STUDIA INFORMATICA 2010
Volume 31 Number 2A (89)

Anna KOWALCZYK-NIEWIADOMY, Adam PELIKANT
Politechnika Łódzka, Instytut Mechatroniki i Systemów Informatycznych

# PRZETWARZANIE ZAPYTAŃ W METAJĘZYKU NATURALNYM ZA POMOCĄ ALGORYTMÓW ROZMYTYCH

**Streszczenie**. Przedmiotem badań jest pozyskiwanie nieprecyzyjnych informacji z relacyjnych baz danych. Wagę problematyki badań podnosi fakt, iż takie podejście nie jest wspierane przez żaden komercyjny system zarządzania bazami danych. Prezentowane rozwiązanie stanowi nowatorskie podejście w tej dziedzinie, oparte na automatycznym generowaniu funkcji przynależności i przetwarzaniu zapytań w bezkontekstowym metajęzyku.

**Słowa kluczowe**: logika rozmyta, grupowanie, zapytania rozmyte

# FUZZY QUERIES PROCESSING BY MEANS OF META-NATURAL LANGUAGE USING FUZZY ALGORITHMS

**Summary**. This paper presents a novel idea of gaining imprecise information from relational database systems. Concernment of investigation rise fact that such kind of processing is not supported by any commercial database system. These researches illustrate a combination of database technology and fuzzy logic. The final aim is to develop a fuzzy querying system based on meta-natural language.

**Keywords**: fuzzy logic, clustering, grouping, fuzzy queries

## 1. Wstęp

Prezentowane rozwiązanie, które jest tematem tej publikacji, stanowi nowatorskie podejście w dziedzinie pozyskiwania nieprecyzyjnych informacji zgromadzonych w systemach zarządzania bazami danych. Prezentowane rozwiązanie stanowi ważną alternatywę dla powszechnie stosowanego podejścia wyszukiwania danych opartego na logice dwuwartościowej. Celem badań jest opracowanie oprogramowania pozwalającego na formułowanie zapytań do

bazy danych Oracle, stosując składnię zbliżoną do języka naturalnego. Zadanie konwersji oparte zostało na automatycznym generowaniu funkcji przynależności oraz zbiorów rozmytych. Dodatkowym elementem projektowanego systemu jest złożony mechanizm przetwarzania danych wraz z inteligentnym algorytmem automatycznie etykietującym, który umożliwi budowę zapytań w bezkontekstowym metajęzyku naturalnym. Przez bezkontekstowość rozumiana jest automatyzacja interpretacji etykiet, np. niski, wysoki bez globalnego kontekstu nadanego przez eksperta.

W dziedzinie baz danych stosowane dotychczas rozwiązania oparte na teorii zbiorów rozmytych zawierają silne ograniczenia na etapie konstruowania zbiorów rozmytych. Opracowane systemy opierają się na sztywno zdefiniowanych funkcjach przynależności, a zatem wymagają współpracy z ekspertem dziedzinowym. Podstawową ideą opisywanych badań jest założenie możliwości automatycznego określenia funkcji przynależności na podstawie rzeczywistego rozkładu danych, które wykazują naturalną tendencję do niejednorodnej dystrybucji. Zastosowanie algorytmów klasyfikacji bez nauczyciela pozwala na wykrycie takiej nierównomierności, a zatem automatyczne określenie liczby zbiorów rozmytych oraz opisujących je funkcji przynależności. W końcowym efekcie pozwoli to na implementację systemu generującego odpowiedź na niejednoznacznie zadane zapytania.

## 2. Zapytania rozmyte

Problematyka języków zapytań rozmytych stanowi przedmiot badań wielu ośrodków naukowych. Na szczególną uwagę zasługują prace: Patrick Bosc i Olivier Pivert [2], Patrice Buche i Catherine Dervin [3], Yoshikane Takahashi [4], Claudia González, Leonid Tineo [5, 6], a także polskiego zespołu prof. dr hab. Janusz Kacprzyk i dr hab. Sławomir Zadrożny [7]. Warto zauważyć, iż część rozwiązań opartych o teorię zbiorów rozmytych, zawiera ograniczenia już na etapie konstruowania tych zbiorów np.: w postaci z góry założonego stopnia rozmycia. Prezentowane autorskie rozwiązanie, zakłada stworzenie systemu wyszukiwania opartego o przetwarzanie nieprecyzyjnie zadanych zapytań. Z punktu implementacyjnego zadanie to zostanie zrealizowane jako rozszerzenie języka zapytań SQL (w standardzie SQL 92) na serwerze baz danych Oracle 11g. W celu implementacji nieprecyzyjnego definiowania zapytań wprowadzony zostanie mechanizm inteligentnego etykietowania danych (w pełni automatyczny – bez potrzeby korzystania z wiedzy eksperta).

