PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS **nr 468 • 2017**

Taksonomia 28
Klasyfikacja i analiza danych – teoria i zastosowania

ISSN 1899-3192
e-ISSN 2392-0041

Urszula Cieraszewska, Monika Hamerska, Paweł Lula
Uniwersytet Ekonomiczny w Krakowie
e-mails: cieraszu@uek.krakow.pl; hamerskm@uek.krakow.pl; pawel.lula@uek.krakow.pl

WYZNACZANIE PODOBIEŃSTWA ZAWARTOŚCI PUBLIKACJI NAUKOWYCH NA PODSTAWIE OPISÓW W NOTACJI UKD

SIMILARITY EVALUATION OF SCIENTIFIC PUBLICATIONS BASED ON THE ANALYSIS OF UDC EXPRESSIONS

DOI: 10.15611/pn.2017.468.05
JEL Classification: C38, C63

Streszczenie: Uniwersalna klasyfikacja dziesiętna (UKD) jest powszechnie wykorzystywanym systemem klasyfikacji obszarów badawczych. Swoim zasięgiem obejmuje wszystkie obszary wiedzy. UKD jest wykorzystywana głównie do opisu zawartości publikacji naukowych w systemach katalogujących dorobek badawczy. Głównym celem niniejszej pracy jest opracowanie metody wyznaczania podobieństwa opisów zawartości publikacji naukowych, zdefiniowanych za pomocą wyrażeń UKD. Prezentację proponowanej metody poprzedza przedstawienie krótkiej charakterystyki uniwersalnej klasyfikacji dziesiętnej oraz zasad obowiązujących przy tworzeniu wyrażeń złożonych wykorzystujących UKD.

Słowa kluczowe: uniwersalna klasyfikacja dziesiętna, podobieństwo publikacji naukowych, podobieństwo semantyczne.

Summary: Universal Decimal Classification (UDC) is a highly flexible classification system for all kinds of information in any medium. Because of its logical hierarchical arrangement and analytical-synthetic nature, it is suitable for a physical organization of collections as well as document browsing and searching. The main purpose of this paper is to present the algorithm for similarity calculation among UDC expressions. The method proposed here can be used for the automatic evaluation of similarity among scientific publications.

Keywords: universal decimal classification, similarity of scientific publications, semantic similarity.

1. Wstęp

Uniwersalna klasyfikacja dziesiętna jest powszechnie wykorzystywanym narzędziem opisu zawartości publikacji naukowych. Do głównych cech UKD należy zaliczyć:

- uniwersalność,
- hierarchiczność,
- otwartość,
- rozszerzalność,
- możliwość budowy symboli złożonych.

Uniwersalność klasyfikacji wynika z szerokiego zakresu jej zastosowań, obejmującego cały zakres wiedzy. Ma charakter *hierarchiczny*, przejawiający się w podziale wiedzy na obszary, które mogą być dzielone na podobszary. Proces podziału może być wielokrotnie powtarzany i prowadzi do zdefiniowania *drzewa wiedzy* reprezentującego całokształt zagadnień znajdujących się w obszarze zainteresowań współczesnej nauki. *Otwartość* systemu UKD wynika z możliwości włączenia charakterystycznych dla niego rozwiązań do innych rozwiązań wspomagających przetwarzanie informacji dotyczących dorobku naukowego oraz z możliwości włączenie fragmentów innych systemów klasyfikacyjnych do rozwiązań wykorzystujących uniwersalną klasyfikację dziesiętną. *Rozszerzalność* przejawia się w możliwości dodawania nowych elementów do istniejącego zbioru kodów. Nie mniej istotną cechą jest *możliwość tworzenia symboli złożonych* łączących wybrane klasy lub też definiujących ich część wspólną bądź uszczegóławianie opisów poprzez definiowanie warunków ograniczających, pełniących rolę filtrów precyzujących szczegółowe informacje dotyczące publikacji.

Analiza cech uniwersalnej klasyfikacji dziesiętnej i obszarów jej zastosowań pozwoliła autorom niniejszej pracy dostrzec potrzebę wzbogacenia rozważań dotyczących omawianego systemu klasyfikacji o elementy związane z określaniem podobieństwa wyrażeń zdefiniowanych przy wykorzystaniu UKD, a tym samym podobieństwa zawartości opisywanych za ich pomocą publikacji naukowych. Próbą realizacji tak określonego celu badawczego jest zamieszczona w artykule autorska propozycja algorytmu wyznaczania podobieństwa wyrażeń zapisanych przy wykorzystaniu uniwersalnej klasyfikacji dziesiętnej.

