METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XI/2, 2010, str. 120–129

INTEGRACJA METOD I DANYCH W OTWARTYM SYSTEMIE WIELOASPEKTOWEJ ANALIZY PORÓWNAWCZEJ

Tomasz Dudek
Zakład Logistyki i Informatyki AM
e-mail: t.dudek@am.szczecin.pl

Streszczenie: Narzędziem wspomagającym wieloaspektową analizę porównawczą są systemy informatyczne, które dysponują metodami analizy i bazują na danych opisujących właściwości porównywanych obiektów. Metody i użyte w nich dane mają charakter heterogeniczny. Ze względu na przedmiot analizy wymagają one integracji. Dodatkowo w celu zwiększenia funkcjonalności i otwartości takich systemów na nowe metody analizy i nowe źródła danych niezbędna jest ich adaptacja. W artykule zaprezentowano koncepcję takiego systemu oraz metodę integracji opartą na ontologiach, ilustrując je przykładowymi zastosowaniami.

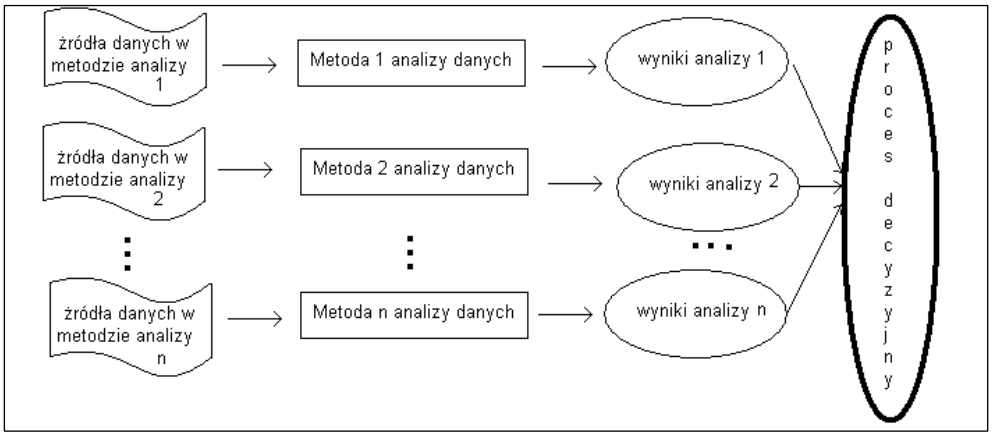
Słowa kluczowe: integracja danych, system wieloaspektowej analizy danych, ontologie, system otwarty, system wspomagania decyzji

WPROWADZENIE

Analiza jest zasadniczym elementem procesu decyzyjnego. W praktyce i teorii wspomagania decyzji istnieje wiele różnych metod analizy, w tym również analizy porównawczej. Metody te wykorzystują różne algorytmy. Każda metoda analizy, wspomagająca procesy decyzyjne wymaga adekwatnych do jej algorytmu danych uważanych za dane źródłowe (wejściowe, zasilające, wstępne). Stosowane w różnych metodach wieloaspektowej analizy porównawczej dane źródłowe mają najczęściej odmienny, heterogeniczny charakter. Istotą danych heterogenicznych jest ich różnorodność nie tylko, co do wartości, ale i również odmienność w obszarze typów, struktur, związków zachodzących miedzy nimi a także źródeł ich pochodzenia. Wyniki analiz realizowanych różnymi metodami często są nieporównywalne. Mają odmienną postać, typ, rodzaj, punkt odniesienia, etc. Oznacza to, że w procesie podejmowania decyzji, decydent wypracowuje decyzję, posługując się wynikami, uzyskanymi z wielu różnych metod. Wyniki te są

efektem przekształceń heterogenicznych źródeł danych. Opisana sytuacja została schematycznie zilustrowana na rysunku 1.

