XIX Seminarium

ZASTOSOWANIE KOMPUTERÓW W NAUCE I TECHNICE' 2009
Oddział Gdański PTETiS
Referat nr 13

MODEL KOMPONENTU INTERNETOWEGO DLA USŁUG SIECIOWYCH

Jerzy KACZMAREK[1]

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

Streszczenie: Problem znalezienia skutecznych metod poszukiwania danych w Internecie wynika z nadmiaru tych danych oraz braku odpowiedniej struktury ułatwiającej ich selekcję. W artykule przedstawiono model danych internetowych w postaci komponentów, który może ułatwić poszukiwanie informacji. Komponent danych internetowych zawiera metadane opisujące jego zawartość oraz metody pozwalające na identyfikację jego struktury i treści w nim zawartych za pomocą usług sieciowych. Dokonano przeglądu istniejących rodzajów metadanych przeznaczonych dla wiedzy dziedzinowej. Pokazano, że poprzez analizę metadanych systemy zarządzania treścią mogą dynamicznie opisywać zbiory danych za pomocą hierarchicznie uporządkowanych taksonomii.

Słowa kluczowe: metadane, taksonomia, komponent internetowy

1. WSTĘP

Ilość danych zgromadzonych w Internecie i wyspecjalizowanych bibliotekach cyfrowych jest ogromna, choć trudna do oszacowania. Wynika to z faktu ciągłego tworzenia nowych danych, jak również z tego, że wiedza dotychczas gromadzona w formie tradycyjnej, na przykład papierowej, jest zamieniana na postać cyfrową.

Znalezienie skutecznych metod przechowywania, poszukiwania i prezentacji danych cyfrowych to jedno z ważniejszych zagadnień współczesnej informatyki. Można wyróżnić dwa kierunki poszukiwań. Jeden związany jest z metodami analizy treści zbioru cyfrowego, a drugi opiera się na wykorzystaniu metadanych opisujących zawartość zgromadzonej w zbiorze wiedzy dziedzinowej. Analiza treści dokumentu poprzez wyspecjalizowane oprogramowanie pozwala na statystyczną ocenę częstości występowania słów kluczowych, popularności strony WWW czy indeksowanie. Organizacja oparta na metadanych przypomina poszukiwanie książki w bibliotece na podstawie kart katalogowych. Ten sposób organizacji danych nie jest łatwo wykorzystać w Internecie z uwagi na różnorodność danych cyfrowych, ich formatów oraz treści dziedzinowych, jakie zawierają. W artykule przedstawiono model komponentu danych internetowych, który stanowi zamkniętą autonomiczną jednostkę. Zawiera dane, metadane oraz metody, które umożliwiają jego publikowanie i funkcjonowanie w ramach usług sieciowych. Proponowane rozwiązanie stanowi pewną alternatywę dla współczesnych mechanizmów wyszukiwania danych w Internecie.

2. METADANE DANYCH DZIEDZINOWYCH

Metadane to zbiór danych o danych. Danymi mogą być dokumenty cyfrowe, książki, artykuły, utwory muzyczne, fotografie czy dokumentacja medyczna. Bardzo trudno poprawnie opisać metadanymi zasoby cyfrowe w oparciu o jeden standard, ponieważ inne informacje są ważne dla publikacji książkowej, a inne dla utworu muzycznego. Z tych względów powstało kilkadziesiąt różnych standardów metadanych przeznaczonych dla różnej wiedzy dziedzinowej. Informacje o świecie, w którym żyjemy, są dzielone na obszary o wspólnych cechach, dla których budowane są odpowiednie struktury metadanych. Do dziedzin, które posiadają poprawnie wykonane i powszechnie używane standardy metadanych można zaliczyć bibliotekarstwo, archiwizację dokumentów, multimedia, e-edukację, medycynę i inne. Należy podkreślić, że wiele z tych standardów tworzonych jest przy udziale bardzo poważnych organizacji i instytucji, takich jak NASA, W3C czy The Library of Congress USA.