2. Uniwersalna klasyfikacja dziesiętna

Postępująca specjalizacja nauki na przełomie XIX i XX w. wymusiła reorganizację stosowanych klasyfikacji dokumentacyjnych [Sosińska-Kalata 2002, s. 153]. Zainteresowanie specjalistów z zakresu klasyfikacji piśmiennictwa zostało ukierunkowane na systemy terminologiczne, typologiczne i systematyki obiektowe. Głównym zadaniem było wypracowanie narzędzi opisu wąskich zagadnień wyróżnionych w poszczególnych dziedzinach i dyscyplinach. Starano się również stworzyć narzędzia umożliwiające precyzyjne przedstawianie szczegółowych, często interdyscyplinarnych tematów dokumentów za pomocą wyrażeń złożonych. Dużą wagę przywiązywano również do zapewnienia rozszerzalności systemów klasyfikacji o nowe, nieuwzględnione dotychczas zagadnienia, przy czym rozbudowa istniejących rozwiązań nie mogła prowadzić do naruszenia spójności dotychczasowego systemu.

W nurcie tych działań Paul Otlet i Henri La Fontaine postanowili przygotować uniwersalną klasyfikację dziesiętną (UKD). Zdecydowali się nadać jej układ systematyczny i zapewnić podział piśmiennictwa adekwatny do stopnia szczegółowości tematów klasyfikowanych dokumentów. UKD została pomyślana jako klasyfikacja dokumentacyjna, która mogłaby stać się międzynarodowym standardem rzeczowego opracowania i narzędziem wyszukiwania dokumentów ze wszystkich dziedzin wiedzy niezależnie od języka i miejsca ich publikacji oraz języka i specjalności poszukujących informacji użytkowników.

UKD została oparta na klasyfikacji dziesiętnej Dewaya (KDD). Od KDD odróżnia ją rozbudowanie w dwóch kierunkach: w kierunku umiędzynarodowienia i dostosowania stopnia szczegółowości podziału do potrzeb opisu mikrodokumentów.

UKD jest tzw. klasyfikacją mieszaną – częściowo monohierarchiczną z elementami fasetyzacji o notacji dziesiętnej, w której z jednej klasyfikacji nadrzędnej można wyodrębnić 10 klas podrzędnych, każda z nich oznaczona jest jedną cyfrą arabską, dopisaną do symbolu klasy nadrzędnej. W tej notacji obowiązuje zasada tzw. delimitacji sekwencji trójcyfrowych – po każdych trzech cyfrach zostaje umieszczona kropka.

Elementarnymi jednostkami leksykalnymi UKD są symbole główne i pomocnicze. Pierwsze są składniowo samodzielne, dzięki czemu mogą tworzyć zdania UKD.

Znajdujące się na najwyższym poziomie symbole główne UKD obejmują:

0 Dział ogólny
1 Filozofia. Psychologia
2 Religia. Teologia
3 Nauki społeczne. Prawo. Administracja
4 *dział pusty*
5 Matematyka. Nauki przyrodnicze
6 Nauki stosowane. Medycyna. Nauki techniczne. Rolnictwo
7 Sztuka. Rozrywki. Sport
8 Językoznawstwo. Nauka o literaturze. Literatura piękna
9 Archeologia. Prehistoria. Geografia. Biografie. Historia

Sposób podziału przedstawionych wyżej obszarów na podobszary przedstawiony zostanie na przykładzie działu: *5 Matematyka. Nauki przyrodnicze.*

5 Matematyka. Nauki przyrodnicze

51 MATEMATYKA

Matematyka dyskretna

51-7 Badania i metody matematyczne w innych dziedzinach nauki
51-8 Gry i rozrywki matematyczne
510 Podstawy i ogólne zasady matematyki
510.21 Filozofia matematyki
510.22 Teoria mnogości

Teoria zbiorów. Zbiory rozmyte

510.5 Algorytmy i funkcje obliczalne

zob. też

004.021 Algorytmy

510.6 Logika matematyczna

zob. też

164 Logistyka. Rachunek logiczny

Symbole pomocnicze pełnią funkcje uzupełniające wobec symbolu głównego do wyrażenia dodatkowych cech obiektu albo pozwalają łączyć ze sobą różne obiekty wskazywane przez symbole główne. Wśród symboli pomocniczych wyróżnia się:

1. symbole poddziałów wspólnych – można łączyć je z dowolnym symbolem głównym (język, forma, rasa, narodowość i grupa etniczna, czas),

2. symbole poddziałów analitycznych – wyrażają cechy specyficzne dla obiektów pewnej dziedziny i uznane są za charakterystyczne dla wielu lub wszystkich takich obiektów (wskaźniki -1/-9, .01/.09, ’0/’9),

3. symbole poddziałów syntetycznych – tworzone są z części symboli głównych pewnego działu, wskazujących różne charakterystyki obiektów, które w dziale tym są sklasyfikowane (‘),

4. znaki łączące (relatory) – służą do wyrażania związków zachodzących między pojęciami, tematami lub przedmiotami treści dokumentów, wyrażonymi za pomocą symboli głównych:

- + (plus) i / (kreska ukośna) – poszerzają zakres symbolu, łączą symbole odnoszące się do zagadnień równorzędnych, niezależnych lub takich, między którymi związek jest bardzo luźny,
- : (dwukropek) i :: (dwukropek podwójny) – znak łączy symbole proste lub rozwinięte, między którymi występuje zależność semantyczna; symbole połączone : podlegają inwersji, a :: nie podlegają inwersji,
- [ ] grupują dwa lub więcej symboli prostych lub rozwiniętych połączonych znakiem + lub : , które jako całość znajdują się w pewnej relacji do symboli umieszczonych poza nawiasem.

