PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 47 2009
TAKSONOMIA 16
Klasyfikacja i analiza danych – teoria i zastosowania

Paweł Lula, Grażyna Paliwoda-Pękosz

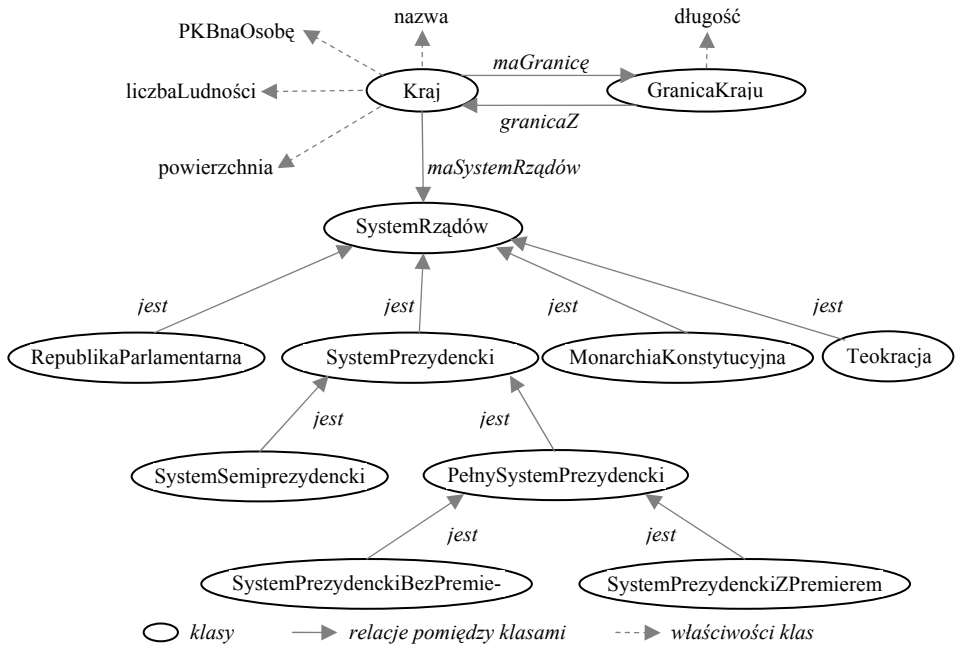
Uniwersytet Ekonomiczny w Krakowie

PODOBIEŃSTWO SEMANTYCZNE W ANALIZIE DANYCH PRZEKROJOWYCH

1. Wstęp

W większości analiz statystycznych stosowana jest tabelaryczna forma reprezentacji informacji dotyczących przetwarzanego zbioru obiektów, w której zwykle opisywanym obiektom odpowiadają wiersze tabeli, jej kolumny reprezentują zaś cechy opisujące badane jednostki. Zasadniczą zaletą tego sposobu organizacji informacji jest łatwość ich przetwarzania. Za jej wady należy uznać trudności dotyczące przedstawienia zróżnicowanego znaczenia poszczególnych cech oraz prezentacji relacji istniejących między obiektami. Metodą rozwiązania przedstawionych problemów może być zastosowanie ontologii będącej sformalizowanym opisem pewnego fragmentu rzeczywistości [Gruber 1995]. Ontologia opisuje pojęcia (klasy), związki pomiędzy klasami (relacje) oraz obiekty (tzw. instancje klas). Można zdefiniować dowolne związki między obiektami (obiekty i związki między nimi są reprezentowane przez graf, w zależności od potrzeb może on być skierowany, nieskierowany, ważony). Związki między pojęciami (klasami) mają charakter hierarchiczny. Każda klasa charakteryzowana jest przez zbiór właściwości. Zachodzi mechanizm dziedziczenia właściwości klas.

Na rysunku 1 zaprezentowano fragment przykładowej ontologii opisującej kraje. Zdefiniowano w niej trzy podstawowe klasy: *Kraj*, *GranicaKraju*, *SystemRządów*. Klasa S*ystemRządów* ma dodatkowo zdefiniowaną hierarchiczną strukturę podklas. Klasy są powiązane z podklasami relacją *jest* (np. *SystemSemiprezydencki* jest podklasą klasy *SystemPrezydencki*, a *SystemPrezydencki* jest podklasą klasy *SystemRządów*). Klasa *Kraj* powiązana jest z klasą *SystemRządów* relacją *maSystemRządów*, klasa *Kraj* powiązana jest z klasą *GranicaKraju* relacją *maGranicę*, natomiast *GranicaKraju* jest powiązana z klasą *Kraj* relacją *granicaZ*. W ujęciu formalnym relacja *granicaZ* stanowi relację odwrotną do *maGranicę*. Dla klasy *Kraj* zdefiniowano właściwości: *nazwa, PKBnaOsobę, liczbaLudności, powierzchnia*. Dla klasy *GranicaKraju* zdefiniowano właściwość: *długość*.



