XXII Seminarium

ZASTOSOWANIE KOMPUTERÓW W NAUCE I TECHNICE' 2012
Oddział Gdański PTETiS
Referat nr 15

METODA WYBORU INFORMACJI Z DEDYKOWANYCH ZBIORÓW DANYCH

Jerzy KACZMAREK[1], Michał WRÓBEL[2]

1. Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska
   tel: (58) 347 26 82 fax: (58) 347 27 27 e-mail: jkacz@eti.pg.gda.pl
2. Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska
   tel: (58) 347 10 37 fax: (58) 347 27 27 e-mail: wrobel@eti.pg.gda.pl

Streszczenie: Poszukiwanie skutecznych metod wyboru informacji wynika z obserwowanego obecnie nadmiaru danych. W artykule jest opisana metoda GQM (Goal, Question, Metric) wykorzystywana w informatyce między innymi do budowy wielowymiarowej funkcji jakości oprogramowania. Opiera się ona na jawnym zdefiniowaniu celu wyboru danych z uwzględnieniem punktu widzenia użytkownika. W artykule wykazano, że metoda ta może być również wykorzystywana do poszukiwania i wyboru danych z dedykowanych zbiorów informacji dziedzinowych. Metoda pozwala na ograniczenie zbioru danych w ściśle określony sposób z uwzględnieniem cech dziedziny, celu, w jakim dane będą wykorzystywane i konsekwencji wynikających z odrzucenia pozostałych danych. Metoda ta może być również stosowana do wyboru i budowy zorganizowanych w formie taksonomii zbiorów danych opisanych w języku naturalnym. Należy przypuszczać, że skuteczne metody wyboru danych będą w przyszłości powszechnie wykorzystywane w wyszukiwaniu i prezentacji informacji.

Słowa kluczowe: wyszukiwanie danych, GQM, taksonomia.

1. WSTĘP

Obserwowany obecnie nadmiar informacji publikowanych w tradycyjnej papierowej formie, czy też w formie elektronicznej jest wyzwaniem do poszukiwania mechanizmów selekcji i prezentacji danych. Problem nie jest łatwy do rozwiązania, o czym można się przekonać poszukując danych w ogromie informacji zgromadzonej w Internecie. Metody indeksacji, poszukiwanie na podstawie słów kluczowych, opis danych poprzez uzgodnione metadane czy budowa dedykowanych zbiorów danych dziedzinowych są mało skuteczne, ale powszechnie stosowane z braku lepszych rozwiązań.

Zagadnienie wyboru informacji z niezorganizowanego zbioru danych można przeanalizować na dużym poziomie uogólnienia. Zgromadzone na przykład w Internecie ogromne ilości danych nie są problemem, problemem jest poszukiwanie określonego typu danych w określonym celu przez konkretnego użytkownika. Po ich znalezieniu dane stają się użyteczną dla poszukującego informacją.

Punkt widzenia użytkownika przy wyborze danych nie jest prosty do zdefiniowania i przekazywania w systemach wyszukiwawczych. Niestety często nie jest prosty do precyzyjnego określenia nawet dla samego użytkownika.

Złożoność problemu wynika z różnorodności punktów widzenia i wielkiej liczby potencjalnych użytkowników. Tego typu problemy decyzyjne występują również w innych dziedzinach nauki, a rozwiązania tam stosowane mogą być wykorzystane do poszukiwania informacji w dużych zbiorach danych.

Jedną z metod, którą można zastosować do wyboru danych jest metoda oparta na celu, perspektywie i środowisku znana pod nazwą GQM (Goal Question Metric). Stosowana jest od dawna do wyboru atrybutów w wielowymiarowej przestrzeni jakości. Polega ona na jawnym podaniu celu wyboru atrybutów, co implikuje pytania i określa, jakie mają być metryki. Metoda w pewnym sensie jest bezwzględna, należy określić cel, co chce się udowodnić i dobrać takie atrybuty, które potwierdzą postawioną z góry tezę. Metoda uwzględnia jednak ważną kwestię, że różne punkty widzenia implikują różną ocenę rzeczy i zmieniają dobór danych do ich oceny. Wybór atrybutów zmienia się w zależności od celu, punktu widzenia czy typu użytkownika dobierającego dane. Metoda wynika z pewnego podejścia praktycznego stosowanego często w informatyce, że lepszy jest brak informacji, niż ich nadmiar.

