STUDIA INFORMATICA 2010
Volume 31 Number 2A (89)

Marek MIŁEK, Bożena MAŁYSIAK-MROZEK, Dariusz MROZEK
Politechnika Śląska, Instytut Informatyki

HURTOWNIA DANYCH ROZMYTYCH: PODSTAWY TEORETYCZNE I PRAKTYCZNE ASPEKTY UŻYCIA[1]

Streszczenie. Współczesne systemy analityczne coraz częściej sięgają po nowe sposoby analizy danych oparte na rozmytym wnioskowaniu i przetwarzaniu informacji, która nie zawsze jest reprezentowana w sposób precyzyjny. W niniejszym artykule zaprezentowano nowatorski, w pełni funkcjonalny, opracowany i zrealizowany przez autorów system hurtowni danych rozmytych (FDW, *Fuzzy Data Warehouse*). Hurtownia danych rozmytych stanowi repozytorium danych, które przechowuje zarówno dane precyzyjne, jak i dane rozmyte oraz pozwala na klasyczne i rozmyte przetwarzanie zgromadzonych w niej danych. W artykule zebrano najważniejsze cechy funkcjonalne systemu FDW oraz wykonanej przez autorów aplikacji analitycznej FDW Browser, należącej do klasy narzędzi eksploracji danych Fuzzy-OLAP.

Słowa kluczowe: hurtownie danych, zbiory rozmyte, systemy wspomagania decyzji, filtrowanie, grupowanie i agregacja danych

A FUZZY DATA WAREHOUSE: THEORETICAL FOUNDATIONS AND PRACTICAL ASPECTS OF USAGE

Summary. Modern analytical tools increasingly make use of new ways of data analysis that base on fuzzy reasoning and fuzzy processing of information. In the paper, we present a Fuzzy Data Warehouse system (FDW), which we have designed and developed. Fuzzy Data Warehouse (FDW) is a data repository, which contains fuzzy data and allows a fuzzy processing of the data. In the paper, we focus on the most important functional features of the FDW system and our newly developed FDW Browser, which is an analytical application adhering to the Fuzzy-OLAP class of data exploration tools.

Keywords: data warehouse, fuzzy sets, fuzzy logic, decision support systems

[1] Praca naukowa finansowana ze środków na naukę w latach 2009-2011 jako projekt rozwojowy nr O R00 0068 07. Tytuł projektu: „Projekt i demonstrator technologii systemu wspomagania działań operacyjno-procesowych dla obronności i bezpieczeństwa państwa".

1. Wprowadzenie

Od wielu lat rynek oprogramowania rozwija się w coraz większym tempie. Dla specjalistów z branży informatycznej zjawisko to nie jest zaskakujące. Nie tylko rozwój przemysłu, ale także prywatne wymagania rzeszy użytkowników na całym świecie nadają impuls do tworzenia ogromnej liczby oprogramowania różnego rodzaju. W ostatnich latach trend ten wyraźnie się powiększa. Wiele firm nie wyobraża sobie swojej działalności bez wykorzystania komputera z odpowiednim oprogramowaniem, które najczęściej tworzone jest na zamówienie pod kątem charakterystyki firmy, aby sprostać w pełni jej oczekiwaniom. Jednak samo gromadzenie i przetwarzanie ściśle wyselekcjonowanej informacji nie jest niczym szczególnym w obliczu potencjału, jaki drzemie w źródle informacji jako całości. Dane gromadzone przez wiele lat stanowią udokumentowany i niezaprzeczalny dowód działalności, będący wiarygodnym źródłem, nadającym się do przetworzenia przez odpowiednie narzędzia analityczne. Surowe dane, często składowane w różnym formacie, nie stanowią jeszcze użytecznej informacji. Dopiero usystematyzowanie i grupowanie danych w odpowiedni logiczny schemat stanowi realne źródło wiedzy. Samo usystematyzowanie informacji często nie jest jeszcze wystarczającym elementem pozwalającym wysnuć określone wnioski z analizowanych danych. Istnieją jednak algorytmy przetwarzające informacje do postaci pozwalającej zaobserwować jej charakterystykę pod innym kątem. Algorytmy te wchodzą w zakres grupy narzędzi analitycznych nazywanych narzędziami eksploracji danych i odkrywania wiedzy *(*ang. *data mining)* i często są one elementem systemów wspomagających podejmowanie decyzji *(*DSS, ang. *decision support systems)* [1].

Współczesne systemy analityczne są bardzo rozbudowanymi aplikacjami pozwalającymi wyciągać szereg interesujących wniosków. Podobnie jak w przypadku standardowych aplikacji biznesowych, opartych na przetwarzaniu transakcyjnym OLTP (ang. *OnLine Transaction Processing*), również systemy analityczne OLAP (ang. *OnLine Analytical Processing*) często projektowane są pod kątem określonego problemu i dla konkretnej firmy [1], [2]. Istnieją także uniwersalne narzędzia pozwalające skonstruować odpowiednie środowisko analityczne. Tradycyjne systemy przeprowadzają operacje między innymi na danych numerycznych, które precyzyjnie wyrażają określone wielkości fizyczne. Jednak w pewnych okolicznościach można uznać, że analizowanie danych liczbowych w dokładny sposób nie jest konieczne. Innymi słowy występują sytuacje, w których kryteria wyszukiwania i przetwarzania informacji niekoniecznie muszą być podane precyzyjnie, aby dostarczyć użytecznej wiedzy.

Logika rozmyta i pojęcie zbioru rozmytego [3] wprowadzają taki sposób wnioskowania, dzięki któremu tradycyjne pojmowanie informacji zostaje w jasny i klarowny sposób zastąpione podejściem bardziej ogólnym. Każde zdanie, wyrażenie, czy warunek w sensie logicznym nie jest rozpatrywane tylko w kategorii prawdy lub fałszu, lecz także może być klasyfi-

kowane wartością pośrednią. Warto zauważyć, że odpowiada to bardziej naturalnemu rozumowaniu, w którym często nie jest możliwa lub wskazana jednoznaczna ocena. Logika rozmyta od dawna jest wykorzystywana m.in. w systemach sterowania oraz innych zastosowaniach technologicznych, gdzie logika dwuwartościowa nie jest wystarczająca. Od pewnego czasu można również zauważyć tendencję do stosowania logiki rozmytej w systemach baz danych [4-8] i systemach analizy danych [9-12]. Daje to możliwość rozmytego przetwarzania danych, doprowadzając do wyciągania bardziej ogólnych wniosków z danych zgromadzonych w systemach baz danych.