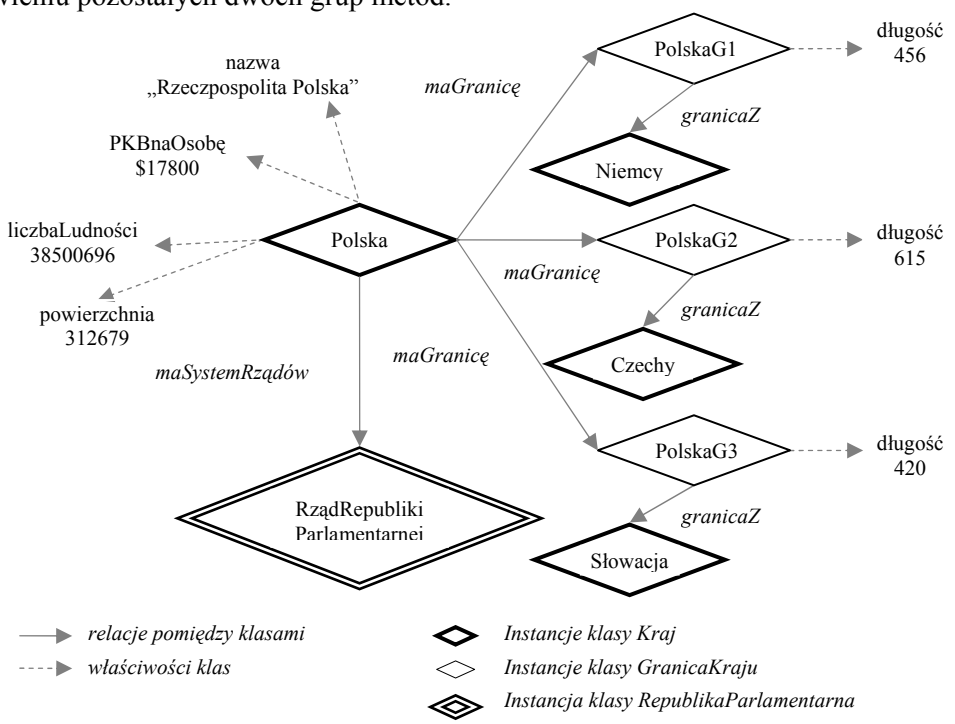
Rys. 1. Fragment przykładowej ontologii opisującej kraje
Źródło: opracowanie własne.

Na rysunku 2 zaprezentowano fragment opisu Polski z wykorzystaniem ontologii zaprezentowanej na rys. 1. Polska jest instancją klasy *Kraj*, zatem można dla niej określić właściwości takie, jak dla klasy *Kraj*. Dla klasy *GranicaKraju* zdefiniowano 3 instancje: PolskaG1, PolskaG2 i PolskaG3, opisujące granice Polski, odpowiednio z Niemcami, Czechami i ze Słowacją. Polska jest republiką parlamentarną i dlatego też jest powiązana relacją *maSystemRządów* z instancją klasy *RepublikaParlamentarna*: *RządRepublikiParlamentarnej*.

Ontologia pozwala definiować różne klasy obiektów, ich własności (o dowolnym stopniu złożoności) oraz różnorodne rodzaje związków mogących zaistnieć pomiędzy określonymi obiektami. Niestety wraz ze wzrostem możliwości w zakresie opisu rzeczywistości zwiększają się problemy z przeprowadzaniem obliczeń. Obiekty nie są już reprezentowane przez wektory, ale przez struktury drzewiaste umiejscowione w przestrzeni zdefiniowanej przez ontologię. W trakcie obliczeń należy uwzględnić nie tylko wartości cech, ale również ich znaczenie, kontekst i wzajemne związki. W analizie danych szczególne znaczenie ma wyznaczanie podobieństwa lub odległości pomiędzy obiektami. W przypadku zastosowania opisu obiektów bazującego na ontologii relacje te określane są jako podobieństwo lub odległość semantyczna.