Rysunek 1. Wspomaganie decyzji wynikami wieloaspektowej analizy



Źródło: opracowanie własne

W celu zwiększenia efektywności i usprawnienia procesu podejmowania decyzji niezbędnym wydaje się system, w którym udostępnia się wiele metod analizy oraz metod prezentacji i wizualizacji wyników. Danymi źródłowymi w takim systemie powinny być heterogeniczne dane zasilające dostępne w systemie metody. W informatyce taki problem nazywa się integracją metod i danych w jednym systemie, tzw. systemie zintegrowanym. Ważną cechą takiego systemu powinna być również jego otwartość zarówno na nowe metody analizy, które mogą pojawić się w przyszłości jak również na nowe źródła danych, wynikające z nowych metod analizy. W artykule zaprezentowano architekturę takiego systemu wraz z metodą integracji opartą na ontologiach. Omówiono również przykładowe zastosowanie systemu i zaimplementowaną w tym systemie zasadę integracji danych i metod.

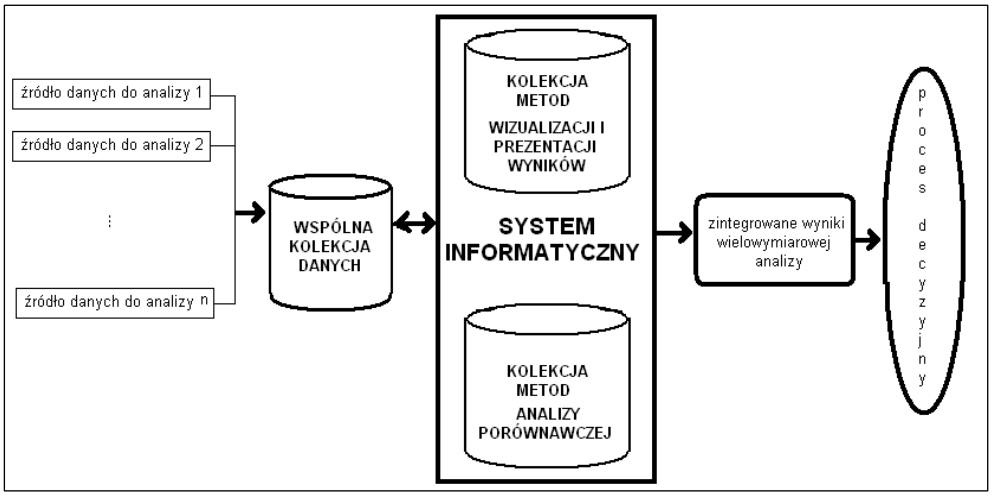
OGÓLNA KONCEPCJA ZINTEGROWANEGO SYSTEMU WSPOMAGAJĄCEGO WIELOASPPEKTOWĄ ANALIZĘ PORÓWNAWCZĄ

Idea integracji w jednym systemie różnych metod wieloaspektowej analizy porównawczej oraz metod prezentacji i wizualizacji wyników tych analiz jest zgodna z rysunkiem 2.

Istotnym elementem zaprezentowanej na rysunku 2 koncepcji systemu wspomagającego wieloaspektową analizę danych w procesie decyzyjnym jest kolekcja danych źródłowych. Ze względu na konieczność dostępu do tej kolekcji,

systemy powinny być oparte na bazach danych i systemach zarządzających tymi bazami.

Rysunek 2. Integracja danych i metod w wieloaspektowej analizy porównawczej, wspomagającej proces podejmowania decyzji



Źródło: opracowanie własne

Systemy takie z natury gromadzą heterogeniczne dane i coraz częściej są wyposażane w algorytmy analitycznego przetwarzania danych (ang. On Line Analytical Processing – OLAP). Możliwości tych systemów są jednak ograniczone i to zarówno w obszarze integracji heterogenicznych źródeł danych jak i w dziedzinie dostępnych algorytmów przetwarzania analitycznego. Integracja heterogenicznych źródeł danych w bazie danych jest możliwa na etapie projektowania systemu. W fazie eksploatacji bazy danych nie ma możliwości wprowadzenia nowych struktur danych, bez zmiany systemu, bez zmiany jego oprogramowania i struktury bazy danych.

Głównym zadaniem obsługi heterogeniczności metod (funkcji) i danych w systemach informatycznych jest budowa takiej zintegrowanej infrastruktury (struktur danych, funkcji, modułów, komponentów, itp.), która umożliwi decydentom korzystającym z wyników analizy prawidłowe użytkowanie systemu przy zachowaniu tej różnorodności. Integrację w dowolnym systemie informatycznym można osiągnąć poprzez integrację danych oraz integrację metod (funkcji) realizowanych w systemie. Wdrażanie technologii informatycznych, integrujących funkcje w dowolnym systemie nawet w dobie rozproszenia systemów jest wsparte wieloma rozwiązaniami zarówno praktycznymi jak i teoretycznymi. Trudniejszym problemem wydaje się integracja danych.