## 3. Prace badawcze

Rozdział prezentuje postęp prac badawczych, które w przyszłości pozwolą na formułowanie zapytań do bazy danych Oracle stosując składnię zbliżoną do języka naturalnego.

### 3.1. Automatyczne generowanie zbiorów rozmytych

Podstawowym założeniem opisywanych badań jest pełna automatyzacja procesu generowania funkcji przynależności i zbiorów rozmytych. Zadanie to zostało zrealizowane poprzez automatyczne dopasowanie stopnia rozmycia do dystrybucji analizowanych danych wejściowych. Zakładając tendencje do niejednorodnej dystrybucji danych gromadzonych w bazie, w pracy wykorzystano wielowymiarowe algorytmy klasteryzacji rozmytej. W pracy wykorzystano znane algorytmy grupowania rozmytego „k-means" oraz „mountain clustering". Zadanie grupowania znane również pod pojęciem analizy skupień lub klasteryzacji (ang. *clustering*), polega na podziale zbioru danych wejściowych (wartości atrybutów) na podzbiory, w taki sposób, aby wartości atrybutów zaliczane do tej samej grupy (kategorii) były do siebie bardziej podobne, niż wartości należące do innych grup (kategorii). Wymienione algorytmy zwracają informacje o stopniu przynależności każdego elementu do każdej z kategorii. Zagadnienie grupowania jest obecnie wykorzystywane w wielu dziedzinach nauki, a do jej głównych zastosowań bliskich tematyce baz danych można zaliczyć:

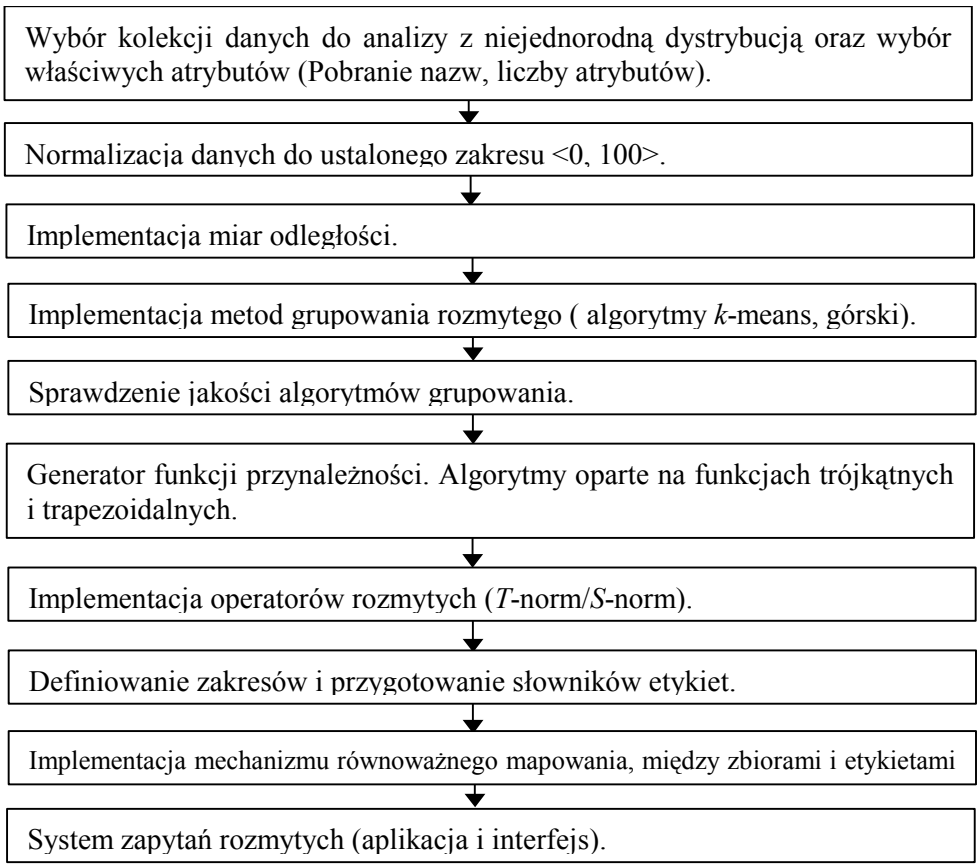
- Eksplorację danych (ang. *Data mining*).
- Wyszukiwanie informacji (ang. *Information Retrieval, Text Mining*).
- Grupowanie danych w problemie harmonogramowania.
- Segmentację obrazu (ang. *Image segmentation*).