Biorąc pod uwagę strukturę symboli UKD, można wyróżnić:

- symbol prosty – symbol główny lub pomocniczy, do którego nie dołączono żadnego innego symbolu, np. *511 Teoria liczb ; 517 Analiza matematyczna*,
- symbol rozwinięty – symbol, który zawiera jeden symbol główny oraz co najmniej jeden symbol poddziałów pomocniczych, np. *519.2(03) Rachunek prawdopodobieństwa. Statystyka matematyczna – słowniki*,
- symbol złożony – symbol składający się z co najmniej dwóch symboli prostych lub rozwiniętych, powiązanych odpowiednim znakiem łączącym, wskazującym na relacje zachodzące między tymi symbolami, np. *51+53 Matematyka i fizyka; 519.86:336 Matematyka finansowa; 004::336 Informatyka finansowa; 657::339.1 Rachunkowość handlowa.*

3. Podobieństwo zakresów tematycznych publikacji

W opracowanym przez autorów niniejszej pracy algorytmie sposób wyznaczania podobieństwa zakresów tematycznych publikacji uzależniony jest od charakteru porównywanych dokumentów i będzie realizowany w inny sposób dla dokumentów *jednotematycznych* oraz *wielotematycznych*. Również *warunki ograniczające,* występujące w symbolach rozwiniętych, będą mieć wpływ na wyznaczane mierniki podobieństwa. Przedstawione powyżej kwestie zostaną omówione w dalszej części bieżącego punktu.

W dalszej części opisu użyte zostaną następujące pojęcia i symbole:

- *drzewo wiedzy* jest strukturą hierarchiczną, pokazującą podział wiedzy na obszary tematyczne w sposób przyjęty w klasyfikacji UKD. Korzeń drzewa reprezentuje całość wiedzy, zaś elementy potomne korzenia odpowiadają symbolom głównym wyróżnionym w klasyfikacji UKD. Dalsza struktura drzewa odpowiada kolejnym podklasom wyróżnionym w systemie;
- *klasa* odpowiada symbolowi głównemu klasyfikacji UKD i jest reprezentowana przez jeden węzeł w drzewie wiedzy, na przykład klasa 330.162 odpowiada zakresowi tematycznemu Etyka gospodarcza. Etyka biznesu;
- *współczynnik podobieństwa Jaccarda wektorów o elementach niezerowych* $\boldsymbol{x} = [x_1, x_2, \dots, x_n]$ i $\boldsymbol{y} = [y_1, y_2, \dots, y_n]$ wyznaczany jest według wzoru:

$$sim_J(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{n} min(x_i, y_i)}{\sum_{i=1}^{n} max(x_i, y_i)}; \quad (1)$$

- *problem optymalnego przyporządkowania* dla dwóch zbiorów $A = \{a_1, a_2, \dots, a_{nA}\}$ i $B = \{b_1, b_2, \dots, b_{nB}\}$ polega na znalezieniu takiego sposobu przyporządkowania obiektów $a_i \leftrightarrow b_j$, dla którego suma mierników podobieństwa pomiędzy łączonymi elementami będzie największa. Popularną metodą rozwiązywania tego typu problemu jest algorytm węgierski [Kuhn 1955, s. 83-97].

3.1. Podobieństwo publikacji jednotematycznych

Przez pojęcie *„publikacji jednotematycznej”* rozumiemy pracę opisaną za pomocą symbolu UKD zawierającego odwołanie do jednej klasy pochodzącej z drzewa wiedzy (np. *330.162* odpowiadającego klasie *Etyka gospodarcza. Etyka biznesu*).

Przyjmijmy, że porównywane są dwie publikacje opisane odpowiednio przez klasy K_1 i K_2. Sposób wyznaczania podobieństwa tak opisanych publikacji opisuje algorytm 1.