W niniejszym artykule zaprezentowano opracowany i zaimplementowany przez autorów system Hurtowni Danych Rozmytych (FDW, ang. *Fuzzy Data Warehouse*) oraz aplikację FDW Browser, umożliwiającą wielowymiarową analizę danych na wzór analitycznych systemów OLAP. Głównym atutem aplikacji jest możliwość analizowania danych rozmytych, czyli takich, których wartość nie jest określona w precyzyjny sposób. Program pozwala również na analizowanie danych dokładnych wykorzystując do tego celu metody aproksymacyjnego przetwarzania danych, wzbogacając w ten sposób możliwości prowadzenia analiz. Aplikacja FDW Browser posiada również funkcje umożliwiające tworzenie środowiska pracy charakterystycznego dla hurtowni danych.

2. Podstawy teoretyczne

W niniejszym rozdziale zostaną przedstawione definicje pojęć z zakresu logiki i arytmetyki rozmytej, a także pojęcia z zakresu hurtowni danych stosowane w pracy.

Liczba rozmyta jest typu L-R, jeżeli są zdefiniowane dla niej funkcje odniesienia (bazowe): L (lewostronna) i R (prawostronna) oraz skalary $\alpha > 0$, $\beta > 0$ i m – wartość modalna. Skalary α, β są nazywane odpowiednio rozrzutem lewo- lub prawostronnym [13].

Symbolicznie liczba rozmyta typu L-R jest reprezentowana przez trójkę (m, α, β).

Ogólnie można także powiedzieć, że liczba rozmyta typu L-R jest zbiorem rozmytym A, określonym na uniwersum liczb rzeczywistych, którego funkcja przynależności opisana jest następującym wyrażeniem:

$$\mu_A = \begin{cases} L\left(\dfrac{m-x}{\alpha}\right) & dla\ x < m \\ 1 & dla\ x = m, \\ R\left(\dfrac{x-m}{\beta}\right) & dla\ x > m \end{cases} \tag{1}$$

gdzie m, α, $\beta \in \Re$.

Przedział rozmyty jest typu L-R, jeżeli istnieją dla niego funkcje odniesienia (bazowe): L (lewostronna) i R (prawostronna) oraz rozrzuty lewo- i prawostronne $\alpha > 0$, $\beta > 0$ i wartości *m, n - gdzie* $m < n$, określające przedział wartości modalnych (m, n).

Przedział rozmyty zapisujemy za pomocą czwórki (m, n, α, β) [13].

Ogólnie można powiedzieć, że przedział rozmyty typu L-R, jest zbiorem rozmytym A, określonym na uniwersum liczb rzeczywistych, którego funkcja przynależności opisana jest następującym wyrażeniem:

$$\mu_A = \begin{cases} L\left(\dfrac{m-x}{\alpha}\right) dla & x < m \\ 1 \quad dla\, m \le x \le n, \\ R\left(\dfrac{x-n}{\beta}\right) dla & x > n \end{cases} \tag{2}$$

gdzie m, n, α, $\beta \in \Re$.

Dla liczb i przedziałów rozmytych typu L-R zdefiniowano wiele operacji arytmetycznych. W pracy skoncentrowano się tylko na tych, które najczęściej są stosowane w hurtowniach danych, a więc na tych, które są niezbędne do wykonywania operacji agregacji danych rozmytych.

2.1. Dodawanie liczb rozmytych typu L-R

Jeśli liczby rozmyte A_1 i A_2 przedstawione są w postaci trójek:

$$A_1 = (m_{A1},\ \alpha_{A1},\ \beta_{A1}),\ A_2 = (m_{A2},\ \alpha_{A2},\ \beta_{A2}), \tag{3}$$

a ich suma w postaci:

$$A_1 + A_2 = (m_{A1+A2},\ \alpha_{A1+A2},\ \beta_{A1+A2}), \tag{4}$$

to zachodzą następujące zależności [14]:

$$m_{A1+A2} = m_{A1} + m_{A2}, \tag{5}$$

$$m_{A1+A2} - \alpha_{A1+A2} = (m_{A1} - \alpha_{A1}) + (m_{A2} - \alpha_{A2}), \tag{6}$$

$$m_{A1+A2} + \beta_{A1+A2} = (m_{A1} + \beta_{A1}) + (m_{A2} + \beta_{A2}), \tag{7}$$

przedstawione na rys. 1.

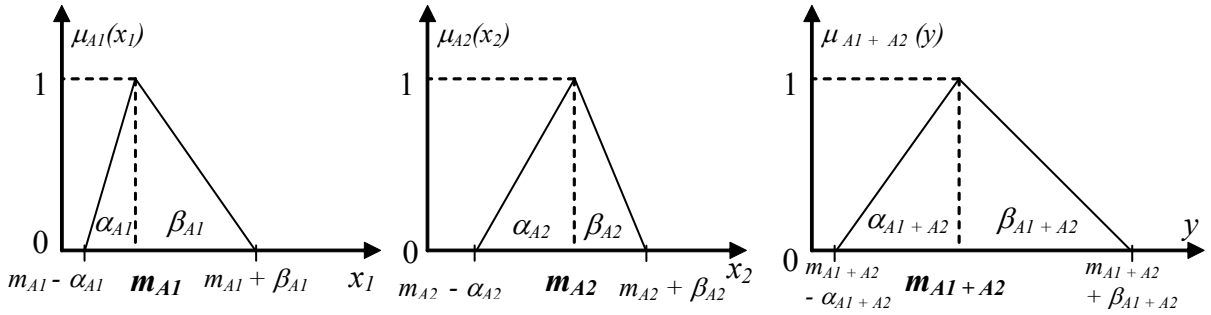
Na podstawie tych zależności można obliczyć parametry sumy liczb rozmytych $(A_1 + A_2)$:

$$\alpha_{A1+A2} = \alpha_{A1} + \alpha_{A2}, \tag{8}$$

$$\beta_{A1+A2} = \beta_{A1} + \beta_{A2}, \tag{9}$$

zatem suma w reprezentacji L-R ma następującą postać [14]:

$$A_1 + A_2 = (m_{A1+A2},\ \alpha_{A1+A2},\ \beta_{A1+A2}) = (m_{A1} + m_{A2},\ \alpha_{A1} + \alpha_{A2},\ \beta_{A1} + \beta_{A2}). \tag{10}$$



