XX Jubileuszowe Seminarium

ZASTOSOWANIE KOMPUTERÓW W NAUCE I TECHNICE' 2010
Oddział Gdański PTETiS
Referat nr 13

WYKORZYSTANIE TAKSONOMII DO INTEGRACJI DANYCH W ZASOBACH INTERNETU

Jerzy KACZMAREK

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: Rozproszony zbiór danych internetowych można zintegrować i efektywnie zorganizować wykorzystując możliwości usług sieciowych i taksonomii. W artykule przedstawiono wyniki pomiarów nakładu pracy niezbędnej do budowy usług sieciowych publikujących zorganizowane zbiory danych. Omówiono zasady ręcznej i automatycznej budowy taksonomii. Przeanalizowano problemy optymalizacji takiej struktury oraz korzyści z kolorowania nazw wyróżnionych węzłów taksonomii. Wykazano, że przy automatycznej budowie hierarchicznej klasyfikacji można wykorzystać metadane opisujące zasoby internetowe.

Słowa kluczowe: usługi sieciowe, taksonomia, metadane, Internet

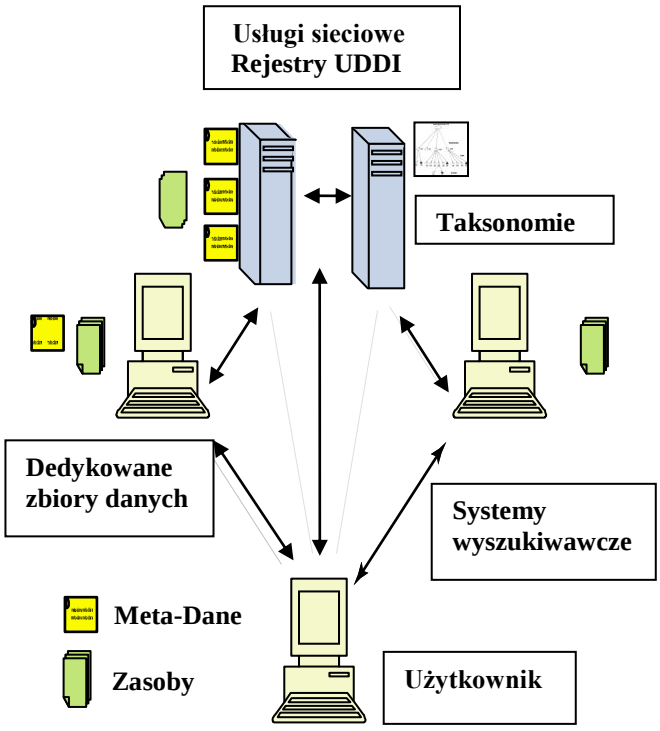
1. WSTĘP

Sieć Internet stała się niemal nieograniczonym zbiorem danych. Zbiór ten jest jednak nadmiarowy, heterogeniczny, wykonany w różnych technologiach, jest rozproszony i nieuporządkowany, a ponadto zawiera informacje nie zawsze wiarygodne. Powstaje problem teoretyczny i praktyczny jak stworzyć metody do wyszukiwania, selekcji oraz prezentacji pewnych grup danych, istotnych z punktu widzenia użytkownika. Trzeba podkreślić, że przetwarzanie tych danych różni się od klasycznych metod stosowanych w bazach danych, które budowane są według przyjętego schematu relacyjnego czy obiektowego. Ta zasadnicza różnica wynika z częstej zmiany struktury przetwarzanych danych, które mogą być wybierane z zasobów Internetu poprzez zapytania użytkownika do wyszukiwarek. Ten podzbiór danych internetowych trzeba poddać selekcji i zaprezentować użytkownikowi w zrozumiałej formie.

