

# PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

# RESEARCH PAPERS

of Wrocław University of Economics

Nr 384

**Taksonomia 24**

**Klasyfikacja i analiza danych –  
teoria i zastosowania**

Redaktorzy naukowi

Krzysztof Jajuga

Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu  
Wrocław 2015

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Tytuł dofinansowany ze środków Narodowego Banku Polskiego  
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Informacje o naborze artykułów i zasadach recenzowania  
znajdują się na stronie internetowej Wydawnictwa  
[www.pracnaukowe.ue.wroc.pl](http://www.pracnaukowe.ue.wroc.pl)  
[www.wydawnictwo.ue.wroc.pl](http://www.wydawnictwo.ue.wroc.pl)

Publikacja udostępniona na licencji Creative Commons  
Uznanie autorstwa-Użycie niekomercyjne-Bez utworów zależnych 3.0 Polska  
(CC BY-NC-ND 3.0 PL)



© Copyright by Uniwersytet Ekonomiczny we Wrocławiu  
Wrocław 2015

**ISSN 1899-3192** (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)  
**e-ISSN 2392-0041** (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)  
**ISSN 1505-9332** (Taksonomia)

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:  
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu  
tel./fax 71 36 80 602; e-mail:[econbook@ue.wroc.pl](mailto:econbook@ue.wroc.pl)  
[www.ksiegarnia.ue.wroc.pl](http://www.ksiegarnia.ue.wroc.pl)

Druk i oprawa: TOTEM

## Spis treści

Wstęp.....	9
<b>Krzysztof Jajuga, Józef Pociecha, Marek Walesiak:</b> 25 lat SKAD.....	15
<b>Beata Basiura, Anna Czapkiewicz:</b> Symulacyjne badanie wykorzystania entropii do badania jakości klasyfikacji.....	25
<b>Andrzej Bąk:</b> Zagadnienie wyboru optymalnej procedury porządkowania liniowego w pakiecie <code>pllord</code> .....	33
<b>Justyna Brzezińska:</b> Analiza klas ukrytych w badaniach sondażowych.....	42
<b>Grażyna Dehnel:</b> Rejestr podatkowy oraz rejestr ZUS jako źródło informacji dodatkowej dla statystyki gospodarczej – możliwości i ograniczenia ..	51
<b>Sabina Denkowska:</b> Wybrane metody oceny jakości dopasowania w <i>Propensity Score Matching</i> .....	60
<b>Marta Dziechciarz-Duda, Klaudia Przybysz:</b> Zastosowanie teorii zbiorów rozmytych do identyfikacji pozafiskalnych czynników ubóstwa.....	75
<b>Iwona Foryś:</b> Potencjał rynku mieszkaniowego w Polsce w latach dekonjunktury gospodarczej.....	84
<b>Eugeniusz Gatnar:</b> Statystyczna analiza konwergencji krajów Europy Środkowej i Wschodniej po 10 latach członkostwa w Unii Europejskiej.....	93
<b>Ewa Genge:</b> Zaufanie do instytucji publicznych i finansowych w polskim społeczeństwie – analiza empiryczna z wykorzystaniem ukrytych modeli Markowa.....	100
<b>Alicja Grześkowiak:</b> Wielowymiarowa analiza uwarunkowań zaangażowania Polaków w kształcenie ustawiczne o charakterze pozaformalnym.....	108
<b>Monika Hamerska:</b> Wykorzystanie metod porządkowania liniowego do tworzenia rankingu jednostek naukowych.....	117
<b>Bartłomiej Jefmański:</b> Zastosowanie modeli IRT w konstrukcji rozmytego systemu wag dla zmiennych w zagadnieniu porządkowania liniowego – na przykładzie metody TOPSIS.....	126
<b>Tomasz Józefowski, Marcin Szymkowiak:</b> Wykorzystanie uogólnionej miary odległości do porządkowania liniowego powiatów województwa podkarpackiego w świetle funkcjonowania specjalnej strefy ekonomicznej Euro-Park Mielec.....	135
<b>Krzysztof Kompa:</b> Zastosowanie testów parametrycznych i nieparametrycznych do oceny sytuacji na światowym rynku kapitałowym przed kryzysem i po jego wystąpieniu.....	144
<b>Mariusz Kubus:</b> Rekurencyjna eliminacja cech w metodach dyskryminacji....	154