Algorytm 1. Wyznaczanie podobieństwa publikacji jednotematycznych

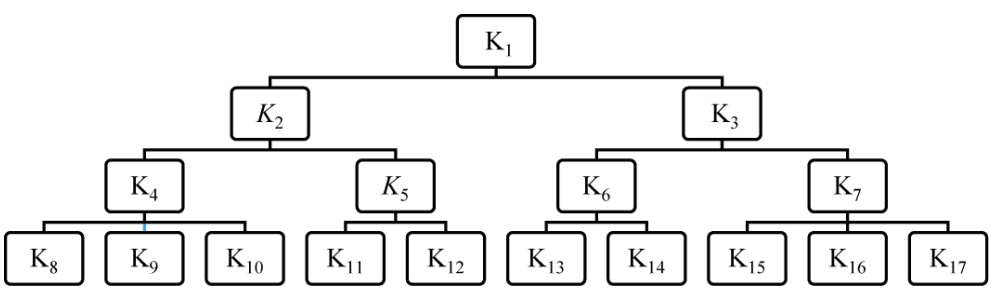
Dane wejściowe: K_1, K_2, DW (drzewo wiedzy)
Dane wyjściowe: sim (miara podobieństwa tematyki publikacji)
Krok 1: Wyznacz minimalne poddrzewo pDW drzewa DW zawierające klasy K_1 *i* K_2.

Krok 2: Przypisz korzeniowi poddrzewa pDW współczynnik równy 1. Przyjmij, że korzeń poddrzewa jest elementem bieżącym.
Krok 3: Każdemu elementowi podrzędnemu przypisz wartość równą $\frac{r}{n_r}$, gdzie r jest wartością przechowywaną w elemencie bieżącym, zaś n_r jest liczbą potomków elementu bieżącego.
Krok 4: Sposób postępowania opisany w kroku 3 zastosuj do przypisania współczynników wszystkim elementom poddrzewa pDW.
Krok 5: Dokonaj normalizacji współczynników poddrzewa pDW. W tym celu podziel wartość każdego współczynnika przez sumę wszystkich współczynników.
Krok 6: Utwórz poddrzewo pDW_1 reprezentujące pierwszą z porównywanych publikacji. Drzewo pDW_1 ma strukturę zgodną z pDW i posiada współczynniki zgodne ze współczynnikami z pDW w części odpowiadającej klasie K_1 i wszystkim klasom podrzędnym. Pozostałym węzłom przypisane zostają współczynniki zerowe.
Krok 7: Utwórz poddrzewo pDW_2 reprezentujące drugą publikację. Zastosuj algorytm analogiczny do opisanego w kroku 6, przy czym niezerowe współczynniki odpowiadać będą części poddrzewa pDW_2 obejmującej klasę K_2 i wszystkie klasy podrzędne.
Krok 8: Umieść wartości wszystkich współczynników poddrzew pDW_1 i pDW_2 odpowiednio w wektorach wDW_1 i wDW_2 w kolejności zgodnej z kolejnością poziomów węzłów odczytywanych od strony lewej do prawej.
Krok 9: Oblicz: $sim_J(wDW_1, wDW_2)$. Wyznaczony współczynnik jest miarą podobieństwa pomiędzy zakresami tematycznymi rozpatrywanych publikacji.

Działanie algorytmu zilustrowane zostanie przykładem.

Przykład 1

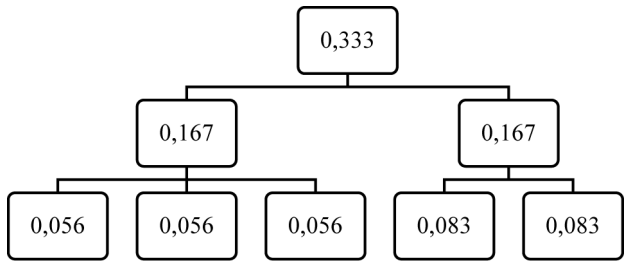
Rozważmy drzewo wiedzy o strukturze przedstawionej na rys. 1. Załóżmy, że pierwsza z porównywanych publikacji opisana jest za pomocą klasy K_2, zaś drugą opisuje klasa K_5.



Rys. 1. Przykładowe drzewo wiedzy wraz z klasami opisującymi porównywane publikacje

Źródło: opracowanie własne.

Stosując operacje opisane w krokach 2–5 algorytmu 1, wyznaczono znormalizowane współczynniki dla minimalnego poddrzewa zawierającego klasy opisujące publikacje (rys. 2).



Rys. 2. Poddrzewo służące do porównania publikacji opisanych za pomocą klas K_2 i K_5

Źródło: opracowanie własne.

