Adaptatywne serwery WWW

Marek Wojciechowski, Maciej Zakrzewicz
Politechnika Poznańska, Instytut Informatyki
ul. Piotrowo 3a, 60-965 Poznań
e-mail: {marek,mzakrz}@cs.put.poznan.pl

Abstrakt. Adaptatywne serwery WWW wykorzystują analizę plików logu w celu automatycznej transformacji zawartości i struktury udostępnianych dokumentów. W rezultacie, serwer WWW samodzielnie „dopasowuje" się do oczekiwań użytkownika, „odgadując" jego intencje. W artykule przedstawiono dostępne metody zautomatyzowanej analizy plików logu oraz stosowania znalezionych trendów i korelacji w dynamicznej transformacji dokumentów WWW..

1. Wprowadzenie

Projektowanie struktury zawartości serwera WWW jest w ogólności problemem złożonym i trudnym. Projektanci podejmują zadanie takiego opracowania wyglądu i powiązań pomiędzy dokumentami WWW, aby były one czytelne i łatwe w nawigacji dla użytkowników. Na problem projektowania struktury zawartości serwera WWW mają wpływ następujące czynniki:

1. *Różni użytkownicy* mogą korzystać z serwera WWW celu znalezienia *innych informacji*.

   Przykładowo, jeżeli do sklepu internetowego przyłącza się użytkownik młody, to dobrze byłoby, aby strona główna zawierała informacje o najnowszych grach komputerowych. Kiedy jednak przesłania strony głównej zażąda użytkownik starszy, wtedy wskazane byłoby umieszczenie na niej informacji o najnowszych pozycjach książkowych i płytach z muzyką klasyczną.

2. W *różnych momentach czasowych*, jeden użytkownik może poszukiwać *innych informacji*.

   Przykładowo, użytkownik internetowego biura podróży będzie w okresie zimowym zainteresowany dokumentami WWW zawierającymi informacje o kurortach narciarskich w Alpach, natomiast w okresie letnim, ten sam użytkownik życzyłby sobie prezentacji dokumentów WWW opisujących wczasy w basenie Morza Śródziemnego.

3. Zawartość serwera WWW *rozrasta się* wraz z upływem czasu.

   Gdy do istniejącego systemu dodawane są nowe dokumenty WWW, projektant musi podjąć odpowiedzialną decyzję o tym, w których z dotychczasowych dokumentów umieścić łączniki do nowej części systemu.

Rozwiązaniem technicznym, które adresuje przedstawione problemy jest *personalizacja* zawartości serwerów WWW. Personalizacja polega na wykorzystywaniu wiedzy o preferencjach użytkowników do dynamicznego dostosowywania wyglądu i struktury przesyłanych dokumentów. Dzięki temu, każdy użytkownik może otrzymywać inny obraz zawartości i struktury dokumentów tego samego serwera. Wiedza o preferencjach użytkowników może być pozyskiwana *jawnie*, poprzez dostarczenie użytkownikom formularzy i narzędzi o charakterze konfiguracyjnym, bądź *niejawnie* – w wyniku obserwacji stylu ich dotychczasowej interakcji z serwerem. Wiele stosowanych dziś rozwiązań w zakresie personalizacji zawartości serwerów WWW silnie bazuje na informacjach uzyskanych od użytkownika w sposób jawny. Taka forma pozyskiwania wiedzy cechuje się dużą subiektywnością i spotyka się z niechęcią użytkowników, „zmuszanych" do wypełniania dodatkowych formularzy i ankiet. Ponadto, tak budowane profile użytkowników posiadają charakter statyczny i z upływem czasu ulegają degradacji.

W ostatnich latach coraz większą uwagę przyciągają metody personalizacji zawartości serwerów WWW poprzez niejawne obserwowanie trendów w zachowaniach użytkowników WWW. W pracy

