyć się do wyników stanowiących realizację pewnej próby losowej. Dla zilustrowania problemu niech posłuży nam przykład dotyczący określenia zanieczyszczenia pewną substancją wody w jeziorze. Jak nietrudno zauważyć stężenie substancji zanieczyszczającej nie musi być jednorodne w calym zbiorniku: może ono przyjmować różne wartości w poszczególnych miejscach (ukształtowanie brzegu, występująca flora) i na różnych głębokościach. Nie jesteśmy również w stanie dokonać pomiarów w każdym punkcie, badanie nasze ograniczamy zatem do pewnej próby losowej. W takim przypadku interesującą nas wielkość staramy się opisać za pomocą atrybutu wielowymiarowego, jakim jest zestaw pomiarów dokonanych w różnych losowo dobranych miejscach. Dobór właściwej metody agregacji jest, oczywiście, sprawą kluczową. Ograniczenie się w konstrukcji zapytania tylko do średniej z dostępnych pomiarów nie wydaje się w tym przypadku rozsądnym podejściem. Trzeba mieć na uwadze, że wartość średniej jest zależna od konkretnej próby losowej i będzie przyjmować różne wartości w konkretnych realizacjach. Widzimy więc, że wnioski dotyczących zanieczyszczenia jeziora wyciągnięte drogą prostej indukcji, mogą w konsekwencji okazać się mylące. Przy tak sformułowanym

zadaniu i tego typu danych niezbędne jest wprowadzenie do zapytania elementów bardziej zaawansowanego wnioskowania statystycznego.

Nasza koncepcja rozwiązania tego problemu jest bardzo naturalna i polega na zastosowaniu w takiej sytuacji testu do weryfikacji hipotez o nieznanym parametrze (np. średniej). Chcąc konstruować elastyczne zapytania rozmyte, proponujemy zastosowanie do tego celu testów dla hipotez rozmytych, omówionych w rozdz. 4. Stosując testy dla hipotez rozmytych jesteśmy w stanie określić stopień przekonania o słuszności danego twierdzenia, co w języku budowy zapytań odpowiada obliczaniu stopnia spełnienia zapytania. Efektu tego nie uzyskamy stosując podejście tradycyjne.

Załóżmy, ze interesują nas informacje dotyczące jezior o *niezbyt dużej głębokości*. Zapytanie rozmyte w tym przypadku przybierze następującą postać:

SELEC jezioro FROM baza - o - jeziorach WHERE jezioro = *niezbyt duża głębokość*.

Przyjmijmy, że przez termin rozmyty *niezbyt duża głębokość* będziemy rozumieć głębokość *około 25m*. Funkcja przynależności tego terminu wyraża się wzorem:

$$\mu_H(\theta) = \begin{cases} \dfrac{\theta - 20}{5} & \text{dla} \quad 20 \le \theta \le 25 \\ \dfrac{30 - \theta}{5} & \text{dla} \quad 25 \le \theta \le 30 \\ 0 & \text{dla} \quad \theta \notin (20,30) \end{cases} \tag{4}$$

Ponieważ dane, którymi dysponujemy dla każdego z jezior występujących w bazie danych, stanowią próbę losową złożoną z pomiarów wykonanych w kilku losowo wybranych miejscach jeziora, chcąc wnioskować o tym, czy średnia głębokość danego zbiornika spełnia nasze zapytanie, musimy odwołać się do odpowiedniego testu statystycznego. Gdyby nasze zapytanie było postawione w sposób ostry, tzn. „czy średnia głębokość jeziora wynosi 25m?", wówczas stosowne byłoby posłużenie się testem do weryfikacji hipotezy $H : \theta = 25$ przy alternatywie $K : \theta \neq 25$, przy czym parametr θ oznaczałby w tym wypadku wartość średnią. Zakładając, że zamieszczone w bazie pomiary stanowią n-wymiarową próbę losową $X_1, \ldots, X_n$ z rozkładu normalnego, do weryfikacji tak postawionej hipotezy należałoby użyć klasyczny test Studenta. Algorytm podejmowania decyzji (na zadanym poziomie istotności α) wyglądałby następująco:

– jeżeli $\overline{X} \in \left(25 - t_{1-\alpha/2}^{[n-1]} \frac{s}{\sqrt{n}}, 25 + t_{1-\alpha/2}^{[n-1]} \frac{s}{\sqrt{n}}\right)$, to przyjąć H

– jeżeli $\overline{X} \notin \left(25 - t_{1-\alpha/2}^{[n-1]} \frac{s}{\sqrt{n}}, 25 + t_{1-\alpha/2}^{[n-1]} \frac{s}{\sqrt{n}}\right)$, to odrzucić H na rzecz K,

przy czym $t_{1-\alpha/2}^{[n-1]}$ oznacza kwantyl rozkładu t-Studenta rzędu $1-\frac{\alpha}{2}$ o $n-1$ stopniach swobody, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ jest średnią arytmetyczną z próby, zaś $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$ jest odchyleniem standardowym z próby.