Wymienione w 8 kroku algorytmu wektory opisujące porównywane publikacje przyjmują postać:

$$wDW_1 = [0{,}333; 0{,}167; 0{,}167; 0{,}056; 0{,}056; 0{,}056; 0{,}083; 0{,}083]$$

oraz:

$$wDW_2 = [0; 0; 0{,}167; 0; 0; 0; 0{,}083; 0{,}083].$$

Realizując obliczenia określone w 9 kroku algorytmu uzyskujemy miarę podobieństwa zawartości publikacji:

$$sim_J(wDW_1, wDW_2) = 0{,}333.$$

3.2. Podobieństwo publikacji wielotematycznych opisanych za pomocą symboli typu A+B

Opis publikacji wielotematycznych typu A+B składa się z dowolnej liczby identyfikatorów klas połączonych za pomocą operatora + (plus):

$$opis = K_1 + K_2 + \cdots + K_{LK}.$$

Klasy K_i występujące w opisie publikacji mają równorzędny charakter i nie muszą być ze sobą powiązane.

Proponowany algorytm wyznaczania podobieństwa pomiędzy opisanymi w ten sposób publikacjami jest zmodyfikowaną wersją metody przedstawionej w poprzednim punkcie. Główna zmiana dotyczy uwzględnienia w trakcie obliczeń minimalnego poddrzewa zawierającego wszystkie klasy K_i występujące w opisie obu publikacji.

Algorytm 2. Wyznaczanie podobieństwa publikacji wielotematycznych typu A+B

Dane wejściowe: DW, $doc_I = K_1^I + K_2^I + \cdots + K_{LK1}^I$, $doc_{II} = K_1^{II} + K_2^{II} + \cdots + K_{LK2}^{II}$
Dane wyjściowe: sim

Krok 1: Wyznacz minimalne poddrzewo pDW drzewa DW zawierające klasy: $K_1^I, K_2^I, \ldots, K_{LK1}^I, K_1^{II}, K_2^{II}, \ldots, K_{LK2}^{II}$.

Kroki 2, 3, 4, 5: jak w **algorytmie 1**.

Krok 6: Utwórz poddrzewo pDW_1 reprezentujące pierwszą z porównywanych publikacji. Drzewo pDW_1 ma strukturę zgodną z pDW i posiada współczynniki zgodne ze współczynnikami z pDW w części odpowiadającej klasom $K_1^I, K_2^I, \ldots, K_{LK1}^I$ i wszystkim klasom podrzędnym. Pozostałym węzłom przypisane zostają współczynniki zerowe.

Krok 7: Utwórz poddrzewo pDW_2 reprezentujące drugą publikację. Zastosuj algorytm analogiczny do opisanego w kroku 6, przy czym niezerowe współczynniki odpowiadać będą części poddrzewa pDW_2 obejmującej klasy $K_1^{II}, K_2^{II}, \ldots, K_{LK2}^{II}$ i wszystkie klasy podrzędne.

Kroki 8, 9: jak w **algorytmie 1**.

3.3. Podobieństwo publikacji wielotematycznych opisanych za pomocą symboli typu A:B

W bieżącym punkcie rozpatrywane będą publikacje, których opis przyjmuje postać:

$$opis = K_1 : K_2 : \ldots : K_{LK}.$$

Użycie w opisie symbolu dwukropka wskazuje na istnienie relacji semantycznej pomiędzy klasami K_i. Przyjmuje się, że udział w zawartości publikacji tematyki reprezentowanej przez poszczególne klasy występujące w opisie jest identyczny. Z tego powodu proponowany wskaźnik podobieństwa powinien weryfikować fakt zbieżności zbiorów wszystkich klas występujących w opisach publikacji. W przedstawionej propozycji do opisu i rozwiązania tak postawionego problemu wykorzystano model opisujący problem optymalnego przyporządkowania.

Algorytm 3. Wyznaczanie podobieństwa publikacji wielotematycznych typu A:B

Dane wejściowe: DW, $doc_I = K_1^I : K_2^I : \ldots : K_{LK1}^I$, $doc_{II} = K_1^{II} : K_2^{II} : \ldots : K_{LK2}^{II}$

Dane wyjściowe: *sim*

Kroki 1, 2, 3, 4, 5: jak w **algorytmie 2.**

Krok 6: Utwórz poddrzewa $pDW_1^I, pDW_2^I, \ldots, pDW_{LK1}^I$ reprezentujące pierwszą z porównywanych publikacji. Drzewo pDW_i^I ma strukturę zgodną z pDW i posiada współczynniki zgodne ze współczynnikami z pDW w części odpowiadającej klasie K_i^I i wszystkim klasom podrzędnym. Pozostałym węzłom przypisane zostają współczynniki zerowe.

Krok 7: Utwórz poddrzewa $pDW_1^{II}, pDW_2^{II}, \ldots, pDW_{LK2}^{II}$ reprezentujące drugą z porównywanych publikacji. Drzewo pDW_i^{II} ma strukturę zgodną z pDW i posiada współczynniki zgodne ze współczynnikami z pDW w części odpowiadającej klasie K_i^{II} i wszystkim klasom podrzędnym. Pozostałym węzłom przypisane zostają współczynniki zerowe.