<b>Marta Kuc:</b> Wpływ sposobu definiowania macierzy wag przestrzennych na wynik porządkowania liniowego państw Unii Europejskiej pod względem poziomu życia ludności .....	163
<b>Paweł Lula:</b> Kontekstowy pomiar podobieństwa semantycznego .....	171
<b>Iwona Markowicz:</b> Model regresji Feldsteina-Horioki – wyniki badań dla Polski .....	182
<b>Kamila Migdał-Najman:</b> Ocena wpływu wartości stałej Minkowskiego na możliwość identyfikacji struktury grupowej danych o wysokim wymiarze .....	191
<b>Małgorzata Misztal:</b> O zastosowaniu kanonicznej analizy korespondencji w badaniach ekonomicznych.....	200
<b>Krzysztof Najman:</b> Zastosowanie przetwarzania równoległego w analizie skupień .....	209
<b>Edward Nowak:</b> Klasyfikacja danych a rachunkowość. Rozważania o relacjach .....	218
<b>Marcin Pelka:</b> Adaptacja metody <i>bagging</i> z zastosowaniem klasyfikacji pojęciowej danych symbolicznych.....	227
<b>Józef Pocięcha, Mateusz Baryła, Barbara Pawelek:</b> Porównanie skuteczności klasyfikacyjnej wybranych metod prognozowania bankructwa przedsiębiorstw przy losowym i nielosowym doborze prób .....	236
<b>Agnieszka Przedborska, Małgorzata Misztal:</b> Wybrane metody statystyki wielowymiarowej w ocenie jakości życia słuchaczy uniwersytetu trzeciego wieku .....	246
<b>Wojciech Roszka:</b> Konstrukcja syntetycznych zbiorów danych na potrzeby estymacji dla małych domen .....	254
<b>Aneta Rybicka:</b> Połączenie danych o preferencjach ujawnionych i wyrażonych .....	262
<b>Elżbieta Sobczak:</b> Poziom specjalizacji w sektorach intensywności technologicznej a efekty zmian liczby pracujących w województwach Polski ....	271
<b>Andrzej Sokołowski, Grzegorz Harańczyk:</b> Modyfikacja wykresu radarowego .....	280
<b>Marcin Szymkowiak, Marek Witkowski:</b> Wykorzystanie mediany do klasyfikacji banków spółdzielczych według stanu ich kondycji finansowej ..	287
<b>Justyna Wilk, Michał B. Pietrzak, Roger S. Bivand, Tomasz Kossowski:</b> Wpływ wyboru metody klasyfikacji na identyfikację zależności przestrzennych – zastosowanie testu <i>join-count</i> .....	296
<b>Dorota Witkowska:</b> Wykorzystanie drzew klasyfikacyjnych do analizy zróżnicowania płac w Niemczech .....	305
<b>Artur Zaborski:</b> Analiza niesymetrycznych danych preferencji z wykorzystaniem modelu punktu dominującego i modelu grawitacji.....	315

## Summaries

<b>Krzysztof Jajuga, Józef Pocięcha, Marek Walesiak:</b> XXV years of SKAD	24
<b>Beata Basiura, Anna Czapkiewicz:</b> Simulation study of the use of entropy to validation of clustering.....	32
<b>Andrzej Bąk:</b> Problem of choosing the optimal linear ordering procedure in the p_llord package.....	41
<b>Justyna Brzezińska-Grabowska:</b> Latent class analysis in survey research...	50
<b>Grażyna Dehnel:</b> Tax register and social security register as a source of additional information for business statistics – possibilities and limitations.....	59
<b>Sabina Denkowska:</b> Selected methods of assessing the quality of matching in Propensity Score Matching .....	74
<b>Marta Dziechciarz-Duda, Klaudia Przybysz:</b> Applying the fuzzy set theory to identify the non-monetary factors of poverty.....	83
<b>Iwona Foryś:</b> The potential of the housing market in Poland in the years of economic recessions.....	92
<b>Eugeniusz Gatnar:</b> Statistical analysis of the convergence of CEE countries after 10 years of their membership in the European Union.....	99
<b>Ewa Genge:</b> Trust to the public and financial institutions in the Polish society – an application of latent Markov models.....	107
<b>Alicja Grześkowiak:</b> Multivariate analysis of the determinants of Poles' involvement in non-formal lifelong learning .....	116
<b>Monika Hamerska:</b> The use of the methods of linear ordering for the creating of scientific units ranking.....	125
<b>Bartłomiej Jefmański:</b> The application of IRT models in the construction of a fuzzy system of weights for variables in the issue of linear ordering – on the basis of TOPSIS method .....	134
<b>Tomasz Józefowski, Marcin Szymkowiak:</b> GDM as a method of finding a linear ordering of districts of Podkarpackie Voivodeship in the light of the operation of the Euro-Park Mielec special economic zone .....	143
<b>Krzysztof Kompa:</b> Application of parametric and nonparametric tests to the evaluation of the situation on the world financial market in the pre- and post-crisis period.....	153
<b>Mariusz Kubus:</b> Recursive feature elimination in discrimination methods ...	162
<b>Marta Kuc:</b> The impact of the spatial weights matrix on the final shape of the European Union countries ranking due to the standard of living.....	170
<b>Paweł Lula:</b> The impact of context on semantic similarity.....	181
<b>Iwona Markowicz:</b> Feldstein-Horioka regression model – the results for Poland.....	190

