

# PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

# RESEARCH PAPERS

of Wrocław University of Economics

Nr 328

**Taksonomia 23**

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

Redaktorzy naukowci

Krzysztof Jajuga, Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu  
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:

[www.ibuk.pl](http://www.ibuk.pl), [www.ebscohost.com](http://www.ebscohost.com),

w Dolnośląskiej Bibliotece Cyfrowej [www.dbc.wroc.pl](http://www.dbc.wroc.pl),

The Central and Eastern European Online Library [www.ceeol.com](http://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](http://www.wydawnictwo.ue.wroc.pl)

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

Kopiowanie i powielanie w jakiegokolwiek formie  
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu  
Wrocław 2014

**ISSN 1899-3192** (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)

**ISSN 1505-9332** (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

## Spis treści

<b>Wstęp</b> .....	11
<b>Małgorzata Rószkiewicz</b> , Wykorzystanie metaanalizy w budowaniu modelu pomiarowego w przypadku braku niezmienniczości zasad pomiaru na przykładzie pomiaru zadowolenia z życia.....	13
<b>Elżbieta Sobczak</b> , Harmonijność inteligentnego rozwoju regionów Unii Europejskiej .....	21
<b>Ewa Roszkowska, Renata Karwowska</b> , Analiza porównawcza województw Polski ze względu na poziom zrównoważonego rozwoju w roku 2010.....	30
<b>Tadeusz Kufel, Magdalena Osińska, Marcin Błażejowski, Paweł Kufel</b> , Analiza porównawcza wybranych filtrów w analizie synchronizacji cyklu koniunkturalnego.....	41
<b>Marcin Salamaga</b> , Próba konstrukcji tablic „wymierania scenicznego” spektakli operowych na przykładzie Metropolitan Opera.....	51
<b>Iwona Foryś</b> , Wykorzystanie analizy dyskryminacyjnej do typowania rynków podobnych w procesie wyceny nieruchomości niemieszkalnych .....	59
<b>Jerzy Korzeniewski</b> , Selekcja zmiennych w klasyfikacji – propozycja algorytmu .....	69
<b>Sabina Denkowska</b> , Testowanie wielokrotne przy weryfikacji wieloczynnikowych modeli proporcjonalnego hazardu Coxa.....	76
<b>Ewa Chodakowska</b> , Teoria równań strukturalnych w klasyfikacji zmiennych jawnych i ukrytych według charakteru ich wzajemnych oddziaływań .....	85
<b>Iwona Konarzewska</b> , Model PCA dla rynku akcji – studium przypadku .....	94
<b>Katarzyna Wójcik, Janusz Tuchowski</b> , Dobór optymalnego zestawu słów istotnych w opiniach konsumentów na potrzeby ich automatycznej analizy	106
<b>Aleksandra Łuczak</b> , Zastosowanie metody AHP-LP do oceny ważności determinant rozwoju społeczno-gospodarczego w jednostkach administracyjnych .....	116
<b>Aleksandra Witkowska, Marek Witkowski</b> , Klasyfikacja pozycyjna banków spółdzielczych według stanu ich kondycji finansowej w ujęciu dynamicznym .....	126
<b>Adam Depta</b> , Zastosowanie analizy korespondencji do oceny jakości życia ludności na podstawie kwestionariusza SF-36v2 .....	135
<b>Marek Lubicz, Maciej Zięba, Konrad Pawelczyk, Adam Rzechonek, Marek Marciniak, Jerzy Kołodziej</b> , Indukcja reguł dla danych niekompletnych i niezbalansowanych: modele klasyfikatorów i próba ich zastosowania do predykcji ryzyka operacyjnego w torakochirurgii .....	146

