

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

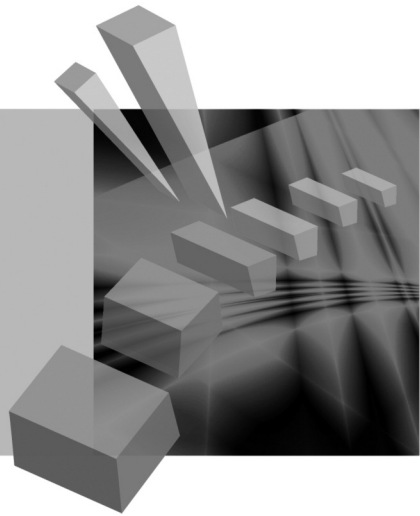
RESEARCH PAPERS

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 Pocięcha,
Miroslaw 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](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

Kopowanie i powielanie w jakiegokolwiek 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

Wstęp	13
Stanisława Bartosiewicz , Jeszcze raz o skutkach subiektywizmu w analizie wielowymiarowej	17
Andrzej Sokolowski , Q uniwersalna miara odległości	22
Eugeniusz Gatnar , Jakość danych w systemach statystycznych banków centralnych (na przykładzie NBP)	31
Marek Walesiak , Pomiar odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej – strategię postępowania.....	39
Krzysztof Jajuga, Marek Walesiak , XXV lat konferencji taksonomicznych – fakty i refleksje	47
Józef Pocięcha, Barbara Pawelek , Model SEM w analizie zagrożenia bankructwem przedsiębiorstw w świetle koniunktury gospodarczej – problemy teoretyczne i praktyczne	50
Paweł Lula , Uczące się systemy pozyskiwania informacji z dokumentów tekstowych	58
Ewa Roszkowska , Zastosowanie metody TOPSIS do wspomaganie procesu negocjacji.....	68
Andrzej Młodak , Sąsiedztwo obszarów przestrzennych w ujęciu fizycznym oraz społeczno-ekonomicznym – podejście taksonomiczne	76
Andrzej Bąk , Modele kategorii nieuporządkowanych w badaniach preferencji	86
Jacek Kowalewski , Zintegrowany model optymalizacji badań statystycznych.....	96
Jan Paradysz, Karolina Paradysz , Obszary bezrobocia w Polsce – problem benchmarkowy.....	106
Tomasz Szubert , W co grać, aby jak najmniej przegrać? Próba klasyfikacji systemów gry w zakładach bukmacherskich.....	116
Izabela Szamrej-Baran , Klasyfikacja krajów UE ze względu na ubóstwo energetyczne	126
Sylwia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , Analiza dojazdów do pracy za pomocą modelu grawitacji.....	135
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Minimum egzystencji a czynniki warunkujące skłonność do korzystania z pomocy społecznej. Klasyfikacja gospodarstw domowych	144
Hanna Dudek , Subiektywne skale ekwiwalentności – analiza na podstawie danych o satysfakcji z osiągniętych dochodów	153