<b>Kamila Migdal-Najman:</b> The assessment of impact value of Minkowski's constant for the possibility of group structure identification in high dimensional data.....	199
<b>Małgorzata Misztal:</b> On the use of canonical correspondence analysis in economic research.....	208
<b>Krzysztof Najman:</b> The application of the parallel computing in cluster analysis.....	217
<b>Edward Nowak:</b> Data classification and accounting. A study of correlations	226
<b>Marcin Pelka:</b> The adaptation of bagging with the application of conceptual clustering of symbolic data.....	235
<b>Józef Pociecha, Mateusz Baryła, Barbara Pawelek:</b> Comparison of classification accuracy of selected bankruptcy prediction methods in the case of random and non-random sampling technique.....	244
<b>Agnieszka Przedborska, Małgorzata Misztal:</b> Selected multivariate statistical analysis methods in the evaluation of the quality of life of the members of the University of the Third Age.....	253
<b>Wojciech Roszka:</b> Construction of synthetic data sets for small area estimation.....	261
<b>Aneta Rybicka:</b> Combining revealed and stated preference data.....	270
<b>Elżbieta Sobczak:</b> Specialization in sectors of technical advancement vs. effects of workforce number changes in Poland's voivodships.....	279
<b>Andrzej Sokółowski, Grzegorz Harańczyk:</b> Modification of radar plot.....	286
<b>Marcin Szymkowiak, Marek Witkowski:</b> Classification of cooperative banks according to their financial situation using the median.....	295
<b>Justyna Wilk, Michał B. Pietrzak, Roger S. Bivand, Tomasz Kossowski:</b> The influence of classification method selection on the identification of spatial dependence – an application of join-count test.....	304
<b>Dorota Witkowska:</b> Application of classification trees to analyze wages disparities in Germany.....	314
<b>Artur Zaborski:</b> Asymmetric preference data analysis by using the dominance point model and the gravity model.....	323

## **Paweł Lula**

Uniwersytet Ekonomiczny w Krakowie

e-mail: pawel.lula@uek.krakow.pl

---

# **KONTEKSTOWY POMIAR PODOBIENSTWA SEMANTYCZNEGO**

---

**Streszczenie:** Miary podobieństwa semantycznego pozwalają wyrazić w sposób ilościowy stopień zgodności znaczenia przypisywanego konceptom występującym w ontologii opisującej rozpatrywany obszar wiedzy. W pracy założono, że podobieństwo semantyczne ma charakter kontekstowy, co oznacza, że zgodność znaczeniowa pojęć jest uzależniona od zakresu tematycznego obszaru, na tle którego dokonywane jest porównanie. Celem niniejszej pracy jest analiza podobieństwa semantycznego w sytuacji, w której zmianie ulega kontekst analizy. Pracę rozpoczyna prezentacja najpopularniejszych miar podobieństwa semantycznego. Następnie pokazano, w jaki sposób przy obliczaniu podobieństwa można uwzględnić informację dotyczącą kontekstu. W kolejnej części pracy przedstawiono zależność zachodzącą pomiędzy zmianą kontekstu a wartością podobieństwa semantycznego.

**Słowa kluczowe:** podobieństwo semantyczne, ontologiczna reprezentacja wiedzy, analiza dokumentów tekstowych.

DOI: 10.15611/pn.2015.384.18

## **1. Wstęp**

Praca poświęcona jest zagadnieniu wyznaczania podobieństwa semantycznego pomiędzy konceptami wchodzącymi w skład ontologii. Zagadnienie to wielokrotnie pojawiają się jako zasadniczy temat prac o charakterze teoretycznym i aplikacyjnym. Za szczególnie interesujący obszar zastosowań tego typu miar należy zaliczyć automatyczną analizę dokumentów tekstowych, w której na podstawie ontologii wyznaczyć można podobieństwo pomiędzy pojęciami występującymi w tekstach, a następnie – po odpowiednim zagregowaniu miar podobieństwa pomiędzy konceptami – można wyznaczyć miarę podobieństwa semantycznego pomiędzy dokumentami.

W niniejszym artykule przyjęto, że wiedza opisywana jest za pomocą ontologii uwzględniającej jedynie zależności hierarchiczne pomiędzy wyróżnionymi koncep-

tami. Natomiast kontekst rozumiany będzie jako poddrzewo wyróżnione w przyjętej ontologii. Brak jawnego zdefiniowania kontekstu powoduje, że jego rolę odgrywa cała ontologia opisująca rozpatrywany fragment wiedzy.

## 2. Pomiar podobieństwa semantycznego

Rozpatrywane miary podobieństwa semantycznego pomiędzy konceptami podzielić można na dwie grupy:

1) miary podobieństwa oparte na długości ścieżki w ontologii. Do tej grupy zaliczymy miary:

- a) Rady,
- b) Wu i Palmera,
- c) Leacocka i Chodorowa;

2) miary podobieństwa semantycznego oparte na teorii informacji. Do tej grupy należą miary:

- a) Resnika,
- b) Jiang i Conratha,
- c) Lina.

### 2.1. Miary podobieństwa semantycznego oparte na długości ścieżki łączącej koncepty

Rada, Mili, Bicknell i Blettner w [1989] proponują w charakterze miary odległości pomiędzy konceptami wykorzystać długość najkrótszej ścieżki pomiędzy nimi:

$$dist_{RMBB}(c_1, c_2) = len(c_1, c_2). \quad (1)$$

W celu unormowania przedstawionej miary odległości do przedziału  $[0; 1]$  można zastosować wzór:

$$dist_{RMBB}^{NORM}(c_1, c_2) = \frac{len(c_1, c_2)}{2 \times D}, \quad (2)$$

gdzie  $D$  jest maksymalną głębokością w drzewie.