Rys. 1. Dodawanie liczb rozmytych typu L-R
Fig. 1. Addition of L-R type fuzzy numbers

2.2. Dzielenie liczby rozmytej typu L-R przez wartość ostrą (dokładną)

W zaimplementowanym systemie zastosowano dzielenie liczby rozmytej typu L-R przez wartość dokładną w procesie wyznaczania wartości średniej ze zbioru liczb rozmytych. Autorzy zaadaptowali zależności, które zachodzą w przypadku dzielenia przez siebie dwóch liczb rozmytych.

Dzielnik (wartość dokładna) potraktowany został jako szczególny przypadek liczby rozmytej, w której rozrzut lewostronny i prawostronny jest równy 0 ($\alpha_B = 0,\ \beta_B = 0$). Zatem w przypadku dzielenia liczby rozmytej A przez liczbę dokładną B zachodzą następujące zależności [14]:

$$m_{A/B} = m_A/m_B, \tag{11}$$

$$m_{A/B} + \beta_{A/B} = (m_A + \beta_A)/m_B, \tag{12}$$

$$m_{A/B} - \alpha_{A/B} = (m_A - \alpha_A)/m_B\,. \tag{13}$$

2.3. Operatory logiczne

W pracy zdefiniowano również operatory większości i mniejszości, które są używane w procesie wyznaczania wartości maksymalnej i minimalnej w zbiorze liczb rozmytych.

Operator większości jest zdefiniowany na podstawie porównywania ze sobą poszczególnych parametrów funkcji trapezowej. Najpierw porównywane są wartości modalne. Jeśli jednak są one równe, sprawdzana jest relacja rozrzutów prawostronnych, a w przypadku ich równości, sprawdzana jest relacja rozrzutów lewostronnych.

Operator mniejszości jest wyznaczany analogicznie, lecz w przypadku równości wartości modalnych obu liczb należy najpierw sprawdzić relacje rozrzutów lewostronnych, a w przypadku ich równości – rozrzutów prawostronnych.

2.4. Operacje na przedziałach rozmytych

Wszystkie z przedstawionych operacji mogą być także realizowane na przedziałach rozmytych typu L-R. Na przykład, dla dwóch przedziałów rozmytych L-R A_1 i A_2, które są reprezentowane odpowiednio przez czwórki:

$$A_1 = (m_{A1}, n_{A1}, \alpha_{A1}, \beta_{A1}), A_2 = (m_{A2}, n_{A2}\ \alpha_{A2}, \beta_{A2}), \tag{14}$$

suma tych przedziałów jest wyznaczana zgodnie z następującym wyrażeniem [14]:

$$A_1 + A_2 = (m_{A1}, n_{A1}, \alpha_{A1}, \beta_{A1}) + (m_{A2}, n_{A2}, \alpha_{A2}, \beta_{A2}) =$$
$$= (m_{A1} + m_{A2}, n_{A1} + n_{A2}, \alpha_{A1} + \alpha_{A2}, \beta_{A1} + \beta_{A2}). \tag{15}$$

Przedstawione pojęcia z dziedziny teorii zbiorów rozmytych stanowią podstawę dla zbudowanej przez autorów hurtowni danych rozmytych. Pozwoliły one również autorom na określenie podstawowych pojęć używanych w artykule.

Definicja 1. Hurtownia danych rozmytych (FDW, od ang. *Fuzzy Data Warehouse*) jest repozytorium danych, ustanowionym głównie do celów raportowych, które przechowuje zarówno dane precyzyjne, jak i dane rozmyte, oraz pozwala na klasyczne i rozmyte przetwarzanie zgromadzonych w niej danych.

Podobnie jak w tradycyjnych systemach hurtowni danych baza hurtowni danych rozmytych powinna być zaprojektowana w odpowiedni sposób, np. w postaci schematu gwiazdy (ang. *star*) lub schematu płatka śniegu (ang. *snowflake*).

Definicja 2. FOLAP (Fuzzy-OLAP, ang. *Fuzzy OnLine Analytical Processing*) jest technologią przetwarzania danych dokładnych i rozmytych, zorientowaną na szybki odczyt, raportowanie i wielowymiarową analizę danych, która w swoim działaniu wykorzystuje elementy teorii zbiorów rozmytych.

Do podstawowych operacji realizowanych przez systemy FOLAP, poza tradycyjnymi sposobami przetwarzania danych, będziemy zaliczać m.in. rozmyte grupowanie danych dokładnych i rozmytych, filtrowanie tych danych przy użyciu rozmytych predykatów, agregację danych rozmytych, rozmyte filtrowanie agregatów, operacje arytmetyczne i logiczne na liczbach rozmytych, konwersję danych dokładnych do postaci rozmytej, możliwość opisu różnych dziedzin danych za pomocą zmiennych lingwistycznych.

3. Podstawowe cechy systemu hurtowni danych rozmytych

Zaprojektowany i zaimplementowany system hurtowni danych rozmytych posiada cechy znanych systemów analitycznych i umożliwia dodatkowo wykonywanie operacji z udziałem rozmytych danych. Projektowany system umożliwia zdefiniowanie własnego środowiska

pracy dla użytkownika, w którym można wykonywać przewidziane przez niego operacje. Aplikacja analityczna korzysta z usług serwera bazy danych w celu uzyskania wymaganych informacji do analizy. W kolejnych podrozdziałach zostaną przedstawione najważniejsze cechy, jakie wzięto pod uwagę podczas projektowania systemu i jakie zapewnia system. Drugi podrozdział zawiera opis elementów systemu, których działanie opiera się na zasadach teorii zbiorów rozmytych, a które w większości dotyczyć będą przetwarzania na serwerze bazy danych.