Krok 8: Stosując wielokrotnie **algorytm 1**, utwórz macierz podobieństwa pomiędzy drzewami opisującymi porównywane publikacje o strukturze:

$$\boldsymbol{P} = \begin{bmatrix} sim_J(pDW_1^I, pDW_1^{II}) & \ldots & sim_J(pDW_1^I, pDW_{LK2}^{II}) \\ \ldots & \ldots & \ldots \\ sim_J(pDW_{LK1}^I, pDW_1^{II}) & \ldots & sim_J(pDW_{LK1}^I, pDW_{LK2}^{II}) \end{bmatrix}$$

Krok 9: Biorąc pod uwagę dane zawarte w macierzy ***P***, rozwiąż problem optymalnego przyporządkowania klas $K_1^I, K_2^I, \dots, K_{LK1}^I$ i klas $K_1^{II}, K_2^{II}, \dots, K_{LK2}^{II}$ w sposób zapewniający maksymalizację podobieństwa pomiędzy przyporządkowywanymi klasami.

Krok 10: Wyznacz wartość średnią z podobieństw odpowiadających przypisanym sobie klasom. Wyznaczoną w ten sposób wartość przyjmij w charakterze miernika podobieństwa pomiędzy porównywanymi publikacjami.

Działanie algorytmu zostanie zaprezentowane na przykładzie.

Przykład 2

Przykład ilustrujący działanie algorytmu 3 będzie bazować na drzewie wiedzy przedstawionym w punkcie 3.1. Załóżmy, że celem obliczeń jest porównanie dwóch publikacji zdefiniowanych w następujący sposób:

$$publ_1 = K_4 : K_5 : K_3,$$

$$publ_2 = K_2 : K_3.$$

Podobieństwa pomiędzy poddrzewami pDW_i^I i pDW_j^{II} wyznaczone zgodnie z algorytmem 1 przedstawia tabela 1.

Tabela 1. Podobieństwa pomiędzy poddrzewami reprezentującymi zakresy tematyczne porównywanych publikacji. Wyróżnione wartości odpowiadają przyporządkowanym klasom (krok 9 algorytmu 3)

		Publikacja $publ_1$		
		K_2	K_3	K_0
Publikacja $publ_2$	K_4	***0,3333***	0,0000	0,3333
	K_5	0,3333	0,0000	***0,3333***
	K_3	0,0000	***1,0000***	1,0000

Źródło: obliczenia własne.

Warto zauważyć, że liczba klas opisujących każdą z publikacji jest różna. W celu jej wyrównania wprowadzona została w opisie publikacji $publ_2$ sztuczna klasa K_0. Mierniki podobieństwa pomiędzy klasą K_0 a klasami opisującymi publikację $publ_1$ wyznaczono jako wartości maksymalne z wartości występujących w poszczególnych wierszach. Zastosowanie takiego rozwiązania pozwala powiązać nadmiarową klasę z opisu $publ_1$ z najbardziej zbliżoną do niej klasą z opisu $publ_2$, niezależnie od jej przypisania do innej klasy charakteryzującej pierwszą publikację. Sposób przyporządkowania klas został zaznaczony poprzez wyróżnienie w tabeli 1 odpowiadających im mierników podobieństwa.

Z formuły opisanej w kroku 10 algorytmu wynika, że podobieństwo pomiędzy rozpatrywanymi publikacjami wynosi:

$$sim(publ_1, publ_2) = \frac{0{,}3333 + 0{,}3333 + 1}{3} = 0{,}5556.$$

3.4. Podobieństwo publikacji wielotematycznych opisanych za pomocą symboli typu A::B

Podobnie jak w przypadku opisów typu A:B, również zapis typu:

$$opis = K_1 :: K_2 :: \ldots :: K_{LK}$$

wskazuje na istnienie relacji semantycznych pomiędzy wskazanymi klasami. Jednakże w tym przypadku każda kolejna klasa wymieniona na liście ma coraz mniejszy udział w treści publikacji. Zdefiniowany w ten sposób opis nie podlega inwersji, gdyż zmiana kolejności klas spowodowałaby zmianę udziału zagadnień odpowiadających poszczególnym identyfikatorom w zawartości publikacji. Zasady obowiązujące w klasyfikacji UKD nie precyzują reguły pozwalającej na liczbowe ujęcie znaczenia poszczególnych składowych. W niniejszej pracy przyjmujemy, że udział każdej z klas określony jest przez elementy wektora $\boldsymbol{\alpha}$ o elementach z przedziału $[0; 1]$ uporządkowanych malejąco:

$$\boldsymbol{\alpha} = [\alpha_1\ \alpha_2\ \ldots\ \alpha_{LK}].$$

Dysponując tak określonymi danymi, podobieństwo zawartości dwóch publikacji obliczyć można w sposób określony w algorytmie 4.