Joanicjusz Nazarko, Ewa Chodakowska, Marta Jaročka , Segmentacja szkół wyższych metodą analizy skupień <i>versus</i> konkurencja technologiczna ustalona metodą DEA – studium komparatywne.....	163
Ewa Chodakowska , Wybrane metody klasyfikacji w konstrukcji ratingu szkół.....	173
Bartosz Soliński , Sektor energetyki odnawialnej w krajach Unii Europejskiej – klasyfikacja w świetle strategii zarządzania zmianą.....	182
Krzysztof Szwarz , Klasyfikacja powiatów województwa wielkopolskiego ze względu na sytuację demograficzną.....	192
Elżbieta Gołata, Grażyna Dehnel , Rejestry administracyjne w analizie przedsiębiorczości.....	202
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień , Wykorzystanie metod taksonomicznych w prognozowaniu wskaźników rentowności banków giełdowych w Polsce.....	212
Katarzyna Dębowska , Modelowanie upadłości przedsiębiorstw przy wykorzystaniu metod dyskryminacji i regresji.....	222
Alina Bojan , Wykorzystanie metod wielowymiarowej analizy danych do identyfikacji zmiennych wpływających na atrakcyjność wybranych inwestycji.....	231
Justyna Brzezińska , Analiza logarytmiczno-liniowa w badaniu przyczyn umieralności w krajach UE.....	240
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka , Analiza klas ukrytych w badaniach satysfakcji studentów.....	247
Bartłomiej Jefmański , Pomiar opinii respondentów z wykorzystaniem elementów teorii zbiorów rozmytych i środowiska R.....	256
Julita Stańczuk , Porównanie rezultatów wielostanowej klasyfikacji obiektów ekonomicznych z wykorzystaniem analizy dyskryminacyjnej oraz sieci neuronowych.....	265
Jerzy Krawczuk , Skuteczność metod klasyfikacji w prognozowaniu kierunku zmian indeksu giełdowego S&P500.....	275
Anna Czapkiewicz, Beata Basiura , Symulacyjne badanie wpływu zaburzeń na grupowanie szeregów czasowych na podstawie modelu Copula-GARCH.....	283
Radosław Pietrzyk , Ocena efektywności inwestycji funduszy inwestycyjnych z tytułu doboru papierów wartościowych i umiejętności wykorzystania trendów rynkowych.....	291
Aleksandra Witkowska, Marek Witkowski , Zastosowanie metody Panzara-Rosse’a do pomiaru poziomu konkurencji w sektorze banków spółdzielczych.....	306
Marcin Pelka , Podejście wielomodelowe z wykorzystaniem metody <i>boosting</i> w analizie danych symbolicznych.....	315
Justyna Wilk , Analiza porównawcza oprogramowania komputerowego w klasyfikacji danych symbolicznych.....	323

Tomasz Bartłomowicz, Justyna Wilk , Zastosowanie metod analizy danych symbolicznych w przeszukiwaniu dziedzinowych baz danych.....	333
Kamila Migdał-Najman , Propozycja hybrydowej metody grupowania opartej na sieciach samouczących	342
Dorota Rozmus , Porównanie dokładności taksonomii spektralnej oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i>	352
Krzysztof Najman , Grupowanie dynamiczne z wykorzystaniem samouczących się sieci GNG	361
Małgorzata Misztal , Wpływ wybranych metod uzupełniania brakujących danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych w przypadku zbiorów danych o niewielkiej liczebności – ocena symulacyjna	370
Mariusz Kubus , Zastosowanie wstępnego uwarunkowania zmiennej objaśnianej do selekcji zmiennych.....	380
Barbara Batóg, Jacek Batóg , Wykorzystanie analizy dyskryminacyjnej do identyfikacji czynników determinujących stopę zwrotu z inwestycji na rynku kapitałowym	387
Katarzyna Wójcik, Janusz Tuchowski , Analiza porównawcza miar podobieństwa tekstów opartych na macierzy częstości i tekstów opartych na wiedzy dziedzinowej	396
Iwona Staniec , Analiza czynnikowa w identyfikacji obszarów determinujących doskonalenie systemów zarządzania w polskich organizacjach	406
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawelczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Analiza porównawcza wybranych technik eksploracji danych do klasyfikacji danych medycznych z brakującymi obserwacjami	416
Iwona Foryś , Wykorzystanie analizy log-liniowej do wyboru czynników determinujących atrakcyjność cenową mieszkań w obrocie wtórnym na przykładzie lokalnego rynku mieszkaniowego.....	426
Ewa Genge , Analiza skupień oparta na mieszankach uciętych rozkładów normalnych.....	436
Jerzy Korzeniewski , Ocena efektywności metody uśredniania zmiennych i metody Ichino selekcji zmiennych w analizie skupień	444
Andrzej Dudek , SMS – propozycja nowego algorytmu analizy skupień	451
Artur Mikulec , Metody oceny wyniku grupowania w analizie skupień.....	460
Małgorzata Machowska-Szewczyk , Algorytm klasyfikacji rozmytej dla obiektów opisanych za pomocą zmiennych symbolicznych oraz rozmytych	469
Artur Zaborski , Analiza PROFIT i jej wykorzystanie w badaniu preferencji	479
Karolina Bartos , Analiza skupień wybranych państw ze względu na strukturę wydatków konsumpcyjnych obywateli – zastosowanie sieci Kohonena	488

Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Klasyfikacja gospodarstw domowych ze względu na bodźce do zawierania umowy o ubezpieczenie z wykorzystaniem modeli zmiennych jakościowych .	496
Izabela Kurzawa , Zastosowanie modelu LA/AIDS do badania elastyczności cenowych popytu konsumpcyjnego w gospodarstwach domowych w relacji miasto–wieś	505
Aleksandra Łuczak, Feliks Wysocki , Metody porządkowania liniowego obiektów opisanych za pomocą cech metrycznych i porządkowych	513
Agnieszka Sompolska-Rzechuła , Porównanie klasycznej i pozycyjnej taksonomicznej analizy zróżnicowania jakości życia w województwie zachodniopomorskim	523
Joanna Banaś, Małgorzata Machowska-Szewczyk , Ocena intensywności wykorzystania skrzynek poczty elektronicznej za pomocą uporządkowanego modelu probitowego	532
Iwona Bąk , Segmentacja gospodarstw domowych emerytów i rencistów pod względem wydatków na rekreację i kulturę	541
Aneta Becker , Zastosowanie metody ANP do porządkowania województw Polski pod względem dynamiki wykorzystania ICT w latach 2008-2010	552
Katarzyna Dębowska , Klasyfikacja sektorów ze względu na ich kondycję finansową przy użyciu metod wielowymiarowej analizy statystycznej	562
Anna Domagała , Propozycja metody doboru zmiennych do modeli DEA (procedura kombinowanego doboru w przód).....	571
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Analiza statystyczna w badaniach zapotrzebowania na usługi teleinformatyczne sieci łączności ruchomej	580
Hanna Gruchociak , Konstrukcja estymatora regresyjnego dla danych o strukturze dwupoziomowej.....	590
Tomasz Klimanek, Marcin Szymkowiak , Zastosowanie estymacji pośredniej uwzględniającej korelację przestrzenną w opisie niektórych charakterystyk rynku pracy	601
Jarosław Lira , Prognozowanie opłacalności produkcji żywca wieprzowego w Polsce	610
Christian Lis , Wykorzystanie metody klasyfikacji w ocenie konkurencyjności portów południowego Bałtyku	619
Beata Bieszk-Stolorz, Iwona Markowicz , Wykorzystanie wielomianowego modelu logitowego do oceny szansy podjęcia pracy przez bezrobotnych .	628
Lucyna Przezbórska-Skobiej, Jarosław Lira , Przestrzeń agroturystyczna Polski i ocena jej atrakcyjności.....	637
Paweł Ulman , Model rozkładu wydatków a funkcje popytu.....	646
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Zastosowanie metod analizy statystycznej w badaniach mięczaków	655