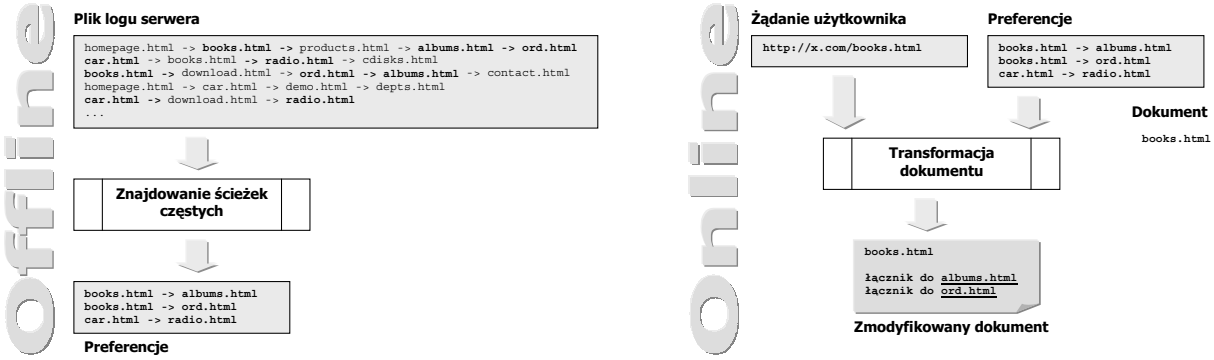
[PE97], zaproponowano termin *adaptatywne serwery WWW* (adaptive web sites), opisujący *serwery WWW, które automatycznie ulepszają swoją zawartość i organizację na podstawie obserwacji ścieżek dostępów użytkowników*. Idea adaptatywnych serwerów polega na analizie plików logu serwera, wyławianiu z nich statystycznych korelacji pomiędzy pobieranymi dokumentami lub pracującymi użytkownikami, a następnie wykorzystywaniu znalezionych korelacji do budowy struktury dokumentów WWW, wysyłanych użytkownikom. W tym artykule opisujemy stan nauki i technologii w zakresie metod konstrukcji adaptatywnych serwerów WWW.

2. Automatyczna adaptacja serwera WWW

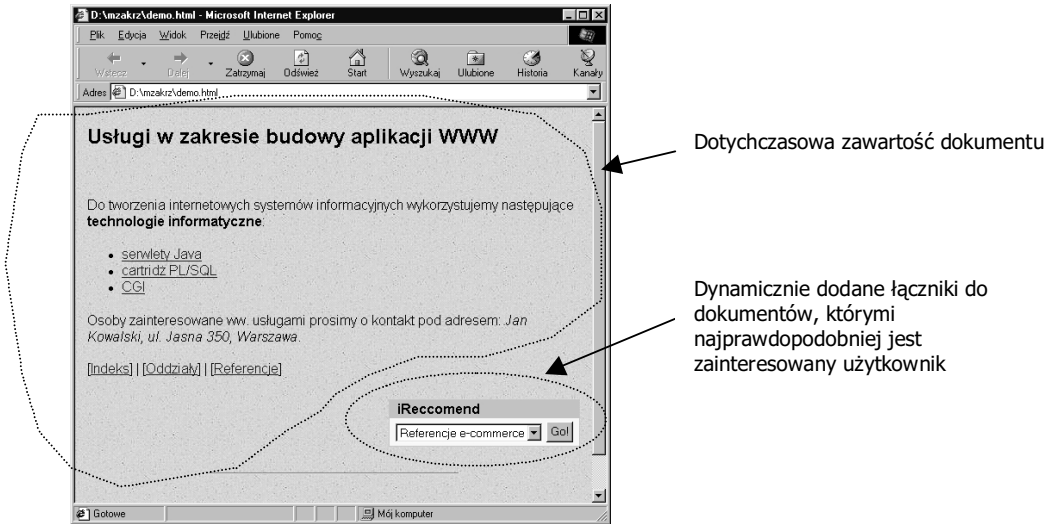
Proces adaptacji serwera WWW przebiega w dwóch fazach:

1. Offline: wykorzystanie pliku logu serwera do znalezienia i pogrupowania najczęstszych ścieżek nawigacyjnych użytkowników. Faza ta realizowana jest asynchronicznie względem połączeń użytkowników, np. w odstępach tygodniowych lub miesięcznych.
2. Online: wykorzystywanie znalezionych grup ścieżek nawigacyjnych do tworzenia *dynamicznych rekomendacji* dla użytkowników, czyli zbioru łączników do dokumentów, którymi ci użytkownicy będą najprawdopodobniej (statystycznie) zainteresowani. Faza ta jest realizowana podczas obsługi każdego żądania użytkownika.

Przedstawmy przykład prostej adaptacji serwera WWW, zilustrowany na rysunku 1. Serwer WWW został odwiedzony przez pięciu użytkowników, których pełne ścieżki nawigacyjne zostały zapisane w pliku logu. W pierwszej fazie adaptacji (offline) wykonywana jest analiza pliku logu i znalezione zostają następujące ścieżki częste: `books.html -> albums.html`, `books.html -> ord.html`, `car.html -> radio.html`. Każda z tych ścieżek pojawiła się w 40% odwiedzin opisanych w pliku logu i w związku z tym będą one traktowane przez nas jako preferencje dla innych użytkowników. W drugiej fazie (online), nowy użytkownik wysyła do serwera żądanie przesłania dokumentu WWW (`books.html`). Serwer pobiera dokument z dysku i przegląda znaleziony wcześniej zbiór ścieżek częstych – wynika z niego, że użytkownicy, którzy pobierali dokument `books.html`, byli później zainteresowani dokumentami `albums.html` i `ord.html`. W związku z tym, w celu ułatwienia nawigacji, do dokumentu `books.html` dynamicznie dodawane są łączniki do powyższych dokumentów. Tak zmodyfikowany dokument trafia do użytkownika (rysunek 2).