<b>Małgorzata Misztal</b> , Wybrane metody oceny jakości klasyfikatorów – przegląd i przykłady zastosowań.....	156
<b>Anna M. Olszewska</b> , Wykorzystanie wybranych metod taksonomicznych do oceny potencjału innowacyjnego województw .....	167
<b>Iwona Bąk</b> , Porównanie jakości grupowań powiatów województwa zachodniopomorskiego pod względem atrakcyjności turystycznej.....	177
<b>Agnieszka Kozera, Joanna Stanisławska, Romana Głowicka-Wołoszyn</b> , Segmentacja gospodarstw domowych według wydatków na turystykę zorganizowaną.....	186
<b>Agnieszka Wałęga</b> , Podejście syntetyczne w analizie spójności ekonomicznej gospodarstw domowych.....	196
<b>Joanna Banaś, Małgorzata Machowska-Szewczyk, Bożena Mroczek</b> , Zastosowanie analizy korespondencji do badania wpływu elektrowni wiatrowych na jakość życia ludności .....	205
<b>Joanna Banaś, Krzysztof Małecki</b> , Klasyfikacja punktów pomiarów ankietowych kierowców na granicy Szczecina z wykorzystaniem zmiennych symbolicznych.....	214
<b>Aneta Becker</b> , Wykorzystanie informacji granularnej w analizie wymagań rynku pracy.....	222
<b>Katarzyna Cheba, Joanna Holub-Iwan</b> , Wykorzystanie analizy korespondencji w segmentacji rynku usług medycznych.....	230
<b>Adam Depta, Iwona Staniec</b> , Identyfikacja czynników decydujących o jakości życia studentów łódzkich uczelni.....	238
<b>Katarzyna Dębowska, Jarosław Kilon</b> , Reguły asocjacyjne w analizie wyników badań metodą Delphi.....	247
<b>Anna Domagała</b> , O wykorzystaniu analizy głównych składowych w metodzie <i>Data Envelopment Analysis</i> .....	254
<b>Alicja Grześkowiak</b> , Analiza wykluczenia cyfrowego w Polsce w ujęciu indywidualnym i regionalnym.....	264
<b>Anna M. Olszewska, Anna Gryko-Nikitin</b> , Pomiar postrzegania jakości kształcenia uczelni wyższej na danych porządkowych z wykorzystaniem środowiska R.....	273
<b>Karolina Paradysz</b> , Hierarchiczna metoda grupowania powiatów jako podejście benchmarkowe w ocenie bezrobocia według BAEL-u w wybranych typach małych obszarów .....	282
<b>Radosław Pietrzyk</b> , Porównanie metod pomiaru efektywności zarządzania portfelami funduszy inwestycyjnych.....	290
<b>Agnieszka Przedborska, Małgorzata Misztal</b> , Wybrane metody statystyki wielowymiarowej w ocenie skuteczności terapeutycznej głębokiej stymulacji elektromagnetycznej u pacjentów z chorobą zwyrodnieniową stawów.....	299

<b>Wojciech Roszka, Marcin Szymkowiak</b> , Podejście kalibracyjne w statystycznej integracji danych .....	308
<b>Iwona Skrodzka</b> , Zastosowanie wybranych metod klasyfikacji do analizy kapitału ludzkiego krajów Unii Europejskiej .....	316
<b>Agnieszka Stanimir</b> , Wielowymiarowa analiza czynników sprzyjających włączeniu społecznemu .....	326
<b>Dorota Strózik, Tomasz Strózik</b> , Przestrzenne zróżnicowanie poziomu życia w województwie wielkopolskim.....	334
<b>Izabela Szamrej-Baran</b> , Identyfikacja przyczyn ubóstwa energetycznego w Polsce przy wykorzystaniu modelowania miękkiego.....	343
<b>Janusz Tuchowski, Katarzyna Wójcik</b> , Klasyfikacja obiektów w systemie Krajowych Ram Kwalifikacji opisanych za pomocą ontologii .....	353
<b>Aleksandra Matuszewska-Janica</b> , Grupowanie krajów Unii Europejskiej ze względu na poziom feminizacji sektorów gospodarczych .....	361
<b>Monika Rozkrut, Dominik Rozkrut</b> , Identyfikacja strategii innowacyjnych przedsiębiorstw usługowych w Polsce .....	369

## Summaries

<b>Małgorzata Rószkiewicz</b> , The use of meta-analysis in building the measurement model in case of the absence of measurement invariance on the example of measuring of life satisfaction.....	20
<b>Elżbieta Sobczak</b> , Harmonious smart growth of European Union regions.....	29
<b>Ewa Roszkowska, Renata Karwowska</b> , The comparative analysis of Polish voivodeships with respect to sustainable development in 2010.....	40
<b>Tadeusz Kufel, Magdalena Osińska, Marcin Błażejowski, Paweł Kufel</b> , Comparative analysis of chosen filters in business cycles analysis .....	50
<b>Marcin Salamaga</b> , The attempt of construction of the life tables for opera works on the example of the Metropolitan Opera .....	58
<b>Iwona Foryś</b> , Using discriminant analysis to select similar markets in non-residential property valuation process.....	68
<b>Jerzy Korzeniewski</b> , Variable selection in classification – algorithm proposal .....	75
<b>Sabina Denkowska</b> , Multiple testing in the verification process of multifactorial Cox proportional hazards models .....	84
<b>Ewa Chodakowska</b> , The theory of structural equations modelling in the classification of observed variables and latent constructs according to the character of their relationship.....	93
<b>Iwona Konarzewska</b> , Modelling stock market by PCA factor model – case study .....	105