Algorytm 4. Wyznaczanie podobieństwa publikacji wielotematycznych typu A::B

Dane wejściowe: DW, $publ_I = K_1^I :: K_2^I :: \cdots :: K_{LK1}^I$, $publ_{II} = K_1^{II} :: K_2^{II} :: \cdots :: K_{LK2}^{II}$, $\boldsymbol{\alpha}^I = [\alpha_1^I, \alpha_2^I, \ldots, \alpha_{LK1}^I]$, $\boldsymbol{\alpha}^{II} = [\alpha_1^{II}, \alpha_2^{II}, \ldots, \alpha_{LK2}^{II}]$

Dane wyjściowe: *sim*

Kroki 1-8: jak w **algorytmie 3**.

Krok 9: Dokonaj przekształcenia macierzy $\boldsymbol{P}$, uwzględniające wartości współczynników $\boldsymbol{\alpha}^I$ *i* $\boldsymbol{\alpha}^{II}$*:*

$$\widetilde{\boldsymbol{P}} = \begin{bmatrix} \frac{\alpha_1^I + \alpha_1^{II}}{2} \times sim_J(pDW_1^I, pDW_1^{II}) & \ldots & \frac{\alpha_1^I + \alpha_{LK2}^{II}}{2} \times sim_J(pDW_1^I, pDW_{LK2}^{II}) \\ \ldots & \ldots & \ldots \\ \frac{\alpha_{LK1}^I + \alpha_1^{II}}{2} \times sim_J(pDW_{LK1}^I, pDW_1^{II}) & \ldots & \frac{\alpha_{LK1}^I + \alpha_{LK2}^{II}}{2} \times sim_J(pDW_{LK1}^I, pDW_{LK2}^{II}) \end{bmatrix}$$

Krok 10: Biorąc pod uwagę dane zawarte w macierzy $\widetilde{\boldsymbol{P}}$, rozwiąż problem optymalnego przyporządkowania klas $K_1^I, K_2^I, \ldots, K_{LK1}^I$ *i klas* $K_1^{II}, K_2^{II}, \ldots, K_{LK2}^{II}$ w sposób zapewniający maksymalizację podobieństwa pomiędzy przyporządkowywanymi klasami.

Krok 11: Wyznacz wartość średnią z podobieństw odpowiadających przypisanym sobie klasom. Wyznaczoną w ten sposób wartość przyjmij w charakterze miernika podobieństwa pomiędzy porównywanymi publikacjami.

Proponowany algorytm różni się od algorytmu 3 jedynie sposobem konstruowania macierzy podobieństwa pomiędzy poddrzewami opisującymi poszczególne zagadnienia poruszane w publikacjach. Wartość pierwotnych współczynników podobieństwa zmniejszana jest proporcjonalnie do średniej ze współczynników określających znaczenie porównywanych klas w zawartości publikacji.

3.5. Podobieństwo publikacji opisanych za pomocą symboli rozwiniętych

Symbole rozwinięte zawierają, oprócz identyfikatorów klas, symbole pomocnicze określające cechy publikacji. Tego typu zapisy traktować można jako nałożone na publikację warunki ograniczające, dotyczące języka, formy publikacji, rasy, narodowości i grupy etnicznej oraz czasu. Symbole pomocnicze, definiujące wspomniane warunki w poddrzewie wiedzy reprezentującym publikację, są przypisywane do jednej lub wielu klas. Przypisanie warunku ograniczającego do identyfikatora klasy powoduje jego odziedziczenie przez wszystkie klasy potomne.

Przy porównywaniu opisów zawartości publikacji uwzględnienie podobieństwa warunków ograniczających może nastąpić na etapie liczenia wartości współczynnika Jaccarda dla wektorów zawierających współczynniki odpowiednich drzew. Odpowiedni wzór przyjmuje postać:

$$sim_J(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{n} s_i^{OGR} min(x_i, y_i)}{\sum_{i=1}^{n} max(x_i, y_i)}, \tag{2}$$

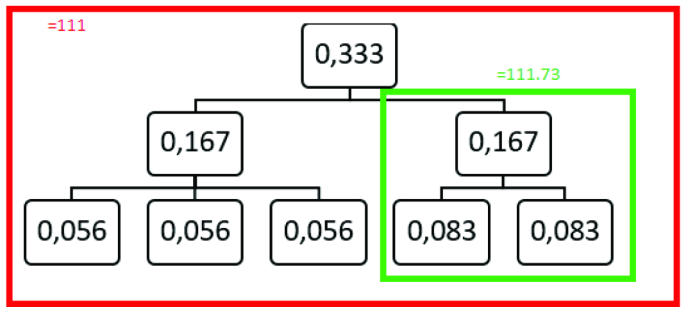
gdzie s_i^{OGR} jest podobieństwem warunków ograniczających przypisanych do i-tego węzła w drzewie. Sposób wyznaczania wartości s_i^{OGR} powinien być zdefiniowany dla każdego typu warunku. W przypadku przypisania do danej klasy więcej niż jednego typu warunku ograniczającego podobieństwo należy wyznaczyć niezależnie dla każdego z typów, a uzyskane wartości należy następnie pomnożyć przez siebie.