Rys. 1. Przykładowy proces adaptacji serwera WWW



Rys. 2. Przykłady dokumentów WWW wzbogaconych o dynamiczne rekomendacje

2.1 Faza Offline algorytmu

Struktura pliku logu

Informacje o dostępach do serwera WWW zapisywane są w logu. Dla każdego dostępu do pojedynczego pliku znajdującego się na serwerze, w logu pojawia się nowy zapis. Jednakże ilość informacji pamiętana w związku z danym dostępem może być różna w przypadku różnych serwerów WWW. Aby umożliwić tworzenie uniwersalnych narzędzi służących do analizy logu, pojawiły się próby standaryzacji jego formatu. Dzisiaj można założyć, że przeważająca większość serwerów WWW generuje pliki logu zgodne z formatem znanym pod nazwą *Common Logfile Format* [L95]. Nie jest to jednak w pełni obowiązujący standard, gdyż niektóre serwery pamiętają również pewne dodatkowe informacje (standard *XLF*). *Common Logfile Format* przewiduje, że zapis w logu powinien mieć następującą postać:

```
remotehost rfc931 authuser [date] "request" status bytes
```

W powyższym formacie pole *remotehost* oznacza nazwę lub adres IP komputera, z którego nastąpiło odwołanie. Pole *rfc931* zawiera nazwę użytkownika na danym komputerze (*ang. logname*). Pole *authuser* zawiera informację o tym, za kogo użytkownik się podaje. Pole [*date*] informuje o tym kiedy nastąpiło odwołanie (data i czas). Pole "*request*" zawiera żądanie przesłane do serwera w takiej formie, w jakiej wygenerował je klient. Obejmuje ono na ogół typ operacji i nazwę pliku, do którego nastąpiło odwołanie, wraz ze ścieżką dostępu. Pole *status* zawiera zwracany klientowi kod statusu, zgodnie z protokołem HTTP wykorzystywanym w usłudze WWW. Długość zawartości przesyłanego dokumentu pamiętana jest w polu *bytes*. Przykład zawartości pliku logu serwera WWW przedstawiono na rysunku 3.

```
154.11.231.17 - - [13/Jul/2000:20:42:25 +0200] "GET / HTTP/1.1" 200 1673

154.11.231.17 - - [13/Jul/2000:20:42:25 +0200] "GET /apache_pb.gif HTTP/1.1" 200 2326
```

```
192.168.1.25 - - [13/Jul/2000:20:42:25 +0200] "GET /demo.html HTTP/1.1" 200 520

192.168.1.25 - - [13/Jul/2000:20:42:25 +0200] "GET /books.html HTTP/1.1" 200 3402

160.81.77.20 - - [13/Jul/2000:20:42:25 +0200] "GET / HTTP/1.1" 200 1673

154.11.231.17 - - [13/Jul/2000:20:42:25 +0200] "GET /car.html HTTP/1.1" 200 2580

192.168.1.25 - - [13/Jul/2000:20:42:25 +0200] "GET /cdisk.html HTTP/1.1" 200 3856

10.111.62.101 - - [13/Jul/2000:20:42:25 +0200] "GET /new/demo.html HTTP/1.1" 200 971
```

Rys. 3. Przykładowy plik logu serwera WWW.