W literaturze [Pankowski, 2001] znane są klasyczne metody integracji danych dla różnych metod (w tym również dla metod wieloaspektowej analizy

danych). Należy do nich metoda integracji wirtualnej oraz integracja poprzez materializację.

Obie te koncepcje integracji danych są trudne implementacyjnie szczególnie wówczas, gdy w systemie pojawi się nowe, dotychczas nieużywane heterogeniczne źródło danych, wynikające choćby z nowej metody analizy. Wynika z tego, że w takim przypadku koncepcje te nie zapewniają otwartości systemu na nowe metody analizy, jakie pojawią się w przyszłości w systemie.

Dlatego innym ważnym obok integracji aspektem systemu wspomagającego wieloaspektową analizę porównawczą w procesie decyzyjnym jest także możliwość zautomatyzowania procesu adaptacji systemu informatycznego do nowych pojawiających się metod analizy i związanych z nimi źródeł danych.

METODA INTEGRACJI Z UŻYCIEM ONTOLOGII

Załóżmy, że w systemie w systemie wspomagającym wieloaspektową analizę porównawczą dostępnych jest *n* źródeł danych symbolicznie oznaczonych jako

$$Z_1, Z_2, ..., Z_n \tag{1}$$

odpowiadających *n* metodom analizy, oznaczonych odpowiednio jako

$$M_1, M_2, ..., M_n \tag{2}$$

W celu zintegrowania danych i metod w jednym systemie wspomagającym wieloaspektową analizę porównawczą w procesie decyzyjnym należy utworzyć wspólny słownik (terminologię) danych. Jest to możliwe dzięki ontologiom lub ontologii. Ontologie klasyfikują wszystkie sfery odpowiadające konkretnym pojęciom, dostarczają kompletnego opisu zjawisk oraz dowodów wymaganych do wykazania prawidłowości twierdzeń. Klasyfikacja taka powinna uwzględniać wszystkie możliwe jednostki, byty czy zjawiska. Ontologie definiują pojęcia i relacji między nimi. Dzięki ontologiom możliwa jest komunikacja bez konieczności operowania na wspólnej bazie wiedzy, oraz na wcześniej zdefiniowanych słownikach. Umożliwiają także efektywny dostęp do informacji zawartych w wielu odrębnych repozytoriach, przez co integrują dane, funkcjonalności, platformy, dziedziny, itp. Użycie ontologii, jako translatorów pojęciowych ułatwia rozwiązanie problemu, jakim jest semantyczna różnorodność danych. Pozwalają także na opis zasobów, które często nie są wiernymi odpowiednikami rzeczywistych obiektów. Zastosowania ontologii wykazują większą swobodę w pozyskiwaniu i utrzymywaniu różnorodnych danych (informacji) dzięki niezależności i niezmienności interpretacyjnej.

Z każdym źródłem danych, użytym w dowolnej metodzie wieloaspektowej analizy danych związana jest określona terminologia (określone nazewnictwo), która wyszczególnia pojęcia używane do opisu uniwersum oraz związki zachodzące pomiędzy tymi pojęciami. Terminologię buduje się poprzez określenie zbioru aksjomatów równoważności oraz aksjomatów podrzędności. Aksjomaty

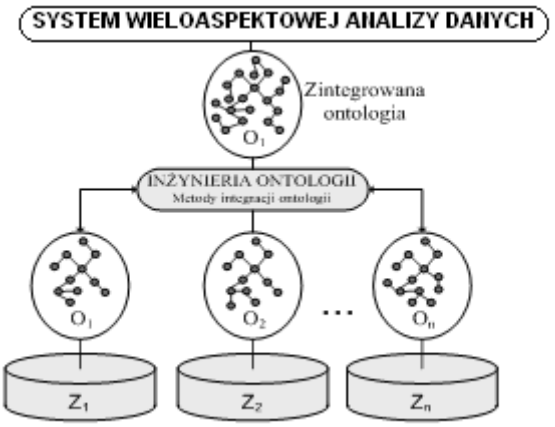
równoważności między konceptami (pojęciami) A i B, symbolicznie zapisane formułą postaci $A \equiv B$, są stwierdzeniami, z których wynika, że zakresy konceptów A i B są równe zaś aksjomaty podrzędności postaci $A \subset B$, stwierdzają, że zakres konceptu A zawiera się w zakresie konceptu B.