Porównywanie obiektów opisanych przez jedną ontologię można rozpatrywać w aspekcie porównywania obiektów reprezentowanych przez wektory cech, po-

równywania hierarchicznej struktury pojęć uwzględnianych w ontologii oraz porównywania relacji pomiędzy obiektami. W pierwszym przypadku można zastosować miary odległości pomiędzy wektorami, których opis można znaleźć np. w [*Elektroniczny podręcznik…* 2006]. W artykule uwaga zostanie skupiona na omówieniu pozostałych dwóch grup metod.



Rys. 2. Opis Polski z wykorzystaniem ontologii
Źródło: opracowanie własne na podstawie CIA: [*The world factbook*… 2009].

Struktura artykułu jest następująca: w podpunkcie 2 artykułu omówione zostaną metody porównywania obiektów wynikające z hierarchii klas, w podpunkcie 3 zaprezentowane będą miary określające podobieństwo obiektów wynikające z relacji pomiędzy obiektami. W kolejnym podpunkcie przedstawione zostaną wyniki grupowania dla przykładowych miar. Artykuł zakończy podsumowanie oraz zarys kierunków dalszych badań.

2. Podobieństwo obiektów wynikające z hierarchii klas

Do obliczania podobieństwa uwzględniającego hierarchię klas, do jakich należą obiekty, można wykorzystać miary bazujące na Wordnecie [*Wordnet: an elec-*

tronic... 1998][1]. W uproszczeniu można powiedzieć, że Wordnet stanowi hierarchiczną strukturę leksykalnych pojęć zdefiniowanych przez tzw. synsety, czyli zbiory synonimów. W pracy [Budanitsky, Hirst 2001] zaprezentowano miary oparte na Wordnecie, których autorami są:

– **Hirst-St-Onge** (rel_{HS}) − uwzględnia długość ścieżki łączącej klasy, do jakich należą porównywane obiekty, oraz zmiany kierunku ścieżki:

$$rel_{HS}(c_1, c_2) = C - pl(c_1, c_2) - k*d, \tag{1}$$

gdzie: *u, v* – porównywane obiekty, c_1, c_2 – klasy ontologii, do jakich należą porównywane obiekty, *pl* – długość ścieżki pomiędzy klasami c_1, c_2, *d* – liczba zmian kierunku ścieżki, *C, k* – stałe, w [Budanitsky, Hirst 2006] przyjmuje się $C = 8$ *i* $k = 1$.

– **Leacock-Chodorow**

$$sim_{LC}(c_1, c_2) = -\frac{\log pl(c_1, c_2)}{2D}, \tag{2}$$

gdzie: *D* – głębokość drzewa,

– **Resnik**

$$sim_R(c_1, c_2) = -\log\left(p\left(lso(c_1, c_2)\right)\right), \tag{3}$$

gdzie: $lso(c_1, c_2)$ – najbliższy wspólny przodek klas c_1, c_2, $p(c)$ – prawdopodobieństwo/częstotliwość napotkania instancji klasy *c* (w określonym korpusie),

– **Lin**

$$sim_L(c_1, c_2) = \frac{2\log\left(p\left(lso(c_1, c_2)\right)\right)}{\log\left(p(c_1)\right) + \log\left(p(c_2)\right)}, \tag{4}$$

– **Jiang-Conrath**

$$dist_{JC}(c_1, c_2) = 2\log\left(p\left(lso(c_1, c_2)\right)\right) - \left(\log\left(p(c_1)\right) + \log\left(p(c_2)\right)\right). \tag{5}$$

Obliczanie powyższych miar zaprezentowane będzie na przykładzie obliczania podobieństwa między Polską i Francją w przestrzeni krajów Unii Europejskiej (liczba wszystkich krajów branych pod uwagę wynosi zatem 27). Polska jest republiką parlamentarną (*RepublikaParlamentarna*), natomiast system rządzenia we Francji można zaklasyfikować jako semiprezydencki (*SystemSemiprezydencki*). Po uwzględnieniu głębokości drzewa (D) równej 3 (zob. rys. 1) oraz faktu, iż najbliższym wspólnym przodkiem dla systemów rządzenia obu państw jest węzeł *Sys-*

[1] Prace nad wersją angielską Wordnetu rozpoczęły się na Uniwersytecie Princeton w 1985 r. (http://wordnetprinceton.edu/). Obecnie tworzone są wersje Wordnetu także w innych językach, w tym w języku polskim (http://plwordnet.pwr.wroc.pl/main/?lang=pl).