Wśród licznych standardów należy wyróżnić Dublin Core [1]. Jest to standard ogólnego przeznaczenia, który ma możliwości rozszerzania zbioru atrybutów i może być użyty do opisu wielu różnego typu danych cyfrowych. Podstawowy zbiór atrybutów składa się z 15 elementów. Posiada trzy modele dla zasobu, opisu treści i słownictwa. Każdy zasób może być też opisywany przez dowolny zbiór par atrybut–wartość. Jest to wspólna cecha wszystkich standardów. Metadane w formie zbioru atrybutów pozwalają na opis dowolnych danych jednak standard jest ogólnie przyjętą w danej dziedzinie strukturą i może być wykorzystany przez systemy zarządzania treścią przeznaczone dla tej dziedziny. Dublin Core wykorzystuje zalety obiektowości. Każdy atrybut może być powiązany z innym z atrybutem czy klasą a klasa może być powiązana z podklasą. Standard Dublin Core jest jednym z najbardziej popularnych standardów, ponieważ ze względu na jego uniwersalność może być stosowany nie tylko w bibliotekach cyfrowych, ale również do opisu dowolnych typów danych. Prostota i rozszerzalność standardu daje ogromne możliwości dostosowania schematów opisu do potrzeb użytkowników. W tabeli 1 przedstawiono niektóre standardy metadanych wraz z instytucjami, które je stworzyły.

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

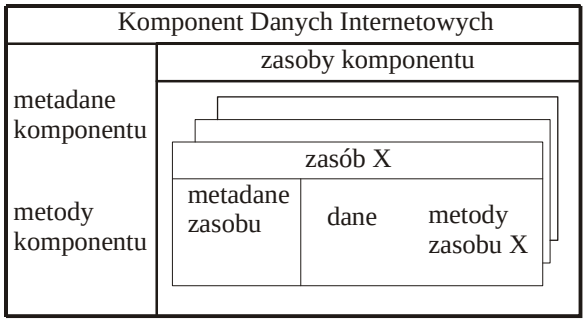
Tablica 1. Standardy dla wiedzy dziedzinowej

Dziedzina	Nazwa standardu	Twórca	Typ zasobu
Archiwizacja	Dublin Core	Dublin Core Metadata Initiative	Różne
	MARC 21	The Library of Congress	Dane bibliograficzne
	ONIX	Association of American Publishers and EDItEUR	Dane o publikacjach
	MODS	The Library of Congress	Dane bibliograficzne
	PREMIS	The Library of Congress USA	Dane konserwatorskie do zarządzania zasobami, bibliograficzne
	METS	The Library of Congress USA	Standard do kodowania Dane, bibliograficzne
	EAD	The Library of Congress USA	Standard kodowania opisów archiwalnych
Strony Web	HTML MetaTags	W3C	Strony internetowe
Nauka	LOM	Learning Technology Standards Commitee IEEE	Obiekty edukacyjne
Kultura i sztuka	CDWA	Art Information Task Force	Dzieła sztuki
	VRA Core	Visual Resource Association	
	Object ID	Council for Prevention of Art Theft	
Geografia	CSDGM	Federal Geographic Data Committee	Dane geoprzestrzenne
	FGDC		
Multimedia	MPEG 7	Moving Picture Experts Group	Dane audiowizualne
	DIG 35	Digital Imaging Group	Obrazy cyfrowe
	NISO Z39.87	National Information Standard Org.	
	JPEG 2000	Joint Photograph Experts Goup	
Medycyna	MeSH	National Library of Medicine	Dane medyczne

3. KOMPONENT DANYCH INTERNETOWYCH

W dziedzinie zarządzania treścią w Internecie występują pewne problemy z definicją używanych pojęć. Powszechnie wykorzystywane jest pojecie obiektu internetowego rozumiane jako dowolny zasób internetowy posiadający swoją identyfikację w postaci URI (ang. Uniform Resource Identifier). Jest to bardzo duże uproszczenie często jednak krytykowane. Dla przykładu w dziedzinie e-edukacji używa się pojęcia obiektu edukacyjnego, który może nie posiadać cech dydaktycznych, a także nie ma cech obiektowości [2]. Rozwiązaniem tego problemu byłoby powszechne wykorzystywanie komponentów edukacyjnych mających tożsamość oraz zachowanie i projektowanych w oparciu o zasady obiektowości [3].

Dla potrzeb technologii usług sieciowych można wprowadzić pojęcie Komponentu Danych Internetowych, którego strukturę przedstawiono na rys.1



Rys.1 Model Komponentu Danych Internetowych

Komponent danych internetowych jest zbiorem ustrukturalizowanych danych cyfrowych. Zawiera przede wszystkim metadane, opisujące jego strukturę oraz te jego cechy, które są istotne dla systemów wyszukiwawczych.

