XX Jubileuszowe Seminarium

ZASTOSOWANIE KOMPUTERÓW W NAUCE I TECHNICE' 2010
Oddział Gdański PTETiS
Referat nr 11

INTEGRACJA USŁUG SIECIOWYCH Z UWZGLĘDNIENIEM POZIOMU WIARYGODNOŚCI ICH DOSTAWCÓW

Adam Łukasz KACZMAREK[1]

1. WETI, Politechnika Gdańska, ul. G. Narutowicza 11/12, 80-952 Gdańsk
tel: 58 347 27 19 fax: 58 347 22 22 e-mail: adam.l.kaczmarek@eti.pg.gda.pl

Streszczenie: Artykuł porusza temat wiarygodności danych pobieranych z usług sieciowych. Przedstawiona została metoda oceny wiarygodności takich danych opierająca się na czterech metrykach: powszechności informacji, niezależności źródła informacji, prestiżu źródła oraz doświadczenia ze współpracy ze źródłem. Metoda ta ma zastosowanie przy integracji usług sieciowych pochodzących od wielu różnych dostawców. Metoda pozwala na automatyczną ocenę poziomu wiarygodności na podstawie informacji dotyczących pochodzenia danych (ang. data provenance).

Słowa kluczowe: wiarygodność danych, usługi sieciowe

1. WSTĘP

Usługi sieciowe (ang. Web Services) są dostępnymi za pośrednictwem sieci Internet fragmentami oprogramowania, z których korzystać mogą aplikacje komputerowe [1]. Aplikacje używające takich usług tworzone są zgodnie ze standardami architektury SOA (Service-oriented architecture). Usługi sieciowe mogą dostarczać aplikacjom wymaganych przez nie danych lub wykonywać na potrzeby aplikacji różnego rodzaju funkcje. Każda usługa ma ściśle określone przeznaczenie i funkcjonalność. Informacje dotyczące charakterystyki usługi i sposobu komunikowania się z nią stanowią interfejs usługi. Dzięki niemu aplikacje mają możliwość określenia, czy usługa jest dla nich odpowiednia oraz w jaki sposób z niej korzystać, jeśli będą ją używać.

Ważnym problemem występującym w aplikacjach tworzonych zgodnie z architekturą SOA jest integracja różnego rodzaju usług sieciowych w ramach jednej aplikacji [2]. Konieczne jest zapewnienie, że wykonywanie każdej z usług nie powoduje konfliktu z wykonywaniem innej usługi. Konflikty takie pojawiać się mogą na przykład w skutek niezgodności protokołów komunikacyjnych lub stosowania różnych ich wersji. Przy integracji usług pojawić się mogą również problemy wynikające z zapewnienia bezpieczeństwa danych. Ponadto pojawiają się błędy w danych wynikające z zakłóceń komunikacji, niepoprawnego funkcjonowania usług oraz ich czasowej niedostępności.

2. ROZWIĄZYWANIE PROBLEMÓW INTEGRACJI USŁUG SIECIOWYCH

Metody rozwiązywania wielu rodzajów problemów pojawiających się podczas integracji usług sieciowych poruszone zostały w ramach projektu Reist [3]. Metody te wywodzą się głównie z technik zapewniania bezpieczeństwa danych w połączeniu z metodami tolerowania błędów (ang. Fault Tolerance). Zapewnienie bezpieczeństwa danych polega na zastosowaniu metod szyfrowania oraz podpisów cyfrowych. Oprócz tego, w przypadku, gdy występują problemy z niepoprawnymi danymi pochodzącymi z usług, stosowane są scenariusze działań mające na celu zidentyfikowanie oraz wykluczanie usług dostarczających nieodpowiednich danych. Wprowadza się także ograniczenia dotyczące czasu, w jakim oczekiwane są dane z usług.

