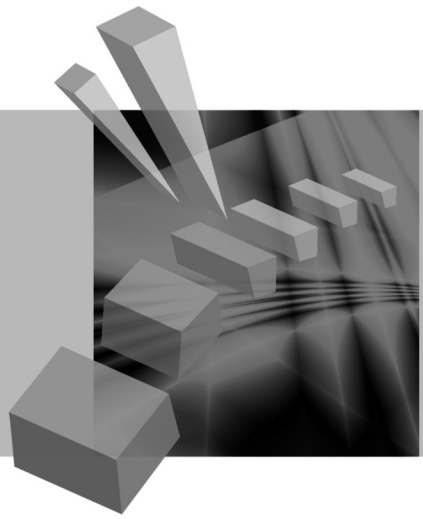
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pociecha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon http://kangur.uek.krakow.pl/
bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

Kopiowanie i powielanie w jakiejkolwiek formie
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu
Wrocław 2012

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Tomasz Bartłomowicz, Justyna Wilk
Uniwersytet Ekonomiczny we Wrocławiu

ZASTOSOWANIE METOD ANALIZY DANYCH SYMBOLICZNYCH W PRZESZUKIWANIU DZIEDZINOWYCH BAZ DANYCH

Streszczenie: Treścią artykułu jest propozycja wykorzystania metodologii analizy danych symbolicznych w filtrowaniu dziedzinowych baz danych. Proponowane rozwiązanie, obok zmiennych klasycznych, uwzględnia zmienne symboliczne, które opisują obiekty bez utraty informacji. Ponadto w rozwiązaniu wykorzystano znormalizowaną miarę Ichino-Yaguchiego dla danych symbolicznych. W opinii autorów połączenie to umożliwia przeszukiwanie baz danych na podstawie wszystkich możliwych kryteriów bez względu na rodzaj występujących zmiennych. W praktyce oznacza to rozszerzenie możliwości występujących w dostępnych powszechnie mechanizmach filtrowania dziedzinowych baz danych, co zostało zaprezentowane na przykładzie przeszukiwania ofert nieruchomości.

Słowa kluczowe: filtrowanie baz danych, analiza danych symbolicznych, pomiar odległości, oferty nieruchomości.

1. Wstęp

Przeszukiwanie rozumiane głównie jako filtrowanie[1] dziedzinowych baz danych to jedno z podstawowych zagadnień informatyki polegające na selekcji zbioru danych względem pewnych zmiennych (cech, atrybutów) charakterystycznych dla każdego z elementów tego zbioru[2]. Czynność ta realizowana niejednokrotnie intuicyjnie stanowi podstawową praktykę wszędzie tam, gdzie ilość prezentowanego materiału przekracza możliwości jego pełnego poznania. Przykładem mogą być tutaj katalogi biblioteczne, serwisy aukcyjne, portale społecznościowe.

Podstawą przeszukiwania jest najczęściej zestaw kryteriów, które powinny spełniać wyszukane obiekty[3]. Ze względu na brak sprecyzowanych preferencji, chęć

[1] Oprócz filtrowania równie popularną czynnością jest sortowanie rekordów bazy danych.
[2] Opis na podstawie internetowego słownika języka polskiego, http://sjp.pwn.pl.
[3] Istnieje możliwość przeszukiwania baz danych na zasadzie negacji, tj. wyszukiwania rekordów, które nie spełniają określonych kryteriów.

porównania kilku obiektów, tudzież ich specyfikę informacje zamieszczone w bazach danych lub dostępne kryteria wyszukiwania są niejednokrotnie mało precyzyjne lub złożone. Mogą mieć formę nie tylko pojedynczej kategorii lub wartości liczbowej, ale także zbioru kategorii, przedziału wartości lub struktury udziałowej. Tego rodzaju dane w literaturze przedmiotu określane są jako dane symboliczne, a metodami umożliwiającymi ich analizę są metody analizy danych symbolicznych (por. [Bock, Diday 2000; Diday, Noirhomme-Fraiture 2008; Billard, Diday 2006; Gatnar, Walesiak 2011]).