<b>Katarzyna Wójcik, Janusz Tuchowski</b> , Selection of the optimal set of relevant words in consumers opinions in the context of the opinion mining ..	115
<b>Aleksandra Łuczak</b> , Application of AHP-LP to the evaluation of importance of determinants of socio-economic development in the administrative units .....	125
<b>Aleksandra Witkowska, Marek Witkowski</b> , A dynamic approach to the ranking of cooperative banks by their financial condition .....	134
<b>Adam Depta</b> , Application of correspondence analysis for the measurement of quality of life – questionnaire SF-36v2 based research .....	145
<b>Marek Lubicz, Maciej Zięba, Konrad Pawelczyk, Adam Rzechonek, Marek Marciniak, Jerzy Kołodziej</b> , Classification rules extraction for missing and imbalance data: models of classifiers and initial results in the rules-based thoracic surgery risk prediction.....	155
<b>Małgorzata Misztal</b> , Selected methods for assessing the performance of classifiers – an overview and examples of applications.....	166
<b>Anna M. Olszewska</b> , The application of selected quantitative methods to the evaluation of voivodeship innovation level potential.....	176
<b>Iwona Bąk</b> , The comparison of the quality of groupings of poviats of West Pomeranian Voivodeship in terms of tourism attractiveness .....	185
<b>Agnieszka Kozera, Joanna Stanisławska, Romana Głowicka-Wołoszyn</b> , Household segmentation with respect to the expenditure on organized tourism.....	195
<b>Agnieszka Wałęga</b> , Synthetic approach in the analysis of economic coherence of households .....	204
<b>Joanna Banaś, Małgorzata Machowska-Szewczyk, Bożena Mroczek</b> , Using the correspondence analysis to examine the impact of wind turbines on the quality of life.....	213
<b>Joanna Banaś, Krzysztof Małecki</b> , Classification of measurement survey points of drivers on the boundary of Szczecin using symbolic variables...	221
<b>Aneta Becker</b> , The use granular information in the analysis of the requirements of the labor market.....	229
<b>Katarzyna Cheba, Joanna Hołub-Iwan</b> , The application of the correspondence analysis of patients segmentation on the medical service market .....	237
<b>Adam Depta, Iwona Staniec</b> , Identification of the factors that determine the quality of students life at universities in Lodz.....	246
<b>Katarzyna Dębkowska, Jarosław Kilon</b> , Association rules in the analysis of research results the Delphi method .....	253
<b>Anna Domagała</b> , About using Principal Component Analysis in Data Envelopment Analysis .....	263
<b>Alicja Grześkowiak</b> , Analysis of the digital divide in Poland at the individual and regional level .....	272

<b>Anna M. Olszewska, Anna Gryko-Nikitin</b> , Assessment of perception of quality of teaching at an institution of higher learning based on the ordinal data with the utilization of R environment.....	281
<b>Karolina Paradysz</b> , The hierarchical method of grouping poviats as a benchmark approach in the assessment of unemployment by BAEL in selected types of small areas .....	289
<b>Radosław Pietrzyk</b> , Comparison of methods of measuring the performance of investment funds portfolios.....	298
<b>Agnieszka Przedborska, Małgorzata Misztal</b> , Selected multivariate statistical analysis methods in the evaluation of efficacy of deep electromagnetic stimulation in patients with degenerative joint disease .....	307
<b>Wojciech Roszka, Marcin Szymkowiak</b> , A calibration approach in statistical data integration .....	315
<b>Iwona Skrodzka</b> , Application of some methods of classification to the analysis of human capital in the European Union.....	325
<b>Agnieszka Stanimir</b> , Multivariate analysis of social inclusion factors.....	333
<b>Dorota Strózik, Tomasz Strózik</b> , Spatial differentiation of the standard of living in Great Poland Voivodeship .....	342
<b>Izabela Szamrej-Baran</b> , Identification of fuel poverty causes in Poland using soft modelling .....	352
<b>Janusz Tuchowski, Katarzyna Wójcik</b> , Classification of objects in the National Classification Framework described by the ontology.....	360
<b>Aleksandra Matuszewska-Janica</b> , Clustering of European Union states taking into consideration the levels of feminization of economic sectors..	368
<b>Monika Rozkrut, Dominik Rozkrut</b> , Identification of service sector innovation strategies in Poland.....	379