Podniesienie niezawodności aplikacji korzystających z usług sieciowych jest możliwe dzięki tworzeniu lepszych aplikacji, jak również dzięki dostarczaniu lepszej jakości usług. W celu tworzenia usług charakteryzujących się wysokim poziomem niezawodności opracowana została architektura Whisper opisana przez Cardoso [4]. Rozwiązanie problemów tolerowania błędów osiągnięte zostało dzięki przyjęciu, że struktura sieci jest dynamiczna, co pozwala dostosowywać dostępność usług odpowiednio do aktualnie występującego stanu sieci. Ponadto zastosowany został zdecentralizowany model udostępniania usługi, przez co usługa sieciowa nie jest całkowicie zależna od poprawności działania pojedynczego serwera, na którym została uruchomiona.

Mimo rozwoju mechanizmów podnoszących wiarygodność usług sieciowych, konieczne jest stosowanie w korzystających z tych usług aplikacjach metod rozwiązywania problemów wynikających z nieprawidłowego działania usług. Do takich problemów należy sytuacja, w której przekazywane są przez usługi dane nieprawidłowe lub sprzeczne. Szczególnie problematyczne jest to, że dane mogą być błędne, mimo że zostały przekazane zgodnie ze stosowanymi protokołami komunikacji. Istnieje wiele powodów występowania takich danych. Opracowywane do tej pory metody tolerowania błędów skupiają się głównie na przyczynach wynikających z błędów podczas komunikacji oraz ingerencji osób trzecich. Jednak błędy pojawić się mogą nawet wtedy, gdy te

Recenzent: Dr inż. Agnieszka Landowska – Wydział Elektroniki, Telekomunikacji i Informatyki Politechnika Gdańska

przyczyny nie występują. Błędy mogą być spowodowane tym, że dane są zdezaktualizowane lub są one prawdziwe jedynie w ściśle określonym kontekście ich stosowania. Ponadto niepoprawność danych może wynikać z nierzetelności źródła tych danych i dostarczania informacji bez należytej weryfikacji ich poprawności.

W przypadku, gdy aplikacja bazuje na danych dostarczanych z usług, przyjęcie przez nią za prawdziwe błędnych danych prowadzi w konsekwencji do niepoprawnego jej działania i podważania jej wiarygodności. Podstawowym sposobem weryfikacji prawdziwości informacji jej sprawdzanie jej w różnych źródłach. Tak na przykład agencje prasowe kierują się zasadą, że informacja jest uznawana za prawdziwą, jeśli została ona potwierdzona przez dwa niezależne źródła. W przypadku danych pochodzących z serwisów internetowych możliwe jest sprawdzanie w różnych usługach sieciowych danych odnoszących się do tej samej informacji, a następnie porównywanie tych danych. Wymaga to jednak stosowania metod rozwiązywania problemów pojawiających się wtedy, gdy dane pozyskane z różnych źródeł są ze sobą sprzeczne. Gdyby takie metody nie były stosowane, to występowanie sprzecznych danych prowadziłoby na podstawie klasycznych zasad logiki do podważenia prawdziwości wszystkich zgromadzonych danych. Zgodnie z zasadą nazywaną *ex falso quodlibet*, ze sprzeczności wynika dowolnego rodzaju zdanie logiczne.

Możliwe jest zastosowanie kilku rodzajów rozwiązań problemu sprzeczności danych. Jednym z nich są metody oparte na szacowaniu prawdopodobieństwa prawdziwości danych. Do takich metod zaliczyć można teorię Dempster-Shafera [5]. Rozwiązane problemu sprzeczności jest również możliwe dzięki zastosowaniu innych niż klasyczne zasad logiki [6]. Zasady takie mogą wywodzić się z logiki klasycznej. Przykładowo, w skład takich zasad mogą wchodzić wszystkie prawa logiki boolowskiej, oprócz prawa polegającego na tym, że koniunkcja zdania logicznego oraz zaprzeczenia tego zdania jest zawsze zdaniem fałszywym. Możliwe jest również zastosowanie innego podejścia do rozwiązywania problemu sprzeczności danych polegającego na tym, że część danych uznawana jest za błędną i jest wykluczana.