Summaries

Stanisława Bartosiewicz , The effects of subjectivism in multivariate analysis revisited.....	21
Andrzej Sokółowski , Q universal distance measure	30
Eugeniusz Gatnar , Data quality in central banks' statistical systems (NBP example)	38
Marek Walesiak , Distance measures for ordinal data – strategies of proceedings.....	46
Krzysztof Jajuga, Marek Walesiak , XXV years of taxonomic conferences – some facts and remarks.....	49
Józef Pocięcha, Barbara Pawelek , General SEM model in researching corporate bankruptcy and business cycles – theoretical and practical problems.....	57
Paweł Lula , Learning-based systems of information extraction from textual resources	67
Ewa Roszkowska , The application of the TOPSIS method to support the negotiation process	75
Andrzej Młodak , Neighborhood of spatial areas in the physical and socio-economic context – a taxonomic approach.....	85
Andrzej Bąk , Models for unordered categories in preference analysis.....	95
Kowalewski Jacek , An integrated model of optimizing statistical surveys	105
Jan Paradysz, Karolina Paradysz , Areas of unemployment in Poland – benchmark problem	115
Tomasz Szubert , How to play to lose the least? Classification of systems in sports bets	125
Izabela Szamrej-Baran , Classification of EU member states in view of fuel poverty	134
Sylvia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , An attempt to use the gravity model in the analysis of commuters.....	143
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Subsistence minimum versus factors influencing tendency to benefit from social care. Classification of households	152
Hanna Dudek , Subjective equivalence scales – analysis based on data about satisfaction with incomes.....	162
Joanicjusz Nazarko, Ewa Chodakowska, Marta Jarocka , Segmentation of universities using cluster analysis versus technological competitors determined by the DEA method – a comparative study	172
Ewa Chodakowska , Selected methods of classification in schools' rating.....	181
Bartosz Soliński , Renewable energy sector in the European Union – classification in the light of change management strategy	191
Krzysztof Szwarz , Classification of Wielkopolska voivodeship due to the demographic situation	201

Elżbieta Gołata, Grażyna Dehnel , Administrative registers in business analysis.....	211
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień , Application of taxonomic methods in forecasting the profitability ratios of listed banks in Poland.....	221
Katarzyna Dębowska , Modeling bankruptcy of firms by using discrimination and regression methods.....	230
Alina Bojan , Identification of variables which influence attractiveness of given investments with the usage of multivariate analysis.....	239
Justyna Brzezińska , Log-linear analysis in the study of mortality in EU.....	246
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka , Latent class analysis in student satisfaction surveys.....	254
Bartłomiej Jefmański , The respondent's opinions measurement in the R program with an application of fuzzy sets theory.....	264
Julita Stańczuk , A comparison of the results of multistate classification of economic objects using discriminant analysis and artificial neural networks.....	274
Jerzy Krawczuk , Effectiveness of classification methods in S&P500 stock index direction changes forecasting.....	282
Anna Czapkiewicz, Beata Basiura , The simulation study of the utility of the Copula-GARCH models for clustering financial time series.....	290
Radosław Pietrzyk , Timing and selectivity in mutual funds performance measurement.....	305
Aleksandra Witkowska, Marek Witkowski , Use of the Panzar-Rosse method to assess of the competition level in the cooperative banks sector.....	314
Marcin Pelka , Ensemble learning with the application of <i>boosting</i> in symbolic data analysis.....	322
Justyna Wilk , Comparative study of symbolic data classification software.....	332
Tomasz Bartłomowicz, Justyna Wilk , Application of symbolic data analysis methods for domain database searching.....	341
Kamila Migdał-Najman , A proposal of hybrid clustering method based on self-learning networks.....	351
Dorota Rozmus , Comparison of accuracy of spectral clustering and cluster ensembles stability based on bagging idea.....	360
Krzysztof Najman , A dynamic grouping based on self-learning GNG networks.....	369
Małgorzata Misztal , Influence of data imputation methods on the results of object classification using classification trees in the case of small data sets – simulation assessment.....	379
Mariusz Kubus , The application of pre-conditioning of explanatory variable for feature selection.....	386
Barbara Batóg, Jacek Batóg , Application of discriminant analysis to the identification of factors determining the rate of return on the capital market.....	395