Treścią artykułu jest propozycja wykorzystania metodologii analizy danych symbolicznych do filtrowania zawartości dziedzinowych baz danych. W szczegółowym zakresie w artykule zostały zaprezentowane format i źródła danych symbolicznych w zastosowaniach bazodanowych oraz proponowany mechanizm przeszukiwaniu baz danych z wykorzystaniem metodologii analizy danych symbolicznych. Całość proponowanego rozwiązania ilustruje przykład selekcji ofert rynku nieruchomości.

2. Definicja problemu – dane symboliczne

Filtrowanie baz danych wiąże się z określeniem kryteriów wyszukiwania i znalezieniem obiektów spełniających te kryteria. Kryteria wyszukiwania oraz cechy obiektów bazodanowych mogą mieć formę danych klasycznych bądź symbolicznych.W ujęciu klasycznym obiekty opisywane są za pomocą zmiennych, których realizacjami są pojedyncze wartości liczbowe, np. rok produkcji samochodu lub kategorie, np. płeć.

W ujęciu symbolicznym obiekty mogą być również charakteryzowane zmiennymi o realizacjach w postaci:

- przedziałów liczbowych (skokowych lub ciągłych, rozłącznych lub nierozłącznych) ze zbioru liczb rzeczywistych, np. zakres cenowy produktu, zakres wiekowy osoby;
- zbiorów kategorii (nominalnych lub porządkowych) bądź zbiorów wartości (o charakterze metrycznym bądź przedziałów liczbowych), np. słowa kluczowe w publikacji, zastosowanie produktu;
- struktur procentowych, tj. zbiorów kategorii (nominalnych lub porządkowych) lub wartości (metrycznych lub przedziałów liczbowych) z przypisanymi współczynnikami wagowymi, np. skład produktu.

Uwzględnia się także występowanie logicznych (hierarchicznych lub taksonomicznych) relacji między zmiennymi, np. marka i model samochodu.

W artykule przewodnie jest spostrzeżenie, iż filtrowanie informacji zawartych w dziedzinowych bazach danych odbywa się głównie z wykorzystaniem kryteriów w postaci zmiennych klasycznych. Dla obiektów w postaci nieruchomości jest to zazwyczaj cena nieruchomości, tudzież inne metryczne cechy, jak powierzchnia nieruchomości, liczba pokoi, piętrowa lokalizacja lokalu itp. Rzadziej na potrzeby filtrowania wykorzystywane są zmienne klasyczne mierzone na słabych skalach pomiaru. Przykładem może być tutaj lokalizacja nieruchomości występująca jako zmienna mierzona na skali nominalnej.

W przypadku zmiennych symbolicznych złożona struktura danych w ujęciu symbolicznym sprawia, iż nie można w sposób bezpośredni stosować metod statystycznych opracowanych dla danych w ujęciu klasycznym (por. [Everitt, Dunn 2001; Hair i in. 2006]). Oznacza to brak mechanizmów filtrowania baz danych na podstawie kryteriów wyszukiwania o charakterze zmiennych symbolicznych, a w przypadku wyszukiwania nieruchomości brak możliwości uwzględnienia m.in. takiej zmiennej występującej w postaci listy kategorii, jaką jest uzbrojenie terenu. Realizacjami tej zmiennej nie są pojedyncze wartości, lecz zbiór kategorii. Oznacza to możliwość występowania dowolnej kombinacji możliwych kategorii dla tej zmiennej, np.: „prąd, woda, gaz", w sytuacji gdy dla innej nieruchomości dowolne z wymienionych czynników nie występują lub są zastępowane innymi.