3.1. Własności systemu w zakresie hurtowni danych i analizy danych

W niniejszym rozdziale omówiono elementy dotyczące wykonanej przez autorów artykułu aplikacji analitycznej FDW Browser, należącej do klasy aplikacji Fuzzy-OLAP, współpracującej z hurtownią danych rozmytych i umożliwiającej wielowymiarową analizę danych i raportowanie.

3.1.1. Projekt analityczny

Aplikacja analityczna FDW Browser umożliwia użytkownikowi systemu rozpoczęcie pracy od utworzenia projektu. Głównym celem projektu jest możliwość zdefiniowania środowiska pracy. Projekt może być zapisywany na dysku tak, aby można było go otworzyć w kolejnych sesjach działania aplikacji. W ramach projektu zapisywane są również informacje na temat połączenia z bazą danych oraz zdefiniowanych struktur danych, które mają służyć w procesach analizy danych. Na rysunku 2 przedstawiono okno aplikacji analitycznej FDW Browser prezentującej strukturę hurtowni danych. Panel w lewej części okna zawiera kolejne elementy projektu analitycznego.

3.1.2. Źródło danych

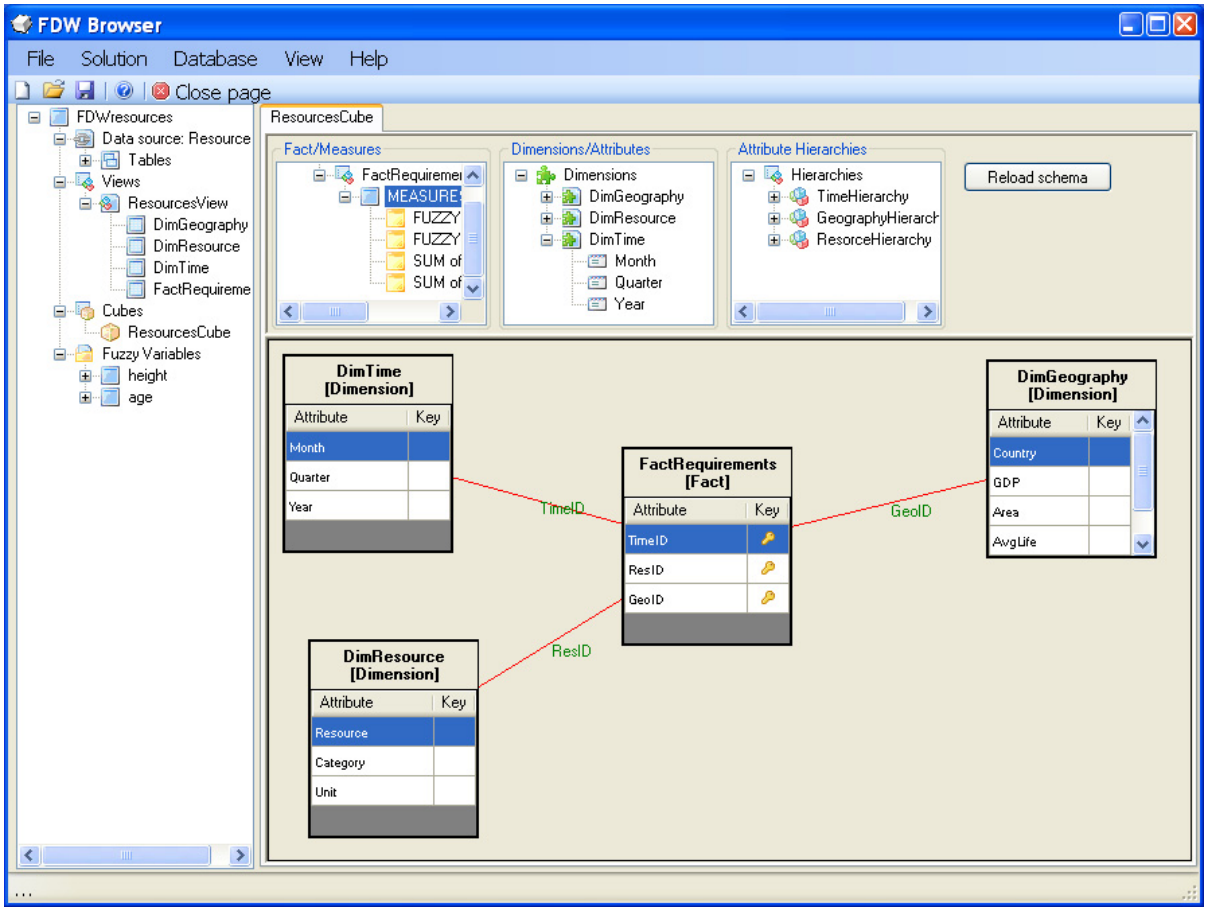
Głównym zadaniem użytkownika w czasie definiowania projektu jest określenie źródła danych. W oknie definiowania połączenia użytkownik ma możliwość określenia nazwy serwera oraz danych uwierzytelniających, które umożliwią dostęp do wybranej bazy danych.

3.1.3. Widoki źródła danych

Zaraz po zdefiniowaniu projektu analitycznego użytkownik ma możliwość przejrzenia schematu bazy danych (rys. 2). Zazwyczaj schematy bazy danych są rozbudowane, więc użytkownik ma możliwość zdefiniowania podzbiorów, określanych mianem widoków źródła danych, które następnie chciałby brać pod uwagę w operacjach analitycznych. W aplikacji FDW Browser użytkownik ma możliwość zdefiniowania dowolnej liczby widoków źródła danych.

3.1.4. Kostka danych

Głównym elementem poddawanym analizie jest struktura kostki danych (rys. 2). Element ten jest definiowany na podstawie tabel w źródłowej bazie danych i dla uproszczenia jest zawsze definiowany w obszarze wybranego widoku źródła danych. Podobnie jak w przypadku widoków danych, użytkownik może definiować dowolną liczbę kostek danych.



