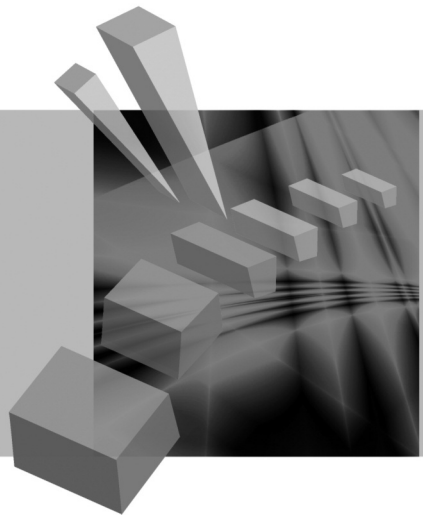
**PRACE NAUKOWE**
Uniwersytetu Ekonomicznego we Wrocławiu
**RESEARCH PAPERS**
of Wrocław University of Economics

**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



**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

**Katarzyna Wójcik, Janusz Tuchowski**
Uniwersytet Ekonomiczny w Krakowie

# ANALIZA PORÓWNAWCZA MIAR PODOBIEŃSTWA TEKSTÓW OPARTYCH NA MACIERZY CZĘSTOŚCI I TEKSTÓW OPARTYCH NA WIEDZY DZIEDZINOWEJ

**Streszczenie:** Zasadniczym celem niniejszej pracy jest próba oceny przydatności znanych z literatury miar podobieństwa tekstów bazujących na macierzy częstości do tych bazujących na wiedzy dziedzinowej w postaci ontologii. W kolejnych punktach artykułu przedstawione zostały najpierw dokumenty tekstowe, które były porównywane w badaniu, a następnie wybrane miary podobieństwa oparte ma macierzy częstości i ich analiza symulacyjna. W dalszej części zaprezentowana została ontologia wykorzystana w badaniach oraz wyniki przeprowadzonej analizy symulacyjnej miar opartych na wiedzy reprezentowanej przez tę ontologię. Na tej podstawie została podjęta próba oceny przydatności tych miar.

**Słowa kluczowe:** miara, podobieństwo, macierz częstości, *text mining*, ontologia.

## 1. Wstęp

Jednym z najważniejszych problemów pojawiających się przy eksploracyjnej analizie danych tekstowych jest wybór sposobu wyrażenia podobieństwa pomiędzy tekstami. W literaturze prezentowane są dwa zasadnicze podejścia do rozwiązania tego problemu.

Pierwsze z nich bazuje na reprezentacji częstotliwościowej dokumentów (macierzy zawierającej informacje o liczbie wystąpień poszczególnych słów w dokumencie). Drugie podejście do obliczania podobieństwa pomiędzy tekstami zakłada wykorzystanie wiedzy dziedzinowej. Do reprezentacji wiedzy dziedzinowej najczęściej stosuje się ontologie.

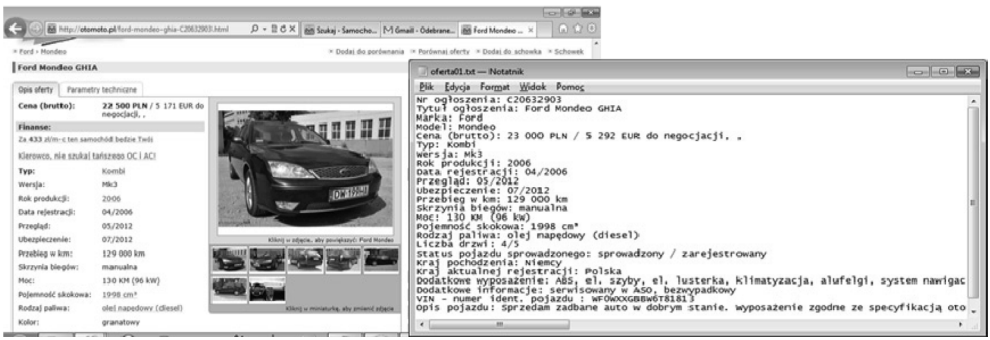
Zasadniczym celem pracy jest prezentacja znanych z literatury miar podobieństwa tekstów uwzględniających wiedzę dziedzinową oraz ich analiza symulacyjna. Wyniki analizy zostały również porównane z wynikami uzyskanymi dla miar opartych na macierzy częstości oraz ze znanym stopniem podobieństwa badanych tekstów.