**Janusz Tuchowski, Katarzyna Wójcik**

Uniwersytet Ekonomiczny w Krakowie

---

## **KLASYFIKACJA OBIEKTÓW W SYSTEMIE KRAJOWYCH RAM KWALIFIKACJI OPISANYCH ZA POMOCĄ ONTOLOGII**

---

**Streszczenie:** W artykule podjęto próbę wykorzystania wiedzy dziedzinowej do klasyfikacji obiektów występujących w Krajowych Ramach Kwalifikacji. Pierwszym etapem było zdefiniowanie ontologii pozwalającej na opis efektów kształcenia oraz programów kształcenia na poziomie szkolnictwa wyższego. W oparciu o tę ontologię przeprowadzono analizę programów kształcenia, proponując miarę podobieństwa programów kształcenia (na poziomie przedmiotów). Model został zbudowany z wykorzystaniem języka RDF/OWL, a do jego analizy użyto algorytmów zapisanych w języku Java. Jako dane posłużyły rzeczywiste informacje pochodzące z systemu Krajowych Ram Kwalifikacji wdrożonego na Uniwersytecie Ekonomicznym w Krakowie.

**Słowa kluczowe:** ontologia, data-mining, podobieństwo, taksonomia, sieć semantyczna.

### **1. Wstęp**

Krajowe Ramy Kwalifikacji (KRK) to podstawowy element reformy szkolnictwa w Polsce [Rozporządzenie 2011]. Pozwalają one na definiowanie efektów kształcenia, które powinny zostać zrealizowane na poszczególnych poziomach kształcenia. Efekty kształcenia tworzą strukturę hierarchiczną reprezentującą wiedzę, umiejętności oraz kompetencje społeczne. Definiują one cele procesu kształcenia, a do ich realizacji służą programy kształcenia. Głównym celem KRK jest zwiększenie przejrzystości systemów szkolnictwa, w tym szkolnictwa wyższego, a co za tym idzie – mobilności osób uczących się, sposobów wydawania dyplomów oraz ułatwień w uznawaniu kwalifikacji absolwentów. System zapewnia również jednolity sposób opisanie kwalifikacji zdobytych na każdym etapie uczenia się. W szkolnictwie wyższym opisy kształcenia, oferowanego studentom przez uczelnię, sformułowane zostały za pomocą wymagań, jakim powinien sprostać student po ukończeniu nauki w ramach danego cyklu kształcenia. Istotnym elementem w poprawie jakości procesu dydaktycznego staje się opracowanie systemu modelującego szeroko rozumiane programy kształcenia (plany studiów, przedmioty itp.).



Głównym celem artykułu jest klasyfikacja obiektów występujących w KRK (w tym przypadku przedmiotów). Zastosowane podejście wykorzystuje wiedzę dziedzinową reprezentowaną przez ontologię.

## 2. Metodologia badań

Zasoby informacyjne przygotowywane z myślą o bezpośrednim wykorzystaniu przez człowieka występują najczęściej w postaci tekstowej, liczbowej oraz graficznej. Są one prezentowane w bardzo przystępnej formie dla człowieka, ale taka postać nie jest dogodna do zautomatyzowania procesów przetwarzania.

Problem pogodzenia tych, w dużym stopniu sprzecznych, wymogów dotyczących postaci zasobów informacyjnych preferowanych przez człowieka oraz przez systemy komputerowe jest częściowo rozwiązywany przez stopniowe wzbogacanie tekstowej reprezentacji danych metainformacjami ułatwiającymi automatyczne przetwarzanie.

Wykorzystanie wiedzy dziedzinowej w procesie analizy danych zwiększa dokładność algorytmów wyszukujących czy też porównujących. Zmiany widoczne w sposobie udostępniania informacji określić można jako zastępowanie tekstowej reprezentacji danych (dokumenty) przez sieci semantyczne.

### 2.1. Sieć semantyczna i ontologie

Sieć semantyczna (*Semantic Web*) to sieć zawierająca elementy posiadające swoje znaczenie, zrozumiałe nie tylko dla ludzi, ale także dla aplikacji. Sama idea umożliwia automatyzację procesów przetwarzania wiedzy wraz z późniejszym wnioskowaniem pozwalającym dostarczyć odbiorcom inteligentnych usług. Najczęściej wykorzystywanymi definicjami związanymi z koncepcją sieci semantycznej są takie sformułowania, jak „dane czytelne dla maszyn” [Feigenbaum i in. 2007], „inteligentni agenci, „rozproszona baza danych”, „automatyczna infrastruktura” [Berners-Lee i in., 2001] czy też „adnotacje”.