Skoro jednak w przeszukiwaniu bazy danych chcemy posłużyć się wspomnianym powyżej zapytaniem rozmytym, musimy w tym celu przejść do testowania hipotezy rozmytej $H : \theta \in \Theta_H$ wobec alternatywy $K : \theta \notin \Theta_H$, gdzie Θ_H jest zbiorem rozmytym odpowiadającym wyrażeniu „około 25m" o funkcji przynależności $\mu_H(\theta)$ danej wzorem (4). Zgodnie z tym, co podaliśmy w rozdz. 3, zadanie weryfikacji tej hipotezy rozmytej sprowadzi się do rozwiązania rodziny zadań testowych dla hipotez nierozmytych

$$\{H_{\theta_0} : \theta = \theta_0 \quad \text{vs} \quad K_{\theta_0} : \theta \neq \theta_0 \mid \theta_0 \in \text{supp}\Theta_H\},$$

gdzie $\text{supp}\Theta_H = [20,30]$. Korzystając z def. 1 można pokazać, że odpowiedni test będzie miał następującą funkcję przynależności

$$\mu_\psi(0) = \mu_T(\overline{X}) = \begin{cases} 0 & dla \quad \overline{X} < 20 - \gamma \\ \frac{\overline{X} - 20 + \gamma}{5} & dla \quad 20 - \gamma \le \overline{X} < 25 - \gamma \\ 1 & dla \quad 25 - \gamma \le \overline{X} \le 25 + \gamma \\ \frac{30 + \gamma - \overline{X}}{5} & dla \quad 25 + \gamma < \overline{X} \le 30 + \gamma \\ 0 & dla \quad \overline{X} > 30 + \gamma \end{cases} \tag{5}$$

$$\mu_\psi(1) = 1 - \mu_\psi(0) = 1 - \mu_T(\overline{X})$$

gdzie $\gamma = t_{1-\alpha/2}^{[n-1]} \frac{s}{\sqrt{n}}$.

Jak wspomniano w rozdz. 4, $\mu_\psi(0)$ może być interpretowane jako stopień przekonania o słuszności przyjęcia rozważanej hipotezy rozmytej H, podczas

gdy $\mu_{\psi}(1)$ wskazuje stopień przekonania o słuszności odrzucenia tej hipotezy.

Wracając do zagadnienia tworzenia rozmytego zapytania w rozważanym problemie wyszukiwania w bazie danych jezior o „niezbyt dużej głębokości", możemy już teraz określić sposób obliczania stopnia spełnienia owego zapytania przez dowolny rekord. Otóż jeżeli rekord zawierać będzie realizację próby losowej pomiarów głębokości danego jeziora, stopień S spełnienia naszego zapytania równy będzie

$$S = \mu_T(\overline{X}),$$

czyli wartości funkcji przynależności danej wzorem (5) w punkcie $\overline{X}$, tzn. dla średniej arytmetycznej z próby opisanej tym rekordem.

6. Zakończenie

W niniejszej pracy przedstawiliśmy prosty model zapytania rozmytego przeznaczonego do wyszukiwania informacji w bazie danych zawierającej dane obciążone niepewnością o charakterze losowym. Metoda budowy takiego zapytania łączy w sobie elementy konstrukcji zwykłego zapytania rozmytego z testowaniem hipotez rozmytych. Zaproponowane w tym podejściu połączenie logiki rozmytej ze statystyką pozwala uchwycić nieprecyzyjność percepcji i rozumowania człowieka wraz z losowością, będącą nieuniknioną składową otaczającego nas świata. Zaimplementowanie procedur umożliwiających tworzenie tego typu zapytań pozwoli użytkownikom zrealizować w praktyce coraz powszechniejszą koncepcję Zadeha [21, 23] „obliczeń na słowach", a zarazem będzie sprzyjać eliminacji błędu metodologicznego polegającego na ignorowaniu niepewności o charakterze losowym, zawartej często w przetwarzanych danych.

Literatura

[1] Arnold B.F., 1995: Statistical tests optimally meeting certain fuzzy requirements on the power function and on the sample size, *Fuzzy Sets and Systems* 75, 365-372.

[2] Arnold B.F., 1996: An approach to fuzzy hypothesis testing, *Metrika* 44, 119-126.

[3] Arnold B.F., 1998: Testing fuzzy hypotheses with crisp data, *Fuzzy Sets and Systems* 94, 323-333.

[4] Bouchon-Meunier B., Yao J., 1991: Linguistic modifiers and imprecise categories, J. *Intell. Inf. Syst.*

[5] Delgado M., Verdegay J.L., Vila M.A., 1985: Testing fuzzy-hypotheses. A Bayesian approach, W: *Approximate Reasoning in Expert Systems*, Gupta M.M., Kandel A., Bandler W., Kiszka J.B. (red.), North-Holland.