Metodę taką, a zwłaszcza jej sformalizowane podejście, można zastosować w wielu innych przypadkach do wyboru adekwatnych do celu danych z pewnego większego zbioru.

W artykule przedstawiono trzy przykłady w formie studiów przypadku. Rozpatrzone zostaną możliwości stosowania metody do wyboru danych edukacyjnych, do budowy taksonomii danych literaturowych w zakresie pewnej dziedziny, oraz do wyboru danych w języku naturalnym opisujących wybrany wycinek rzeczywistości.

W dobie semantycznego Internetu, gdy budowane są różnego rodzaju taksonomie dla wiedzy dziedzinowej, czy ontologie opisujące wybrane wycinki rzeczywistości podejście proponowane przez metodę GQM może przyczynić się do budowy takich zbiorów danych, które będą przydatne większej grupie użytkowników, a nie tylko indywidualnym twórcom takich zbiorów. Należy przypuszczać, że skuteczne metody wyboru danych są i będą w przyszłości powszechnie wykorzystywane ponieważ ułatwiają również graficzną prezentację informacji.

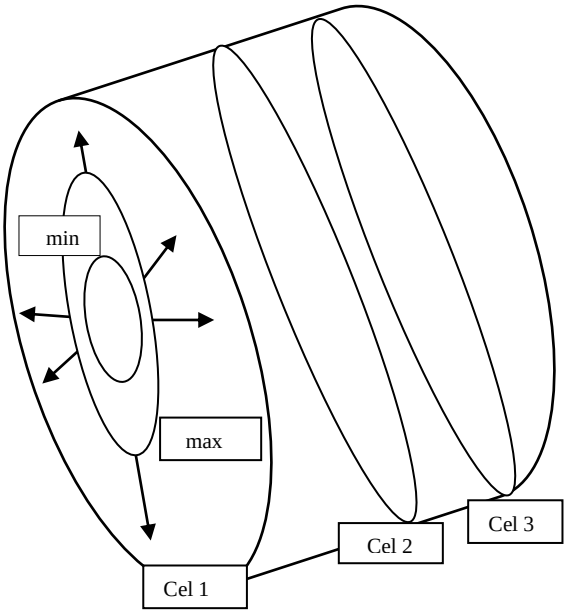
Recenzent: Prof. dr hab. inż. Kazimierz Goczyła – Wydział Elektroniki, Telekomunikacji i Informatyki Politechnika Gdańska

2. METODA GOAL QUESTION METRIC - GQM

W informatyce podobnie jak w innych dziedzinach występują problemy z pomiarem jakości produktów, procesów czy pracowników. Niezdefiniowane pojecie jakości jest pojęciem pustym, subiektywnym i niemierzalnym [1]. Subiektywność ocen jakości i ich niejednoznaczność prowadzi do konfliktów pomiędzy producentami a klientami produktów informatycznych. W rezultacie wprowadzenie pomiarów i ilościowych miar jakości napotykało na trudności lub było wręcz eliminowane. Normy jakości określiły jawnie, że producent i użytkownik powinni zapisać w kontrakcie, w jaki sposób rozumieją jakość i jakie jej atrybuty będą mierzone.

Próbę rozwiązania problemu wyboru atrybutów jakości podjęli V. Basili i D. Rombach z University of Maryland w 1988 roku. Opracowali metodę cel – pytania – metryki, zwaną GQM (ang. Goal Question Metric) [2,3].