W kolejnych rozdziałach pracy najpierw opisane zostaną badane dokumenty. Następnie krótko scharakteryzowane zostaną wybrane miary podobieństwa tekstów

bazujące na macierzy częstości. W kolejnych krokach omówiona zostanie ontologia wykorzystana w badaniach oraz system opracowany do analizy skupień bazującej na ontologiach. Pracę zakończą wnioski oraz dalsze plany badawcze.

## 2. Dokumenty tekstowe wykorzystane w badaniu

W symulacji wzięte zostały pod uwagę 22 dokumenty tekstowe zawierające teksty ogłoszeń z ofertami sprzedaży osobowych samochodów używanych. W przypadku ofert sprzedaży, porównując ogłoszenia, porównuje się oferowane produkty. Wybierając ogłoszenia, skupiono się na kilku wybranych modelach samochodów podobnej klasy. Wśród ofert znalazła się oferta kontrolna (powielone ogłoszenie 1 ze zmienionym jedynie numerem ogłoszenia) oraz oferta znacznie różniąca się od pozostałych. Ogłoszenia wybrane do badania pochodziły z najpopularniejszego polskiego serwisu Otomoto.pl.




**Rys. 1.** Przykładowa oferta sprzedaży samochodu używanego w serwisie Otomoto.pl oraz jej treść wyciągnięta z tego serwisu

Źródło: http://otomoto.pl/ford-mondeo-ghia-C20632903.html.

Ogłoszenia pobrane ze strony internetowej zostały przekształcone do formatu tekstowego przy wykorzystaniu opensource'owej biblioteki Javy Jsoup. Wyciągnięte zostały tylko dane związane z ofertą. Pominięto graficzne składowe ogłoszenia, reklamy i inne elementy witryny internetowej niezwiązane z ofertami (rys. 1).

Po operacji ekstrakcji danych otrzymano dokumenty tekstowe w częściowo ustrukturyzowanej formie. Prawa część rys. 1 przedstawia treść przykładowej oferty znajdującej się po jego lewej stronie wyciągniętą z serwisu internetowego. Tak przygotowane dokumenty zostały następnie wykorzystane w badaniu symulacyjnym.

## 3. Miary podobieństwa tekstów oparte na macierzy częstości

W tej części pracy przedstawione zostaną pokrótce miary podobieństwa tekstów bazujące na macierzy częstości oraz etapy badania symulacyjnego dotyczące tych miar.

### 3.1. Macierz częstości

$$A = \begin{matrix} \\ \text{wyrazy} \end{matrix} \overset{\text{dokumenty}}{\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{122} \\ a_{21} & a_{22} & \cdots & a_{222} \\ \vdots & \vdots & \ddots & \vdots \\ a_{11101} & a_{11102} & \cdots & a_{11122} \end{bmatrix}_{1110 \times 22}} \quad , \tag{1}$$

gdzie: $A$ – macierz częstości,
$a_{ij}$ – liczba wystąpień $i$-tego ($i = 1, \ldots, 1110$) wyrazu w $j$-tym ($j = 1, \ldots, 22$) dokumencie (w przypadku macierzy binarnej wystąpienie danego wyrazu w dokumencie ($a_{ij} = 1$) lub jego brak ($a_{ij} = 0$)).

Macierz częstości to macierz, której kolumny reprezentują dokumenty, a wiersze − wyrazy (wzór 1). Wartości wewnątrz macierzy częstości w jej wersji podstawowej odzwierciedlają liczbę wystąpień konkretnego słowa w danym dokumencie, a w wersji binarnej wszystkie wartości niezerowe są zamieniane na 1.