Podsumowując, należy zauważyć, iż dostępne, głównie na stronach internetowych, mechanizmy filtrowania baz danych są co najmniej niewystarczające, a nawet – co stanowi tezę niniejszego artykułu – nieadekwatne do zawartości dziedzinowych baz danych. W przekonaniu autorów, z racji występowania wspomnianych zmiennych symbolicznych, dedykowane rozwiązania w postaci „klasycznych" mechanizmów filtrowania uszczuplają możliwości filtrowania baz danych. Tym samym celem artykułu jest prezentacja rozwiązania pozwalającego wymienione wady wyeliminować, a przynajmniej w wysokim stopniu je ograniczyć.

3. Proponowane rozwiązanie

Rozwiązanie postawionego w artykule problemu polega na zaproponowaniu sposobu filtrowania dziedzinowych baz danych z wykorzystaniem porządkowania obiektów na podstawie odległości od obiektu wzorca z uwzględnieniem odpowiedniej miary podobieństwa. Warto podkreślić, iż ze względu na obiekty opisane m.in. zmiennymi symbolicznymi w rozwiązaniu zdecydowano się wykorzystać znormalizowaną miarę odległości Ichino-Yaguchiego U_3. Konstrukcję miar odległości danych symbolicznych zaprezentowano m.in. w pracach: [Bock, Diday 2000, s. 165-185; Diday, Noirhomme-Fraiture 2008, s. 126-129; Wilk 2005; 2006; Malerba i in. 2001; Malerba, Esposito, Monopoli 2002].

Miara Ichino-Yaguchiego U_3 w swojej konstrukcji wykorzystuje operatory połączenia „$\oplus$" oraz przekroju „$\otimes$". Operator połączenia jest definiowany jako (por. [Bock, Diday 2000, s. 170-171]):

$$v_{ik} \oplus v_{jk} = \begin{cases} \min\{\underline{v_{ik}}, \underline{v_{jk}}\}, \max\{\overline{v_{ik}}, \overline{v_{jk}}\}] & \text{zmienna w postaci przedziału liczbowego,} \\ v_{ik} \cup v_{jk} & \text{zmienna w postaci zbioru kategorii,} \end{cases}$$

gdzie: i, j – i-ty, j-ty obiekt,

v_{ik}, v_{jk} – realizacja k-tej zmiennej dla odpowiednio i-tego i j-tego obiektu,

$\underline{v_{ik}}, \underline{v_{jk}}$ $\left(\overline{v_{ik}}, \overline{v_{jk}}\right)$ – dolne (górne) końce przedziału k-tej zmiennej, odpowiednio dla obiektów i oraz j.

Operator przekroju dla wszystkich typów zmiennych definiowany jest jako:

$$v_{ik} \otimes v_{jk} = v_{ik} \cap v_{jk} .+.$$

W strukturze miary U_3 wyróżnia się odległości składowe oraz odległość agregatową. Odległości obiektów względem *k*-tej zmiennej wyznacza się w następujący sposób (por. [Bock i Diday 2000]):

$$d_{ijk} = \frac{\mu(v_{ik} \oplus v_{jk}) - \mu(v_{ik} \otimes v_{jk}) + \gamma v(v_{ik}, v_{jk})}{\mu(V_k)}, \tag{1}$$

gdzie: $v(v_{ik}, v_{jk}) = 2\mu(v_{ik} \otimes v_{jk}) - \mu(v_{ik}) - \mu(v_{jk})$,

γ – parametr, $\gamma \in [0; 0,5]$,

$\mu(v_{ik})$ – rozpiętość przedziału *k*-tej zmiennej dla *i*-tego obiektu lub liczba kategorii *k*-tej zmiennej dla *i*-tego obiektu,

$\mu(v_{ik} \oplus v_{jk})$ – wartość bezwzględna z różnicy wartości minimalnej i maksymalnej wyrażenia $v_{ik} \oplus v_{jk}$ lub liczba kategorii dla v_{ik} oraz v_{jk},

$\mu(v_{ik} \otimes v_{jk})$ – wartość bezwzględna części wspólnej przedziałów v_{ik} i v_{jk} lub liczba wspólnych kategorii dla v_{ik} oraz v_{jk},

$\mu(V_k)$ – rozpiętość zbioru realizacji *k*-tej zmiennej $\mu(V_k) = \left|\max\{\overline{v_k}\} - \min\{\underline{v_k}\}\right|$ lub liczba wszystkich kategorii *k*-tej zmiennej.