Szczególnym zastosowaniem klasteryzacji jest automatyczne wykrywanie skupień mające na celu generowanie funkcji przynależności. Szczegółowy opis zagadnienia grupowania, jak również wykorzystanych w pracy algorytmów, znajduje się w publikacjach [8] i [13].

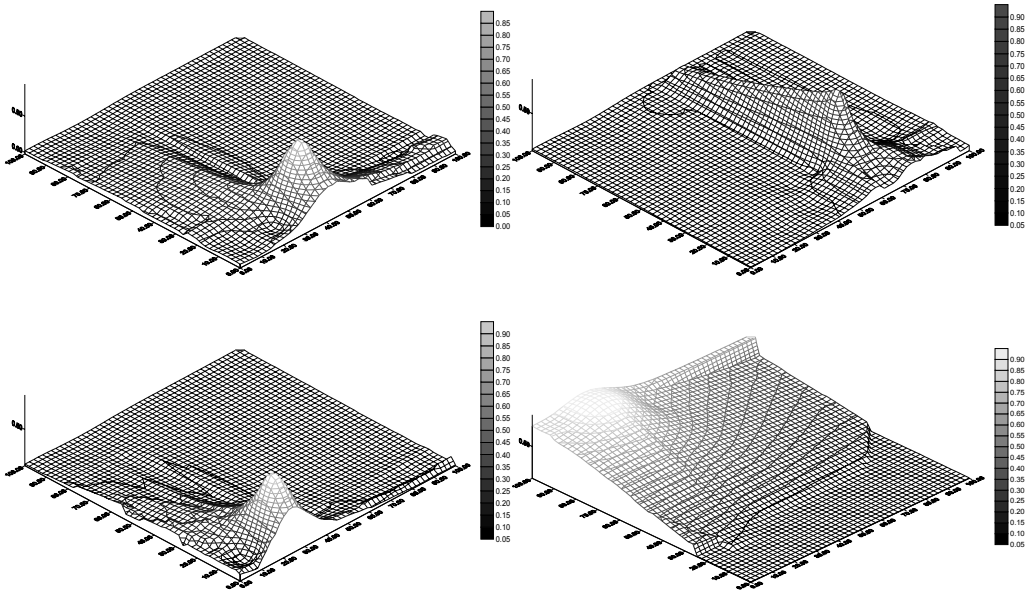
Zaimplementowane algorytmy przetwarzania rozmytego oraz relacyjnego zostały w pełni umiejscowione na serwerze bazy danych Oracle 11g, co centralizuje i przyśpiesza obliczenia na danych. Dodatkowo eliminuje to potrzebę transferu danych między serwerem bazy danych, a końcówkami klienckimi. Kod wykonywalny algorytmów przetwarzania zaimplementowany został w postaci kodu JSP (Java Stored Procedures) enkapsulowanego w postaci pakietów PL/SQL. Przeważająca część tak przygotowanej funkcjonalności opakowana zostanie do postaci pakietu PL/SQL (wrapper JSP), który stanowić będą API systemu rozmytego wyszukiwania.