Jej głównym założeniem jest jawnie zdefiniowany cel wyboru zbioru atrybutów. Trudności i konflikty pojawiają się bowiem w momencie, gdy zdefiniowany model jakości ma być wspólny dla większej liczby podmiotów, a nie tylko dla jego twórcy. Ten problem ma charakter ogólny i będzie dotyczył innych rozpatrywanych dalej przypadków. Na rysunku 1. przedstawiono pewną przestrzeń mierzalnych atrybutów określających na przykład jakość programu komputerowego. Atrybutów jakości może być bardzo dużo, nawet kilka tysięcy. Powstaje problem, które z nich wybrać do pomiaru i oceny jakości. Wybór to przecięcie płaszczyzną tej przestrzeni atrybutów w zdefiniowany sposób i uzyskanie w płaszczyźnie przecięcia wielowymiarowej funkcji jakości. Na każdej osi musi się znajdować wymiar w jakim dana wielkość jest mierzona, skala określająca zakres pomiaru oraz dwa punkty określające granice akceptacji tej wielkości mierzalnej, uzgodnione przez producenta i konsumenta w procesie negocjacji. Akceptowana jakość produktu występuje wtedy, gdy wszystkie atrybuty znajdują się w granicach akceptacji. Warto podkreślić, że przy ocenie jakości produktów informatycznych przyjmuje się założenie, że wszystkie atrybuty są ortogonalne, niezależne od siebie. Zależności między atrybutami, nawet jeśli wydaje się intuicyjnie, że istnieją, są trudne do określenia.



Rys. 1. Budowa wielowymiarowej funkcji jakości

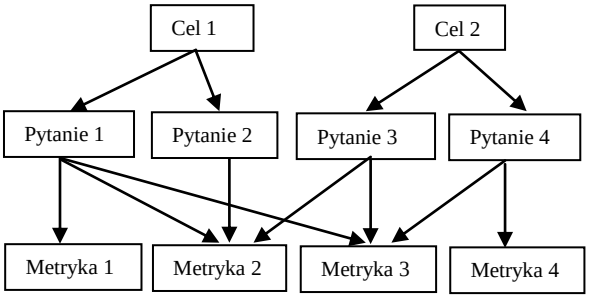
Metoda GQM składa się z trzech poziomów: konceptualnego, w którym określa się cel pomiarów, operacyjnego, w którym zadawane są pytania o istotne charakterystyki produktu oraz ilościowego, w którym dokonuje się pomiarów atrybutów. Warto podkreślić, że niektóre atrybuty są atomowe, jednostkowe, jednoznacznie mierzalne, a inne są złożone i wymagają często ocen eksperta. Po uzyskaniu danych dokonywana jest ocena, czy wszystkie atrybuty znajdują się w granicach akceptacji. W ramach metody można opracować pewien szablon dla definiowania celu w postaci następującej :

Analizować	***(co?)***
z zamiarem	***(dlaczego?)***
w odniesieniu do	***(jaki atrybut?)***
z punktu widzenia	***(czyjego?)***
w środowisku	***(w jakim kontekście?)***

Stosując szablon w dziedzinie nieinformatycznej, na przykład do oceny masła, można zdefiniować z pozoru kontrowersyjny cel, który jednak oddaje istotę metody GQM, jej bezwzględne ukierunkowania na cel.

Szablon celu może mieć postać: przeanalizować masło z zamiarem wykazania, że jest lepsze od margaryny, w odniesieniu do wybranych atrybutów zgodnych z celem - co nie jest trudne do wykonania - z punktu widzenia producenta masła w środowisku rynku polskiego.

Oczywiście zamiana słowa masło na margarynę jest możliwa, choć atrybuty wybrane do udowodnienia tego nowego celu będą inne. Zadawane pytania mogą być typu jak dużo, jak dobrze, długo itp. Dany atrybut jakości może występować w kilku celach, a nawet wynikać z tego samego pytania, zwłaszcza, jeśli jest atrybutem złożonym z kilku atrybutów atomowych, co ilustruje rysunek 2.