Rys. 2. Okno aplikacji analitycznej FDW Browser prezentujące strukturę kostki danych
Fig. 2. FDW Browser analytical application with a structure of cube exposed in the central pane

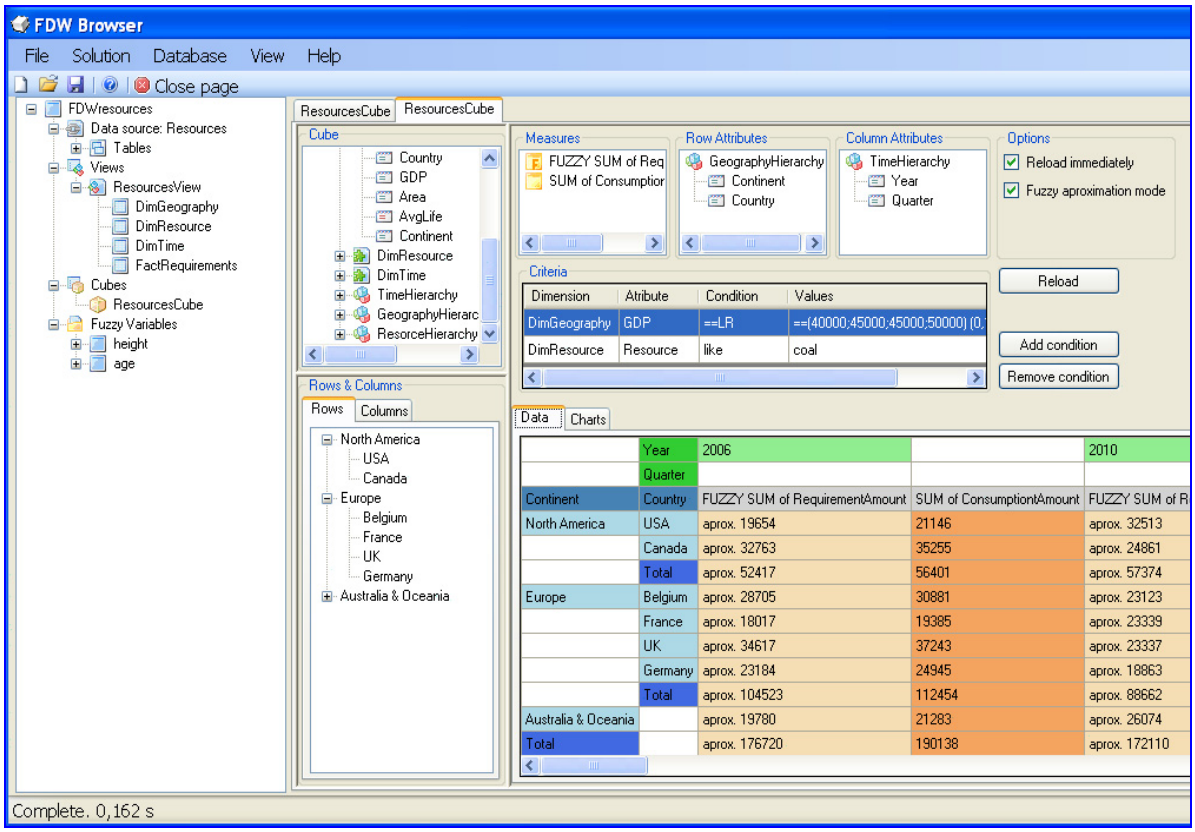
3.1.5. Kreator kostki danych

Aplikacja analityczna FDW Browser jest wyposażona w kreator kostki danych. Kreator ten ułatwia i skraca czas potrzebny na zdefiniowanie struktur wymaganych podczas analizy. Kolejne etapy definiowania kostki zakładają:

- Wybór widoku źródła danych.
- Określenie funkcji, jakie mają spełniać poszczególne tabele (fakty i wymiary).
- Wybór atrybutów, możliwość zdefiniowania hierarchii atrybutów.
- Wybór miar, możliwość zdefiniowania grup miar.

3.1.6. Przeglądanie kostki danych

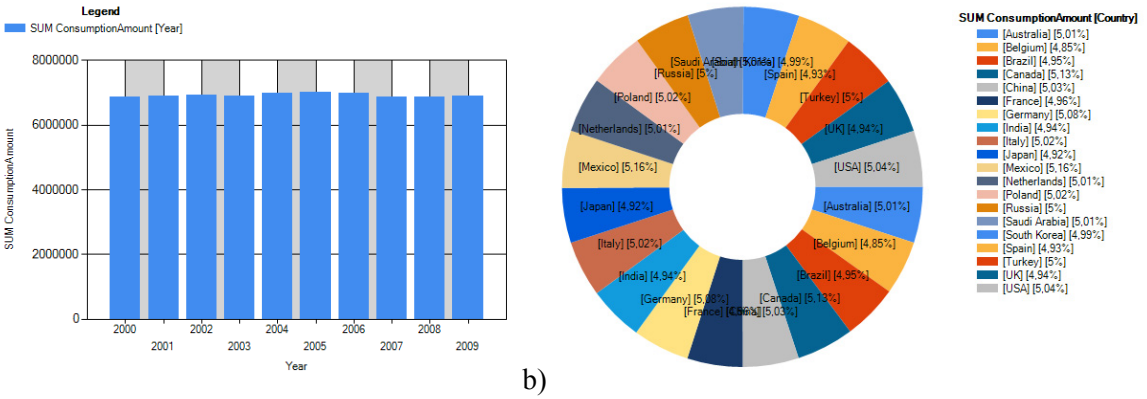
W widoku przeglądania kostki danych użytkownik posiada dowolność w wybieraniu określonych miar, które zamierza analizować. Wybór atrybutów lub całych hierarchii atrybutów jest także dowolny. W przypadku wyboru kilku atrybutów lub hierarchii atrybutów, widok analizowanej kostki danych umożliwia zwijanie oraz rozwijanie wybranych poziomów uszczegółowienia wraz z podsumowaniami na określonych poziomach oraz podsumowaniami na przekroju każdego z wymiarów (rys. 3). Użytkownik ma także możliwość zdefiniowania kryteriów filtrujących dane. Kryteria te mogą mieć zarówno charakter klasyczny, jak również mogą zawierać elementy logiki rozmytej (panel *Criteria*, rys. 3).