Wówczas można dla tych źródeł danych (odnosi się to również do metod analizy) opracować odpowiednie ontologie wg zasady, że O_i jest ontologią dla źródła Z_i *(i=1, ..., n).*

$$O_1, O_2, ..., O_n \tag{3}$$

Z możliwości odwzorowania każdego heterogenicznego źródła danych w metodzie analizy w ontologię wynika wniosek, że problem integracji źródeł danych sprowadza się do konieczności łączenia ze sobą ontologii. Jeśli dostępne są w systemie służącym do wieloaspektowej analizy danych odpowiednie źródła $Z_1 ... Z_n$, a dla nich zbudowano ontologie $O_1, ..., O_n$ (gdzie n oznacza ilość dostępnych źródeł danych do metod dostępnych w systemie), to możliwa jest ich integracja (integracja ontologii odpowiadających tym źródłom danych). Koncepcję takiej integracji przedstawiono schematycznie na rysunku 3.

Rysunek 3. Koncepcja integracji danych w systemie wspomagania wieloaspektową analizę porównawczą



Źródło: opracowanie własne

Wynika z niej, że integracja heterogenicznych źródeł danych oparta na ontologiach jest procesem dwuetapowym. W etapie pierwszym każde integrowane źródło danych jest odwzorowywane w tzw. ontologię cząstkową, a w etapie drugim następuje integracja wielu ontologii cząstkowych w jedną ontologię, tzw. ontologię globalną. Możliwość i dokładność odwzorowania dowolnego źródła danych została praktycznie zweryfikowana licznymi implementacjami w ramach inżynierii ontologii [Dudek T., 2008].

Podstawą tworzenia ontologii O_i dla dowolnego źródła heterogenicznych danych Z_i, (gdzie $i=1, 2, ..., n$) jest opracowanie sieci semantycznej konceptów w postaci grafu G_i. Graf ten jest siecią semantyczną konceptów i stanowi graficzną reprezentację wiedzy dotyczącej wybranego zakresu dziedziny związanej ze źródłem Z_i. Dlatego często nazywa się go grafem powiązań konceptów. Węzłami grafu G_i są koncepty, zaś gałęzie (łuki) grafu charakteryzuje się relacją zachodzącą między konceptami. Wówczas

$$G_i, =< ZK_i, RK_i > \tag{4}$$

gdzie symbolem ZK_i oznaczono zbiór węzłów – konceptów a symbolem RK_i odpowiedni zbiór relacji między węzłami.

Aby metoda integracji danych była użyteczna ważnym zagadnieniem jest wcześniejsze określenie zasad realizacji procesu budowy (modelowania) ontologii dla danych źródłowych (wejściowych) dowolnej metod analizy porównawczej. Do korzyści, jakie można osiągnąć stosując ontologie należy redukcja pojęciowej i terminologicznej złożoności źródeł wykorzystywanych w analizie, ujęcie i śledzenie powiązań między pojęciami, uwidocznienie złożoności związków pomiędzy pojęciami oraz uniknięcie wieloznaczności lub niejasności ontologicznej. Wprowadzając ontologie, których zadaniem jest określenie zależności między pojęciami, osiąga się wymianę danych między metodami.

Integracja ontologii wielu $O_1, ..., O_n$ to proces iteracyjny, w którym początkowo integruje się dwie ontologie a następnie wynik tej operacji integruje się z kolejną ontologią, odpowiadającą kolejnemu źródłu danych i kolejnej metodzie analizy. Aby zintegrować dwie dowolne ontologie O_1 i O_2 zdefiniowane zgodnie ze wzorem (5)

$$O_1 =< T_1, G_1 > \ , \ O_2 =< T_2, G_2 > \tag{5}$$

z dokładnością do terminologii T_1 i T_2 oraz sieci semantycznych konceptów wyrażonych w formie grafów G_1 i G_2 odpowiednio dla tych ontologii, można:

- Przekształcić ontologię O_1 w ontologię O_2,
- Odwzorować ontologię O_1 w ontologię O_2 lub odwrotnie,
- Połączyć ontologię O_1 z O_2 tworząc ontologię OG z odpowiadającą jej terminologią i siecią semantyczną, zgodną ze wzorem (6).