Należy rozwiązać problem podziału zbioru danych zaczerpniętych z Internetu na pewne grupy o wspólnych cechach. Trzeba uwzględnić typy danych, dziedziny rzeczywistości, których dotyczą oraz wziąć pod uwagę przy budowie algorytmów do podziału tych danych, takie pojęcia jak podobieństwo, klasyfikacja, grupowanie czy kategoryzacja [1]. Powstaje również pytanie, czy podział danych na grupy ma być automatyczny czy ręczny. Ręczny podział przez użytkownika wymaga nakładów pracy, ale dzięki inteligencji użytkownika unika się wielu błędów przy klasyfikacji, które wynikają z trudności w automatycznym rozumieniu tekstów. Ostatnim etapem przetwarzania jest prezentacja danych użytkownikowi, w której jak zostanie przedstawione w artykule można wykorzystać taksonomię jako skuteczny i zrozumiały sposób grupowania danych.

2. INTEGRACJA DANYCH INTERNETOWYCH

Rozwój technologii usług sieciowych daje możliwości budowy systemów, które pozwalają na wybór i integrację danych zgromadzonych w Internecie, co ilustruje rysunek 1.



Rys.1. Powiązania między zasobami Internetu

Podstawowym zbiorem danych prezentowanym użytkownikowi po zadaniu pytania do wybranego systemu wyszukiwawczego jest często ogromna liczba adresów internetowych, związana semantycznie ze słowami tworzącymi zapytanie. Istnieją liczne metody porządkujące takie dane, oparte o popularność adresów, częstość występowania słów tworzących zapytanie w dokumentach czy mechanizmy podpowiedzi dla dalszych poszukiwań. Takie tradycyjne poszukiwania mają swoje ograniczenia wynikające z nadmiaru danych do przetwarzania.

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

Budowa dedykowanych zbiorów danych zapewnia większą integrację informacji dotyczącą pewnej dziedziny wiedzy czy techniki. Taki sposób organizacji danych wymaga jednak uzgodnienia standardów opisu danych dziedzinowych, poprzez metadane oraz wykonania dodatkowej pracy przy budowie systemów do poszukiwania, przetwarzania i prezentacji informacji.

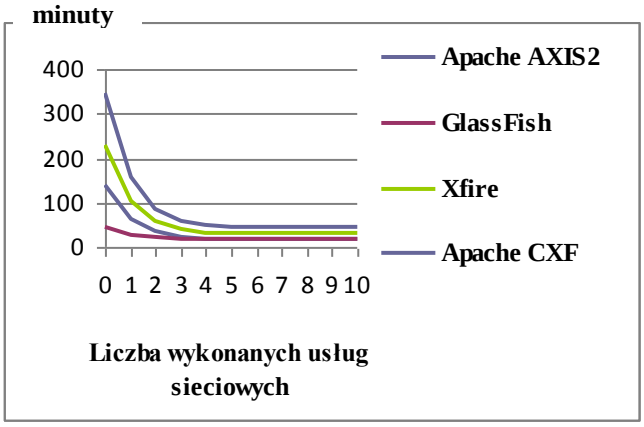
Rozwój usług sieciowych daje duże możliwości publikowania przez dostawcę informacji o swoich danych w ogólnie dostępnych rejestrach. Przeglądanie takich rejestrów przypomina poszukiwanie informacji w książkach telefonicznych. W rejestrach mogą być zawarte zarówno dokumenty jak również adresy internetowe wskazujące na dane rozproszone. Rejestry mogą być dedykowane dla danej dziedziny i wówczas mogą zawierać zarówno dane jak również ich opisy w postaci metadanych. Zbiory rejestrów mogą przekazywać informacje między sobą, co w sposób istotny zwiększa bezpieczeństwo i integracje danych. Rejestry pozwalają na budowę taksonomii hierarchicznie zorganizowanych struktur danych.

Należy zauważyć, że w systemach działających w technologii usług sieciowych nie muszą znajdować się wszystkie dane, które są w nich opublikowane. Umieszczać można jedynie opis usługi, metadane obiektów internetowych oraz adresy internetowe pozwalają na dostęp do zasobów [2]. Zwiększa to integrację danych zgromadzonych w rozproszonych zasobach Internetu.

Budowa takich usług sieciowych do prezentacji dedykowanych danych dziedzinowych ma jednak pewne wady. Wymaga wiedzy informatycznej oraz pewnego nakładu pracy przy organizacji danych. Należy dokonać opisu danych metadanymi, scalić metody, treści dokumentów i zintegrować obiekty internetowe.