3. TEORIA DEMPSTER-SHAFERA

Problem interpretacji sprzecznych ze sobą danych jest przedmiotem teorii Demster-Shaffera [5] będącej uogólnieniem klasycznego rachunku prawdopodo-bieństwa opartego na twierdzeniu Bayesa. Definiuje się w niej pojęcie przekonania (ang. belief) oraz wiarygodności (ang. plausibility). Poziomy przekonania oraz wiarygodności dotyczące pewnej informacji obliczane są na podstawie danych podawanych przez różne źródła informacji. Maksymalna możliwa wartość poziomu przekonania i poziomu wiarygodności jest równa jeden, podobnie jak maksymalna wartość prawdopodobieństwa zajścia dowolnego zdarzenia. W teorii Dempster-Shafera w szczególnie charakterystyczny sposób obliczane są wartości poziomów przekonania. W przypadku, gdy podawane są przez różne źródła informacji dane sprzeczne dotyczące informacji na ten sam temat, określany jest poziom przekonania odnoszący się do tego, że dana jest prawdziwa oraz poziom przekonania, że dana jest nieprawdziwa. Poziomy te przyjmują, podobnie jak wielkości prawdopodobieństwa, wartości od 0 do 1. W przypadku klasycznego rachunku prawdopodobieństwa suma prawdopodobieństwa tego, że dana jest prawdziwa oraz tego, że jest ona nieprawdziwa wynosi 1. Jednak w teorii Demster-Shafera poziomy przekonania nie sumują się do jedności, pozostawiając pewien zakres odnoszący się do tego, że prawdziwość danych nie jest możliwa do rozstrzygnięcia.

W teorii Dempster-Shafera definiowane jest również pojęcie wiarygodności. Wartość poziomu wiarygodności dotycząca tego, że pewna dana jest prawdziwa, jest równa odjęciu od jedności poziomu przekonania odnoszącego się do tego, że dana ta jest nieprawdziwa. Wartości poziomów wiarygodności, podobnie jak poziomów przekonania, przyjmują wartości od 0 do 1.

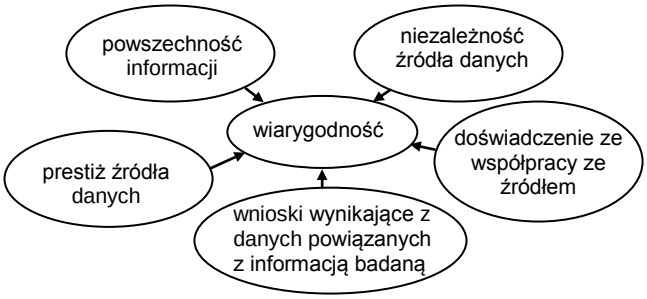
Teoria Dempster-Shafera stosowana jest przede wszystkich przy interpretacji odczytów sensorów. Dane pochodzące z sensorów, na przykład sensorów ruchu, mogą dostarczać sprzecznych informacji. W szczególności informacje te dotyczyć mogą tego, czy na obserwowanym obszarze występuje ruch, czy też nie. Teoria Dempster-Shafera pozwana na integrację odczytów wielu sensorów i interpretowanie wyników przez sensory podawanych. Teoria ta została również zaadaptowana na potrzeby usług sieciowych [7].

4. IDENTYFIKOWANIE NIEWIARYGODNYCH ŹRÓDEŁ INFORMACJI

Oprócz metod bazujących na wyznaczaniu prawdopodobieństwa prawdziwości danych, problem sprzeczności rozwiązać można dzięki stosowaniu nieboolowskich zasad logiki. Jednak przeprowadzanie wnioskowania na podstawie takich zasad niepotrzebnie komplikuje przetwarzanie danych, gdyż częstym powodem występowania sprzeczności jest to, że część danych jest błędna.

W przypadku pozyskiwania danych z różnych źródeł nie można mówić o tym, że uzyskuje się prawdziwe dane. Ze źródeł informacji nie są pozyskiwane dane, które na pewno są prawdziwe, lecz pewne tezy (ang. claims) podawane przez te źródła [8]. Sprzeczności występujące wśród tez podawanych przez różne źródła wynikać mogą z tego, że przynajmniej jedno z tych źródeł jest niewiarygodne i podało błędne dane. Rozwiązaniem problemu sprzeczności danych jest wtedy pominięcie danych pochodzących z niewiarygodnego źródła.