Rys. 3. Okno analizy danych w aplikacji analitycznej FDW Browser
Fig. 3. Data analysis in the FDW Browser analytical application

3.1.7. Wykresy

Dodatkowym elementem umożliwiającym analizę danych są wykresy przedstawiające wyniki obliczeń. Dla każdej kombinacji atrybutu oraz miary istnieje możliwość wyświetlenia danych w postaci wykresu. W przypadku wykresu słupkowego atrybuty stanowiące podstawę wyznaczenia grup zajmują oś poziomą, natomiast obliczenia podsumowań oś pionową wykresu (rys. 4a). Drugim typem wykresu jest wykres kołowy, który umożliwia przeglądanie wyników podsumowań wraz z podaniem udziału procentowego poszczególnych grup na tle całości (rys. 4b).



a) b)

Rys. 4. Prezentacja danych w postaci wykresów: a) słupkowych, b) kołowych
Fig. 4. Presentation of data in the form of: a) bar chart, b) pie chart

3.2. Własności systemu w zakresie logiki rozmytej

Teoria zbiorów rozmytych wprowadza nowy rodzaj informacji, który jest nietypowy dla znanych powszechnie systemów zarządzania bazami danych. W zakresie tej teorii istnieje pojęcie liczby rozmytej, czyli liczby, która opisuje w nieprecyzyjny sposób pewną wartość. W systemie został zaimplementowany sposób reprezentacji liczb rozmytych wraz z elementami umożliwiającymi przeprowadzanie przekształceń na tego typu liczbach.

3.2.1. Typ liczby rozmytej

W celu gromadzenia w hurtowni danych, obok danych precyzyjnych, również danych rozmytych, w systemie zarządzania bazą danych zdefiniowano typ reprezentujący tego rodzaju dane. System umożliwia zatem przechowywanie danych rozmytych w hurtowni danych, a także w obszarze systemu zarządzania bazą danych udostępnia funkcje i procedury, dzięki którym możliwe jest wykonywanie różnych operacji na danych rozmytych.

Dane o charakterze rozmytym przechowywane w hurtowni danych rozmytych mogą być prezentowane na dwa różne sposoby:

- Jako czwórka (l, m, n, p), gdzie: (m, n) jest przedziałem wartości modalnych, $l=m-\alpha$, $p=n+\beta$, oraz $\alpha>0$, $\beta>0$ to odpowiednio lewy i prawy rozrzut. Czwórka (l, m, n, p) określa parametry trapezowej funkcji przynależności dla liczby rozmytej. Trójkątna funkcja przynależności jest traktowana, jako specjalny przypadek funkcji trapezowej, gdzie $m=n$.
- Jako zmienna lingwistyczna postaci *approx. x*, gdzie x jest średnią wartości m i n.

Poniżej przedstawiono przykładowe wartości rozmytej miary *RequirementAmount*, opisującej wielkość zapotrzebowania na zasoby naturalne w obu reprezentacjach.

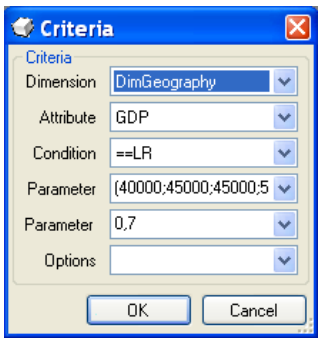
```
RequirementAmount          RequirementAmount
-------------------------- ------------------
(2416;2517;2617;2717)      approx.2567
(976;1017;1017;1057)       approx.1017
(1571;1637;1697;1762)      approx.1667
```

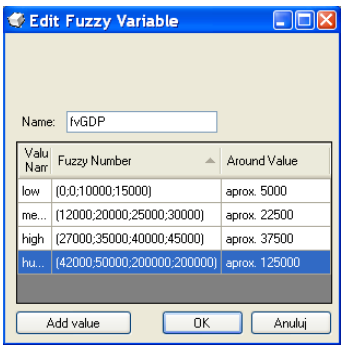
3.2.2. Logika i arytmetyka rozmyta

Wraz z typem rozmytym w systemie zarządzania bazą danych został zdefiniowany zestaw operacji logicznych i arytmetycznych dla danych rozmytych. Operacje logiczne umożliwiają przeprowadzanie takich działań, jak porównywanie liczb rozmytych z innymi liczbami rozmytymi lub dokładnymi. Implementacja arytmetyki liczb rozmytych pozwala na przeprowadzanie podstawowych działań, dzięki którym możliwe jest m.in. utworzenie funkcji agregujących. Operacje te są także możliwe do zrealizowania, gdy jednym ze składników jest liczba rozmyta, a drugim liczba dokładna. W systemie można zatem agregować liczby przechowywane w postaci takiej, jak zaprezentowano w punkcie 3.2.1.

3.2.3. Rozmyte warunki filtrujące

Zdefiniowane w systemie zarządzania bazą danych operacje logiczne z udziałem liczb rozmytych pozwalają na konstruowanie rozmytych warunków filtrujących. System umożliwia definiowanie takich warunków w trakcie wykonywania analiz. Dzięki temu użytkownik ma możliwość określenia w niedokładny sposób kryteriów, jakimi powinien kierować się procesor zapytań w momencie pobierania informacji z bazy danych. Zarówno dla danych rozmytych, jak i danych dokładnych istnieje możliwość zdefiniowania rozmytych warunków filtrujących. Podczas definiowania rozmytych kryteriów zapytań możliwe jest określenie wartości stopnia przynależności, który stanowi minimalny stopień zgodności, z jakim warunek rozmyty powinien być spełniony. Na rysunku 5a przedstawiono sposób definiowania warunku

a) 

b) 

Rys. 5. Okna definiowania: a) warunków filtrujących, b) zmiennej lingwistycznej i jej wartości
Fig. 5. Application window for defining: a) filtering criteria, b) linguistic variable and its values

filtrującego *GDP około 45 000*, pozwalającego wyświetlić dane tych krajów, których produkt krajowy brutto wynosi około 45 tys. dolarów, z minimalnym stopniem zgodności 0,7.