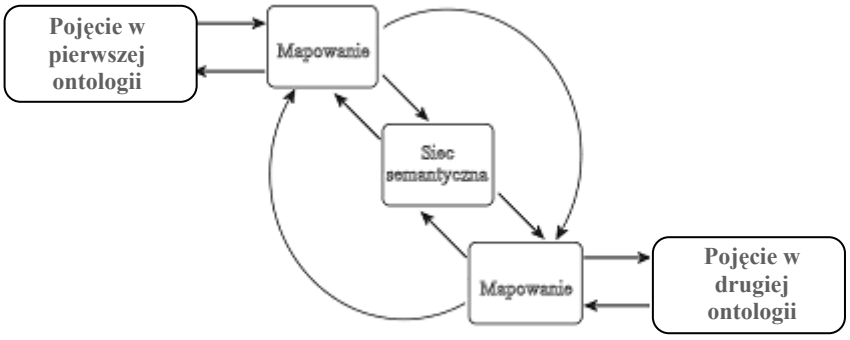
$$OG =< TG, GG > \tag{6}$$

Przekształcenie lub odwzorowanie ontologii O_1 w ontologię O_2 (lub odwrotnie) jest możliwe wówczas, gdy terminologia jednej ontologii zawiera przynajmniej jeden koncept drugiej terminologii. Najczęściej w integracji ontologii stosuje się jednak metodę łączenia i tworzenia nowej ontologii. Łączenie ontologii zwykle realizuje się w dwóch głównych etapach. Pierwszym etapem jest tworzenie terminologii oraz sieci semantycznej danej ontologii. Jeśli ontologia powstanie z połączenia ontologii O_1 i O_2 zdefiniowanych wzorem (4), to należy zdefiniować TT oraz zbudować sieć semantyczną GG posługując się terminologiami T_1, T_2 oraz sieciami G_1 i G_2. Drugim etapem łączenia jest odwzorowanie ontologii OG kolejno w ontologie O_1 i O_2, a następnie ontologie O_1 i O_2 w ontologię OG. Tak

opracowana ontologia *OG*, powstała z połączenia ontologii O_1 i O_2 (zwanych podrzędnymi), nosi nazwę ontologii nadrzędnej. Połączenie dwóch a w rezultacie wielu ontologii i ich sieci semantycznych w jedną całość generuje potrzebę określenia spójnej terminologii odpowiedniej dla każdej z dziedzin. Próba stworzenia sieci globalnej, czyli scalenia dwóch (wielu w procesie iteracyjnym) sieci podrzędnych, lokalnych (cząstkowych), wymaga określenia podobieństwa między rozpatrywanymi w nich pojęciami oraz pojęciami z ich otoczenia. Wymagane jest odnalezienie w łączonych sieciach semantycznych pojęć najbardziej do siebie podobnych (np. o tych samych nazwach) i zbadanie stopnia ich podobieństwa. Jeżeli pojęcia są tożsame to możliwe jest przeprowadzenie tzw. procesu mapowania (odwzorowania), rozumianego jako tworzenie map pojęć w kontekście przekazywania wiedzy. Mapy te składają się z zestawu pojęć, które mogą odpowiadać obiektom rzeczywistym, ich klasom lub obiektom abstrakcyjnym a także relacji występujących między tymi pojęciami. Istnieje także możliwość, w której w jednym kontekście pojęcie będzie się odnosić do konkretnego wystąpienia konceptu (obiektu), a w innym do jego klasy.

Mapy pojęć są także określane mianem odwzorowania pojęć jednej ontologii w pojęcia odpowiednie dla drugiej. Przypadek taki zilustrowano schematycznie na rysunku 4.

Rysunek 4. Mapowanie jako odwzorowanie rzeczywistości



Źródło: opracowanie na podstawie [Uschold M i in. 1996]

Istotnym elementem odwzorowania jest tu mechanizm identyfikacji pojęć, umożliwiający łączenie map pojęć należących zarówna do obu ontologii.