Rys.2. Wybór atrybutów dla różnych celów

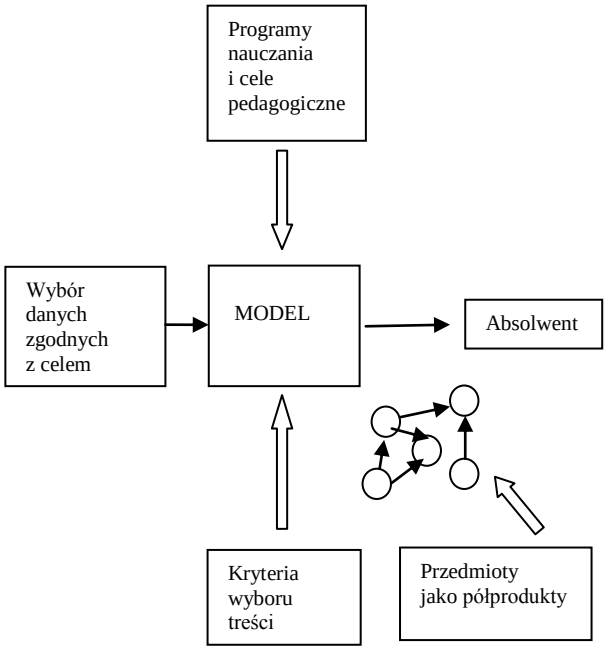
Należy zwrócić uwagę na bezwzględność wyboru atrybutów. Jednoznaczne ukierunkowanie na cel, ta pewna bezwzględność metody GQM, jest też jej zaletą, ponieważ prowadzi do redukcji danych poprzez eliminację danych nie związanych z celem. Tę cechę można wykorzystać w innych dziedzinach zwłaszcza przy wyborze danych z dużych zbiorów danych. Zbiory danych cechuje nadmiarowość, która utrudnia znajdowanie informacji przydatnych dla konkretnego użytkownika. Najlepszym przykładem jest Internet, w którym jest bardzo dużo informacji, ale znalezienie czegoś konkretnego nie jest proste. Poszukiwanie skutecznych metod wyboru informacji, ich selekcji, porządkowania i prezentacji graficznej to obecnie bardzo ważny obszar badawczy współczesnej informatyki. Wydaje się, że metodę GQM można wykorzystać do wyboru informacji z dużych zbiorów danych.

3. WYBÓR DANYCH EDUKACYJNYCH

Jednym z problemów współczesnej edukacji jest wybór danych, które będą nauczane na określonym poziomie kształcenia. Można wykorzystać przedstawiony szablon metody GQM do doboru danych edukacyjnych.

Szablon celu może mieć postać: przeanalizować system edukacyjny z zamiarem poprawy, w odniesieniu do przydatności przekazywanej wiedzy w pracy zawodowej, z punktu widzenia kandydata na studia w środowisku konkretnej uczelni.

W tym celu bardzo istotny jest punkt widzenia, którym może być interes uczącego się, uczelni, pracodawcy, a nawet społeczeństwa. Przydatność wiedzy w pracy zawodowej, a także kompetencje ludzi w danych zawodach są bardzo trudne do zdefiniowania. Z drugiej strony ilość danych, jakich można dostarczać uczącym się, jest praktycznie nieograniczona.



Rys.3. Model procesu edukacji

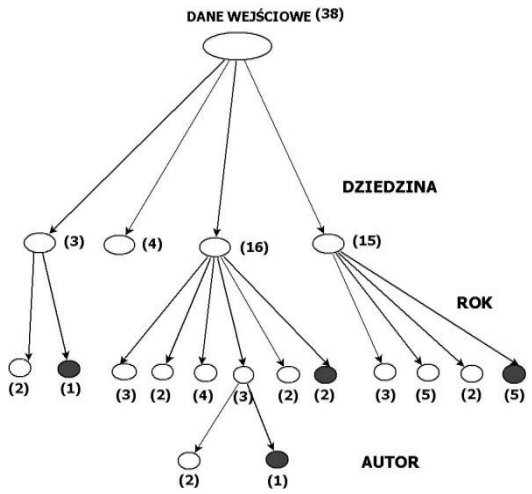
Na rysunku 3. przedstawiono pewien uproszczony model doboru treści edukacyjnych. Podstawą wyboru przedmiotów do nauczania jest model edukacyjny, zależny od planowanej sylwetki absolwenta inny dla inżyniera, lekarza czy aktora. Ale model tylko z pozoru jest oczywisty, o czym świadczą liczne zmiany organizacji studiów i programów nauczania. Metoda GQM zastosowana w tym przypadku wnosi pewne nowe spojrzenie na to zagadnienie. Zdefiniowanie celu i punktu widzenia jednoznacznie i bezwzględnie definiuje dane, jakie powinny być wybrane, eliminując dane zbędne i powtarzające się. Pokazuje również możliwe błędy przy budowie systemu edukacji, kiedy wybiera się obiekty edukacyjne niezgodne ze zdefiniowanym celem.