### 3.2. Miary podobieństwa tekstów

W badaniu wzięto pod uwagę dwie miary podobieństwa stosowane w odniesieniu do tekstów. Dokonując wyboru miar podobieństwa do badania, uwzględniono wyniki badań przeprowadzonych przez dra Dariusza Borratyńskiego [2009] oraz wyniki własnych badań [Wójcik 2010].

Do badania wybrano odległość kątową (wzór 2) oraz odległość Jaccarda (wzór 3) [Deza, Deza 2009]. Obydwie te miary są znormalizowane.

$$d_1(X,Y) = 1 - \frac{\sum_{k=1}^{n} x_k y_k}{\sqrt{\sum_{k=1}^{n} x^2{}_k \sum_{k=1}^{n} y^2{}_k}}, \tag{2}$$

$$d_2(X,Y) = 1 - \frac{\sum_{k=1}^{n} x_k y_k}{\sum_{k=1}^{n} x_k^2 + \sum_{k=1}^{n} y_k^2 - \sum_{k=1}^{n} x_k y_k}, \tag{3}$$

gdzie: $X$, $Y$ – dokumenty, kolumny macierzy częstości,
$d(X, Y)$ – odległość pomiędzy dokumentami $X$ i $Y$,
$k$ – numery wyrazów, wiersze macierzy częstości,
$x_k, y_k$ – liczba wystąpień $k$-tego słowa w dokumentach $X$, $Y$, elementy macierzy częstości na przecięciu $k$-tego wiersza i kolumn $X$ i $Y$.

Dużą zaletą znormalizowanych miar jest łatwość przekształcenia miary odległości na miarę podobieństwa. Wykorzystany do tego może zostać wzór (4). Miary znormalizowane są również łatwiejsze w interpretacji. Można je wyrazić jako wartość procentową, co bezpośrednio może zostać przełożone na stwierdzenie, że dokument $X$ jest podobny do dokumentu $Y$ w $100 \cdot s(X,Y)\%$.

$$s(X,Y) = 1 - d(X,Y), \tag{4}$$

gdzie $s(X,Y)$ – podobieństwo pomiędzy dokumentami $X$ i $Y$.

### 3.3. Wstępne przetwarzanie dokumentów

Przed przystąpieniem do symulacyjnej analizy przydatności wybranych miar podobieństwa tekstów należy poddać dokumenty wstępnemu przetwarzaniu. W tym celu zostały one połączone w korpus. Następnie wszystkie litery zostały zamienione na małe, usunięto interpunkcję i białe znaki oraz usunięto słowa znajdujące się na tzw. stopliście [Feinerer, Hornik, Meyer 2008]. Zarówno w badaniu symulacyjnym, jak i w przetwarzaniu wstępnym wykorzystano język R, a szczególnie pakiet **tm**.

### 3.4. Badanie symulacyjne

Na podstawie tak przygotowanych dokumentów utworzona została macierz częstości w dwóch wersjach: podstawowej i binarnej. Jak wskazują wcześniejsze badania i literatura [Borratyński 2009], dla miary kątowej lepsze wyniki daje wykorzystanie podstawowej wersji macierzy częstości, a dla miary Jaccarda wersji binarnej. Ponieważ teksty były zbliżonej długości, nie było konieczne dodatkowe ważenie macierzy częstości wagą uwzględniającą liczbę wyrazów w poszczególnych dokumentach.

Utworzona macierz ma 1110 wierszy (terminy) i 22 kolumny (dokumenty). Rzadkość macierzy wynosi 88%, co oznacza, że 88% wszystkich wartości w macierzy to 0.

Na tak przygotowanych macierzach częstości przeprowadzono dalsze badania.