Aby proces mapowania pary ontologii O_1 i O_2 mógł być uznany za prawidłowy, należy dla każdego konceptu ontologii O_1 znaleźć odpowiadający mu, podobny semantycznie koncept lub zbiór konceptów z ontologii O_2 (i odwrotnie).

Utrudnieniem tego procesu może być wiele czynników. Wśród nich wymienić należy niejednolitą przestrzeń stosowania pojęć (stosowanie odmiennych języków opisu semantycznego), niejednolite znaczenie pojęć (różne poziomy

szczegółowości pojęć), różnorodność semantyczną pojęć (stosowanie odmiennych języków naturalnych lub klasyfikacji pojęć) a także niejednolitą reprezentację relacji między pojęciami. W wielu przypadkach proces mapowania oznacza konieczność odnajdowania odpowiedników dla wszystkich pojęć jednej ontologii wśród pojęć drugiej ontologii.

Równie częstym, jest także proces łączenia dwóch odrębnych dziedzinowo i semantycznie repozytoriów wiedzy. Łączenie to można osiągnąć poprzez określenie stopnia podobieństwa pojęć. Umożliwi to określenie pozycji tych pojęć w sieci semantycznej konceptów. Zgodnie z literaturą [Doerr M. 2001] można tu skorzystać z równoważności dwóch pojęć: zupełnej (oba mapowane pojęcia są według eksperta tożsame) i niezupełnej (oba pojęcia posiadają część wspólną) a także z równoważności częściowej (jedno pojęcie jest nadrzędne w stosunku do drugiego) lub złożonej (występuje wówczas, gdy zależności zachodzą dla większej ilości pojęć niż dwa rozważane).

Możliwe jest również użycie nie tylko przekształceń konceptów ontologii globalnej w koncepty lokalne i odwrotnie, ale również dopasowanie konceptów i pojęć do ich synonimów. Należy wówczas zastosować odpowiednie współczynniki zgodności, dopasowania pojęć i synonimów. O wartościach współczynników i użyciu synonimów każdorazowo będzie decydował ekspert.

Proces odwzorowania O_i ontologii cząstkowych w tzw. ontologię globalną OG można funkcjonalnie zrealizować poprzez odwzorowanie matematyczne zgodne ze wzorem:

$$F_i : \ O_i \rightarrow OG \tag{7}$$

gdzie symbolem O_i oznaczono ontologie cząstkowe odpowiedniego źródła danych (i = 1, 2, …, n). Na podobnych zasadach, zgodnie ze wzorem (8) można zdefiniować również odwzorowanie odwrotne, gdzie i ma takie znaczenie jak we wzorze (7).

$$G_i : \ OG \rightarrow O_i \tag{8}$$

Omówione metody integracji dwóch, a w rezultacie wielu ontologii, szczegółowo opisano w literaturze [Dudek T., 2008].

OTWARTY CHARAKTER SYSTEMU

Stosując metodę integracji danych w systemie wspomagającym wieloaspektową analizę porównawczą przy użyciu ontologii można uzyskać otwarty charakter systemu i zrealizować postulat jego adaptacyjności do nowych źródła danych i nowych metod analizy.

Gdy powstaje nowa metoda analizy oraz nowe heterogeniczne źródło danych, to winna być dla danych źródłowych (wejściowych) tej metody zbudowana ontologia O_{new} skojarzona z tym źródłem danych, tzw. ontologia cząstkowa. Można ją powiązać z dotychczas istniejącą ontologią globalną OG na

podobnych zasadach jak dokonano tego w przypadku integracji ontologii $O_1,O_2,..,O_n$ w ontologię OG omówione powyżej. Dowodzi to otwartości systemu.

Aby postulaty otwartości zostały jednak osiągnięte, to system wspomagający wieloaspektową analizę porównawczą powinien być wyposażony w moduł tworzenia ontologii dla dowolnego źródła danych oraz moduł integracji dwóch dowolnych ontologii. Moduły te można z powodzeniem zaimplementować w systemie informatycznym wspomagającym wieloaspektową analizę porównawczą na zasadach podobnych jak w oprogramowaniu OntoStudio.