Architektura sieci semantycznej to wielowarstwowa hierarchia wykorzystująca pokaźną liczbę standardów i technologii informatycznych zarówno już dostępnych (URI, XML, N3, RDF, OWL, SPARQL) jak i dopiero opracowywanych (RIF, inteligentni agenci, certyfikaty).

Sama sieć semantyczna obejmuje kilka zagadnień, między innymi:

- a) definicje pojęć i obiektów (w tym również złożonych),
- b) opis relacji pomiędzy pojęciami i obiektami,
- c) sposób reprezentacji wiedzy dziedzinowej przy wykorzystaniu ontologii,
- d) mechanizmy wnioskowania.

Opisana przez sieć semantyczną wiedza opiera się głównie na definicjach różnego rodzaju konceptów wykorzystywanych w danej dziedzinie (np. dane geopolityczne: koncept Kraj). Koncepty, inaczej nazywane pojęciami, mogą tworzyć po-

między sobą taksonomiczną hierarchię, zazwyczaj zbudowaną na zasadzie specjalizacji/generalizacji (np. koncept Region jest specjalizacją konceptu Kraj). Na bazie tak zdefiniowanych pojęć, które stanowią tylko abstrakcyjną specyfikację, możliwe jest tworzenie konkretnych obiektów (np. obiekt Polska zbudowany w oparciu o koncept Kraj).

Pomiędzy pojęciami i obiektami zachodzą relacje określające powiązania pomiędzy konkretnymi elementami. Najczęściej wykorzystywanym typem relacji jest relacja pomiędzy dwoma obiektami. Przykładowo obiekt Polska jest w relacji z obiektem Niemcy (relacja *graniczy*). Inne rodzaje relacji to między innymi relacje pomiędzy różnymi pojęciami (np. relacja między konceptem Kraj i Organizacja – relacja *jest Członkiem*) oraz relacje pomiędzy obiektami i pojęciami.

Wiedza domenowa przedstawiana jest głównie w postaci ontologii. Samo pojęcie ontologii zostało zaczerpnięte z filozofii, gdzie oznacza między innymi analizę pojęć i idei w celu ustalenia, co istnieje oraz jakie związki zachodzą pomiędzy istniejącymi elementami. Na potrzeby informatyki termin ontologii oznacza: „formalną specyfikację konceptualizacji pewnego obszaru wiedzy” [Gruber 1993]. Rozwijając termin podstawowy, ontologię można traktować jako reprezentację pewnej dziedziny wiedzy, na którą składa się zapis zbiorów pojęć i relacji między nimi. Pojęcia mogą mieć właściwości w postaci atrybutów, a instancje są traktowane jako reprezentacja obiektów rzeczywistych [Lula, Paliwoda-Pękosz 2008].

Przy projektowaniu ontologii wykorzystywane są metody kategoryzacji i hierarchizacji. Pewnym pojęciom abstrakcyjnym i grupom obiektów, mającym wspólne cechy, przyporządkowywane są nazwy (w ten sposób tworzone są klasy). Uzyskane klasy umieszczane są w strukturze hierarchicznej.

## 2.2. Podobieństwo obiektów

Podstawowym zagadnieniem rozpatrywanym w analizie danych, dostępnych w sieciach semantycznych, jest problem obliczenia podobieństwa lub odległości pomiędzy badanymi pojęciami czy też obiektami.

Porównywanie obiektów opisanych przez sieci semantyczne można rozpatrywać w aspekcie porównywania obiektów reprezentowanych przez wektory cech, porównywania hierarchicznej struktury pojęć uwzględnianych w ontologiach oraz porównywania relacji pomiędzy obiektami. Na potrzeby obliczenia podobieństwa całkowitego zostały zdefiniowane trzy rodzaje podobieństwa [Maedche, Zacharias 2001]:

- podobieństwo strukturalne (taksonomiczne) (*TS*) – podobieństwo obiektów opierające się na ich przynależności do hierarchii konceptów,
- podobieństwo relacyjne (*RS*) – podobieństwo obiektów na bazie ich relacji z innymi obiektami,
- podobieństwo atrybutów (*AS*) – podobieństwo obiektów związane z rodzajem i wartościami tych atrybutów.

Do obliczenia podobieństwa strukturalnego oraz relacyjnego zostały wykorzystane głównie miary podobieństwa przeznaczone dla grafów oraz drzew. Miary te bazują głównie na odległości edycyjnej, maksymalnym wspólnym podgrafie, minimalnym wspólnym nadgrafie oraz modelu przestrzeni wektorowej.