Odległość agregatową obiektu *i*-tego oraz *j*-tego wyznacza się w następujący sposób:

$$d_{ij} = \left[\sum_{k=1}^{p} (\omega_k d_{ijk})^{\lambda}\right]^{1/\lambda}, \tag{2}$$

gdzie: $d_{ij} \in [0,1]$,

λ – parametr, $\lambda \geq 1$,

ω_k – waga *k*-tej zmiennej $(k = 1, ..., p)$, $\omega_k > 0$, $\sum_{k=1}^{p} \omega_k = 1$.

Sposób zastosowania znormalizowanej miary Ichino-Yaguchiego U_3 do rozwiązania postawionego w artykule problemu filtrowania dziedzinowych baz danych prezentuje poniższy przykład.

4. Filtrowanie dziedzinowych baz danych na przykładzie przeszukiwania ofert niezabudowanych nieruchomości gruntowych

W proponowanym przykładzie zawartość dziedzinowej bazy danych stanowi 38 wariantów (profilów) niezabudowanych nieruchomości gruntowych (działek budowlanych) opisanych zestawem ośmiu cech (atrybutów) w postaci następujących zmiennych klasycznych: cena nieruchomości, powierzchnia działki, lokalizacja oraz nasłonecznienie nieruchomości. W przypadku zmiennych symbolicznych są to: funkcja w miejscowym planie zagospodarowania przestrzennego, forma władania nieruchomością, uzbrojenia terenu oraz rodzaj dojazdu do nieruchomości. Należy zauważyć, iż każdą ze zmiennych charakteryzują określone realizacje (por. tab. 1). W przypadku ceny oraz powierzchni działki są to pojedyncze wartości liczbowe; w przypadku lokalizacji oraz nasłonecznienia nieruchomości są to wybrane, pojedyncze kategorie. W odniesieniu do lokalizacji są to następujące strefy: „centralna", „śródmiejska" lub „peryferyjna", w przypadku nasłonecznienia nieruchomości – wybrane, pojedyncze kategorie: „bardzo słoneczna" lub „słoneczna".

Tabela 1. Cechy niezabudowanych nieruchomości gruntowych na jeleniogórskim rynku nieruchomości z oferty wybranego biura pośrednictwa w obrocie nieruchomościami

Nazwa cechy	Rodzaj cechy	Możliwe realizacje
Cena nieruchomości	ilorazowa	42.000-1.457.000 zł
Powierzchnia działki	ilorazowa	437-49.970 m^2
Lokalizacja nieruchomości	nominalna	centralna
		śródmiejska
		peryferyjna
Funkcja w miejscowym planie zagospodarowania przestrzennego	symboliczna (lista kategorii)	mieszkaniowa
		usługowa
		rolna
Forma władania	symboliczna (lista kategorii)	własność
		użytkowanie wieczyste
		hipoteka
Uzbrojenie terenu	symboliczna (lista kategorii)	kanalizacja
		prąd
		siła
		woda
		gaz
Dojazd do nieruchomości	symboliczna (lista kategorii)	droga asfaltowa
		droga utwardzona
		droga nieutwardzona
Nasłonecznienie działki	nominalna	bardzo słoneczna
		słoneczna

Źródło: opracowanie własne.