Opracowane zostały metody służące identyfikowa-niu i wykluczaniu błędnych danych pojawiających się w sieci semantycznej (ang. Semantic Web) oraz strukturach sieci wiedzy (ang. knowledge grid) [9]. Metoda opisana w publikacji [9] polega na ocenie wiarygodności danych na podstawie pięciu metryk: powszechności informacji (ang. information commonality), niezależności źródła informacji (ang. source independence), prestiżu źródła (ang. prestige of the source), doświadczenia ze współpracy ze źródłem (ang. experience with the source) oraz wniosków wynikających z innych pozyskanych informacji (ang. conclusions from related information).



Rys. 1. Metryki służące do oceny wiarygodności danych

Stosowana w omawianej metodzie metryka powszechności polega na zwiększaniu szacowanej wiarygodności informacji, gdy znajdowane są kolejne źródła podające tę informację. Metryka niezależność źródła, brana też pod uwagę w tej metodzie, opiera się na tym, że za bardziej wiarygodną uznawana jest informacja podawana przez dwa niezależne od siebie źródła, niż taka, która jest podawana przez dwa źródła w przypadku, gdy jedno z tych źródeł przekazuje informacje pozyskane z drugiego źródła. Występuje również metryka prestiżu źródła polegająca na tym, że zwiększana jest wiarygodność informacji pochodzących ze znanych i uznawanych instytucji. Metryka doświadczenia ze współpracy ze źródłem związana jest z utratą zaufania do źródła w przypadku wystąpienia sytuacji, w której źródło podało błędną informację. Ostatnią braną pod uwagę metryką jest metryka wniosków. W metryce tej w celu weryfikacji wiarygodności informacji brane są pod uwagę inne informacje związane z informacją badaną. System pobierający informacje może posiadać pewien zasób danych wskazujący na to, czy nowa informacja jest prawdziwa czy nie.

Z pomocą wyżej przedstawionej metody możliwe jest wykluczenie danych powodujących w przypadku, gdy takie sprzeczności wystąpiły. Spośród wszystkich zgromadzonych danych za błędne uznawane są te dane, które prowadzą do sprzeczności, a jednocześnie posiadają niski poziom wiarygodności.

5. WIARYGODNOŚĆ DANYCH Z USŁUG SIECIOWYCH

Podobnie, jak w wyżej opisany sposób, przeprowadzać można ocenę wiarygodności danych pozyskiwanych z usług sieciowych. Ocena polegałaby wówczas na zastosowaniu metryk wyznaczających poziom wiarygodności. Wiarygodność taka różniłaby się od wiarygodności określanej na podstawie teorii Demster-Shaffera tym, że wiarygodność byłaby wyznaczana na podstawie wartości metryk, a poziomy prawdopo-dobieństwa prawdziwości danych nie byłyby określane.

Przy korzystaniu z usług sieciowych nie jest konieczne pobieranie danych jedynie z jednego źródła. Korzystanie z usług sieciowych charakteryzuje się tym, że istnieje możliwość wyboru spośród wielu podobnych do siebie usług oferowanych przez różnych dostawców. W celu zwiększenia wiarygodności danych możliwe jest pobieranie tego samego rodzaju danych z wielu różnych źródeł informacji. Dzięki temu można z większą pewnością określać wiarygodność danych niż wtedy, gdy dane pozyskane zostały z jedynie jednego źródła.