Katarzyna Wójcik, Janusz Tuchowski , Comparative analysis of text documents similarity measures based on frequency matrix and based on domain knowledge.....	405
Iwona Staniec , Factor analysis in the identification of areas that determine the improvement of management systems in Polish organizations.....	415
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Comparative analysis of selected data mining approaches to the classification of medical data with missing values (covariates).....	425
Iwona Foryś , The log-linear analysis using to select the factors determining the attractiveness of the price of flats on the secondary market on the example of local housing market.....	435
Ewa Genge , Trimming approach to the mixtures of normal distributions.....	443
Jerzy Korzeniewski , Efficiency assessment of Ichino method and mean value method of selecting variables in cluster analysis.....	450
Andrzej Dudek , SMS – proposal of new clustering algorithm.....	459
Artur Mikulec , Evaluation methods for the grouping result in cluster analysis.....	468
Małgorzata Machowska-Szewczyk , Fuzzy clustering algorithm for objects described by symbolic or fuzzy variables.....	478
Artur Zaborski , PROFIT analysis and its using in the research of preferences.....	487
Karolina Bartos , Cluster analysis of selected countries due to the structure of their citizens' consumer expenditures – the use of Kohonen networks.....	495
Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Classification of households according to the impulses of concluding the insurance contract by means of qualitative variable models.....	504
Izabela Kurzawa , The application of LA/AIDS model to examine price elasticities of demand of households in the urban-rural relationship.....	512
Aleksandra Luczak, Feliks Wysocki , Linear ordering methods of objects described by a set of metric and ordinal characteristics.....	522
Agnieszka Sompolska-Rzechuła , The comparison of the classical and positional taxonomic analysis of the quality of life differentiation in Zachodniopomorskie voivodeship.....	531
Joanna Banaś, Małgorzata Machowska-Szewczyk , Evaluation of intensity of mailboxes using with the ordered probit model.....	540
Iwona Bąk , Segmentation of pensioners and annuitants households in terms of expenditures on recreation and culture.....	551
Aneta Becker , Application of ANP method to organize Polish voivodships in terms of dynamics of the use of ICT in 2008-2010.....	561
Katarzyna Dębowska , The classification of sectors' financial situation using the methods of multivariate statistical analysis.....	570

Anna Domagała , Proposal of a new method for variable selection in DEA models (combined forward stepwise selection method).....	579
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Statistical analysis in demand research of ICT services in mobile networks.....	589
Hanna Gruchociak , Construction of regression estimator for two-level data	600
Tomasz Klimanek, Marcin Szymkowiak , Application of spatial models in indirect estimation of some labor market characteristics	609
Jarosław Lira , Forecasting of hog livestock production profitability in Poland	618
Christian Lis , The utilization of taxonomic methods in the appraisal of competitiveness of south Baltic ports	627
Beata Bieszk-Stolorz, Iwona Markowicz , The application of the multinomial logit model in evaluating employment odds for the unemployed job seekers	636
Lucyna Przezbórska-Skobiej, Jarosław Lira , Agritourism space of Poland and its valuation.....	645
Paweł Ulman , Model of expenses distribution and demand functions.....	654
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Methods of statistical analysis in research of molluscs	663

Katarzyna Wójcik, Janusz Tuchowski

Uniwersytet Ekonomiczny w Krakowie

ANALIZA PORÓWNAWCZA MIAR PODOBIENSTWA 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 na 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.

The image shows a screenshot of a car listing on the Otomoto.pl website. The listing is for a Ford Mondeo GHIA. The left side of the screenshot shows the car's details, including its price (23 500 PLN / 5 171 EUR), type (kombi), and other specifications. The right side shows the extracted text content of the listing, which includes details such as the car's registration date, mileage, and engine type.

Opis oferty

Parametry techniczne

Cena (brutto): 23 500 PLN / 5 171 EUR do negocjacji, ..