W przypadku występujących w przykładzie zmiennych symbolicznych realizacjami jest zbiór kategorii, tj. dowolna kombinacja pojedynczych kategorii dostępnych dla zmiennej. W przypadku funkcji w miejscowym planie zagospodarowania przestrzennego dostępne kategorie to: „mieszkaniowa", „usługowa" oraz „rolna"; dla formy władania nieruchomością – „własność", „użytkowanie wieczyste" i „hipoteka"; w przypadku uzbrojenia terenu – „kanalizacja", „prąd", „siła", „woda" i „gaz"; odnośnie do dojazdu do nieruchomości – „droga asfaltowa", „droga utwardzona" i „droga nieutwardzona" (por. tab. 1). Źródło danych stanowią strony internetowe jeleniogórskich biur pośrednictwa w obrocie nieruchomościami[4].

Przedstawione w artykule rozwiązanie w miejsce filtrowania obiektów (ofert nieruchomości) z wykorzystaniem wyłącznie zmiennych klasycznych, co oznacza automatyczną eliminację kryteriów „symbolicznych", proponuje mechanizm polegający każdorazowo na wykorzystaniu obu rodzajów zmiennych – klasycznych oraz symbolicznych. Połączenie takie oznacza automatycznie brak możliwości zastosowania „klasycznych" metod wielowymiarowej analizy statystycznej na rzecz metodologii analizy danych symbolicznych. Z uwzględnieniem odpowiedniej miary odległości Ichino-Yaguchiego dla danych symbolicznych proponuje się obliczenie odległości od obiektu wzorca do każdego z obiektów bazy danych. W proponowanym rozwiązaniu umożliwia to wyróżnienie obiektów najbardziej do wzorca podobnych, których odpowiednią liczbę uznać należy za wynik przeszukiwania bazy danych.

Na potrzeby prezentacji proponowanego rozwiązania przyjmuje się założenie, iż użytkownika bazy danych (strony internetowej z ofertami nieruchomości) szczególnie interesuje na własność wariant nieruchomości w postaci bardzo słonecznej działki mieszkaniowej w cenie do 300 000 zł o powierzchni między 700 a 3000 m^2, charakteryzujący się „śródmiejską" lub „peryferyjną" lokalizacją z dostępem do „kanalizacji", „prądu", „wody" oraz „gazu", posiadający dojazd w postaci „drogi asfaltowej" i/lub „drogi utwardzonej". Jednocześnie zakłada się, iż użytkownika interesuje pewien zbiór obiektów – działek budowlanych – w przykładzie są to 3 działki budowlane, które spełniają wszystkie kryteria lub jak najwięcej kryteriów wyszukiwania.

Aby możliwe było wyszukanie obiektów spełniających jak najwięcej zadanych kryteriów, w proponowanym rozwiązaniu przeszukanie bazy danych polega na wyznaczeniu macierzy odległości między obiektami z wykorzystaniem wzorów (1) oraz (2), a następnie odpowiednim ich uszeregowaniu. Obliczenia, wykorzystując w tym celu polecenie `dist.SDA` z modułu `symbolicDA`, zrealizowano w środowisku R, otrzymując następujący zestaw wyników dla miary Ichino-Yaguchiego:

```
> library(symbolicDA)
> x<-parse.SO("nieruchomosci")
> d<-dist.SDA(x, type="U_3", gamma=0.2, power=2)
```