Wykonanie systemu komputerowego świadczącego usługi internetowe wymaga również nakładu pracy na stworzenie opis usługi w języku WSDL, umieszczeniu na serwerze, opublikowanie usługi w rejestrze oraz wykonanie taksonomii i graficznych interfejsów użytkownika. Duży nakład pracy związany jest również z aktualizacją danych i ciągłym utrzymaniem systemów informatycznych.

Z tych względów pomimo burzliwego rozwoju usług sieciowych popularność takich rozwiązań jest niezadowalająca. W ramach prowadzonych prac dokonano pomiarów czasu potrzebnego informatykowi na wykonanie systemu usług sieciowych do prezentacji wiedzy dziedzinowej przy różnych technologiach wykonania. System zawierał serwer stron, mechanizm dynamicznie zmiennej taksonomii oraz liczne interfejsy użytkownika. Wyniki pomiarów przedstawiono na rysunku 2.



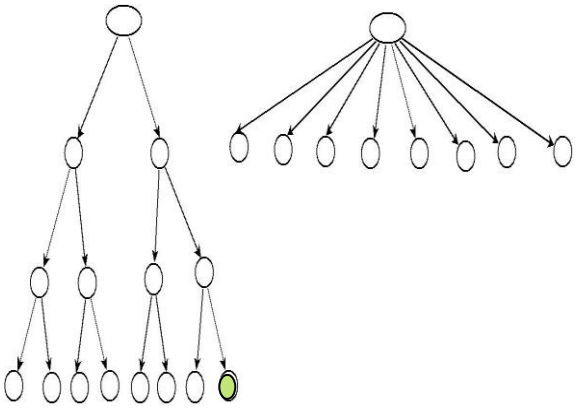
Rys.2. Czas wytworzenia usługi sieciowej

Wykorzystano popularne narzędzia do wytwarzania usług sieciowych, dostępne na zasadach licencji open source takie jak AXIS2, JAX-WS, Apache CXF, oraz XFire.
Można zauważyć na rysunku 2, że decydującym dla czasu wykonania usługi jest doświadczenie programisty liczone w ilości wykonanych systemów, a nie rodzaj zastosowanej technologii. Należy zaznaczyć, że pomiary nie uwzględniają nakładu pracy potrzebnego na utrzymanie systemu. Gdy tymczasem nakłady te mogą być znaczące przy dużej ilości i zmienności danych. Zmienność danych wpływa zasadniczo na kształt i złożoność interfejsów użytkownika, które powinny zawierać informacje o zgromadzonych danych w formie taksonomii.

3. ZASADY BUDOWY TAKSONOMII

Słowo taksonomia pochodzi od greckiego słowa taxinomia, które jest złożeniem słów taksis, oznaczającego układ, porządek, oraz nomos, oznaczającego prawo. Taksonomia zależnie od kontekstu jest rozumiana albo jako określenie hierarchicznej klasyfikacji albo jako definicja reguł leżących u podstaw tej klasyfikacji. Taksonomia przedstawiana jest jako drzewiasta struktura, na szczycie, której znajduje się pojedynczy klasyfikator, korzeń drzewa, odnoszący się do wszystkich obiektów. Węzły poniżej korzenia stanowią dokładniejszą klasyfikację, która dotyczy określonego podzbioru wszystkich klasyfikowanych obiektów. Hierarchiczna klasyfikacja zbioru danych może mieć formę drzewa katalogów podobną do drzewa katalogów w komputerze. Takie drzewo katalogów będzie poprawnie zbudowaną taksonomią, o ile nazwy katalogów odpowiadają treści danych w nich zawartych, a w katalogach znajdować się będą podkatalogi, których położenie jest logicznie uzasadnione strukturą taksonomii.

Przy budowie taksonomii należy unikać błędów dotyczących jej struktury przedstawionych na rysunku 3.