### 3.2. Wielowymiarowe przetwarzanie danych

Kolejne etapy badań prezentuje Rys 1. W pierwszej kolejności ekspert dziedzinowy podejmuje decyzję o liczbie i rodzaju analizowanych atrybutów. Przykładowo analizie poddajemy zbiór danych meteorologicznych. Stan pogody mogą charakteryzować takie atrybuty jak wilgotność, temperatura powietrza, temperatura rosy, całkowity opad, prędkość wiatru, ciśnienie. W dalszej części wartości atrybutów podlegają normalizacji do przedziału <0, 100>. Dzięki temu procesowi eliminujemy problem skali, wartości ujemnych i zapewniamy integralność na poziomie generowania zbiorów rozmytych. Co więcej proces dopasowywania etykiet do zbiorów staje się uniwersalny i nie zależy od danych wejściowych. Ważnym etapem badań jest dobór odpowiedniej metryki i algorytmu grupowania. Jakość grupowania podlega ocenie i weryfikacji, dzięki czemu uzyskujemy poprawnie wygenerowane zbiory rozmyte (Rys.2). Dalszy etap prac ukierunkowany jest na mechanizm etykietowania.




Rys. 1. Etapy badań
Fig. 1. Research steps

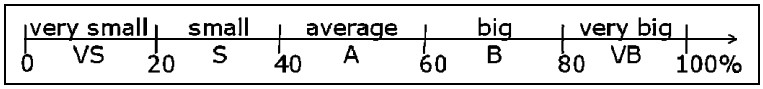


Rys. 2. Przykładowe zbiory rozmyte wygenerowane algorytmem mountain clustering
Fig. 2. Exemplary fuzzy set generated with mountain clustering method

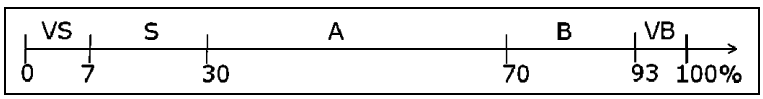
### 3.3. Etykietowanie

Istotnym elementem systemu przetwarzającego zapytania rozmyte jest mechanizm etykietowania. Inteligentny algorytm etykietujący ma na celu prawidłowy przydział etykiet do zbiorów rozmytych uzyskanych uprzednio w procesie automatycznego generowania funkcji przynależności. W procesie etykietowania można wyróżnić następujące kroki:

- Stworzenie dynamicznych słowników etykiet.
- Dobór etykiet do atrybutów (wraz z wyborem rodzaju gramatycznego etykiety).
- Ustalenie przedziałów, w jakich dana etykieta ma się zawierać.
- Wybór metody etykietowania.




Rys. 3. Liniowy przydział etykiet
Fig. 3. Linear labels arrangement




Rys. 4. Ekspercki przydział etykiet
Fig. 4. Expert's labels arrangement

Dynamiczne słowniki etykiet zostały stworzone w taki sposób, aby zapewnić możliwość edycji oraz ich rozszerzenia o kolejne zestawy. Wybierając zestawy etykiet dla poszczególnych atrybutów należy pamiętać o następujących zasadach:

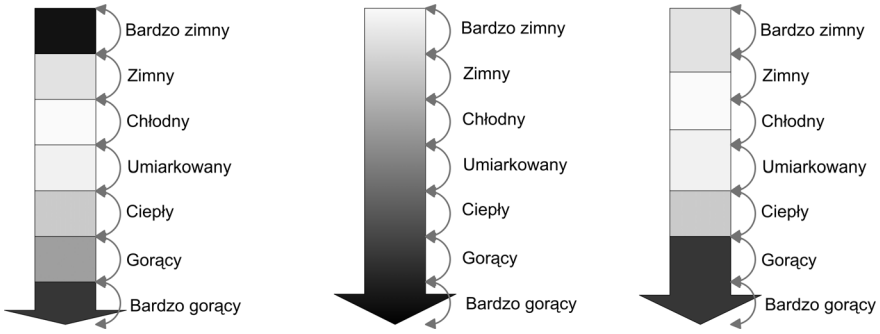
- Każdy atrybut musi mieć przypisany dokładnie jeden zestaw etykiet.

- Zestaw etykiet może być użyty wielokrotnie dla różnych atrybutów.

Ustalając przedziały dla poszczególnych etykiet należy zdecydować o sposobie przydziału zakresów. Przydział ten może mieć charakter liniowy (automatyczny) bądź ekspercki. Na rys.3 i rys. 4 przedstawiono przykładowe przydziały etykiet.