Dodatkowo do obliczenia podobieństwa atrybutów wykorzystane zostały miary związane z wartościami liczbowymi, łańcuchami znaków, tekstami, zbiorami oraz sekwencjami.

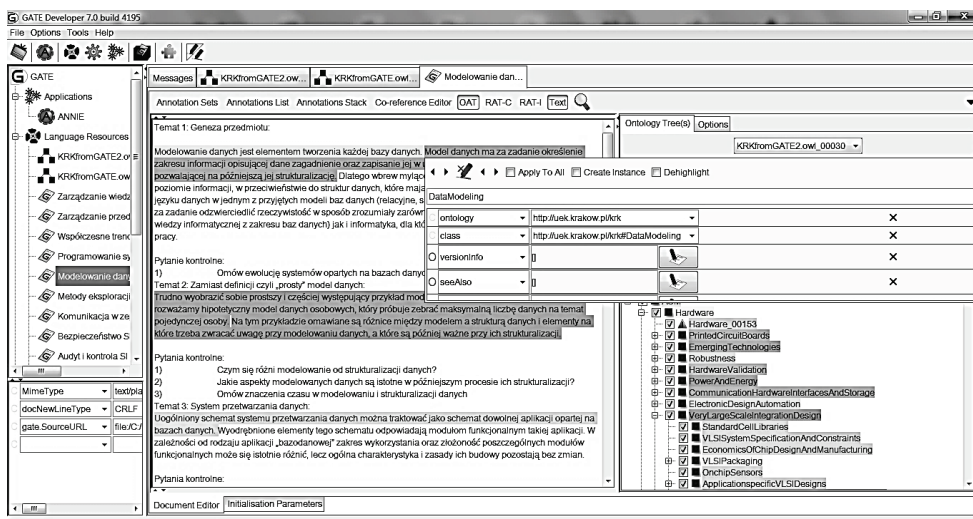
Podobieństwo całkowite  $sim(I_1, I_2)$  pomiędzy dwoma obiektami zostało obliczone na podstawie zagregowania podobieństw cząstkowych [Lula, Paliwoda-Pękosz 2008].

$$sim(I_1, I_2) = f_{agr}(TS(I_1, I_2), RS(I_1, I_2), AS(I_1, I_2)), \quad (1)$$

gdzie  $I_1, I_2$  to obiekty (instancje) brane pod uwagę przy obliczaniu podobieństwa, a  $f_{agr}$  to funkcja agregująca.

### 3. Badania empiryczne

Jako materiał badawczy wykorzystano rzeczywiste dane z kart przedmiotów, pozyskane z systemu Krajowych Ram Kwalifikacji wdrożonego na Uniwersytecie Ekonomicznym w Krakowie. Wybrano dziesięć przedmiotów prowadzonych na kierunku informatyka stosowana. Ekstrakcja danych polegała na wyciągnięciu opisów



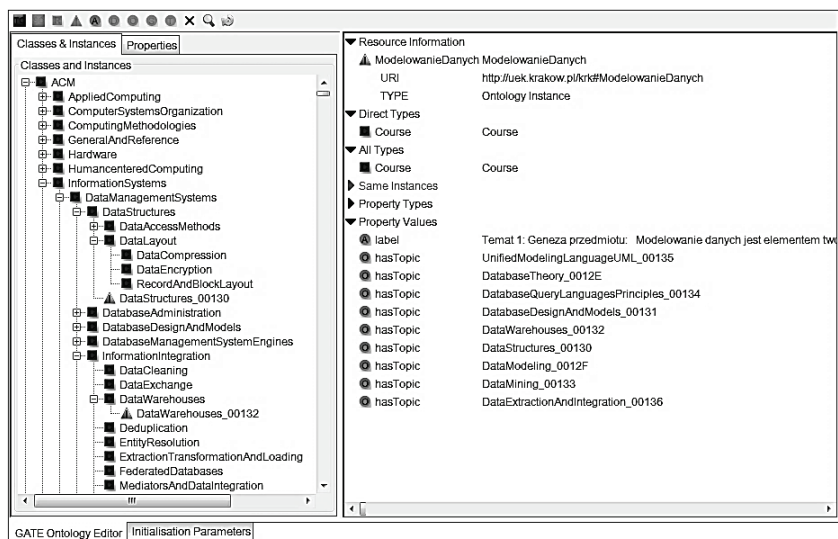
Rys. 1. Proces otagowania danych

Źródło: opracowanie własne – aplikacja GATE.

przedmiotów z plików HTML i zapisu ich w formie czystego tekstu. Wiedza domenowa w postaci ontologii została zbudowana na bazie dostępnego w sieci systemu klasyfikacji pojęć informatycznych ACM<sup>1</sup> (*Association for Computing Machinery*). Zastosowano ręczną konwersję danych ze standardu SKOS (*Simple Knowledge Organization System*) do przyjętego dla ontologii formatu OWL.