3.2.4. Zmienna lingwistyczna oraz jej wartości

System przewiduje możliwość zdefiniowania zmiennych lingwistycznych reprezentujących pewną dziedzinę wartości fizycznych, których rozmyta reprezentacja byłaby pożądana w analizie. Dla każdej ze zmiennych lingwistycznych można zdefiniować jej wartości rozmy-

te, definiując każdą za pomocą liczby rozmytej. Na rysunku 5b przedstawiono sposób definiowania zmiennej lingwistycznej *fvGDP* i jej rozmyte wartości: *low*, *medium*, *high*, *huge* (ozn. odpowiednio: niski, średni, wysoki i ogromny produkt krajowy brutto).

3.2.5. Grupowanie danych rozmytych

Grupowanie liczb rozmytych ze względu na nieprecyzyjny charakter traktuje ten typ danych w sposób specjalny. Tradycyjne grupowanie liczb rozmytych powodowałoby przyporządkowanie do odrębnych grup liczb, które są do siebie zbliżone, lecz nie takie same. Traktowanie takich liczb w sposób dokładny byłoby zatem niepożądane.

W związku z tym w opracowanym systemie zdefiniowano inne sposoby grupowania. Jednym ze sposobów, który aplikacja udostępnia, jest grupowanie lingwistyczne [15], które opiera się na wybranej zmiennej lingwistycznej i prowadzi grupowanie względem jej wartości. Drugim sposobem grupowania jest grupowanie metodą K-średnich (ang. *k-means*) [16]. W zaimplementowanym algorytmie użytkownik definiuje liczbę grup, które zostaną utworzone w wyniku jego działania. Efektem działania algorytmu jest utworzenie określonej liczby grup reprezentowanych przez liczby rozmyte określające środek każdej z grup.

Obie metody grupowania pozwalają grupować dane rozmyte o postaci przedstawionej w punkcie 3.2.1.

3.2.6. Rozmyte grupowanie danych precyzyjnych

W systemie zostały także zaimplementowane metody rozmytego grupowania danych precyzyjnych. Umożliwia to połączenie w grupy podobnych danych, np. ludzi w podobnym wieku, i przeprowadzenie analizy danych zagregowanych dla każdej z powstałych grup. W systemie zaimplementowano grupowanie liczb ostrych metodą K-średnich (ang. *k-means*) [16] i grupowanie względem zmiennej lingwistycznej [15].

3.2.7. Agregacja danych rozmytych

Kluczowe dla przetwarzania analitycznego są agregacje danych. Obejmują one takie operacje, jak wyznaczenie sumy lub średniej dla wybranego zakresu danych. Innymi agregacjami są wyznaczanie wartości minimalnej oraz maksymalnej. W przypadku liczb rozmytych pierwsze dwa agregaty (FSUM i FAVG) są realizowane dzięki arytmetyce liczb rozmytych, a kolejne dwa (FMAX i FMIN) dzięki logice rozmytej. Piątym agregatem jest wyznaczanie liczby analizowanych danych (COUNT). Operacja ta nie wymaga jednak specjalnego traktowania i nie musi być implementowana w szczególny sposób.

3.2.8. Rozszerzenie języka SQL

Mechanizmy związane z obsługą liczb rozmytych są dostępne dla każdego wybranego źródła danych i zostały zaimplementowane jako rozszerzenie języka SQL. Jeśli wybrane źró-

dło danych nie zawiera wspomnianego rozszerzenia, należy je dołączyć za pomocą udostępnionego instalatora.

3.2.9. Konwersja do liczb rozmytych

W wielu przypadkach źródło danych może nie zawierać danych rozmytych. Aplikacja analityczna FDW Browser oferuje funkcję umożliwiającą wprowadzenie tego typu danych. Dane rozmyte można uzyskać na zasadzie konwersji istniejących danych precyzyjnych. Na przykład, posiadając szacunkowe dane o PKB można dokonać ich rozmycia w celu poprawy możliwości ich analizy. Konwersja może dotyczyć dokładnych danych liczbowych lub danych tekstowych. W obu przypadkach konwersja oparta jest na wybranej, zdefiniowanej wcześniej zmiennej lingwistycznej. Dane liczbowe kojarzone są z odpowiednią wartością zmiennej lingwistycznej, reprezentowaną liczbą rozmytą. W przypadku danych tekstowych konwersja polega na odpowiednim skojarzeniu tekstu z wartością lingwistyczną. Użytkownik ma możliwość zdefiniowania reguły analizowania tekstu wejściowego.

4. Podsumowanie

Przedstawiony w niniejszym artykule system hurtowni danych rozmytych stanowi zwieńczenie wieloletnich prac autorów w dziedzinie zastosowania logiki rozmytej w systemach baz danych. O ile w literaturze światowej można znaleźć przykłady systemów, które przetwarzają w sposób rozmyty dane z klasycznych hurtowni danych (m.in. w pracach [9-12]), o tyle w zakresie budowy hurtowni danych rozmytych autorzy nie spotkali się z takimi pracami.

Hurtownie danych rozmytych dają szerokie możliwości przechowywania i przetwarzania danych o nieprecyzyjnym charakterze. Dane tego typu mogą pochodzić z różnych źródeł, m.in. z systemów pomiarowych, ze źródeł, które nie są w stanie dokładnie określić pewnych wartości i mogą je jedynie oszacować, z systemów, w których dane celowo gromadzone są w nieprecyzyjny sposób, aby w przyszłości zapewnić możliwość odszukania informacji potencjalnie spełniającej podane kryterium wyszukiwania z pewnym stopniem podobieństwa. Systemy klasy Fuzzy-OLAP współpracujące z hurtownią danych rozmytych umożliwiają dodatkowo zaawansowaną analizę zgromadzonych danych zarówno rozmytych, jak i danych dokładnych. Udostępniają one bowiem szereg funkcji rozmytego przetwarzania danych w hurtowni, np. rozmytego grupowania danych precyzyjnych i rozmytych, definiowania kryteriów filtrujących zawierających wyrażenia rozmyte, agregacji i prezentacji danych rozmytych i in.