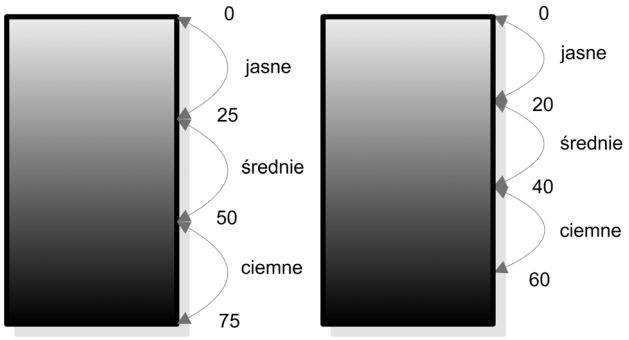
### *3.3.1. Etykietowanie równomierne*

Najprostszym sposobem etykietowania, jest równomierny przydział etykiet bez wykorzystywania informacji o klastrach / grupach, do których dane należą. Niestety metoda działa skutecznie jedynie dla danych, których rozkład nie jest zbytnio równomierny, a najlepsze rezultaty osiąga się, gdy grupy danych praktycznie dopasowują się do przedziałów etykiet. Rys 5. prezentuje najbardziej typowy rozkład danych podlegających etykietowaniu.




Rys. 5. Przypadki rozkładu danych do etykietowania
Fig. 5. Exemplary data distribution used in labeling process

### *3.3.2. Etykietowanie „środkiem ciężkości”*

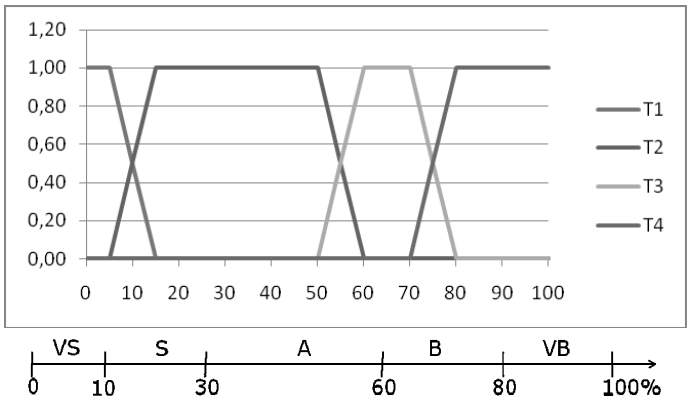



Rys. 6. Przykładowy przydział etykiet metodą „środka ciężkości”
Fig. 6. Exemplary labeling process

Aby ograniczyć wpływ rodzaju danych wejściowych na jakość etykietowania tych danych, został stworzony algorytm, który opiera się nie bezpośrednio na wartości atrybutu, a na stopniu przynależności danego atrybutu do danego klastra (grupy). Przykładowy rozkład etykiet w metodzie „środka ciężkości” ilustruje rys. 6.

Algorytm zakłada obliczenie średniej arytmetycznej wszystkich wartości, które należą do danego klastra i których stopień przynależności jest większy niż próg odcięcia (określony przez eksperta) a następnie dopasowanie właściwej etykiety.

### *3.3.3. Etykietowanie „metodą dopasowania”*




Rys. 7. Dopasowanie zbiorów rozmytych do eksperckiego przydziału etykiet
Fig. 7. Trapezoidal fuzzy sets and expert's labels arrangement

Tabela 1

Porównanie etykietowania „twardego” i „miękkiego”

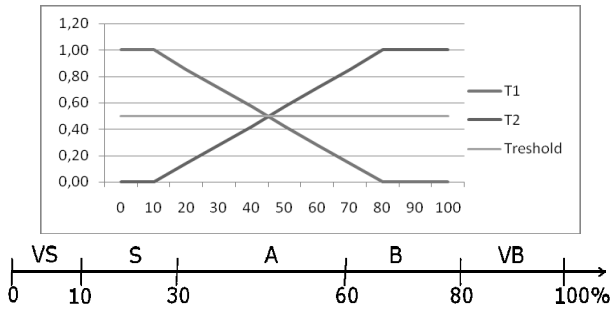
| | **Metoda** | **Przedział** | **Nazwa etykiety** |
|---|---|---|---|
| 1 Trapez | Etykietowanie miękkie | 0, 15 | very small, small |
| | Etykietowanie twarde | 0, 5 | very small |
| 2 Trapez | Etykietowanie miękkie | 5, 60 | very small, small, average |
| | Etykietowanie twarde | 15, 50 | small, average |
| 3 Trapez | Etykietowanie miękkie | 50, 80 | average, big |
| | Etykietowanie twarde | 60, 70 | big |
| 4 Trapez | Etykietowanie miękkie | 70, 100 | big, very big |
| | Etykietowanie twarde | 80, 100 | very big |