EFEKTY ZASTOSOWANIA METODY INTEGRACJI

W celu weryfikacji stosowalności zaprezentowanej metody integracji heterogenicznych źródeł danych w wieloaspektowej analizie posłużono się przykładem analizy porównawczej w zakresie oceny jakości kolekcji programów komputerowych. Ocenę wielu różnych programów zrealizowano trzema różnymi metodami (metodykami) firmy Hewlett-Packard, IBM oraz Motorola. Każdy program komputerowy był poddany ocenie co najmniej jedną z tych metod. Wśród rozważanych programów komputerowych były takie, które zostały poddane tylko jednej z tych metod. Dla pełnej i jednorodnej oceny kolekcji programów komputerowych niezbędne było określenie, który program z tej kolekcji jest oceniony lepiej a który gorzej. Odpowiedź na to pytanie wymagała integracji metod i odpowiednich do nich źródeł danych.

W tym celu dla każdej metody oceny programów komputerowych utworzono ontologię. Każda z tych ontologii operowała innymi konceptami, terminologiami, aksjomatami, sieciami semantycznymi oraz grafami powiązań konceptów. Nazwano je ontologiami cząstkowymi.

W celu przeprowadzenia analizy porównawczej jakości oprogramowania ocenianego tymi trzema różnymi metodami opracowano ontologię globalną, integrującą te trzy metody oceny. Wówczas mimo posiadania różnych ocen dla różnych programów komputerowych, możliwe okazało się porównanie tych ocen w jednej terminologii i na jednej płaszczyźnie odniesienia. Nastąpiło to dzięki integracji ontologii cząstkowych, zbudowanych dla źródeł danych związanych z metodami firmy Hewlett-Packard, IBM oraz Motorola (trzy ontologie cząstkowe) w ontologię globalną. Utworzona ontologia globalna w pełni integrowała ontologie cząstkowe i umożliwiła pełną analizę porównawczą w jednej płaszczyźnie odniesienia nawet wówczas, gdy oceny programów komputerowych były dostępne w obszarze jednej z tych metod. Szczegółowo, ten przykład zastosowań został opisany w pracy [Dudek T., 2008].

Zaprezentowana metoda została również zweryfikowana praktycznie na przykładzie analizy porównawczej jakości kształcenia w szkole wyższej. Wyniki tej weryfikacji zaprezentowano w literaturze [Dudek T., 2006].

PODSUMOWANIE

Ważną cechą danych źródłowych zasilających metody wieloaspektowej analizy porównawczej jest ich różnorodność. Każde takie źródło można opisać odpowiednią ontologią cząstkową i wówczas integracja metod i danych je zasilających sprowadza się jedynie do integracji tychże ontologii w tzw. ontologię globalną. Dzięki zastosowaniu metody integracji opartej na ontologiach możliwe jest dodawanie nowych źródeł danych i nowej metody analizy, a dzięki współdzieleniu słowników oraz odwzorowaniu lokalnych ontologii w ontologię globalną możliwe jest porównanie wyników wielu metod wieloaspektowej analizy.

LITERATURA

Doerr M. (2001) Semantic problems of thesaurus mapping, jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr.

Dudek T. (2006) Integracja danych w systemie oceny jakości kształcenia, Badania operacyjne i systemowe, Szczecin.

Dudek T. (2008) Metoda integracji heterogenicznych źródeł danych w ekspertowym systemie oceny jakości, rozprawa doktorska, Szczecin.

Pankowski T. (2001) Integracja i przetwarzanie heterogenicznych źródeł danych w bazach obiektów częściowo etykietowanych, Materiały III Krajowej Konferencji n.t. Metody i systemy komputerowe w badaniach naukowych i projektowaniu inżynierskim, Kraków.

Uschold M., Gruninger M. (1996) Ontologies: Principles, methods and Applications, Knowledge Engineering Review.

Data integration method for multiaspect comparative analysis system

Abstract: Comparative analysis systems are dedicated to maintenance specific types of different (heterogeneous) data and methods. Throughout the integration technology implementation those systems are capable to achieve that goal. What's more the system is fully alterable and can be adjust to the most sophisticated needs. That kind of ontology based solutions were presented in the article.

Key words: data integration, multiaspect system for data analysis, ontology, open system, data support system