W ostatnim okresie autorzy artykułu prowadzili badania nad budową hurtowni danych rozmytych gromadzącej dane dotyczące globalnych zasobów naturalnych na świecie, gdzie

planowane zapotrzebowanie na zasoby było określane szacunkowo [17]. W obecnej chwili trwają prace nad budową hurtowni danych rozmytych dla systemu wspomagania działań operacyjno-procesowych i prowadzenia analiz kryminalistycznych wspierających obronność i bezpieczeństwo państwa.

BIBLIOGRAFIA

1. Kimball R., Reeves L., Margy R., Thornthwaite W.: The Data Warehouse Lifecycle Toolkit. John Wiley & Sons, 1998.
2. Ponniah P.: Data Warehousing Fundamentals. A Comprehensive Guide for IT Professionals. John Wiley and Sons, 2001.
3. Zadeh L.A.: Fuzzy sets. Information and Control. 1965, 8 (3), s. 338÷353.
4. Tang X., Chen G.: A complete set of fuzzy relational algebraic operators in fuzzy relational databases. Proceedings of the 2004 IEEE International Conference on Fuzzy Systems, 2004, s. 565÷569.
5. Bosc P., Pivert O.: SQLf: A Relational Database Language for Fuzzy Querying. IEEE Transactions on Fuzzy Systems. 1995, Vol. 3, No. 1.
6. Kacprzyk J., Zadrozny S.: SQLf and FQUERY for Access. IFSA World Congress and 20th NAFIPS International Conference, 2001, s. 2464÷2469.
7. Małysiak B.: Fuzzy Values in SQL Queries Submitted to Databases. Studia Informatica. Vol. 24, No. 2A(53), s. 179÷190, Gliwice 2003.
8. Małysiak B., Mrozek D., Kozielski S.: Processing Fuzzy SQL Queries with Flat, Context-Dependent and Multidimensional Membership Functions. Proc. of 4th IASTED International Conference on Computational Intelligence (CI 2005), Calgary, Canada. ACTA Press, 2005, s. 36÷41.
9. Chaudhuri S., Ganjam K., Ganti V., Motwani R.: Robust and efficient fuzzy match for online data cleaning. Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data. San Diego, California, 2003, s. 313÷324.
10. Hua-Yang Lin, Ping-Yu Hsu, Gwo-Ji Sheen: A fuzzy-based decision-making procedure for data warehouse system selection. An International Journal of Expert Systems with Applications. 2007, s. 939÷953.
11. Perez D., Somodevilla M.J., Pineda I.H.: Fuzzy Spatial Data Warehouse: A Multidimensional Model. 8th Mexican International Conference on Current Trends in Computer Science, 2007, s. 3÷9.
12. Fasel D., Zumstein D.: A Fuzzy Data Warehouse Approach for Web Analytics. LNCS, Vol. 5736, sp. 276÷285. Springer, Heidelberg 2009.

13. Bouchon-Meunier B., Yager R.R., Zadeh L.A.: Fuzzy logic and soft computing. Advances in Fuzzy Systems, Application and Theory vol.4, Singapore 1995.
14. Dubois D., Prade H.: Fundamentals of fuzzy sets. Kluwer Academic Publisher, 2000.
15. Małysiak-Mrozek B., Mrozek D., Kozielski S.: Data Grouping Process in Extended SQL Language Containing Fuzzy Elements. Advances in Intelligent and Soft Computing Vol. 59, Springer Verlag GmBH, 2009, s. 247÷256.
16. MacQueen J.B.: Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1967, vol. 1, s. 281÷297.
17. Małysiak-Mrozek B., Mrozek D., Kozielski S.: Processing of Crisp and Fuzzy Measures in the Fuzzy Data Warehouse for Global Natural Resources. LNAI, Springer, Heidelberg 2010, w publikacji.

Recenzent: Prof. dr hab. inż. Jerzy Klamka

Wpłynęło do Redakcji 31 stycznia 2010 r.

Abstract

Data warehouses are special purpose repositories that collect huge volumes of data usually for reporting and querying. These kinds of systems support fast and complex data analysis and constitute a foundation for decision support systems [1], [2].

For the last three decades, we can observe several attempts to incorporate fuzzy processing into database systems. In the work [4], we can find some theoretical considerations about fuzzy relational algebraic operators useful for designing fuzzy query languages. Furthermore, different real-life implementations, like SQLf [5], FQUERY [6], or FuzzySQL [7], [8], show how to extend standard databases to submit queries supporting imprecision, fault-tolerance, and data similarity. This approximate approach to data analysis and processing gives several advantages.

Since data warehouses gather so huge amount of data, it can be desirable to include some fuzziness into these systems in particular IT projects. However, concerning the fuzziness in the context of data processing we have to distinguish two different cases. The first one is fuzzy processing of crisp values that we store in a data warehouse. The processing must cover fuzzy grouping of crisp data, fuzzy filtering of rows and groups. The second case is processing of fuzzy values that we can collect in the data warehouse. If we imagine we have

fuzzy attributes in dimensions of the data warehouse, we have to use grouping by fuzzy attributes while aggregating data. We also need a possibility to filter data according to fuzzy filtering criteria, which is associated with slicing and dicing on dimensions in multidimensional cubes. Moreover, having fuzzy measures (measures defined on columns that store fuzzy data), we must implement the arithmetic of fuzzy numbers in order to aggregate fuzzy data.

In the paper, we present a Fuzzy Data Warehouse system (FDW), which we have designed and developed, recently. We define a Fuzzy Data Warehouse (FDW) as a data repository, which contains fuzzy data and allows a fuzzy processing of the data. In the paper, we focus on the most important functional features of the FDW system and our newly developed FDW Browser, which is an analytical application adhering to the Fuzzy-OLAP class of data exploration tools.

Adresy

Marek MIŁEK: student Politechniki Śląskiej, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, marek.milek@gmail.com .

Bożena MAŁYSIAK-MROZEK: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, bozena.malysiak@polsl.pl .

Dariusz MROZEK: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, dariusz.mrozek@polsl.pl .