Unormowaną miarę odległości można przekształcić do miary podobieństwa:

$$sim_{RMBB}(c_1, c_2) = 1 - \frac{len(c_1, c_2)}{2 \times D}. \quad (3)$$

Natomiast Wu i Palmer [1994] proponują wyznaczanie podobieństwa za pomocą wzoru:

$$sim_{WP}(c_1, c_2) = \frac{2 \times d_{LCS}(c_1, c_2)}{d_{c_1} + d_{c_2}}, \quad (4)$$

gdzie  $d_{c_1}$  i  $d_{c_2}$  są odpowiednio głębokościami konceptów  $c_1$  i  $c_2$ , zaś  $d_{LCS}(c_1, c_2)$  jest głębokością najbliższego wspólnego przodka konceptów  $c_1$  i  $c_2$ . Tak wyznaczona



miara ma charakter unormowany do przedziału  $[0; 1]$ . Przyjmuje ona wartość zero wówczas, gdy najbliższy wspólny przodek jest korzeniem drzewa opisującego ontologię.

Z kolei Leacock i Chodorow w pracy [ ] proponują pomiar podobieństwa poprzez zastosowanie formuły:

$$sim_{LC}(c_1, c_2) = -\log \frac{len(c_1, c_2)}{2 \times D}, \quad (5)$$

gdzie  $D$  jest maksymalną głębokością w przyjętej ontologii. W celu unormowania miary należy zastosować wzór:

$$sim_{LC}^{NORM}(c_1, c_2) = -\frac{1}{\log(2 \times D)} \log \frac{len(c_1, c_2)}{2 \times D}. \quad (6)$$

Warto zauważyć, że formuły podane za pomocą wzorów (5) i (6) nie mogą zostać wyznaczone w przypadku, gdy  $c_1 = c_2$ .

## 2.2. Miary podobieństwa semantycznego oparte na teorii informacji

Miara podobieństwa pomiędzy konceptami zaproponowana w pracy [Resnik 1995] wyznaczana jest jako:

$$sim_{RESNIK}(c_1, c_2) = -\log \left( P(LCS(c_1, c_2)) \right). \quad (7)$$

Z kolei Jiang i Conrath w [Jiang & Conrath 1997] proponują pomiar odległości semantycznej jako:

$$dist_{JC}(c_1, c_2) = 2 \times \log \left( P(LCS(c_1, c_2)) \right) - (\log(P(c_1)) + \log(P(c_2))), \quad (8)$$

co można przekształcić na miarę podobieństwa:

$$sim_{JC}(c_1, c_2) = \frac{1}{2 \times \log(P(LCS(c_1, c_2))) - (\log(P(c_1)) + \log(P(c_2)))}. \quad (9)$$

Stosowanie formuły (9) jest niemożliwe, jeśli  $c_1 = c_2$ . Dlatego też lepszym rozwiązaniem może być:

$$sim_{JC}(c_1, c_2) = \frac{1}{2 \times \log(P(LCS(c_1, c_2))) - (\log(P(c_1)) + \log(P(c_2))) + 1}. \quad (10)$$

Natomiast Lin w pracy [Lin 1998] zaproponował pomiar podobieństwa pomiędzy konceptami poprzez zastosowanie formuły:

$$sim_{LIN}(c_1, c_2) = \frac{2 \times \log(P(LCS(c_1, c_2)))}{\log(P(c_1)) + \log(P(c_2))}. \quad (11)$$

### 2.3. Pomiar podobieństwa semantycznego pomiędzy dokumentami

Rozpatrując zagadnienie podobieństwa semantycznego pomiędzy dokumentami, przyjmijmy, że rozważania będą dotyczyły dokumentów  $D_1$  i  $D_2$  traktowanych jako zbiory konceptów zaczerpniętych z przyjętej ontologii:

$$D_1 = \{c_1^1, c_2^1, \dots, c_m^1\}$$

$$D_2 = \{c_1^2, c_2^2, \dots, c_n^2\}.$$

Obliczenia prowadzące do określenia podobieństwa pomiędzy dokumentami zdefiniowanymi w taki sposób opierają się na wartościach podobieństw pomiędzy parami konceptów  $(c_i^1, c_j^2)$ . Macierz ta przyjmuje postać:

$$\begin{array}{c} c_1^2 \quad c_2^2 \quad \dots \quad c_n^2 \\ \begin{array}{c} c_1^1 \\ c_2^1 \\ \vdots \\ c_m^1 \end{array} \left[ \begin{array}{cccc} s_{11} & s_{12} & \dots & s_{1n} \\ s_{21} & s_{22} & \dots & s_{2n} \\ \dots & \dots & \dots & \dots \\ s_{m1} & s_{m2} & \dots & s_{mn} \end{array} \right], \end{array}$$

gdzie  $s_{ij}$  jest podobieństwem pomiędzy konceptami  $c_i^1$  i  $c_j^2$ .

W pracy [Rada et al. 1989] podobieństwo pomiędzy zbiorami konceptów (dokumentami) określone zostało jako wartość średnia z podobieństw pomiędzy każdą parą konceptów reprezentujących każdy ze zbiorów.

Natomiast autorzy pracy [Wan & Peng 2005] starali się rozwiązać zagadnienie określenia podobieństwa pomiędzy dokumentami poprzez zastosowanie algorytmu optymalnego przyporządkowania konceptów z obu dokumentów.

Jeszcze inną propozycję przedstawiono w pracy [Tuchowski i in. 2011], gdzie do wyznaczenia podobieństwa semantycznego pomiędzy zbiorami konceptów stosuje się uśrednioną wartość z podobieństw pomiędzy każdym z konceptów a najbliższym mu znaczeniowo konceptem z drugiego ze zbiorów.