Rys. 3. Problem głębokości i szerokości taksonomii

Taksonomia nie powinna być zbyt głęboka czy zbyt szeroka, ponieważ utrudnia to budowę interfejsów użytkownika. Takie zjawisko występuje wtedy, gdy węzły tworzone są dla małych ilości danych. Rozwiązaniem może być pozostawianie części danych w formie nieuporządkowanej w węźle wyróżnionym, aż do uzyskania odpowiedniej ilości danych, która uzasadnia stworzenie nowego węzła.

Najbardziej znaną taksonomią jest klasyfikacja roślin, zwierząt i drobnoustrojów występujących w świecie przyrody, zapoczątkowana przez Karola Linneusza, która jest rozbudowywana po dziś dzień.

Taksonomia Linneusza zawiera dla roślin jedynie siedem poziomów o nazwach: królestwo, gromada, klasa, rząd, rodzina, rodzaj, gatunek. Te wzorcowo wykonane taksonomie zawierają informację o kilku milionach istot.

Taksonomię można budować ręcznie z wykorzystaniem wiedzy ekspertów z danej dziedziny i korzystając z inteligencji człowieka, lub automatycznie przez systemy komputerowe realizujące wybrane algorytmy klasyfikujące. Budowę taksonomii można podzielić na następujące etapy;

- Wybieranie danych, które mają być zorganizowane,
- Wydobywanie konceptów i grupowanie ich w klastry,
- Budowę wstępnej taksonomii i badania jej struktury,
- Ocenę możliwości zastosowania znanych taksonomii,
- Budowę własnych interfejsów z taksonomią,
- Administrację i modyfikację danych i taksonomii.

Rejestry usług sieciowych posiadają kilka domyślnych taksonomii, które umożliwiają klasyfikacje w szerokim zakresie terminów, a także można tworzyć własne taksonomie w oparciu o dostarczane technologie.

3.1. Ręczna budowa taksonomii

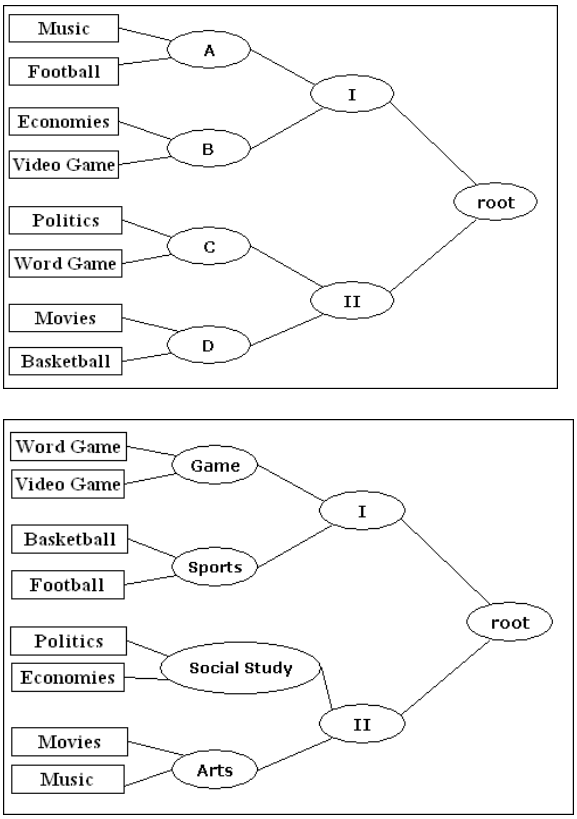
Etapy ręcznej budowy taksonomii mogą być nieco inne niż przedstawione powyżej. Ręcznie można budować taksonomię nie posiadając danych a jedynie przewidując ich powstanie i planując miejsca gdzie będą umieszczane. Taki proces projektowania taksonomii został wybrany przy budowie rozproszonego systemu komputerowego, opartego o usługi sieciowe przeznaczonego dla systemu szkolnictwa podstawowego i średniego. Taksonomia zawiera poziomy; typ szkoły, klasa, przedmiot, semestr, lekcja. Na tych pięciu poziomach można zgromadzić przewidywaną ilość danych, ocenianą na ponad 30 tysięcy obiektów edukacyjnych. Prowadząc badania nad jakością interfejsów użytkownika dla tego typu taksonomii, zauważyłem użyteczność kolorowania węzłów w drzewie taksonomii.