Przygotowane dane zostały wprowadzone do programu GATE<sup>2</sup> (*General Architecture for Text Engineering*) i poddane procesowi otagowania (rys. 1).

Każdy przedmiot stanowił instancję klasy **Course** i wchodził w relację *hasTopic* z wybranymi przez użytkownika instancjami klas ontologii **ACM** (rys. 2).



Rys. 2. Przykładowa instancja klasy Course wraz z drzewem klasyfikacyjnym ACM

Źródło: opracowanie własne – aplikacja GATE.

Zbudowana w ten sposób ontologia końcowa została wprowadzona do autorskiej aplikacji OBCAS (*Ontology Based Clustering Analysis System*) [Tuchowski i in. 2011], wykorzystującej biblioteki SimPack<sup>3</sup> oraz Jena<sup>4</sup>. Zadaniem aplikacji było policzenie podobieństwa taksonomicznego pomiędzy klasami, z którymi badane instancje wchodzi w relacje. Do obliczeń wykorzystano miarę podobieństwa Dekang Lin [Lin 1998]:

$$\text{sim}(C_1, C_2) = \frac{2 \log(P(C_0))}{\log(P(C_1)) + \log(P(C_2))}, \quad (2)$$

<sup>1</sup> <http://www.acm.org/about/class/2012> (3.07.2013).

<sup>2</sup> <http://gate.ac.uk/> (7.07.2013).

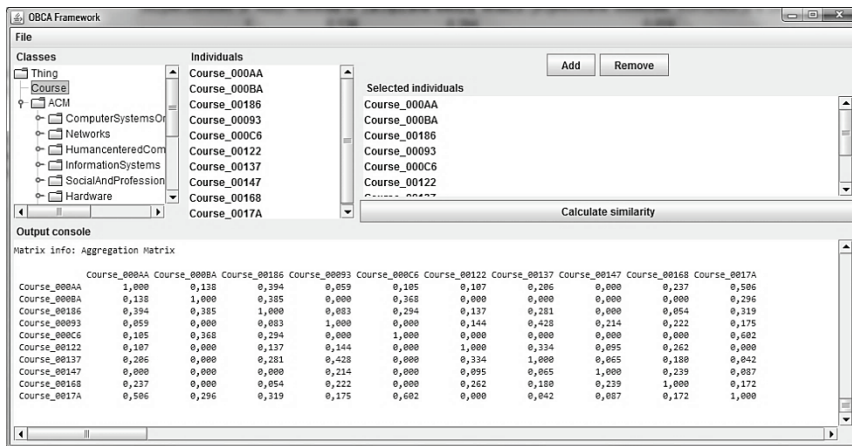
<sup>3</sup> <https://files.ifi.uzh.ch/ddis/oldweb/ddis/research/simpack/> (12-07-2013).

<sup>4</sup> <http://jena.apache.org/> (25-06-2013).

gdzie  $P(C_1)$ ,  $P(C_2)$  to prawdopodobieństwa wystąpienia danej klasy, a  $P(C_0)$  to prawdopodobieństwo wystąpienia najbliższej wspólnej klasy nadrzędnej. Poszczególne prawdopodobieństwa zostały obliczone na podstawie wzoru:

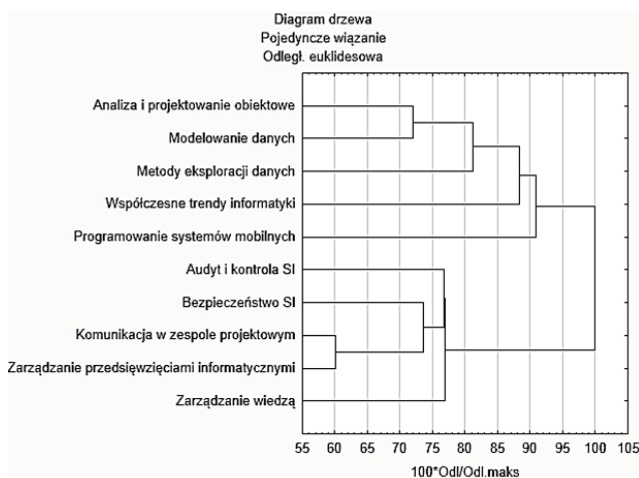
$$P(C) = \frac{1+z}{n} \tag{3}$$

Gdzie  $z$  to liczba dzieci danej klasy, a  $n$  liczba wszystkich klas w ontologii. Jako korzeń przyjęto klasę **ACM**. Wyniki zostały zagregowane do jednej macierzy podobieństwa (rys. 3).