temRządów, dla którego częstotliwość napotkania instancji tej klasy wynosi 1, otrzymujemy następujące wartości miar:

- rel_{HS}(*RepublikaParlamentarna*, *SystemSemiprezydencki*) = 8- 3 - 1*1 = 4,
- sim_{LC}(*RepublikaParlamentarna*, *SystemSemiprezydencki*) = -log(3/(2*3)) = 0,3,
- sim_R(*RepublikaParlamentarna, SystemSemiprezydencki*) =
- -log(p(*SystemRządów*)) = -log(1) = 0,
- sim_L(*RepublikaParlamentarna*, *SystemSemiprezydencki*) = 0,
- $dist_{JC}$(*RepublikaParlamentarna*, *SystemSemiprezydencki*) =
- 2log(1-(log(16/17)+log(3/27)) = 0,59.

3. Podobieństwo obiektów wynikające z istniejących relacji

Podobieństwo pomiędzy obiektami może wynikać z istniejących relacji między obiektami. Rozważmy dwa obiekty: *u* oraz *v*. Niech obiekt *u* pozostaje w związku z obiektami (ma relacje z obiektami) tworzącymi zbiór N_u, zaś obiekt *v* pozostaje w związku z obiektami tworzącymi zbiór N_v. Biorąc pod uwagę istniejące relacje, można stwierdzić, że podobieństwo obiektów *u* oraz *v* jest tym większe, im zbiory N_u oraz N_v są do siebie bardziej podobne.

W niniejszym opracowaniu uwzględnia się wyłącznie przypadek istnienia (lub nieistnienia) relacji. Nie jest brana pod uwagę siła relacji (wyrażona przez wagę skojarzoną z odpowiednią krawędzią w grafie).

W pracy [Batagelj, Mrvar 2008] definiuje się następujące miary niepodobieństwa obiektów uwzględniające fakt istnienia (lub nie) relacji:

$$d_1(u,v) = \frac{|N_u \Delta N_v|}{MD_1 + MD_2}, \tag{6}$$

$$d_2(u,v) = \frac{|N_u \Delta N_v|}{|N_u \cup N_v|}, \tag{7}$$

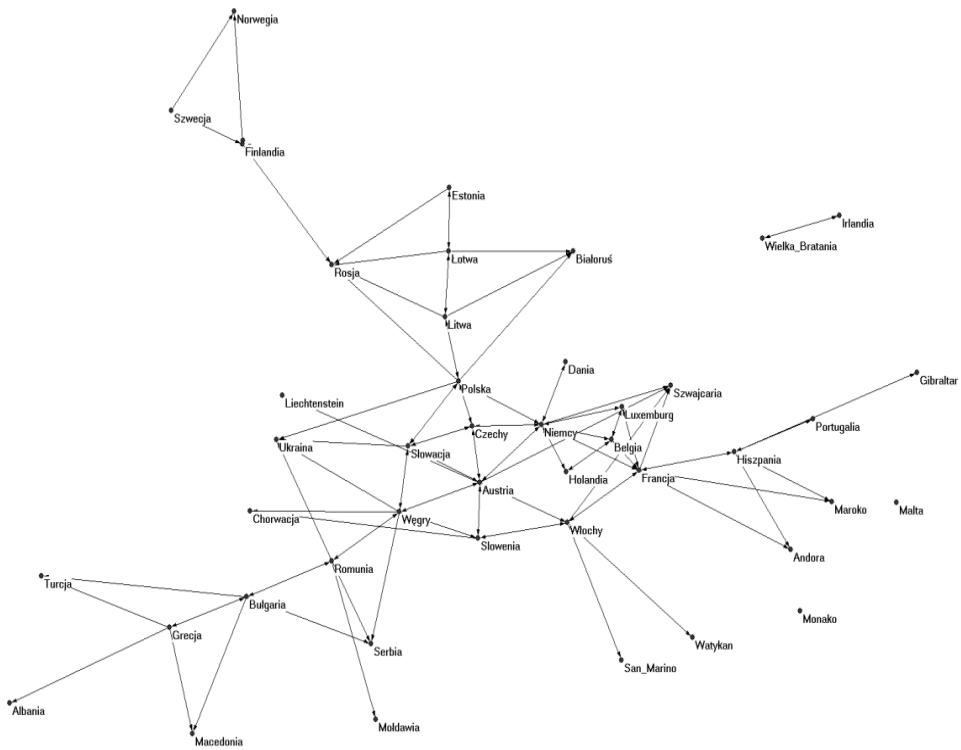
$$d_3(u,v) = \frac{|N_u \Delta N_v|}{|N_u| + |N_v|}, \tag{8}$$

$$d_4(u,v) = \frac{\max(|N_u - N_v|, |N_v - N_u|)}{\max(|N_u|, |N_v|)}, \tag{9}$$

gdzie: *u,v* − porównywane obiekty, N_u, N_v – zbiory, z którymi obiekty *u* i *v* są w relacji, Δ – różnica symetryczna zbiorów, *MD*1, *MD*2− maksymalna (*MD*1) oraz druga pod względem wielkości (*MD*2) wartość stopnia węzła w grafie, którego dotyczą obliczenia.