## 4. Miary podobieństwa tekstów oparte na wiedzy dziedzinowej w postaci ontologii

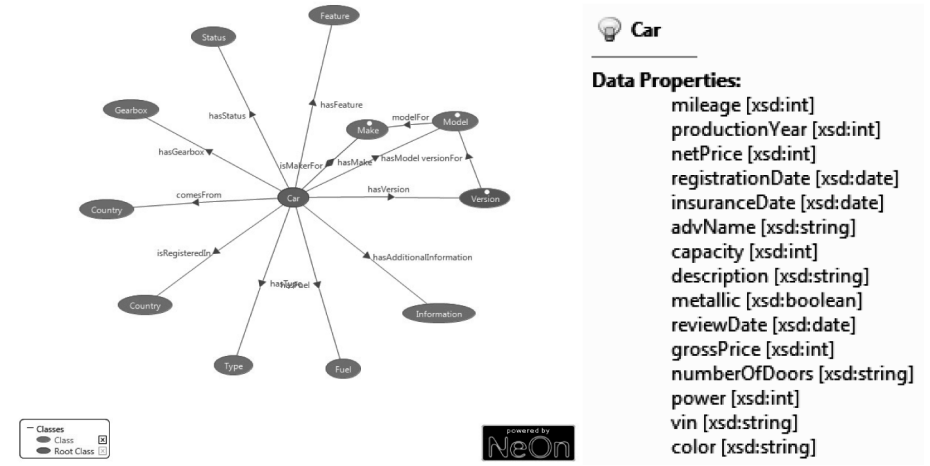
Ta część artykułu poświęcona została omówieniu analizy symulacyjnej miar podobieństwa tekstów opartych na wiedzy dziedzinowej. W pierwszej kolejności przedstawiona zostanie ontologia wykorzystana w badaniu. Następnie omówione zostanie ontologiczne podejście do analizy skupień wykorzystane w badaniu symulacyjnym.

### 4.1. Ontologia

Do reprezentacji wiedzy dziedzinowej wykorzystywane są ontologie. Ontologia w sensie informatycznym to formalna reprezentacja pewnej dziedziny wiedzy, na

którą składa się zapis zbiorów pojęć (*concept*) i relacji między nimi. Pojęcia mogą posiadać również właściwości w postaci atrybutów. Instancje to reprezentacje obiektów rzeczywistych w ontologii [Lula, Paliwoda-Pękosz 2008].

Na podstawie opcji dostępnych przy wyszukiwaniu ofert utworzona została ontologia zawierająca przykładową strukturę ogłoszenia. Jest ona zaprezentowana na rys. 2. Do utworzenia ontologii wykorzystywane były programy Protégé i NeOn Toolkit. Podstawową klasą jest klasa Adverts. Wszystkie pozostałe są jej podklasami. Zostały one połączone relacjami.




**Rys. 2.** Struktura ontologii Cars wykorzystanej w badaniach

Źródło: opracowanie własne w programie NeOn Toolkit.

Kolejnym krokiem po stworzeniu ontologii było zaimportowanie do niej danych dostępnych w postaci tekstu w ogłoszeniach. Do tego zadania wykorzystano program GATE (*General Architecture for Text Engineering*), czyli opensource'owe oprogramowanie służące do przetwarzania i analizy tekstu. Teksty uzyskane w wyniku ekstrakcji danych ze stron internetowych z ogłoszeniami połączono w korpus dokumentów. Następnie wczytano uprzednio utworzoną ontologię zawierającą jedynie definicje klas, relacji i atrybutów. Dokonano również tokenizacji dokumentów, czyli podziału na poszczególne słowa oraz spacje pomiędzy nimi.

W celu oznaczenia w tekście i zaimportowania z dokumentów tekstowych do ontologii danych wykorzystano język JAPE (*Java Annotation Patterns Engine*), pozwalający na wyszukiwanie w tekście wzorców zdefiniowanych na bazie mechanizmu wyrażeń regularnych. Rysunek 3 (z lewej strony) przedstawia przykładową ofertę z rys. 1 wczytaną do programu GATE. Na kolorowo oznaczono frazy pasujące do konkretnych reguł JAPE wymienionych po prawej stronie okna programu.