Na podstawie przedstawionej w rozdziale 4 metody oceny wiarygodności danych dostępnych w strukturach sieci semantycznej i sieci wiedzy sformułować można metryki oceny danych pozyskanych za pomocą usług sieciowych. Do oceny wiarygodności danych w usługach sieciowych zastosowanie mają cztery następujące metryki: powszechności informacji, niezależności źródła informacji, prestiżu źródła oraz doświadczenia ze współpracy ze źródłem. Oznaczając przez *w* poziom wiarygodności pewnej danej podlegającej ocenie, poziom ten jest równy wartości podanej wzorem 1.

$$w = \sum_{i=1}^{Q} \left(N_i P_i D_i \right) \tag{1}$$

gdzie: w – poziom wiarygodności, Q – powszechność informacji, N – niezależność źródła informacji, P – prestiż źródła, D – doświadczenie ze współpracy ze źródłem

Powszechność informacji oznacza, że im więcej źródeł podaje pewną informację, tym większa jest jej wiarygodność. Wiarygodność pojedynczej informacji jest jednak modyfikowana przez wartości innych metryk. Jeśli brana byłaby pod uwagę tylko metryka powszechność informacji, to wiarygodność informacji byłaby równa liczbie źródeł potwierdzających tą informację.

Metryka niezależność źródła polega na tym, że dwa źródła czerpiące dane od siebie nawzajem uznawane są za mniej wiarygodne niż takie, które nie wymieniają ze sobą informacji. Wynika to z tego, że jeśli dane są przykazywane między źródłami, to w przypadku, gdy dane te byłyby nieprawdziwe, dochodziłoby do propagacji danych błędnych. Ustalając wpływ tej metryki na poziom wiarygodności, konieczne jest ustalenie, czy należy uznawać jedno źródło informacji za równie wiarygodne, co dwa zależne od siebie źródła. W pracy [9] przyjęto, że dwa zależne od siebie źródła są bardziej wiarygodne niż jedno źródło niezależne. Ponadto przyjęto, że każde kolejne zależne źródło zwiększa wiarygodność danej pobranej ze źródła w połowie mniejszym stopniu niż źródło poprzednie. Wartość metryki niezależność, oznaczonej literą *N*, gdy liczba źródeł zależnych jest oznaczona literą *z*, równa jest wyrażeniu $(0,5)^z$. Współczynnik *z* jest zwiększany dla każdego kolejnego zależnego źródła. Wartość poziomu wiarygodności zależeć może od kolejności analizowania danych z poszczególnych źródeł z powodu wartości pozostałych metryk. W związku z tym przyjmuje się, że dane analizowane są w kolejności od danych o największym poziomie wiarygodności obliczonym na podstawie pozostałych metryk do danych o poziomie najniższym.

Kolejną metryką jest prestiż źródła. W pracy [9] przyjęte zostały dwa poziomy prestiżu: instytucje charakteryzujące się wysokim prestiżem oraz instytucje pozostałe. Instytucjom o wysokim prestiżu przypisany został poziom 1, natomiast instytucjom pozostałym poziom 0,5. Wartości takie przyjmuje współczynnik *P*, w zależności od rodzaju źródła, z którego dana pochodzi.

Stosowana jest też metryka doświadczenie ze współ-pracy ze źródłem. Służy ona do obliczania wiarygodności danych pochodzących ze źródeł, o których wiadomo, że dawniej podawały błędne dane. Można przypuszczać, że źródła takie będą w przyszłości częściej podawały dane nieprawidłowe. Przyjęto, że podanie przez źródło błędnych danych obniża jego wiarygodność o połowę.

Uwzględniona w pracy [9] metryka wniosków wynikających z innych pozyskanych informacji nie jest stosowana do określania wiarygodności danych. Wynika to z tego, że w odróżnieniu od danych pochodzących z usług sieciowych, dane w sieci semantycznej i sieci wiedzy podawane są w postaci zdań logicznych. Dane takie można w szerokim zakresie analizować, znajdować zachodzące w nich zależności i wyciągać na ich podstawie wnioski. Natomiast dane pozyskiwane z usług sieciowych mogą być zbiorem informacji, których przetwarzanie w formie zależności logicznych jest utrudnione. Z analogicznych powodów przy ustalaniu poziomu wiarygodności danych z usług sieciowych nie są brane pod uwagę dane sprzeczne z danymi ocenianymi. Ocena wiarygodności odbywa się tylko na podstawie danych potwierdzających badaną informację.