Finanso:

Zx 433 atm - c ten samochód będzie Twoi

Kierownic, nie szukał tańszych OC i AC)

Typ: kombi

Wersja: 1803

Rok produkcji: 2006

Data rejestracji: 04/2006

Przebieg: 05/2012

Ubezpieczenie: 07/2012

Przebieg w km: 129 000 km

Skrzynia biegów: manualna

Moc: 130 KM (96 kW)

Pojemność skokowa: 1998 cm³

Rodzaj paliwa: olej napędowy (diesel)

Kolor: granatowy

oferujLot - Notatnik

plik: Edycja Format Widok Pomoc

NR ogłoszenia: C20632903
 Tytuł ogłoszenia: Ford Mondeo GHIA
 Marka: Ford
 Model: Mondeo
 Cena (brutto): 23 000 PLN / 5 292 EUR do negocjacji, ..
 Typ: kombi
 wersja: 1803
 Rok produkcji: 2006
 Data rejestracji: 04/2006
 Przebieg: 05/2012
 Ubezpieczenie: 07/2012
 Przebieg w km: 129 000 km
 Skrzynia biegów: manualna
 Moc: 130 KM (96 kW)
 Pojemność skokowa: 1998 cm³
 Rodzaj paliwa: olej napędowy (diesel)
 Liczba drzwi: 4/5
 status pojazdu sprawdzono: sprawdzony / zarejestrowany
 kraj pochodzenia: niemiecy
 kraj aktualnej rejestracji: Polska
 Dodatkowe wyposażenie: ABS, el. szyby, el. lusterka, klimatyzacja, alufelgi, system nawigacji
 Dodatkowe informacje: serwisowany w ASO, bezwypadkowy
 VIN - numer ident. pojazdu: WF0CWA08183B13
 opis pojazdu: sprzedam zadbane auto w dobrym stanie, wyposażenie zgodne ze specyfikacją oto

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

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{122} \\ a_{21} & a_{22} & \cdots & a_{222} \\ \vdots & \vdots & \ddots & \vdots \\ a_{11101} & a_{11102} & \cdots & a_{111022} \end{bmatrix} \begin{matrix} \text{dokumenty} \\ \\ \\ \end{matrix} \quad \text{wyrazy ,} \quad (1)$$

1110×22

gdzie: A – macierz częstości,

a_{ij} – liczba wystąpień i -tego ($i=1, \dots, 1110$) wyrazu w j -tym ($j=1, \dots, 22$) 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.

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

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

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 $100 \cdot s(X, Y)\%$.

$$s(X, Y) = 1 - d(X, Y), \quad (4)$$

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. stopniście [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 [Boratyński 2009], dla miary kątovej 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

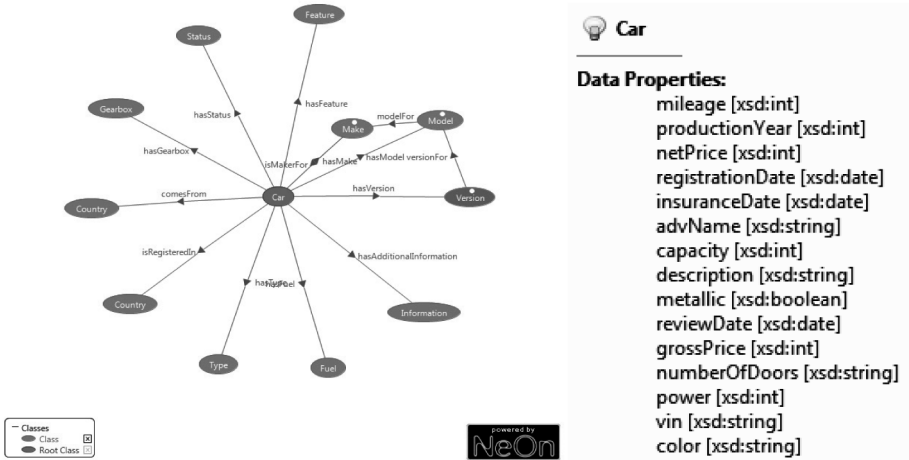
Ta część artykułu poświęcona została omówieniu analizy symulacyjnej miar podobieństwa tekstów opartych na wiedzy dziedzinowej. W pierwszej kolejności przedstawiona zostanie ontologia wykorzystana w badaniu. Następnie omówione zostanie ontologiczne podejście do analizy skupień wykorzystane w badaniu symulacyjnym.