Przedstawione powyżej rozważania zilustruje przykład liczbowy.

Przykład 3

Rozważmy sytuację zbliżoną do tej, która została omówiona w przykładzie 1. Przyjmijmy jednak, że publikacji opisanej za pomocą klasy K_2 towarzyszy dodatkowy zapis w postaci „=111” (oznaczający, że publikacja została przygotowana w brytyjskiej wersji języka angielskiego), zaś w opisie publikacji o tematyce odpowiadającej klasie K_5 występuje zapis „=111.73” (oznaczający zastosowanie amerykańskiej wersji języka angielskiego).

Niech podobieństwo użytych w zapisie warunków ograniczających określone będzie za pomocą tabeli 2.



Rys. 3. Specyfikacja języka dla porównywanych publikacji

Źródło: opracowanie własne.

Tabela 2. Podobieństwa pomiędzy warunkami ograniczającymi, określającymi wersję języka publikacji

	=111	=111.73	brak warunku
=111	1,00	0,90	0,95
=111.73	0,90	1,00	0,95
brak warunku	0,95	0,95	1,00

Źródło: dane umowne.

W wyniku zastosowania wzoru (2) uzyskano:

$$sim_J(wDW_1, wDW_2| = 111; = 111.73) = 0{,}3,$$

można więc zauważyć, że uwzględnienie różnic wynikających z zastosowania odmiennych wersji języka angielskiego w trakcie przygotowywania publikacji nieznacznie zmniejszyło miernik podobieństwa pomiędzy nimi.

4. Zakończenie

W pracy przedstawiono proponowaną przez autorów metodę określania podobieństwa wyrażeń zdefiniowanych w notacji UKD. Zaproponowany algorytm znajduje zastosowanie do analizy podobieństwa wyrażeń prostych, złożonych oraz rozwiniętych.

W celu wyznaczenia podobieństwa algorytm wykorzystuje jedynie informacje o strukturze drzewa wiedzy będącego podstawą notacji UKD. Fakt ten w istotny sposób ułatwia realizację obliczeń, ale równocześnie nie pozwala na uwzględnienie informacji dotyczącej częstości występowania poszczególnych kodów UKD w bazach bibliograficznych, a tym samym ich wartości informacyjnej.

Dalsze badania dotyczyć będą oceny poprawności zaproponowanej metody oraz oceny zgodności miar podobieństwa publikacji naukowych wyznaczonych na podstawie wyrażeń UKD z wynikami uzyskanymi za pomocą innych metod.

Literatura

Broughton V., 2006, *The need for a faceted classification as the basis of all methods of information retrieval*, Aslib Proceedings , vol. 58, iss. 1/2, s. 49-72.

Budanitsky A., Hirst G., 2001, *Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures*, Workshop on WordNet and Other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, PA, s. 29-34.

Colillas M.G., 2011, *UDC on the Internet: Theory and project in evolution for use of indexing and retrieval systems*, IFLA Journal, vol. 37, iss. 4, s. 305-313.

Chatterjee A., 2015, *Universal Decimal Classification and Colon Classification: Their mutual impact*, Annals of Library & Information Studies, vol. 62, iss. 4, s. 226-230.

http://www.udcc.org/.

http://www.udcsummary.info/php/index.php.

Kuhn H.W., 1955, *The Hungarian method for the assignment problem*, Naval Research Logistics Quarterly, vol. 2, s. 83-97.

Lula P., 2009, *Analiza taksonomiczna obiektów opisywanych za pomocą ontologii*, [w:] Pociecha J. (red.), *Współczesne problemy modelowania i prognozowania zjawisk społeczno-gospodarczych*, Uniwersytet Ekonomiczny, Kraków, s. 429-440.

Lula P., Tuchowski J., Wójcik K., 2014, *Similarity between compound objects and its application in recruitment process*, Enterprise in Hardship – Economic, Managerial and Juridical Perspective, Ariccia, s. 9-26.

Rückemann C., 2015, *Knowledge integration for scientific classification and computation*, AIP Conference Proceedings, vol. 1648, iss. 1, s. 1-4.

Sosińska-Kalata B., 1993, *Uniwersalna Klasyfikacja Dziesiętna: podręcznik*, Wydawnictwo SBP, Warszawa.

Sosińska-Kalata B., 2002, *Klasyfikacja*, Nauka – Dydaktyka – Praktyka, nr 52, Wydawnictwo SBP, Warszawa.