Identyfikacja ścieżek nawigacyjnych

Z punktu widzenia analizy istotnymi informacjami w logu serwera WWW są: nazwa lub adres IP komputera, z którego nastąpiło odwołanie, nazwa użytkownika dokonującego odwołania, dokładna data i czas oraz pełna nazwa pliku, którego dotyczyło żądanie. Analiza plików logu polega na znajdowaniu często powtarzających się sekwencji w ścieżkach dostępów użytkowników do serwera WWW lub na grupowaniu użytkowników wykazujących podobne zachowanie. Z tego powodu koniecznym etapem wstępnej obróbki danych zawartych w logu jest grupowanie zapisów dotyczących odwołań tego samego użytkownika. Grupowanie to odbywa się na podstawie adresu IP lub nazwy komputera oraz nazwy użytkownika. Niestety nie zawsze nazwa użytkownika jest znana. Sytuacja taka ma miejsce często w przypadku gdy użytkownik korzysta z systemu operacyjnego, który nie zakłada wielodostępu. Na szczęście fakt, że z komputera pracującego pod kontrolą systemu operacyjnego, który nie jest wielodostępny, może w danej chwili korzystać tylko jeden użytkownik, pozwala traktować odwołania pochodzące z tego samego komputera jako odwołania jednego użytkownika, gdy nazwa użytkownika nie jest znana. Oczywiście powyższe założenie jest poprawne tylko w przypadku odwołań, których czasy zawierają się w okresie odpowiadającym możliwemu czasowi trwania pojedynczej sesji użytkownika. Mechanizm ten nie pozwala więc na identyfikację sekwencji dostępów w ramach wielu sesji użytkownika na przestrzeni np. miesiąca, gdyż z danego komputera może w różnych godzinach korzystać wiele osób.

Ze względu na fakt, że użytkownik może wielokrotnie korzystać z usług danego serwera WWW za każdym razem szukając innych informacji, niekiedy wskazane jest rozbicie sekwencji dostępów danego użytkownika na fragmenty odpowiadające poszczególnym sesjom. Nie jest to jednak zadanie trywialne, gdyż protokół http nie posługuje się pojęciem sesji. Najprostsze rozwiązanie tego problemu polega na wyodrębnianiu sesji użytkowników w oparciu o założenie, że jeśli czas między kolejnymi odwołaniami do serwera jest znacznie dłuższy niż typowy czas przeglądania jednej strony, to odwołania te nastąpiły w ramach dwóch różnych sesji. Alternatywnym rozwiązaniem może być rozszerzenie funkcjonalności serwera o obsługę identyfikatorów sesji na czas zbierania informacji o zachowaniach użytkowników [YJG+96].

Celem analizy plików logu serwera WWW może być znajdowanie częstych ścieżek nawigacji lub znajdowanie grup stron, do których użytkownicy często odwołują się w ramach sesji. W pierwszym przypadku istotne są informacje o wszystkich stronach, do których odwoływał się użytkownik z uwzględnieniem kolejności odwołań. W pozostałych przypadkach może się okazać, że istotne są odwołania tylko do tych stron, których treść zainteresowała użytkownika (strony służące jedynie jako ścieżka dostępu do szukanego dokumentu nie są uwzględniane). W [CMS97] zaproponowano podział odwołań do stron na zorientowane na zawartość i zorientowane na nawigację. Niektóre strony zawierają głównie odnośniki do innych stron, w związku z czym odwołania do nich na pewno będą miały charakter nawigacyjny. Jednakże wiele stron zawiera zarówno treść jak i odnośniki do innych stron. Takie strony mogą różnym użytkownikom służyć do różnych celów. Dlatego rozsądnym kryterium podziału dostępów na zorientowane na nawigację i zawartość wydaje się czas, na jaki użytkownik zatrzymuje się na danej stronie (być może znormalizowany w stosunku do rozmiaru strony). Czas przeglądania danej strony jest obliczany

jako różnica etykiet czasowych dwóch kolejnych zapisów w logu (odpowiadających następnej i bieżącej stronie). W przypadku stron kończących sesję użytkownika przyjmuje się, że dostęp do nich miał miejsce ze względu na ich zawartość, choć oczywiście w konkretnym przypadku wcale nie musi to być prawdą.

Problemy obróbki plików logu

Informacje zawarte w logu mogą być nie tylko niepełne, ale również zafałszowane ze względu na wykorzystywanie serwerów proxy i podręcznej pamięci przeglądarek [P97]. Serwer proxy służy jako „okno na świat" dla wielu komputerów, pozwalając uzyskać dostęp do Internetu użytkownikom na nich pracującym. Zapisy w logu serwera WWW odpowiadające odwołaniom użytkowników komputerów „ukrytych" za serwerem proxy są opisane adresem serwera proxy. W związku z tym fakt, że kilka zapisów w logu dotyczy jednego adresu IP, nie musi wcale oznaczać, iż zapisy te odpowiadają odwołaniom z tego samego komputera. W [PPR96] zaproponowano metodę wykrywania takich sytuacji w oparciu o założenie, że jeśli dane odwołanie dotyczy dokumentu, do którego nie ma łącza w poprzednio żądanym dokumencie, to prawdopodobnie żądania są kierowane przez dwóch różnych użytkowników. Mimo że doświadczenia pokazują [CP95], iż dostęp do kolejnego dokumentu jest najczęściej wynikiem wybrania dostępnego w dokumencie łącza (*ang. hyperlink*) lub powrotem do poprzedniego dokumentu (operacja „*Back*"), wspomniana metoda nie gwarantuje żadnej pewności. Dlatego dla celów identyfikacji użytkowników stosuje się tzw. *cookies* lub dodatkową autoryzację. Cookie jest identyfikatorem generowanym przez serwer i przesyłanym do klienta (przeglądarki) w celu późniejszej identyfikacji użytkownika. Niedoskonałość tego mechanizmu wynika z faktu, że użytkownicy mogą w dowolnej chwili usunąć cookie lub w ogóle zabronić akceptacji cookies. Dodatkowa identyfikacja użytkowników poprzez żądanie wypełnienia formatki rejestracyjnej również wymaga dobrej woli użytkowników, gdyż dane przez nich podawane mogą być przecież fałszywe.