Część środkowa i położona po prawej stronie rys. 3 przedstawia przykładowe ogłoszenie opisane w ontologii. W środku znajduje się graf ilustrujący relacje pomię-


**Rys. 3.** Przykładowe ogłoszenie z oznaczonymi na kolorowo frazami spełniającymi konkretne reguły JAPE (po lewej) oraz opisane w ontologii

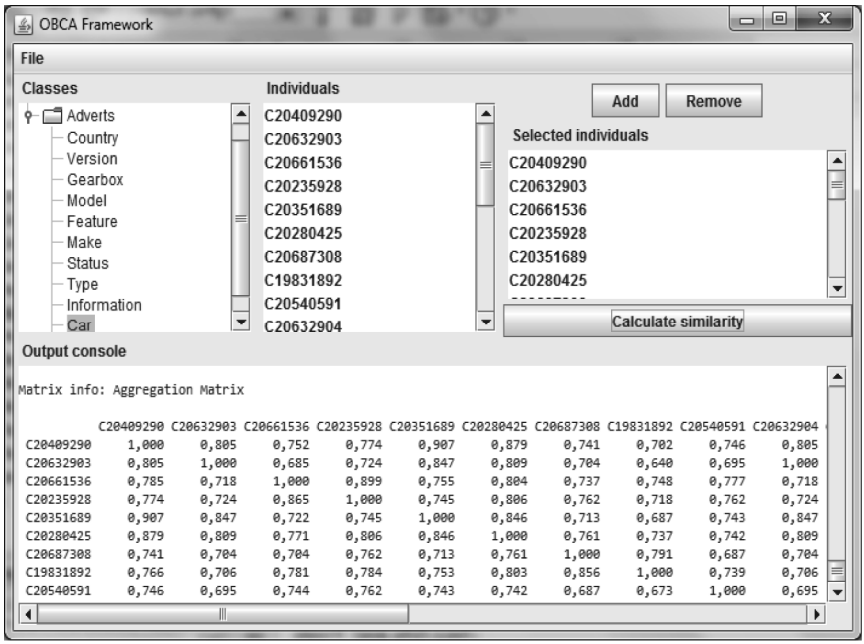
Źródło: opracowanie własne z wykorzystaniem programu GATE i NeOn Toolkit.

dzy ogłoszeniem (samochodem) a tymi jego elementami, które zostały ujęte w ontologii jako instancje klas. Po prawej stronie przedstawiono atrybuty ogłoszenia (samochodu) z konkretnymi ich wartościami.

Ontologię wypełnioną w ten sposób danymi wykorzystano w badaniu symulacyjnym.

### 4.2. Badanie symulacyjne

Do porównania ofert sprzedaży samochodów z wykorzystaniem ontologii wykorzystano system OBCA (*Ontology-Based Cluster Analysis*). OBCA jest to podejście do analizy skupień bazujące na ontologiach. Koncepcja systemu (OBCA Framework) została zaproponowana przez Lulę i Paliwodę-Pękosz [2008]. Jej implementacja została opisana w [Tuchowski i in. 2011].




**Rys. 4.** Okno aplikacji OBCA System

Źródło: opracowanie własne.

Aplikacja została napisana w Javie. Podczas implementacji wykorzystano dwa pakiety Javy. JENA służy do wczytania ontologii i odwoływania się do jej elementów (klasy, instancje, relacje i atrybuty). Natomiast SimPack zawiera interfejsy i klasy poszczególnych miar podobieństwa, które były wykorzystywane w badaniu. Rysunek 4 przedstawia interfejs graficzny aplikacji z wczytaną ontologią wykorzystaną w badaniu i instancjami wybranymi do porównania.

Na podobieństwo całkowite ($sim(I_i, I_j)$) każdych dwóch instancji opisanych w jednej ontologii ($I_i, I_j$) składają się trzy rodzaje podobieństwa: taksonomiczne ($TS(I_i, I_j)$), atrybutów ($AS(I_i, I_j)$) i relacyjne ($RS(I_i, I_j)$). Podobieństwa te są następnie agregowane do jednej wartości. Przedstawia to formuła (5).

$$sim(I_i, I_j) = f_{agr}\left(TS(I_i, I_j), RS(I_i, I_j), AS(I_i, I_j)\right). \tag{5}$$

W badaniach jako funkcja agregująca ($f_{agr}(...)$) wykorzystana została średnia ważona.