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.



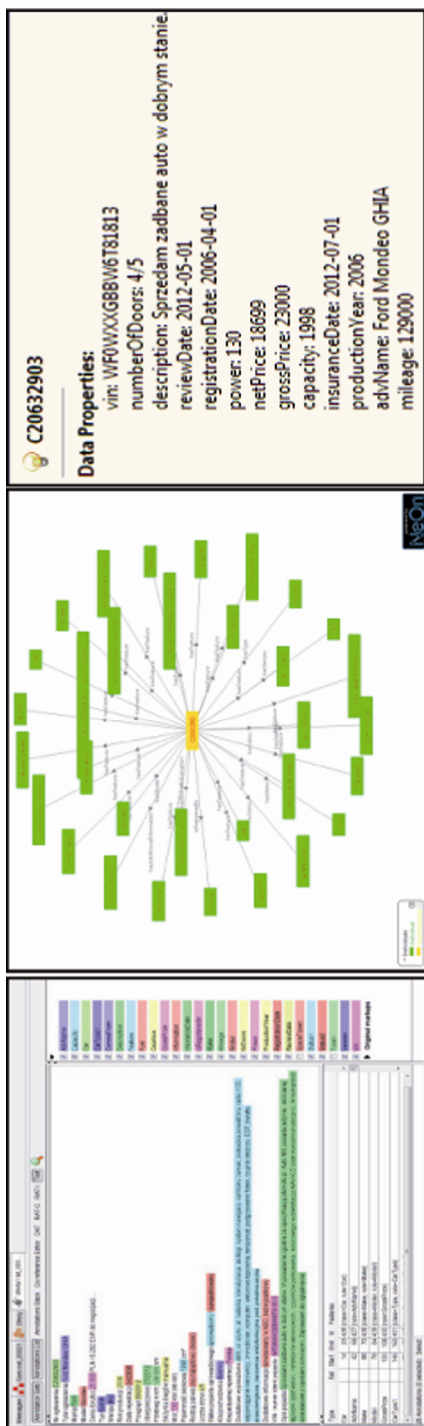
Rys. 2. Struktura ontologii Cars wykorzystanej w badaniach

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 pomiędzy



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

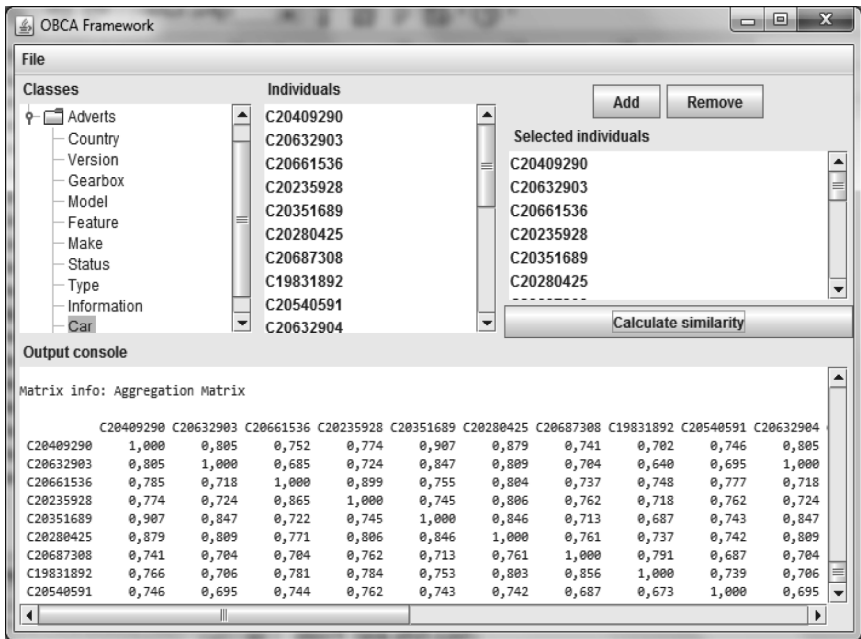
Źródło: opracowanie własne z wykorzystaniem programu GATE i NeOn Toolkit.