Sam proces edukacyjny przypomina proces produkcji wyrobu finalnego z komponentów i może być opisany równaniem produkcji. Pozostaje do rozwiązania problem, kto definiuje cel i model edukacyjny i z czyjego punktu widzenia jest on tworzony i realizowany.

4. WYBÓR DANYCH DZIEDZINOWYCH

Ilość danych zgromadzonych w Internecie jest ogromna. Prowadzone są próby podzielenia tych danych na grupy z punktu widzenia szerszego grona użytkowników. Jedną ze skuteczniejszych metod podziału, jest zdefiniowanie zakresu wiedzy dziedzinowej na podstawie obszaru jej zastosowania. Obszarem zastosowania może być określona dziedzina nauki, muzyka, film, handel, itp.

W obszarze wiedzy dziedzinowej należy uzgodnić sposób opisu prezentowanych danych. Skutecznym sposobem opisu są słowa kluczowe, a zwłaszcza metadane. Wiele dziedzin opracowało i uzgodniło standardy budowy metadanych, co nie było zagadnieniem prostym [4]. Po uzgodnieniu metadanych możliwa jest budowa rozproszonej bazy wiedzy złożonej z dedykowanych serwerów gromadzących zbiory danych z określonego obszaru zainteresowań użytkowników. Prezentacja takich zbiorów może się odbywać w formie taksonomii. Taksonomia jest hierarchicznie uporządkowanym zbiorem danych. Opracowano dotychczas wiele takich taksonomii dla różnych dziedzin wiedzy. Na rysunku 4. przedstawiono przykład taksonomii literatury naukowej.



Rys.4. Taksonomia literatury naukowej

Powszechne wykorzystanie takiego sposobu gromadzenia wiedzy napotyka na trudności związane ze zdefiniowaniem grupy docelowej. Jeżeli buduje taksonomię pojedynczy użytkownik, problem jest trywialny, ale gdy taksonomię ma wykorzystywać wielu użytkowników - można zastosować metodę GQM do wyboru danych.

Szablon celu może mieć postać : przeanalizować zbiór artykułów z zamiarem wyboru w odniesieniu do określonej tematyki z punktu widzenia zdefiniowanej grupy użytkowników w zorganizowanym środowisku uczonych.

W tym wypadku można zauważyć ważną rolę zorganizowanego środowiska w definicji celu. Niektóre konferencje i ośrodki naukowe o dużej tradycji publikują bardzo dobrze wybrane i opisane zbiory artykułów z danej specjalności. Wyboru danych w określonym zakresie tematycznym dokonują ludzie zorientowani w danej tematyce. Takie podejście jest użyteczne z uwagi na nadmiar danych literaturowych o zróżnicowanym poziomie jakości.

5. WYBÓR DANYCH OPISOWYCH

Wybór danych opisowych jest zagadnieniem niezwykle złożonym z uwagi na ilość tych danych i złożoność opisywanej rzeczywistości. Opisem rzeczywistości zajmuje się nauka zwana ontologią. W ontologii ogólnej, której językiem opisu może być język naturalny, przyjmuje się obecnie postulat, że nie ma możliwości stworzenia jednej generalnej ontologii opisującej całą rzeczywistość. Można natomiast zbudować wiele ontologii cząstkowych. Ontologia w sensie informatycznym, to opis obszaru rzeczywistości wykonany w językach możliwych do interpretacji i wnioskowania przez systemy komputerowe takich jak OWL czy RDF. Metoda GQM daje praktyczne wskazówki dla budowy przykładowej ontologii.