Komponentem internetowym może być dowolny zasób danych dostępny w Internecie, taki jak strona czy portal internetowy, komponent programowy, artykuł naukowy, dokument cyfrowy, utwór muzyczny czy artystyczny. Może być tworzony przez rozbudowane serwisy internetowe lub też przez proste programy działające po stronie klienta. W najprostszym przypadku komponent danych internetowych jest to katalog z plikami zawierającymi dokumenty cyfrowe czy dane multimedialne, w którym jeden z plików zawiera metadane opisujące ten katalog w języku XML. Jest to uogólnienie sposobu zarządzania danymi w systemach operacyjnych gdzie plik jest opisany metadanymi takimi jak i-węzeł, które umieszczane są w specjalnie do tego celu przygotowanym obszarze na dysku. Wykonanie prostego systemu ułatwiającego tworzenie metadanych w postaci zbioru par atrybut – wartość lub zgodnie z przyjętym standardem nie jest skomplikowane.

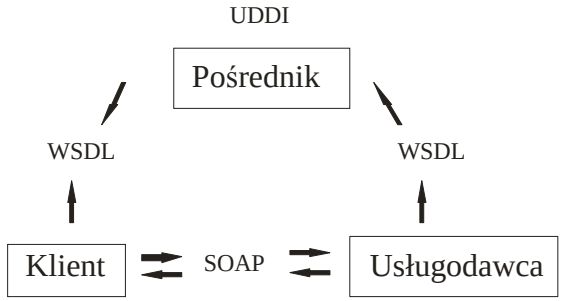
Komponent danych internetowych może być zbiorem innych komponentów i może przypominać drzewo katalogów ze zbiorami plików zawierającymi określony typ danych. Takie dane w postaci zorganizowanego zbioru plików mogą być przetwarzane przez stosunkowo proste systemy zarządzania treścią po stronie klienta.

W bardziej rozbudowanej postaci komponent danych internetowych może być realizowany jako usługa sieciowa i zawierać zarówno dane, jak również zbiór metod, które odpowiedzą na zapytanie klienta o funkcjonalność komponentu. Z punktu widzenia poszukiwania informacji, istotne są metody związane z metadanymi takie jak: pokaż typ metadanych, podaj wybrany atrybut metadanych, pokaż statystykę historii poszukiwań, pokaż wyspecjalizowane biblioteki, w których się znajduje, czy podaj taksonomię, w której został umieszczony. Inne zaproponowane metody mogą dotyczyć przesyłania wybranych danych znajdujących się w komponencie na adres internetowy, który zgłasza zapytanie. Komponent internetowy powinien mieć cechy obiektowości, takie jak tożsamość, zachowanie, hermetyczność czy dziedziczenie. Komponent danych internetowych może być w całości kopiowany do systemów zarządzania treścią lub może przekazywać jedynie dane w nim zawarte.

Opisanie zbioru danych metadanymi w standardzie przeznaczonym jedynie dla danej dziedziny ma tę wadę, że wymusza używanie wyspecjalizowanych dla tej dziedziny systemów zarządzania treścią często konkretnych producentów. Przedstawiona struktura komponentu internetowego uniezależnia dane dziedzinowe od wyspecjalizowanych systemów zarządzania treścią i umożliwia poszukiwanie danych opisanych różnymi typami metadanych. Wykonanie komponentu internetowego wymaga pewnego nakładu pracy, niezbędne jest wypełnienie pól w strukturze metadanych, oraz należy wytworzyć systemy komputerowe do poszukiwania, gromadzenia i prezentacji danych, które w ten sposób są ustrukturalizowane. Niezbędne jest również wykonanie i utrzymanie infrastruktury usług sieciowych. Opracowane dla technologii usług sieciowych standardy takie jak WSDL i SOAP formalizują sposoby komunikacji i przesyłania informacji pomiędzy klientem a usługodawcą [4]. Należy jednak uwzględnić dynamiczny rozwój tej technologii, która staje się coraz bardziej powszechna.

4. ZARZĄDZANIE KOMPONENTAMI

Technologia usług sieciowych (ang. web services) oparta jest na strukturze klient-pośrednik-usługodawca, co ilustruje rys.2.