[4] Główne źródło danych o nieruchomościach stanowiła strona WWW biura pośrednictwa w obrocie nieruchomościami „Nieruchomości Kopczyński", http://www.nkop.pl.

```
      1    2    3    4    5    6    7    8    9   10   11  ...  36   37   38
2  0.41
3  1.01 0.92
4  0.73 0.80 0.80
5  0.86 0.76 0.53 0.80
6  1.01 1.00 0.38 0.80 0.84
7  1.01 1.00 1.13 0.80 1.36 0.92
8  0.82 0.60 0.60 0.84 0.80 0.60 0.80
9  0.82 0.97 1.10 0.38 1.10 1.10 0.97 1.07
10 1.08 0.86 1.26 1.10 1.00 1.36 1.47 1.16 1.03
11 0.67 0.93 0.54 0.80 0.76 0.85 0.85 0.96 0.80 1.25
12 0.98 0.96 0.80 0.39 0.80 0.60 0.80 0.93 0.76 1.29 1.11
13 1.04 0.97 0.26 0.83 0.60 0.45 1.17 0.65 1.12 1.27 0.59 ...
14 1.05 1.23 0.80 0.76 1.10 0.96 0.61 1.13 0.84 1.53 0.60 ...
...
37 0.82 0.96 1.22 0.66 1.10 1.22 1.10 1.19 0.76 0.90 0.97 ... 0.01
38 0.82 0.96 1.22 0.66 1.10 1.22 1.10 1.19 0.76 0.90 0.97 ... 0.01 0.00
39 0.90 1.10 0.80 0.65 0.96 1.03 0.71 1.13 0.65 1.34 0.60 ... 0.85 0.85 0.85.
```

Ze względu na potrzebę uszeregowania profilów nieruchomości od najbardziej do najmniej podobnego względem profilu wzorca w obliczeniach stosuje się dodatkowo polecenie `sort`, co pozwala uzyskać następującą, ostateczną postać wyników:

```
> library(symbolicDA)
> x<-parse.SO("nieruchomosci")
> d<-dist.SDA(x, type="U_3", gamma=0.2, power=2)
> d<-as.matrix(d)
> round(sort(d[1,]), 4)
```

```
     1     29      2     16     11      4     17      8      9     30     31
0.0000 0.4137 0.4142 0.4922 0.6750 0.7257  0.7257 0.8179 0.8179 0.8179 0.8179
    32     33     34     35     36     37     38      5     20     21     22
0.8179 0.8179 0.8179 0.8179 0.8179 0.8179 0.8179 0.8602 0.8602 0.8602 0.8964
    25     39     12      3      6      7     13     27     14     28     10
0.9006 0.9006 0.9764 1.0122 1.0122 1.0123 1.0428 1.0467 1.0470 1.0572 1.0802
    23     15     24     18     19     26
1.1043 1.1441 1.2717 1.2902 1.4325 1.4987.
```

Rezultaty przeszukiwania jednocześnie wskazują, iż wariantami nieruchomości najbardziej podobnymi do profilu wzorcowego (w liczbie 3 sztuk, nie licząc wariantu wzorcowego opatrzonego numerem 1) są profile nr 29, 2 oraz 16. Analiza wariantów nieruchomości wskazuje, iż w przypadku wariantu nr 29 wszystkie cechy mieszczą się w granicach kryteriów wariantu wzorca, w przypadku dwóch pozostałych wariantów różnice dotyczą pojedynczych cech. I tak w przypadku wariantu nr 2 zauważyć można niezgodność w odniesieniu do powierzchni działki, gdzie powierzchnia wynosi „437 m^2" w miejsce wymaganych co najmniej „700 m^2"; w przypadku wariantu nr 16 różnica dotyczy uzbrojenia terenu i polega na występowaniu dodatkowej kategorii w postaci „siły", tj. *de facto* kompletnego uzbrojenia na tle wzorca w postaci kombinacji „kanalizacji, prądu, wody, gazu".

Warto w tym miejscu zauważyć, iż dla przykładu wariant najbardziej odległy (nr 26) od wzorcowego profilu nieruchomości różni się względem wzorca w zakresie 4 cech. Różnice te obejmują następujące realizacje (w odniesieniu do obiektu wzorca): cena – „1 457 000 zł" (wzorzec – maksymalnie „300 000 zł"), powierzchnia – „26 489 m^2" (wzorzec – „3300 m^2"), funkcja w miejscowym planie zagospodarowania przestrzennego – „rolna" (wzorzec – „mieszkaniowa"), uzbrojenie – „prąd" (wzorzec – „kanalizacja, prąd, woda, gaz").