Miara d_1 określa liczbę elementów, które występują w jednym ze zbiorów N_u lub N_v (nigdy w obu jednocześnie). Wartość w mianowniku ma na celu przeskalowanie wartości, ale nie jest to skalowanie miary do przedziału [0,1], mogą się pojawić większe wartości. Pozostałe miary d_2, d_3, d_4 są miarami unormowanymi.

Korzystając z przykładu dotyczącego krajów, rozpatrywać można relacje między obiektami wynikające z sąsiedztwa (posiadania wspólnej granicy). Uwzględniany jest jedynie fakt posiadania wspólnej granicy, a nie jej długość. Na rysunku 3 zaprezentowano fragment grafu przedstawiającego relacje między krajami wynikające z sąsiedztwa.

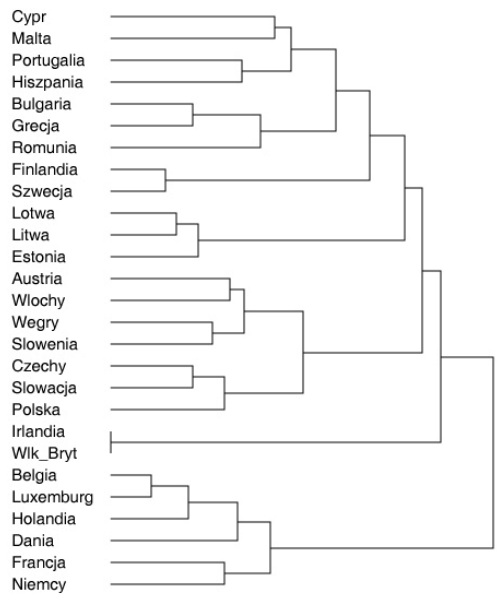


Rys. 3. Relacje pomiędzy krajami wynikające z sąsiedztwa
Źródło: opracowanie własne.

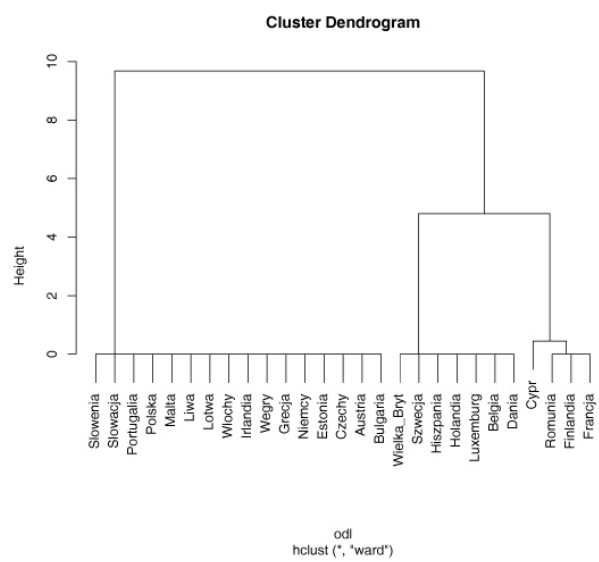
Miara odległości (d_1) oraz miary podobieństwa między Polską i Francją wynikające z sąsiedztwa wynoszą odpowiednio:
d_1(Polska, Francja) = 1,53 d_2(Polska, Francja) = 0,93
d_3(Polska, Francja) = 0,87 d_4(Polska, Francja) = 0,88.

4. Przykładowe wyniki klasyfikacji

Na rysunku 4 przedstawiono wyniki klasyfikacji metodą Warda krajów Unii Europejskiej przy wykorzystaniu miary d_2, która bazuje na podobieństwie obiektów wynikającym z podobieństwa sąsiadów tych obiektów. Wyniki klasyfikacji odpowiadają podobieństwu obiektów związanemu z położeniem geograficznym.