### 3. Uwzględnienie kontekstu w wyznaczaniu podobieństwa semantycznego

W przypadku miary  $dist_{RMBB}^{NORM}$  uwzględnienie kontekstu  $K$  prowadzi do formuły:

$$dist_{RMBB}^{NORM}(c_1, c_2 | K) = \frac{len(c_1, c_2)}{2 \times D_K}, \quad (12)$$

gdzie  $D_K$  jest maksymalną głębokością w poddrzewie reprezentującym kontekst.

Miara podobieństwa Wu i Palmera po uwzględnieniu kontekstu przyjmuje postać:

$$sim_{WP}(c_1, c_2|K) = \frac{2 \times d_{LCS(c_1, c_2)}^K}{d_{c_1}^K + d_{c_2}^K}, \quad (13)$$

w której głębokości elementów liczone są zawsze względem poddrzewa reprezentującego kontekst.

Natomiast miara  $sim_{LC}$  po unormowaniu i uwzględnieniu kontekstu wyraża się formułą:

$$sim_{LC}(c_1, c_2|K) = -\frac{1}{\log(2 \times D_K)} \times \log \frac{len(c_1, c_2)}{2 \times D_K}. \quad (14)$$

Miara podobieństwa Resnika mieści się w przedziale od 0 do  $\log\left(\frac{N}{min}\right)$ , gdzie  $N$  jest liczbą konceptów zidentyfikowanych w korpusie, zaś  $min$  jest liczbą wystąpień najrzadziej występującego konceptu. Stosując ją w odniesieniu do korpusu definiującego kontekst i jednocześnie dążąc do jej unormowania, należy zastosować formułę:

$$sim_{RESNIK}(c_1, c_2|K) = -\frac{1}{\log\left(\frac{N}{min}\right)} \log\left(P_K(LCS(c_1, c_2))\right). \quad (15)$$

Warto zauważyć, że podobieństwo liczone według formuły zaproponowanej przez Resnika uwzględnia jedynie wspólne cechy konceptów, natomiast nie bierze pod uwagę elementów, które je różnicują.

Podobieństwo semantyczne liczone według podejścia, które zaproponowali Jiang i Conrath, może zostać policzone dla wskazanego kontekstu według oryginalnej formuły (przy uwzględnieniu prawdopodobieństw wyznaczonych w odniesieniu do kontekstu):

$$sim_{JC}(c_1, c_2|K) = \frac{1}{2 \times \log(P_K(LCS(c_1, c_2))) - (\log(P_K(c_1)) + \log(P_K(c_2))) + 1}. \quad (16)$$

Również metoda Lina może zostać zastosowana do realizacji obliczeń w odniesieniu do kontekstu:

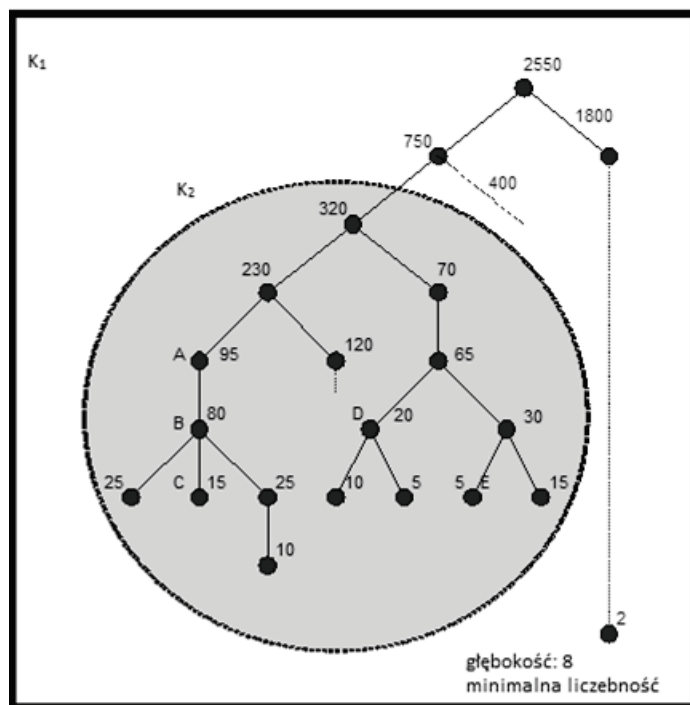
$$sim_{LIN}(c_1, c_2|K) = \frac{2 \times \log(P_K(LCS(c_1, c_2)))}{\log(P_K(c_1)) + \log(P_K(c_2))}. \quad (17)$$

#### 4. Wpływ zmiany kontekstu na podobieństwo

Głównym celem rozważań przedstawionych w niniejszym punkcie jest analiza kształtowania się podobieństwa semantycznego w zależności od zmiany kontekstu.

Opisaną powyżej sytuację opisuje rys. 1. Przedstawia on fragment drzewa opisyującego ontologię. Cała ontologia tworzy kontekst  $K_1$ . Rozważana będzie sytuacja przejścia do kontekstu  $K_2$  (ukazanego w całości na rysunku). Liczby widoczne

przy poszczególnych węzłach określają liczbę wystąpień każdego z konceptów w analizowanym korpusie. Pięciu węzłom przypisano oznaczenia literowe (A, B, C, D, E). Pomiędzy nimi wyznaczone będą miary podobieństwa semantycznego.