Rys.2 Schemat architektury usług sieciowych

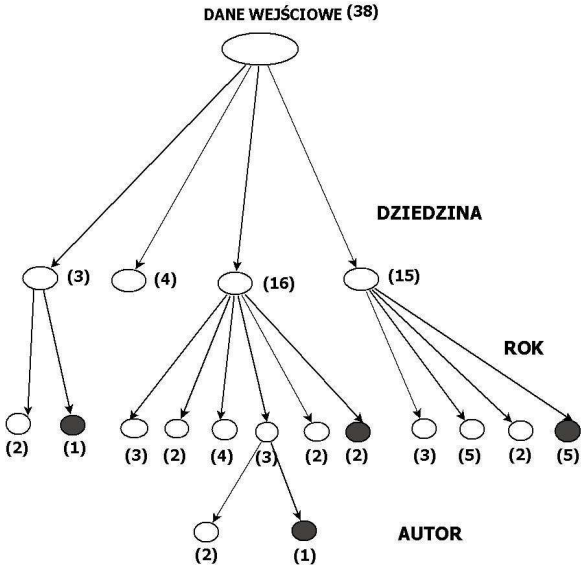
Klient znajduje interesującą go usługę oraz adres internetowy usługodawcy na serwerze pośrednika, jakim może być rejestr usług sieciowych, np. UDDI. Rejestr ten przypomina książkę telefoniczną z danymi o usługodawcy, jego adresem i informacjami o usługach, jakie on oferuje.

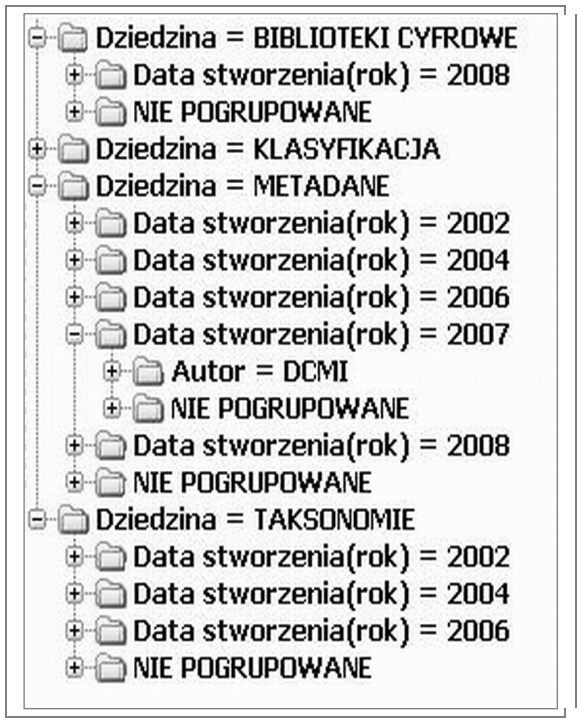
W przypadku komponentów danych internetowych w rejestrze mogą być zapisane typy i wartości metadanych, które opisują zasoby oraz informacje, jakie rodzaje metod oferują komponenty. Wstępne poszukiwanie danych w rejestrach UDDI jest szybsze i skuteczniejsze, a także pozwala na poszukiwanie danych w Internecie w różnych zbiorach dziedzinowych.

Zarządzanie danymi pochodzącymi z Internetu można podzielić na dwa etapy. Pierwszym jest wstępny wybór danych z ogromnego zbioru istniejącego w Internecie. Drugi etap polega na selekcji i porządkowaniu tych danych. Może to się odbywać po stronie klienta lub w wyspecjalizowanych serwisach. Ważnym etapem przetwarzania zgromadzonych danych jest ich prezentacja w przystępnej dla użytkownika formie. Pierwszy etap może być skutecznie realizowany poprzez usługi sieciowe oparte o komponenty internetowe. Po zgromadzeniu pewnego zbioru danych systemy komputerowe działające po stronie klienta lub w dedykowanych dla danej dziedziny serwisach mogą automatycznie dokonać porównania komponentów, ich selekcji i prezentacji graficznej na przykład w postaci hierarchicznej klasyfikacji zwanej taksonomią.