Kolory mogą sygnalizować typ danych, rozróżniać informacje dydaktyczne od organizacyjnych takich jak sylabusy. Inny kolor węzła może oznaczać również ilość dostępnych wersji danej lekcji. Kolor wyróżniony np. czerwony, może oznaczać brak danych i taki węzeł może być jednocześnie ofertą dla nauczycieli by przygotować określone materiały.

Na rysunku 4 przedstawiono wykonaną w ramach prowadzonych badań taksonomię z kolorowaniem węzłów, wykonaną dla systemu szkolnictwa afrykańskiego państwa Malii, w którym językiem urzędowym jest francuski.

Rys.4. Przykład kolorowania węzłów w taksonomii

Ręczna budowa taksonomii pomimo znacznych nakładów pracy niezbędnych do jej wykonania, posiada kilka zalet. Przede wszystkim unika się błędów związanych z niejednoznacznością słów, istnieniem synonimów oraz błędów w strukturze taksonomii wynikających z braku powiązań semantycznych na różnych poziomach taksonomii. Na rysunku 5 przedstawiono klasyczny przykład właściwie i nie właściwie wykonanych taksonomii [3].



Rys. 5. Przykład złej i dobrej taksonomii.

Właściwie wykonana taksonomia ma logiczną budowę na wszystkich poziomach swojej struktury. Taka budowa wymaga rozróżnienia znaczenia słów w sensie semantycznym i ontologicznym.

3.2. Automatyczna budowa taksonomii

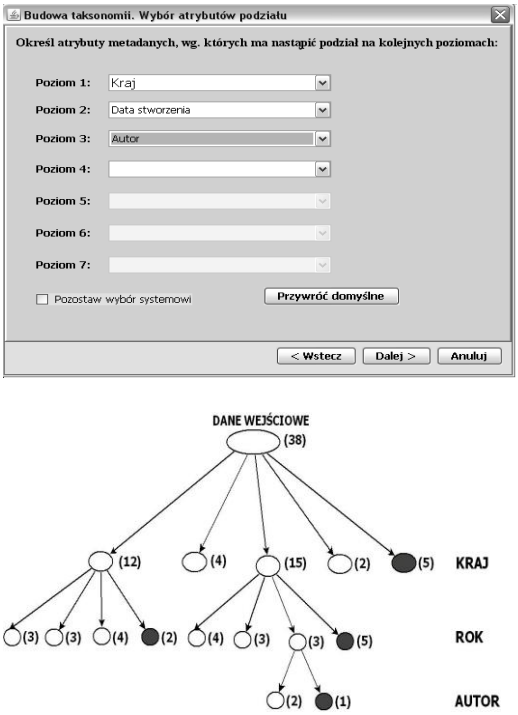
Automatyczna budowa taksonomii przez systemy komputerowe polega na wstępnym wyborze danych internetowych, a następnie na ich analizie i podziale na grupy według ustalonych reguł. Metody automatyczne opierają się na podejściu bottom-up, z dołu do góry i na razie klasyfikują gorzej niż robi to ręcznie człowiek ze swoją intuicją, inteligencją oraz znajomością w szerokim kontekście wiedzy dziedzinowej i języka naturalnego.

Przy automatycznym podziale zbiorów danych, należy uwzględniać typy danych a także kontekst, w jakim one występują. Kiedy dochodzi do dekompozycji danych, ich klasteryzacji, grupowania, klasyfikacji, kategoryzacji czy generalizacji, systemy wyszukiwawcze muszą uwzględniać znaczenie semantyczne słów języka [4]. Przy dużej ilości tych słów, a także niejednoznaczności synonimów złożoność zagadnienia staje się problemem poważnym.

Opracowano liczne metody automatycznego tworzenia taksonomii. Podstawowa różnica między nimi dotyczy wyboru atrybutów zasobu, które są brane pod uwagę przy klasyfikacji oraz rodzajów algorytmów oceny podobieństwa przy grupowaniu zasobów w węzłach taksonomii.