Rys. 1. Przykładowa ontologia

Źródło: opracowanie własne.

W pierwszym etapie prac analizie poddano wersję unormowaną  $dist_{RMBB}^{NORM}$  zdefiniowaną na pomocą wzoru (12). Po przejściu od kontekstu  $K_1$  do kontekstu  $K_2$  odległość zmienia się zgodnie z formułą:

$$\frac{dist_{RMBB}^{NORM}(c_1, c_2 | K_2)}{dist_{RMBB}^{NORM}(c_1, c_2 | K_1)} = \frac{D_{K_1}}{D_{K_2}}. \quad (18)$$

Wzór (18) wskazuje, że stosunek odległości wyrażonej w kontekście  $K_2$  i  $K_1$  jest równy stosunkowi głębokości poddrzew odpowiadających tym kontekstom. Obliczenia dla danych opisanych w rozpatrywanej ontologii przedstawione zostały w tab. 1.

Przedstawione obliczenia potwierdzają proporcjonalną zmianę odległości w wyniku zmiany kontekstu. Warto zauważyć, że po przekształceniu omawianej miary odległości do miary podobieństwa (wzór (3)) relacja podobieństwa pomiędzy znaczeniem konceptów nie zmienia się już w sposób proporcjonalny.

**Tabela 1.** Zmiana odległości semantycznej pomiędzy konceptami w wyniku zmiany kontekstu (metoda rady)

		A	B	C	D	E
A	$K_1$	0,0000	0,0625	0,1250	0,3125	0,3750
	$K_2$	0,0000	0,1000	0,2000	0,5000	0,6000
	$K_2/K_1$	-	1,6000	1,6000	1,6000	1,6000
B	$K_1$		0,0000	0,0625	0,3750	0,4375
	$K_2$		0,0000	0,1000	0,6000	0,7000
	$K_2/K_1$		-	1,6000	1,6000	1,6000
C	$K_1$			0,0000	0,4375	0,5000
	$K_2$			0,0000	0,7000	0,8000
	$K_2/K_1$			-	1,6000	1,6000
D	$K_1$				0,0000	0,1875
	$K_2$				0,0000	0,3000
	$K_2/K_1$				-	1,6000
E	$K_1$					0,0000
	$K_2$					0,0000
	$K_2/K_1$					-

Źródło: obliczenia własne.

Podobieństwo Wu i Palmera zdefiniowane za pomocą wzoru (13) przy zmianie kontekstu zmienia się w następujący sposób:

$$\frac{sim_{WP}(c_1, c_2 | K_2)}{sim_{WP}(c_1, c_2 | K_1)} = \frac{d_{LCS(c_1, c_2)}^{K_2} \times (d_{c_1}^{K_1} + d_{c_2}^{K_1})}{(d_{c_1}^{K_2} + d_{c_2}^{K_2}) \times d_{LCS(c_1, c_2)}^{K_1}} \quad (19)$$

**Tabela 2.** Zmiana podobieństwa semantycznego pomiędzy konceptami w wyniku zmiany kontekstu (metoda Wu i Palmera)

		A	B	C	D	E
A	$K_1$	1,0000	0,8889	0,8000	0,4444	0,4000
	$K_2$	1,0000	0,8000	0,6667	0,0000	0,0000
	$K_2/K_1$	1,0000	0,9000	0,8333	0,0000	0,0000
B	$K_1$		1,0000	0,9091	0,4000	0,3636
	$K_2$		1,0000	0,8571	0,0000	0,0000
	$K_2/K_1$		1,0000	0,9429	0,0000	0,0000
C	$K_1$			1,0000	0,3636	0,3333
	$K_2$			1,0000	0,0000	0,0000
	$K_2/K_1$			1,0000	0,0000	0,0000
D	$K_1$				1,0000	0,7273
	$K_2$				1,0000	0,5714
	$K_2/K_1$				1,0000	0,7857
E	$K_1$					1,0000
	$K_2$					1,0000
	$K_2/K_1$					1,0000

Źródło: obliczenia własne.

W tym przypadku nie istnieje zależność liniowa pomiędzy podobieństwem wyznaczonym przy uwzględnieniu różnych kontekstów. Wniosek ten potwierdzają obliczenia dla przykładowych danych (tab. 2).

W przypadku miary  $sim_{LC}$  zmiana kontekstu powoduje zmianę podobieństwa w następujący sposób:

$$\frac{sim_{LC}(c_1, c_2 | K_2)}{sim_{LC}(c_1, c_2 | K_1)} = \frac{\log \frac{len(c_1, c_2)}{2 \times D_{K_2}} \times \log(2 \times D_{K_1})}{\log(2 \times D_{K_2}) \times \log \frac{len(c_1, c_2)}{2 \times D_{K_1}}} \quad (20)$$

Obliczenia dla przykładowych danych przedstawiono w tab. 3.