Rezultaty działania takiego systemu, zrealizowanego w ramach prac badawczych autora, przedstawiono na rysunku 3. Zbiór przetwarzanych danych obejmował 38 pozycji literaturowych dotyczących sposobów zarządzania danymi cyfrowymi. Pierwszym krokiem w zarządzaniu takim zbiorem jest wybór typu metadanych. Wybrano Dublin Core jako standard ogólnego przeznaczenia. Spośród jego piętnastu podstawowych atrybutów wybrano dla przykładu jedynie trzy. Są to; dziedzina, data i typ publikacji. Kolejność, w jakim system będzie dokonywał selekcji danych można dowolnie zmieniać.

Na rysunku 3 przedstawiono rezultat podziału zbioru artykułów według jednej z możliwych kolejności trzech atrybutów: dziedzina - data utworzenia - typ. Przedstawiono drzewo taksonomii oraz zrzut z ekranu interfejsu użytkownika.





Rys.3 Hierarchiczna klasyfikacja zbioru danych

Zarządzanie danymi przez takie systemy przypomina działania skrupulatnego użytkownika porządkującego swoje zbiory w komputerze. Użytkownik gromadzi zbiór danych, który następnie dzieli na katalogi według przyjętego algorytmu nadając katalogom, co ważne, nazwy odpowiednie do treści zawartych w katalogach.

Tworzy się taksonomię, czyli hierarchiczną klasyfikację, która jest wyświetlana w interfejsie użytkownika jako drzewo katalogów.

Drzewo takie może się rozrastać w zależności od ilości zgromadzonych danych. Przewaga systemów zarządzania treścią nad selekcją w sposób ręczny polega na automatyzacji tych czynności oraz na tym, że podział zbioru danych na taksonomię może się odbywać w sposób dynamiczny, wielokrotnie, przy różnych zasadach podziału.

5. WNIOSKI KOŃCOWE

Podział danych na obszary dziedzinowe, opis ich metadanymi i poszukiwanie informacji przez wyspecjalizowane systemy zarządzania treścią może wyeliminować trudności w poszukiwaniu informacji w nadmiarze danych zgromadzonych w Internecie.

Przedmiotem rozważań niniejszego artykułu są mechanizmy wstępnej selekcji danych, oraz opis zasad i metod działania skutecznych systemów zarządzania treścią. Rozwój technologii usług sieciowych daje możliwości budowy systemów, które trafnie wyszukują informację oraz prezentują ją użytkownikowi w sposób przyjazny i zrozumiały. Jedną z bardzo skutecznych metod prezentacji informacji jest przedstawianie danych w postaci klasyfikacji hierarchicznej. Taka taksonomia może być dynamicznie tworzona poprzez porównywanie komponentów internetowych zawierających dane dziedzinowe opisane metadanymi. Dynamiczna budowa taksonomii jest zagadnieniem złożonym, ponieważ wymaga opracowania skutecznych metod porównywania komponentów

Budowa i utrzymanie systemów informatycznych dla efektywnie zorganizowanych bibliotek cyfrowych wymaga nakładów pracy przy opisie informacji i kosztów przy budowie wyspecjalizowanych systemów zarządzania treścią, ale jest to ważny i przyszłościowy kierunek w rozwoju usług internetowych.

6. BIBLIOGRAFIA

1. Dublin Core Metadata Initiative, http://dublincore.org
2. Friesen N.: Three Objections to Learning Objects and E-learning Standard, Online Education Using Learning Object, Routledge, no.1 pp 59-70, 2004.
3. Kaczmarek J., Landowska A.: Model of Distributed Learning Objects, Interactive Learning Environments Journal, Routledge, no.1, volume 14, pp. 1-16, 2006.
4. Graham S.: Java, Usługi WWW, Helion, 2003.

Wykonano w ramach grantu MNiSW N N516 383534

AN INTERNET COMPONENT MODEL FOR WEB SERVICE

Key-words: metadata, taxonomie, internet component

Effective methods of internet information retrieval are difficult to design because of high redundancy of the information and its unstructured nature. The paper presents a model of internet data components that simplify information search and retrieval. An internet data component supplies both metadata that describe its contents and methods that enable access to stored information. The Web services technology was used to access component methods and retrieve its contents. An overview of existing metadata models of domain knowledge description has been presented. It was shown that metadata analysis is an effective method that enables information management systems to dynamically describe and process information using hierarchically ordered taxonomies.