Rys. 4. Wyniki klasyfikacji krajów Unii Europejskiej dla miary d_2
Źródło: opracowanie własne.



Rys. 5. Wyniki klasyfikacji krajów Unii Europejskiej ze względu na formę rządu (podobieństwo między krajami wyznaczono za pomocą formuły Lina, klasyfikacja za pomocą algorytmu Warda z metryką Euklidesa)
Źródło: opracowanie własne.

Na rysunku 5 zaprezentowano wyniki klasyfikacji krajów Unii Europejskiej uwzględniającej formę rządu. Podobieństwo między krajami wyznaczono za pomocą formuły zaproponowanej przez Lina. W procesie klasyfikacji zastosowano podejście zgodne z propozycją Warda i wykorzystujące metrykę Euklidesa.

5. Podsumowanie i wnioski

W artykule przedstawiono przegląd miar, jakie można wykorzystać do porównywania obiektów opisanych za pomocą jednej ontologii. Miary te charakteryzują się dużym zróżnicowaniem pod względem zarówno podstaw teoretycznych, możliwości interpretacyjnych, jak i możliwości zastosowań.

Część miar opiera się na teorii grafów (np. Hirst-St-Onge, Leacock-Chodorow), natomiast inne bazują na teorii informacji (np. Resnik, Lin). Różne są także możliwości interpretacji tych miar. Unormowane miary podobieństwa (np. Leacock-Chodorow, d_2), przyjmujące wartości z przedziału [0, 1], są łatwe do interpretacji i można je agregować. Miary nieunormowane (np. Hirst-St-Onge, d_1) nie dają, niestety, takich możliwości. Z tego też względu podczas konstrukcji zagregowanych mierników podobieństwa należy brać pod uwagę miary unormowane lub podjąć próby unormowania miar odległości.

Ontologie dają możliwości pełniejszego odzwierciedlenia rzeczywistości i dokonania lepszej charakterystyki danych, ale wielopłaszczyznowość opisu danych wiąże się ze zwiększeniem problemów technicznych i złożoności obliczeniowej, np. z problemem przechowywania ontologii w pamięci komputera.

Dalsze badania zostaną skierowane na konstrukcję zagregowanych miar podobieństwa, które będą uwzględniały zarówno hierarchiczną strukturę ontologii, jaki i podobieństwo obiektów wynikające z relacji tych obiektów z innymi obiektami. Kluczowym zagadnieniem jest wybór miar do agregacji, określenie sposobu agregacji tych miar, a w szczególności wag, jakie będą przypisane różnym rodzajom miar. Podczas badań uwzględnione będą różne dziedziny zastosowań tych miar podobieństwa.

Literatura

Batagelj V., Mrvar A. (2008), *Pajek. Program for Analysis and Visualization of Large Networks*, Ljubljana, 1 April, 2008 Reference Manual, http://vlado.fmf.uni-lj.si/pub/networks/pajek/doc/ pajekman.pdf.

Budanitsky A., Hirst G. (2001), *Semantic distance in WordNet: an experimental, application-oriented evaluation of five measures*, http://ftp.cs.toronto.edu/pub/gh/Budanitsky+Hirst-2001.pdf.

Budanitsky A., Hirst G. (2006), *Evaluating Wordnet-based measures of lexical semantic relatedness*, http://ftp.cs.toronto.edu/pub/gh/Budanitsky+Hirst-2006.pdf.

Elektroniczny podręcznik Statystyki PL (2006), Statsoft, Kraków, http://www.statsoft.pl/textbook/ stathome.html.

Gruber T. (1995), *Toward principles for the design of ontologies used for knowledge sharing*, [w:] *Formal Ontology in Conceptual Analysis and Knowledge Representation*, red. N. Guarino, R. Poli, Kluwer Academic Publishers.

Wordnet: an electronic lexical database (1998), red. Ch. Fellbaum, The MIT Press.

The world factbook (2009), *https://www.cia.gov/library/publications/the-world-factbook, 26.02.2009.*

SEMANTIC RELATEDNESS IN SPATIAL DATA ANALYSIS

Summary

The main objective of this paper was to compare semantic distance and semantic similarity measures for ontology-based data. First, the advantages of an ontology-based object description were discussed. Subsequently, similarity measures and distance measures that are based on the ontology hierarchical structure and relationships between objects were presented. Second, examples of measures for European Union countries were counted and then compared, taking into account their theoretical background, computational aspects and possibilities of interpretations and applications. Finally, directions for further research were outlined.