Podobieństwo taksonomiczne to podobieństwo wynikające z hierarchicznej zależności pomiędzy klasami. Liczone nie dla porównywanych instancji, ale dla klas, do których obiekty te należą. W badaniach wykorzystano miarę Wu-Palmer przedstawioną wzorem (6).

$$s(x, y) = 1 - \frac{2d\left(LPS(x, y)\right)}{d(x) + d(y)}, \tag{6}$$

gdzie: $d(x)$ – odległość węzła $x$ od korzenia,
$LPS(x, y)$ – najbliższy wspólny przodek dla węzłów $x$ i $y$.

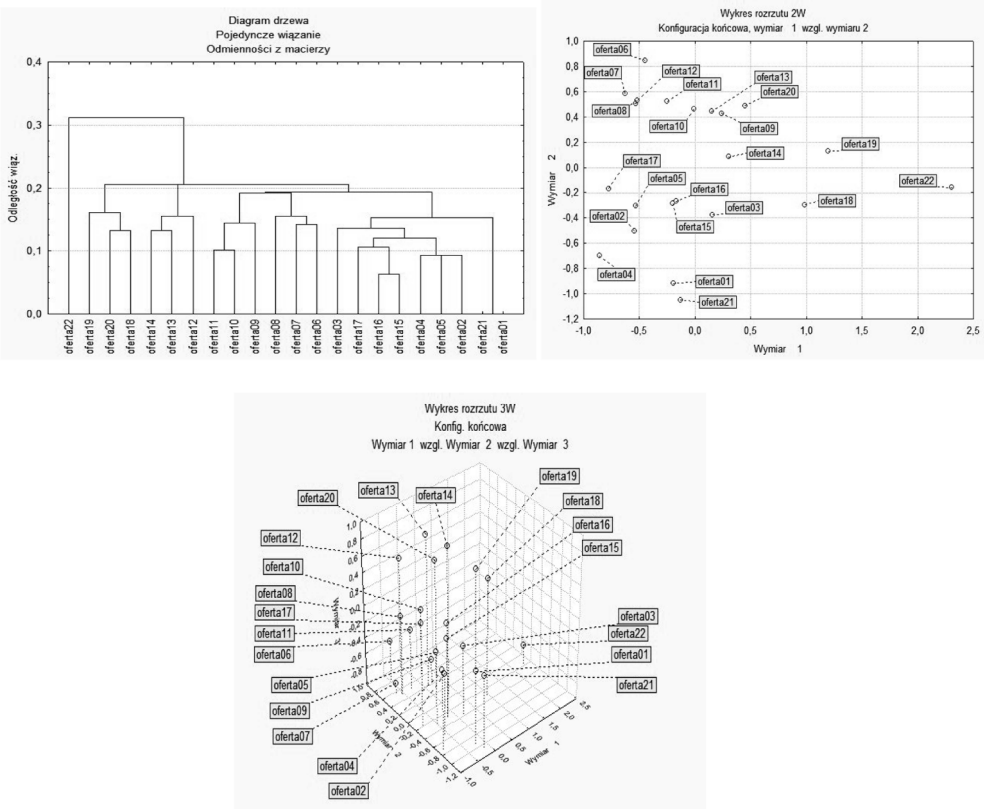
Podobieństwo atrybutów to zagregowane podobieństwo podobieństw liczonych dla każdego atrybutu osobno. Podobieństwa względem konkretnych atrybutów są liczone według różnych wzorów w zależności od typu atrybutu. W chwili obecnej system ma zaimplementowane miary dla łańcuchów znakowych (miara Levensteina – minimalna liczba operacji potrzebnych do zamiany jednego łańcucha tekstowego w drugi), wartości numerycznych (wzór (7)), dat (zamiana na liczbę całkowitą i korzystanie z miar dla tychże) oraz dla wartości logicznych (wzór (8)). Prowadzone są prace nad miarą kątową do dłuższych tekstów

$$s(a, b) = 1 - \frac{|a - b|}{|MAX - MIN|}, \tag{7}$$

$$s(a, b) = \begin{cases} 0 \text{ dla } a \neq b \\ 1 \text{ dla } a = b \end{cases}. \tag{8}$$

Podobieństwo relacyjne to zagregowane podobieństwo taksonomiczne i atrybutów instancji, które z badanymi instancjami wchodzą w relacje. Bazuje ono na podobieństwie relacji, jakie tworzą rozpatrywane obiekty z innymi obiektami.