Szablon celu może mieć postać: wybrać zbiór danych w celu opis budynku wyższej uczelni z różnych punktów widzenia np. turysty, kandydata na studia, pracownika tej uczelni czy strażaka w środowisku społeczności lokalnej.

W tym przypadku występują problemy z perspektywą wyboru danych [5,6]. Jeżeli ontologie buduje twórca dla samego siebie, zagadnienie jest proste, jeżeli ma to być zbiór użytkowników, to konieczne jest uzgodnienie tej perspektywy z jakiej opisywany jest wybrany wycinek rzeczywistości. Jednoznaczne ukierunkowanie na cel prowadzi do redukcji danych i eliminacji danych nie związanych z celem. Brak takiego podejścia powoduje, że budowane są obecnie ontologie o bardzo szerokim zakresie perspektywy. Powstają ontologie, często bardzo złożone, które zmierzają do stworzenia jednej ogólnej ontologii, co jest praktycznie niewykonalne.

Znana ontologia np. CYC *(ang. enCYClopedia)* zawiera 500 000 pojęć i 5 000 000 faktów dotyczących tych pojęć, a słownik WordNet zawiera obecnie 350 000 pojęć. Duże zbiory danych są trudne w wykorzystywaniu i graficznej prezentacji i dlatego podejście proponowane przez metodę GQM wydaje się być podejściem użytecznym praktycznie przy budowie cząstkowych ontologii.

6. WNIOSKI KOŃCOWE

Metoda GQM powstała w celu wyeliminowania niejednoznaczności w rozumieniu znaczenia jakości oprogramowania, z punktu widzenia producenta czy konsumenta. Metoda powstała, by wyeliminować zjawisko dokonywania pomiarów dla pomiarów, ponieważ jest to kosztowne i często bezcelowe, a wyniki pomiaru są często źle interpretowane lub wcale nie interpretowane.

Jak pokazano na kilku wybranych przykładach metoda GQM ma pewien potencjał umożliwiający jej zastosowania w wielu dziedzinach. Należy podkreślić jej bezwzględne ukierunkowanie na cel. Takie podejście jest przydatne praktycznie, prowadzi do znacznej redukcji danych, a tym samym do jasności perspektywy, z której analizuje się opisywany wycinek rzeczywistości. Zmniejszenie ilości danych i ich adekwatny wybór, ułatwiają prezentację i analizę tych danych. Należy przypuszczać, że metoda GQM będzie w przyszłości powszechnie wykorzystywana w doborze i szukaniu informacji.

7. BIBLIOGRAFIA

1. Fenton N.: Software Metric - A Rigorous Approach, Chapman&Hall 1991.
2. Górski J. :Inżynieria oprogramowania, MIKOM, Warszawa 1999.
3. Ince D.: Software Quality Assurance, McGraw-Hill, London 1995.
4. Kaczmarek J.: Model komponentu internetowego dla usług sieciowych, Zaszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskiej, nr.26, Gdańsk 2009.
5. Waloszek W.: Measure for evaluation of structure and semantics of ontologies, Metrology and Measurement Systemy, Vol. 19, 2012.
6. Gangeni A., et al.: Modelling ontology evaluation and validation, Proc. of the 3rd European Semantic Web Conference, Springer-Verlag 2006.

A METHOD OF INFORMATION SELECTION FROM DEDICATED DATA SETS

Key-words: data selection, GQM, taxonomy

Considering the excess of available data, it is necessary to design effective methods of data selection. This paper presents an application of the GQM (Goal, Question, Metric) method in data selection and processing. The GQM method is typically used in software development to define and specify multidimensional software quality functions that assume explicit definition of data selection purpose depending on user or software developer requirements. In this paper, it is shown that the method may also be used to select appropriate data from different dedicated domain data sets. The method enables a user to narrow the set of selected data in a strictly specified manner considering the characteristic of a target domain, the purpose of data selection and consequences of potential data rejection. As shown in the paper, the method may be applied to select data from educational data sets, taxonomy-based literature sets and natural language description of object oriented structures. It is expected that effective methods of data selection will be commonly applied as they improve the process of select and visual data presentation.