**Tabela 3.** Zmiana podobieństwa semantycznego pomiędzy konceptami w wyniku zmiany kontekstu (metoda Leacocka i Chodorowa)

		A	B	C	D	E
A	$K_1$	-	1,0000	0,7500	0,4195	0,3538
	$K_2$	-	1,0000	0,6990	0,3010	0,2218
	$K_2/K_1$	-	1,0000	0,9320	0,7176	0,6271
B	$K_1$		-	1,0000	0,3538	0,2982
	$K_2$		-	1,0000	0,2218	0,1549
	$K_2/K_1$		-	1,0000	0,6271	0,5195
C	$K_1$			-	0,2982	0,2500
	$K_2$			-	0,1549	0,0969
	$K_2/K_1$			-	0,5195	0,3876
D	$K_1$				-	0,6038
	$K_2$				-	0,5229
	$K_2/K_1$				-	0,8660
E	$K_1$					-
	$K_2$					-
	$K_2/K_1$					-

Źródło: obliczenia własne.

Przy zastosowaniu podobieństwa Resnika zmiana podobieństwa semantycznego spowodowana zmianą kontekstu określona jest następująco:

$$\frac{sim_{RESNIK}(c_1, c_2 | K_2)}{sim_{RESNIK}(c_1, c_2 | K_1)} = \frac{\log(P_{K_2}(LCS(c_1, c_2)))}{\log(P_{K_1}(LCS(c_1, c_2)))} \quad (21)$$

Kształtowanie się tej relacji uzależnione jest od prawdopodobieństw przypisanych konceptom pełniącym funkcję korzeni drzew reprezentujących rozpatrywane koncepty. Obliczenia dla przykładowych danych zawiera tab. 4.

Przy zastosowaniu metody Resnika podobieństwo pomiędzy identycznymi konceptami zwykle jest różne od jedności. Przyjmuje wartość jeden jedynie dla konceptu najrzadziej występującego. Jedynie w przypadku miary Resnika zmniejszenie kontekstu może prowadzić do zwiększenia miary podobieństwa.

**Tabela 4.** Zmiana podobieństwa semantycznego pomiędzy konceptami w wyniku zmiany kontekstu (metoda Resnika)

		A	B	C	D	E
<b>A</b>	$K_1$	0,4601	0,4601	0,4601	0,2903	0,2903
	$K_2$	0,2920	0,2920	0,2920	0,0000	0,0000
	$K_2/K_1$	0,6347	0,6347	0,6347	0,0000	0,0000
<b>B</b>	$K_1$		0,4841	0,4841	0,2903	0,2903
	$K_2$		0,3333	0,3333	0,0000	0,0000
	$K_2/K_1$		0,6885	0,6885	0,0000	0,0000
<b>C</b>	$K_1$			0,7182	0,2903	0,2903
	$K_2$			0,7358	0,0000	0,0000
	$K_2/K_1$			1,0245	0,0000	0,0000
<b>D</b>	$K_1$				0,6780	0,5104
	$K_2$				0,6667	0,3833
	$K_2/K_1$				0,9833	0,7509
<b>E</b>	$K_1$					0,8719
	$K_2$					1,0000
	$K_2/K_1$					1,1470

Źródło: obliczenia własne.

Analizując zależność pomiędzy zmianą kontekstu a kształtowaniem się podobieństwa  $sim_{JC}$  zdefiniowanego za pomocą wzoru (8), warto zauważyć, że pomiędzy prawdopodobieństwami wystąpienia konceptu  $c_i$  w dwóch rozpatrywanych kontekstach zachodzi zależność:

$$P_{K_2}(c_i) = a \times P_{K_1}(c_i). \quad (22)$$

Korzystając z zależności (22), wyznaczyć można relację pomiędzy wartościami miar podobieństwa semantycznego przy uwzględnieniu kontekstu  $K_2$  i kontekstu  $K_1$ :

$$\frac{sim_{JC}(c_1, c_2 | K_2)}{sim_{JC}(c_1, c_2 | K_1)} = \frac{\log\left(\frac{10 \times P_{K_1}^2(LCS(c_1, c_2))}{P_{K_1}(c_1) \times P_{K_1}(c_2)}\right)}{\log\left(\frac{10 \times P_{K_2}^2(LCS(c_1, c_2))}{P_{K_2}(c_1) \times P_{K_2}(c_2)}\right)} = \frac{\log\left(\frac{10 \times P_{K_1}^2(LCS(c_1, c_2))}{P_{K_1}(c_1) \times P_{K_1}(c_2)}\right)}{\log\left(\frac{10 \times a^2 \times P_{K_1}^2(LCS(c_1, c_2))}{a \times P_{K_1}(c_1) \times a \times P_{K_1}(c_2)}\right)} = 1. \quad (23)$$

Uzyskany wynik potwierdzają obliczenia przeprowadzone dla rozpatrywanego zbioru danych (tab. 5).

Podobieństwo Lina w wyniku zmiany kontekstu zmienia się w następujący sposób:

$$\frac{sim_{LIN}(c_1, c_2 | K_2)}{sim_{LIN}(c_1, c_2 | K_1)} = \frac{2 \times \log(P_{K_2}(LCS(c_1, c_2))) \times (\log(P_{K_1}(c_1)) + \log(P_{K_1}(c_2)))}{(\log(P_{K_2}(c_1)) + \log(P_{K_2}(c_2))) \times 2 \times \log(P_{K_1}(LCS(c_1, c_2)))}. \quad (24)$$

Wyniki obliczeń dla przykładowych danych przedstawia tab. 6.