Na podstawie powstałej w wyniku obliczeń zagregowanej macierzy podobieństwa przeprowadzono analizę skupień, której efekt można zobaczyć na rys. 5 po lewej stronie oraz skalowanie wielowymiarowe (w środku i po prawej stronie).




**Rys. 5.** Dendrogram przedstawiający wynik hierarchicznej analizy skupień badanych ofert (po lewej) oraz skalowanie wielowymiarowe macierzy podobieństwa do przestrzeni dwu- (w środku) i trójwymiarowej (po prawej)

Źródło: opracowanie własne z wykorzystaniem programu STATISTICA.

## 5. Podsumowanie

W artykule przedstawione zostały pokrótce badania symulacyjne dotyczące oceny przydatności wybranych miar podobieństwa tekstów. Porównując dokumenty, porównywano ich zawartość merytoryczną. W tym przypadku porównywano nie same oferty, ale oferowane w nich samochody.

Nie wszystkie oferty zawierały pełny zestaw cech. Brak pewnych danych wpływa na jakość pomiaru podobieństwa pomiędzy obiektami.

Miary bazujące na ontologii dały lepsze wyniki (bardziej zbliżone do oczekiwanych). W podejściu ontologicznym istnieje możliwość uwzględnienia typu danych (wartości liczbowe, logiczne itp.). Nie ma ograniczenia w postaci traktowania całości jako łańcucha tekstowego.

Rozwiązaniu opartemu na ontologii brak uniwersalności. W badaniach wykorzystano ontologię zbudowaną dla konkretnej dziedziny. Powstaje pytanie, czy nakłady potrzebne na stworzenie ontologii i zapisanie reguł pozwalających na automatyczne uzupełnienie jej danymi nie przewyższają korzyści, które można osiągnąć dzięki temu podejściu.

Innym poważnym problemem przy budowaniu ontologii jest brak standaryzacji w ich tworzeniu. Dla jednej dziedziny może powstać kilka zupełnie różnych ontologii. Struktura ontologii powstaje w wyniku subiektywnego postrzegania danej dziedziny przez twórcę ontologii.

## Literatura

Borratyński D., *Ocena przydatności częstotliwościowej reprezentacji dokumentów w języku polskim,* Rozprawa doktorska, Wydawnictwo UE, Kraków 2009.

Deza M.M., Deza E., *Encyclopedia of Distances,* Springer-Verlag Berlin, Heidelberg 2009.

Feinerer I., Hornik K., Meyer D., *Text Mining Infrastructure in R,* „Journal of Statistical Software", marzec 2008.

Lula P., Paliwoda-Pękosz G., *An Ontology-Based Cluster Analysis Framework.Proceedings of the First International Workshop on Ontology-Supported Business Intelligence,* ACM, New York, NY, USA 2008.

Tuchowski J., Wójcik K., Paliwoda-Pękosz G., Lula P., *OBCAS – Ontology based cluster analysis system,* Sopot 2011.

Wójcik K., *Analiza porównawcza miar podobieństwa tekstów,* [w:] Taksonomia 17, Wydawnictwo UE, Wrocław 2010.

## COMPARATIVE ANALYSIS OF TEXT DOCUMENTS SIMILARITY MEASURES BASED ON FREQUENCY MATRIX AND BASED ON DOMAIN KNOWLEDGE

**Summary:** The main objective of this paper is an attempt of evaluation of usefulness of similarity measures of text documents. Mostly known from literature are the ones based on frequency matrix and those based on domain knowledge represented by ontologies. Firstly the documents that were used in the research are presented. Secondly, chosen measures based on frequency matrix are shortly described. To summarize the first part the simulation analysis based on those measures is presented. Next part of the article is devoted to the results of a simulation analysis achieved when measures based on ontologies are used. On this basis an attempt of evaluation of usefulness of similarity measures of texts is made.

**Keywords:** text mining, similarity, measure, frequency matrix, ontology.