Metryki niezależności, prestiżu oraz doświadczenia ze współpracy ze źródłem nie odnoszą się bezpośrednio do danych, lecz do źródła, z którego dane te pochodzą. Metryki te przyjmować będą te same wartości dla wszystkich danych pochodzących z tego samego źródła. W celu określenia tych metryk konieczne jest posiadanie informacji dotyczących źródła danych. Prowadzone są liczne prace dotyczące metod rejestrowania pochodzenia danych (ang. data provenance) [10]. Dzięki takim metodom możliwe jest wyznaczanie wartości opisywanych metryk.

6. PODSUMOWANIE

Przedstawiona metoda oceny wiarygodności pozwala automatycznie weryfikować wiarygodność danych na podstawie liczby źródeł informacji potwierdzających weryfikowaną daną oraz charakterystyki tych źródeł. Metoda umożliwia zwiększenie niezawodności aplikacji, w których integrowane są usługi sieciowe. Metoda może być też w szerokim stopniu rozwijana, zarówno w kontekście wprowadzania nowych metryk, jak i rozbudowy sposobu wyznaczania metryk już wprowadzonych.

7. PODZIĘKOWANIA

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

8. BIBLIOGRAFIA

1. Newcomer E., Lomow G.: Understanding SOA with Web Services, Addison-Wesley, 2004, ISBN: 978-0321180865
2. Milanovic N, Malek M.: Current Solutions for Web Service Composition, IEEE Internet Computing, vol. 8, no. 6, Nov./Dec. 2004, s. 51-59, ISSN: 1089-7801
3. Strigini L., Neves N., Raynal M., Harrison M., Kaaniche M., von Henke F.: Resilience-Building Technologies: State of Knowledge, Raport Techniczny TR–07–26, ReSIST: Resilience for Survivability in IST, Faculdade de Ciencias da Universidade de Lisboa, Lisboa, Portugal, November 2007
4. Cardoso J.: Semantic Integration of Web services and Peerto-Peer Networks to Achieve Fault-tolerance, IEEE International Conference on Granular Computing, May 2006, s. 796–799, ISBN: 1-4244-0134-8
5. Shafer G.: A mathematical theory of evidence, Princeton University Press, 1976, ISBN: 978-0691100425
6. Schaffert S., Bry F., Besnard P., Decker H., Decker S., Enguix C., Herzig A., Position Paper: Paraconsistent Reasoning for the Semantic Web, Proceedings of Workshop Uncertainty Reasoning for the Semantic Web, Galway, Ireland, 2005
7. Liu J., Xiang Z., Zhu P.: Trust evolvement method of Web service combination based on network behavior, Journal of Central South University of Technology, Vol. 15, No. 4, 2008, s. 558−563, ISSN: 1005-9784
8. Bizer C., Oldakowski R.: Using Context- and Content-Based Trust Policies on the Semantic Web, WWW 2004, New York, NY, USA, ACM, May 2004, s. 228 – 229, ISBN:1-58113-912-8
9. Kaczmarek A. Ł.: Automatic Evaluation of Information Credibility in Semantic Web and Knowledge Grid, WEBIST 2008, Proceesings of the Fourth International Conference on Web Information Systems and Technologies, Vol. 2, Funchal, Madeira - Portugal, 4-7 May 2008, INSTICC, Portugal, 2008, s. 275-278, ISBN: 978-989-8111-27-2
10. Tsai W. T., Wei X., Chen Y., Paul R., Chung J. and Zhang D.: Data provenance in SOA: security, reliability, and integrity, Service Oriented Computing and Applications, vol. 1, no. 4, Springer-Verlag, Dec. 2007, pp. 223-247, ISSN: 1863-2386

WEB SERVICES INTEGRATION CONSIDERING THE LEVEL OF VENDORS BELIEVABILITY

Key-words: data believability, web services

This paper is concerned with the believability of data acquired from web services. In the paper a method for estimating data believability is presented. The estimation is based on four metrics: information commonality, source independence, prestige of the source and experience with the source. Presented method supports the integration of web services provided by various vendors. The method makes possible to automatically determine the level of data believability on the basis of data provenance.