**Tabela 5.** Zmiana podobieństwa semantycznego pomiędzy conceptami w wyniku zmiany kontekstu (metoda Jiang i Conratha)

		A	B	C	D	E
A	$K_1$	1,0000	0,9305	0,5551	0,3661	0,3000
	$K_2$	1,0000	0,9305	0,5551	0,3661	0,3000
	$K_2/K_1$	1,0000	1,0000	1,0000	1,0000	1,0000
B	$K_1$		1,0000	0,5790	0,3564	0,2934
	$K_2$		1,0000	0,5790	0,3564	0,2934
	$K_2/K_1$		1,0000	1,0000	1,0000	1,0000
C	$K_1$			1,0000	0,2830	0,2418
	$K_2$			1,0000	0,2830	0,2418
	$K_2/K_1$			1,0000	1,0000	1,0000
D	$K_1$				1,0000	0,3808
	$K_2$				1,0000	0,3808
	$K_2/K_1$				1,0000	1,0000
E	$K_1$					1,0000
	$K_2$					1,0000
	$K_2/K_1$					1,0000

Źródło: obliczenia własne.

**Tabela 6.** Zmiana podobieństwa semantycznego pomiędzy conceptami w wyniku zmiany kontekstu (metoda Lina)

		A	B	C	D	E
A	$K_1$	1,0000	0,9745	0,7809	0,5101	0,4358
	$K_2$	1,0000	0,9339	0,5682	0,0000	0,0000
	$K_2/K_1$	1,0000	0,9583	0,7276	0,0000	0,0000
B	$K_1$		1,0000	0,8053	0,4995	0,4281
	$K_2$		1,0000	0,6235	0,0000	0,0000
	$K_2/K_1$		1,0000	0,7743	0,0000	0,0000
C	$K_1$			1,0000	0,4158	0,3651
	$K_2$			1,0000	0,0000	0,0000
	$K_2/K_1$			1,0000	0,0000	0,0000
D	$K_1$				1,0000	0,6622
	$K_2$				1,0000	0,4599
	$K_2/K_1$				1,0000	0,6945
E	$K_1$					1,0000
	$K_2$					1,0000
	$K_2/K_1$					1,0000

Źródło: obliczenia własne.

## 5. Zakończenie

Przeprowadzone analizy potwierdzają, że zmiana kontekstu ma wpływ na kształtowanie się miar podobieństwa semantycznego pomiędzy conceptami. Zwykle zawężenie kontekstu prowadzi do zmniejszenia się mierników podobieństwa. Jednakże przeprowadzone badania wskazują, że istnieją również wyjątki od tej zasady.



Unormowana odległość semantyczna rady zmienia się proporcjonalnie do ilorazu głębokości drzew reprezentujących konteksty. Zmianom miary podobieństwa Resnika towarzyszy brak uporządkowania. Jedynie miara Jiang i Conratha i nieunormowana miara rady nie ulegają zmianie przy modyfikacji kontekstu.

## Literatura

- Jiang J. & Conrath D., 1997, *Semantic similarity based on corpus statistics and lexical taxonomy*, [in:] *Proceedings on International Conference on Research in Computational Linguistics*, pp. 19-33.
- Leacock C. & Chodorow M., 1998, *Combining Local Context and WordNet Similarity for Word Sense Identification*, [in:] C. Fellbaum, ed., *WordNet: An Electronic Lexical Database*, MIT Press, pp. 265-283.
- Lin D., 1998, *An Information-Theoretic Definition of Similarity*, [in:] *Proceedings of the Fifteenth International Conference on Machine Learning {ICML} 1998*, Madison, Wisconsin, USA, July 24-27, pp. 296-304.
- Rada R. et al., 1989, *Development and application of a metric on semantic nets*, *IEEE Transactions on Systems, Man and Cybernetics*, pp. 17-30.
- Resnik P., 1995, *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*, [in:] *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1. IJCAI'95*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 448-453, Available at: <http://dl.acm.org/citation.cfm?id=1625855.1625914>.
- Tuchowski J. et al., 2011, *OBCAS – An Ontology-Based Cluster Analysis System*, [in:] S. Wrycza (ed.), *Research in Systems Analysis and Design: Models and Methods*, Springer, Berlin, pp. 106-112.
- Wan X. & Peng Y., 2005, *A new retrieval model based on texttiling for document similarity search*, *J. Comput. Sci. Technol.*, 20(4), pp.552–558. Available at: <http://dblp.uni-trier.de/db/journals/jcst/jcst20.html#WanP05>.
- Wu Z. & Palmer M., 1994, *Verb Semantics and Lexical Selection*.

## THE IMPACT OF CONTEXT ON SEMANTIC SIMILARITY

**Summary:** In the paper the problem of semantic similarity between concepts from ontology is discussed. The analysis is focused on the issue of relationships between context and semantic similarity and tries to show how the adjustment of context changes the measure of similarity. In the first part of the paper the most popular measures of semantic similarity are presented. Next the problem of context involvement is shown. In the empirical part of the paper the results of numerical experiments are discussed. The substantial findings are gathered in the final part of the text. The results of the study may be useful in the area of automatic text analysis.

**Keywords:** semantic similarity, ontology-based approach, text analytics.