Skutecznym sposobem automatycznej budowy struktury taksonomii jest wykorzystanie metadanych obiektów internetowych. Metadane mogą być tworzone przez publikującego w rejestrze informację, dostawcę usługi internetowej lub przez samego użytkownika. Na podstawie metadanych można dynamicznie określać poziomy taksonomii a następnie rozmieszczać posiadane dokumenty w węzłach. Podstawowy warunek, który trzeba spełnić, to jednolitość struktury metadanych i konieczność wypełnienia zawsze dla wszystkich dokumentów tych pól, które są brane pod uwagę przy budowie taksonomii. Nie wypełnienie pola oznacza, że dany dokument nie może być brany pod uwagę i może się znaleźć jedynie w węźle ogólnym, który powinien być wyróżniony innym kolorem.

Na rysunku 6 przedstawiono przykład wykonanego systemu do automatycznej budowy taksonomii na podstawie metadanych. Poziomy taksonomii określane są poprzez słowa języka naturalnego stanowiące metadane. Mogą one być ustalane dowolnie przez użytkownika. Jedną z możliwości jest stosowanie standardowych dla danej dziedziny metadanych. Ważne jest, aby opis informacji w postaci metadanych był jednolity w ramach danego zbioru, na podstawie którego tworzy się taksonomię.



Rys. 6. Budowa taksonomii w oparciu o metadane

4. WNIOSKI KOŃCOWE

Rozproszone zasoby Internetu stały się ważną, bardzo rozbudowaną bazą wiedzy. Wykorzystanie tych informacji nie jest jednak proste z uwagi na nadmiar danych i brak ich uporządkowania. Rozwój usług sieciowych daje nadzieję na poprawę szybkości poszukiwania danych i budowę systemów komputerowych integrujących dane internetowe. Jak wykazano, nakłady pracy niezbędne do wykonania takich systemów maleją wraz z uzyskiwaniem przez programistów doświadczeń w ich budowie.

Prezentacja wyników poszukiwań informacji powinna być dokonywana w postaci interfejsów o dobrej jakości w formie zrozumiałej dla użytkownika. Skuteczną metodą integracji rozproszonych danych internetowych i ich prezentacji użytkownikowi, wydaje się być taksonomia.

Przedmiotem rozważań niniejszego artykułu są problemy, jakie występują przy budowie taksonomii. Zwrócono uwagę na użyteczność wyróżniania kolorem niektórych szczególnych węzłów taksonomii oraz na konieczność zachowania spójności semantycznej i ontologicznej w całej strukturze taksonomii. Wykazano, że wykorzystywanie metadanych znacznie upraszcza mechanizm automatycznej budowy taksonomii.

Ręczna czy automatyczna budowa taksonomii jest zagadnieniem złożonym, ponieważ wymaga opracowania skutecznych metod porównywania danych i określania ich logicznego miejsca w hierarchicznej klasyfikacji.

Praca naukowa finansowana ze środków na naukę w latach 2009-2012 jako projekt badawczy nr N N519 172337.

5. BIBLIOGRAFIA

1. Leppanen M.: Towards an Abstraction Ontology, Information Modelling and Knowledge Bases XVIII, IOS, Press, 2007.
2. Kaczmarek J.: Model komponentu internetowego dla usług sieciowych, Zeszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskiej, nr.26, s.61-64, 2009.
3. Bouquet P. et all.: Bootstrapping semantics on the Web, IW3C2, May 23-26, Edinburgh, Scotland, 2006.
4. Spangler S.: MindMap: Utilizing Multiple Taxonomies and Visualization to Understand a Document Collection, Proc. Of the 35th Hawaii International Conference on System Sciences 2002.

TAXONOMY USE FOR DATA INTEGRATION IN INTERNET RESOURCES

Key-words: Web services, taxonomy, metadata, Internet.

Web services and taxonomies can be used for effective integration and categorization of distributed internet data. The paper presents effort measurements results of work required for development of Web services supplying categorized data sets. Automated and human-driven approaches in development of taxonomies were discussed. Taxonomy structure optimization and advantages of taxonomy nodes coloring were also analyzed. It was proved that metadata can be used for automated and dynamic generation of hierarchical classifications.