Główne założenie w procesie etykietowania „metodą dopasowania”, dotyczy bezkontekstowości. Sformułowania w języku naturalnym mają zazwyczaj charakter subiektywny i często są źródłem niejednoznaczności. Konflikt znaczeniowy można przedstawić na przykładzie pojęcia „wysoka temperatura”. W zależności od rozważanej dziedziny np.: pogoda, gotowanie, wytop metali, to samo określenie będzie dotyczyło zupełnie odrębnych zakresów wartości. Stąd przed przystąpieniem do procesu etykietowania, ważnym etapem jest normalizacja zarówno po stronie zbiorów rozmytych jak i normalizacja po stronie zakresów etykiet. Normalizacja do przedziału <0÷100> pozwala na procentowy przydział etykiet oraz eliminuje problem skali i liczb ujemnych. Rozpatrzymy trzy sposoby etykietowania metodą dopasowania:

- Etykietowanie twarde.
- Etykietowanie miękkie.
- Etykietowanie mieszane.

Głównym założeniem etykietowania twardego jest wykorzystanie wyłącznie całkowitej przynależności elementu do zbioru rozmytego. Część klastra z przynależnością mniejszą niż

1 zostaje pominięta podczas doboru etykiety. W przypadku etykietowania miękkiego, koncepcja zakłada wykorzystanie wszystkich wartości przynależności elementu do zbioru rozmytego większych od 0. Mamy zatem do czynienia z rozszerzeniem etykietowania twardego o elementy z niepełną przynależnością. Obie metody zostały porównane i odpowiednio zilustrowane rys. 7 i tabelą 1.




Rys. 8. Dopasowanie zbiorów rozmytych do eksperckiego przydziału etykiet
Fig. 8. Trapezoidal fuzzy sets and expert's labels arrangement.

Tabela 2

Porównanie etykietowania trzema „metodami dopasowania"

| | **Metoda** | **Przedział** | **Nazwa etykiety** |
|---|---|---|---|
| 1 Trapez | Etykietowanie miękkie | 0, 80 | very small, small, average, big |
| | Etykietowanie twarde | 0, 10 | very small |
| | Etykietowanie mieszane | 0, 45 | very small, small, average |
| | | | |
| 2 Trapez | Etykietowanie miękkie | 10, 100 | small, average, big, very big |
| | Etykietowanie twarde | 80, 100 | very big |
| | Etykietowanie mieszane | 45, 100 | average, big, very big |

Trzecia metoda łączy mechanizm etykietowania twardego i miękkiego. W tym przypadku ważnym elementem zależnym od decyzji eksperta jest wielkość progu odcięcia. Dopasowanie etykiet zakłada wykorzystanie wyłącznie elementów o przynależności większej od wyznaczonego progu odcięcia <μt(x), 1>. Zaproponowane metody zostały porównane i przedstawione na rys. 8 i w tabeli 2.

## 4. Podsumowanie

Wykorzystanie zagadnień logiki i zbiorów rozmytych do opisu skonkretyzowanego problemu na platformie serwera baz danych Oracle potwierdziły zasadność stosowania rozmytego modelowania zjawisk. Na podstawie rzeczywistej dystrybucji atrybutów oraz przy wykorzystaniu algorytmów grupowania rozmytego udało się automatycznie uzyskać zbiory rozmyte. Co więcej zaproponowane przez autorów mechanizmy w pełni zautomatyzowanego inteligentnego etykietowania w przyszłości pozwolą na rozszerzenie języka SQL o możliwość nie-

precyzyjnego definiowania zapytań. Pełna implementacja tak opisanego problemu stanowi wartościowy element w dziedzinie baz danych, co dowodzi zasadności prowadzenia dalszych badań.