Należy podkreślić, iż proponowany mechanizm filtrowania oznacza wyszukanie zadeklarowanej liczby obiektów, które ze wszystkich z dostępnych w bazie danych w największym stopniu odpowiadają ofercie porównywanej. Oznacza to, iż proponowany mechanizm przeszukiwania, co należy uznać za jego zaletę, pozwala bez względu na objętość bazy danych wyszukiwać obiekty najbardziej, choć nie do końca, spełniające kryteria obiektu wzorcowego. Filtrowanie w „klasycznym" rozumieniu tego pojęcia takiej możliwości nie daje.

5. Podsumowanie

Jak to zostało zauważone w artykule, dostępne narzędzia przeszukiwania (filtrowania) bazują głównie na zmiennych klasycznych opisujących obiekty będące przedmiotem baz danych. Przedstawione w artykule rozwiązanie obok zmiennych klasycznych uwzględnia dodatkowo zmienne symboliczne, które opisują obiekty w sposób pełny, bez utraty informacji o obiektach, których dotyczą. W opinii autorów artykułu umożliwia to przeszukiwanie baz danych na podstawie dowolnej, w tym pełnej liczby kryteriów bez względu na rodzaj występujących zmiennych. W praktyce oznacza to bardziej efektywne przeszukiwanie baz danych z uwzględnieniem możliwości do tej pory niewystępujących w dostępnych mechanizmach filtrowania dziedzinowych baz danych.

Literatura

Billard L., Diday E., *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, Wiley, Chichester 2006.

Bock H.H., Diday E. (red.), *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin, Heidelberg 2000.

Diday E., Noirhomme-Fraiture M. (red.), *Symbolic Data Analysis and the Sodas Software*, John Wiley & Sons, Chichester 2008.

Everitt B.S., Dunn G., *Applied Multivariate Data Analysis*, Arnold, London 2001.

Gatnar E., Walesiak M. (red.), *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*, C.H. Beck, Warszawa 2011.

Hair J.F., Black W.C., Babin B.J, Anderson R.E., Tatham R.L., *Multivariate Data Analysis*, Pearson Prentice Hall, New Jersey 2006.

Malerba D., Esposito F., Giovalle V., Tamma V., *Comparing Dissimilarity Measures for Symbolic Data Analysis*, [w:] *New Techniques and Technologies for Statistics and Exchange of Technology and Know-how*, P. Nanopoulos (red.), ETK-NTTS'01 2001.

Malerba D., Esposito F., Monopoli M., *Comparing Dissimilarity Measures for Probabilistic Symbolic Objects*, [w:] *Data Mining III. Series Management Information Systems*, A. Zanasi, C.A. Brebbia, N.F.F. Ebecken, P. Melli (red.), vol. 6, WIT Press, Southampton 2002.

Wilk J., *Miary odległości obiektów opisanych zmiennymi symbolicznymi z wagami*, [w:] Taksonomia 13, *Klasyfikacja i analiza danych – teoria i zastosowania*, K. Jajuga, M. Walesiak (red.), Prace Naukowe Akademii Ekonomicznej we Wrocławiu, Wrocław 2005.

Wilk J., *Problemy klasyfikacji obiektów symbolicznych. Symboliczne miary odległości*, [w:] *Ilościowe i jakościowe metody badania rynku. Pomiar i jego skuteczność*, J. Garczarczyk (red.), ZN AE nr 71, Wydawnictwo AE, Poznań 2006.

APPLICATION OF SYMBOLIC DATA ANALYSIS METHODS FOR DOMAIN DATABASE SEARCHING

Summary: The paper presents the application of symbolic data analysis methods for domain database searching. Because it is better to use more information about objects, the proposition of authors' database searching includes classical and symbolic variables and Ichino-Yaguchi dissimilarity measure. In authors' opinion it means the most effective database filtering. The authors illustrate the presented solution using symbolic data analysis method on an empirical example.

Keywords: database searching, symbolic data analysis, distance measurement, real estate offers.