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 OBICA (*Ontology-Based Cluster Analysis*). OBICA jest to podejście do analizy skupień bazujące na ontologiach. Koncepcja systemu (OBICA Framework) została zaproponowana przez Lułę i Paliwodę-Pękosz [2008]. Jej implementacja została opisana w [Tuchowski i in. 2011].



Rys. 4. Okno aplikacji OBICA 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).

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

W badaniach jako funkcja agregująca ($f_{agr}(\dots)$) 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).

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

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

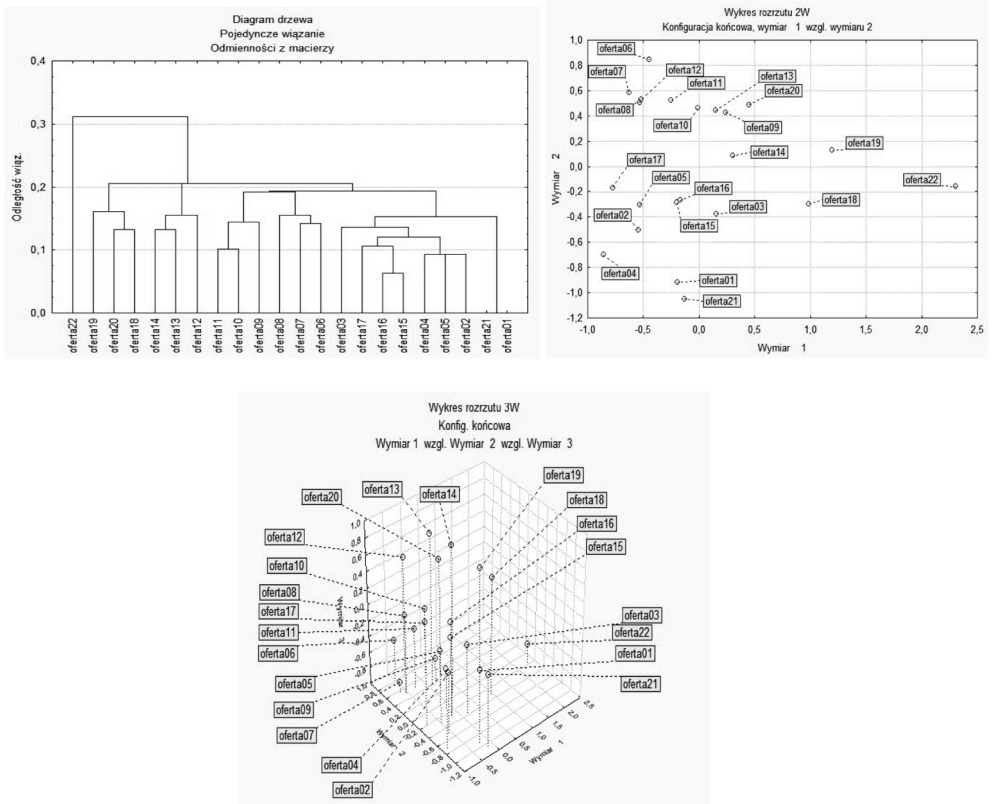
Podobieństwo atrybutów to zagregowane podobieństwo podobieństw liczonych dla każdego atrybutu osobno. Podobieństwa względem konkretnych atrybutów są liczone według różnych wzorów w zależności od typu atrybutu. W chwili obecnej system ma zaimplementowane miary dla łańcuchów znakowych (miara Levensteina – 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

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

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

Podobieństwo relacyjne to zagregowane podobieństwo taksonomiczne i atrybutów instancji, które z badanymi instancjami wchodzi 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

- Boratyński 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.