[6] Dobrzyński W, 1996: Zastosowanie logiki rozmytej w interfejsie logicznym użytkownika do baz danych, *Rozprawa doktorska*, IBS PAN, Warszawa.

[7] Grzegorzewski P., 2000: Testing statistical hypotheses with vague data, *Fuzzy Sets and Systems* 112, 501-510.

[8] Jarke M., 1985: A field evaluation of natural language for data retrieval, IEEE Trans. Software Eng. 11, 93-113.

[9] Kacprzyk J., Zadrożny S., 1994: Fuzzy querying for Microsoft Access. Proceedings of the *Third IEEE Conference on Fuzzy Systems* (Orlando, USA), Vol. 1, 167-171.

[10] Kacprzyk J., Zadrożny S., 1994: Fuzzy queries in Microsoft Access: toward a 'more intelligent' use of Microsoft Windows based DBMSs, Proceedings of the *1994 Second Australian i New Zealand Conference on Intelligent Information Systems - ANZIIS'94* (Brisbane, Australia), 492 - 496.

[11] Kacprzyk J., Zadrożny S., 1995: FQUERY for Access: fuzzy querying for a Windows-based DBMS. In: P. Bosc i J. Kacprzyk (red.) *Fuzziness in Database Management Systems*, Physica-Verlag, Heidelberg, 415 - 433.

[12] Kacprzyk J., Zadrożny S., 1995: Fuzzy queries in Microsoft Access v. 2, Proceedings of (Sao Paolo, Brazil), Vol. II, 341 - 344. *6th International Fuzzy Systems Association World Congress*

[13] Kacprzyk J., Zadrożny S., 1997: Fuzzy queries in Microsoft Access v. 2, in D. Dubois, H. Prade i R.R. Yager (red.): *Fuzzy Information Engineering - A Guided Tour of Applications*, Wiley, New York, 223 - 232.

[14] Kacprzyk J., Zadrożny S., 1997: Implementation of OWA operators in fuzzy querying for Microsoft Access. In: R.R. Yager i J. Kacprzyk (red.) *The Ordered Weighted Averaging Operators: Theory and Applications*, Kluwer, Bostonn, 293 - 306.

[15] Kacprzyk J., Zadrożny S., 1997: Flexible querying using fuzzy logic: An implementation for Microsoft Access, in T. Andreasen, H. Christiansen i H.L. Larsen (red.): *Flexible Query Answering Systems*, Kluwer, Boston, 247-275.

[16] Saade J.J., 1994: Extension of fuzzy hypothesis testing with hybrid data, *Fuzzy Sets and Systems* 63, 57-71.

[17] Saade J.J., Schwarzlander H., 1990: Fuzzy hypothesis testing with hybrid data, *Fuzzy Sets and Systems* 35, 197-212.

[18] Taheri S.M., Behboodian J., 1999: Neyman-Pearson lemma for fuzzy hypotheses testing, *Metrika* 49, 3-17.

[19] Watanabe N., Imaizumi T., 1993: A fuzzy statistical test of fuzzy hypotheses, *Fuzzy Sets and Systems* 53, 167-178.

[20] Zadeh L., 1965: Fuzzy sets, *Information and Control* 8, 338-353.

[21] Zadeh L.A., 1973: Outline of a new approach to the analysis of complex system and decision processes, *IEEE Trans. SMC* 3, 28-44.

[22] Zadeh L.A., 1983: A computational approach to fuzzy quantifiers in natural languages, *Computers Math. Appl.* 9, 149-184.

[23] Zadeh L.A., 1999: Fuzzy logic = computing with words, W: Zadeh L.A., Kacprzyk J. (Red.): *Computing with Words in Information / Intelligent Systems. Part 1. Foundations*, Springer - Physica Verlag, Heidelberg, 3-23.

WYŻSZA SZKOŁA INFORMATYKI STOSOWANEJ I ZARZĄDZANIA

pod auspicjami
Polskiej Akademii Nauk

ZAŁOŻYCIELEM
Wyższej Szkoły Informatyki Stosowanej i Zarządzania
jest
FUNDACJA KRZEWIENIA NAUK SYSTEMOWYCH
powołana z inicjatywy
Prezesa
POLSKIEJ AKADEMII NAUK

FUNDATOREM
Fundacji Krzewienia Nauk Systemowych
jest
POLSKA AKADEMIA NAUK

ORGANEM
sprawującym nadzór jest
MINISTERSTWO EDUKACJI NARODOWEJ

Wyższa Szkoła Informatyki Stosowanej i Zarządzania
prowadzi studia wyższe na kierunkach:
INFORMATYKA
ZARZĄDZANIE I MARKETING

SIEDZIBA
Instytut Badań Systemowych
Polskiej Akademii Nauk
ul. Newelska 6, 01-447 Warszawa

ISBN 83-85847-54-5