## BIBLIOGRAFIA


1. Zadeh L. A.: Fuzzy sets. Information and Control, 1965.
2. Bosc P., Pivert O.: A Relational Database Language for Fuzzy Querying. IEEE Transactions on Fuzzy Systems, Vol. 3, No. 1, February 1995.
3. Buche P., Dervin C.: Fuzzy Querying of Incomplete, Imprecise and Heterogeneously Structured Data in the Relational Model Using Ontologies and Rules. IEEE Transactions on Fuzzy Systems, Vol. 13, No. 3, June 2005.
4. Takahashi Y.: A Fuzzy Query Language for Relational Databases. IEEE Transactions on Systems, Man. And Cybernetics Part, Vol. 21, No. 6, December 1991.
5. González C., Tineo L., Galindo J.: Fuzzy Database Languages Integration using Expressive Power. Fifth International Conference on Fuzzy Systems and Knowledge Discovery.
6. González C., Goncalves M., Tineo L.: A New Upgrade to SQL. Towards a Standard in Fuzzy Databases. 20th International Workshop on Database and Expert Systems Application, IEEE Computer Society, DOI 10.1109/DEXA.2009.35.
7. Kacprzyk J., Zadrozny S.: FQUERY for Access: Fuzzy Querying for Windows-Based DBMS. In Fuzziness in Database Management Systems, P. Bosc and J. Kacprzyk (eds.), Physica-Verlag, 1995, s. 415÷433.
8. Kowalczyk A., Pelikant A.: Fuzzy Clustering in Relational Databases. XII International Conference-System Modelling and Control, 2007.
9. Kowalczyk A., Pelikant A.: Implementation of automatically generated membership functions based on grouping algorithms The International Conference on Computer as a tool 2007.
10. Kowalczyk A., Pelikant A.: Fuzzy queries in relational databases XIII International Conference-System Modelling and Control, 2009 (publikacja pokonferencyjna w JACS 2010).
11. Fraley C., Raftery A.: How many clusters? Which clustering method? Answers via model-based cluster analysis. The Computer Journal 41, 8 (1998), s. 578÷588.
12. Schwartz G.: Estimating the dimension of a model. The Annals of Statistics, 6 (1978).
13. Kowalczyk-Niewiadomy A., Pelikant A.: Zagadnienie grupowania w kontekście budowania zapytań rozmytych. Konferencja BDAS, Ustroń, 2008.

Recenzent: Dr inż. Bożena Małysiak-Mrozek

Wpłynęło do Redakcji 31 stycznia 2010 r.

**Abstract**

This paper, takes into consideration the problem of retrieving ambiguous and imprecise information from relational database system. As the fact, traditional SQL does not provide any essential mechanism for solving such issues. In recent years, fuzzy SQL language that allows making flexible queries has become a very interesting object of research. Unfortunately in most cases the implementation of fuzzy sets theory is based on one constant threshold and depends strictly on experts decision: Patrick Bosc i Olivier Pivert [2], Patrice Buche i Catherine Dervin [3], Yoshikane Takahashi [4], Claudia González, Leonid Tineo [5, 6], prof. dr hab. Janusz Kacprzyk i dr hab. Sławomir Zadrożny [7].

This article presents the novel way of gaining imprecise and incomplete information from database. The idea based on fuzzy clustering methods provides an effective tool for fuzzy sets generation and gaining fuzzy query results from database system automatically. The most important points of the idea are the data normalization, fuzzy clustering algorithms and clustering quality measurement methods. What is more, we are able to generate membership functions automatically by means of clustering. Presented conception of labeling mechanism is to enable getting satisfactory result for query written in meta natural language. It is worth mentioning that the results do not depend on context so the solution is universal. Although some methods and technical concepts need to be extended and optimized, a fuzzy clustering and classification querying approach remains effective.

**Adresy**

Anna KOWALCZYK NIEWIADOMY: Politechnika Łódzka, Instytut Mechatroniki i Systemów Informatycznych ul.Stefanowskiego 18/22 90-924 Łódź, Polska, anna.kowalczykniewiadomy@gmail.com .

Adam PELIKANT: Politechnika Łódzka, Instytut Mechatroniki i Systemów Informatycznych ul.Stefanowskiego 18/22 90-924 Łódź, Polska, apelikan@p.lodz.pl .