Rys. 3. Zagregowana macierz wyników – podobieństwo pomiędzy obiektami

Źródło: opracowanie własne – aplikacja OBCA.



Rys. 4. Otrzymane wyniki badań w formie dendrogramu

Źródło: opracowanie własne.

Otrzymane wyniki pozwoliły określić podobieństwo badanych obiektów (rys. 4).

Na dendrogramie (rys. 4) wyraźnie widać podział pomiędzy przedmiotami miękkimi a bardziej ścisłymi. Podział ten dość dobrze odzwierciedla rzeczywiste podobieństwo pomiędzy porównywanymi przedmiotami.

#### 4. Podsumowanie

Otrzymane wyniki badań pozwalają stwierdzić, że analiza porównawcza z wykorzystaniem ontologii jako bazy wiedzy przynosi wymierne korzyści. Zaproponowane rozwiązanie wymaga dalszych analiz uwzględniających większe zbiory danych wzbogacone o dodatkowe atrybuty badanych obiektów. Istotnym elementem wydaje się budowa jednolitej bazy wiedzy z wszystkich dziedzin nauki oraz przeniesienie etapu oznaczania poszczególnych przedmiotów z aplikacji na osoby tworzące sylabusy do przedmiotów. Ważną cechą wykorzystanego podejścia jest elastyczność wyboru różnych miar podobieństwa związanych z ontologiami.

Zaproponowane rozwiązanie ma wymiar praktyczny i może być wykorzystywane przez uczelnie w procesie tworzenia planów studiów budowanych zgodnie z wytycznymi KRK. Przykładowo pozwala ono na wyeliminowanie nadmiernego powtarzania się treści kształcenia na różnych przedmiotach w ramach jednego kierunku studiów.

#### Literatura

- Berners-Lee T., Hendler J., Lassila O. (2001), *The Semantic Web*, „Scientific American” 284, s. 34-43.
- Feigenbaum L., Herman I., Hongsermeier T., Neumann E., Stephens S. (2007), *The Semantic Web in action*, „Scientific American” 297, s. 64-71.
- Gruber T.R. (1993), *A translation approach to portable ontology specifications*, „Knowledge Acquisition” 5, s. 199-220.
- Lin D. (1998), *An information-theoretic definition of similarity*, Proceedings of the 15th International Conference on Machine Learning, vol. 1, s. 296-304.
- Lula P., Paliwoda-Pękosz G. (2008), *An ontology-based cluster analysis framework*, Proceedings of the first international workshop on Ontology-supported business intelligence – OBI '08 1-6.
- Maedche A., Zacharias V. (2002), *Clustering Ontology-based Metadata in the Semantic Web*, Principles of Data Mining and Knowledge Discovery, Springer, Berlin – Heidelberg, s. 348-360.
- Rozporządzenie (2011), Rozporządzenie Ministra Nauki i Szkolnictwa Wyższego z dnia 2 listopada 2011 r. w sprawie Krajowych Ram Kwalifikacji dla Szkolnictwa Wyższego (Dz.U. nr 253, poz.1520).
- Tuchowski J., Wójcik K., Paliwoda-Pękosz G., Lula P. (2011), *OBCAS – Ontology Based Cluster Analysis System*, Research in Systems Analysis and Design: Models and Methods, Springer Berlin Heidelberg, s. 106-112.

## **CLASSIFICATION OF OBJECTS IN THE NATIONAL CLASSIFICATION FRAMEWORK DESCRIBED BY THE ONTOLOGY**

**Summary:** This paper is an attempt to use domain knowledge to classify objects from the National Qualifications Framework. The first step was to define an ontology allowing to define the education effects and education programs at higher level of education. On the basis of this ontology an analysis of the learning programs was conducted offering education programs similarity measure (at the level of courses). The model was built using the RDF/OWL language, and its analysis was conducted using algorithms written in Java. The data that were used are actual ones from the National Qualifications Framework system implemented at the Cracow University of Economics.

**Keywords:** ontology, data-mining, similarity, taxonomy, semantic web.