Równie istotnym problemem jak identyfikacja użytkowników jest identyfikacja faktycznych odwołań do dokumentów. Ze względu na stosowanie przez przeglądarki pamięci podręcznej, kolejne odwołania danego użytkownika do tego samego dokumentu mogą nie być odnotowane na serwerze, gdyż mogą być zrealizowane przez sprowadzenie dokumentu z pamięci podręcznej przeglądarki a nie z serwera. W sposób znaczący może to zakłócić odkryte ścieżki nawigacji użytkowników. Jeszcze poważniejszy problem wynika ze stosowania pamięci podręcznej przez serwery proxy. Jeśli użytkownik, korzystający z Internetu poprzez serwer proxy, odwołuje się do dokumentu znajdującego się w pamięci podręcznej proxy, serwer WWW może być w ogóle nieświadomy, że dany użytkownik odwoływał się do danego dokumentu. Aby obronić się przed wspomnianymi sytuacjami serwery WWW mogą stosować techniki zapobiegające wykorzystywaniu pamięci podręcznej określane jako *cache-busting*, polegające np. na podawaniu dat z przeszłości jako terminów upłynięcia ważności poszczególnych dokumentów. Tego typu techniki mogą być uciążliwe dla użytkowników, gdyż wydłużają czas odpowiedzi. Z tego względu pojawiły się propozycje, aby zamiast monitorowania wszystkich dostępów do serwera, ograniczyć się tylko do pewnej próbki statystycznej i na jej bazie dokonywać analiz.

Czyszczenie plików logu

Proces wstępnej obróbki danych nie kończy się na identyfikacji odwołań poszczególnych użytkowników. Zapisy w logu dotyczą pojedynczych plików, a nie dokumentów traktowanych jako obiekty złożone. W przypadku dostępu do strony zawierającej np. obrazy, dźwięki lub filmy, w logu znajdzie się zapis dotyczący głównego dokumentu (najczęściej z rozszerzeniem *html* lub *htm*), ale także zapisy związane ze wszystkimi obiektami zagnieżdżonymi w stronie (obrazami, filmami, itp.). Na szczęście charakter pliku można w dużym stopniu wywnioskować z jego rozszerzenia. Przykładowe rozszerzenia nazw plików odpowiadające obiektom zagnieżdżanym w dokumentach to *jpg, jpeg, gif* dla obrazów, *au, wav* dla dźwięków, *avi, mov* dla filmów. Aby dane źródłowe do analiz zawierały tylko informacje o dostępach do istotnych dokumentów, należy poddać plik logu serwera WWW procesowi filtracji, w wyniku którego ignorowane są zapisy dotyczące plików nie będących głównymi dokumentami odpowiadającymi tzw. stronom WWW (*ang. Web page*).

Odkrywanie preferencji użytkowników

Preferencje użytkowników są reprezentowane przez zbiory podobnych najczęściej stosowanych ścieżek nawigacyjnych. W celu znalezienia preferencji, realizowany jest dwufazowy algorytm:

1. Przeszukaj log serwera WWW w celu znalezienia wszystkich najczęściej występujących ścieżek nawigacyjnych.
2. Pogrupuj znalezione ścieżki nawigacyjne, kierując się ich współstosowaniem przez użytkowników (tzn. podobieństwo dwóch ścieżek wynika z tego, iż wielu użytkowników, którzy podążają jedną z nich, podąża również drugą).

2.2. Faza Online algorytmu

Od chwili pierwszego podłączenia się użytkownika do serwera WWW, wszystkie operacje tego użytkownika są rejestrowane w formie tzw. *historii sesji*. Za każdym razem, kiedy użytkownik żąda przesłania dokumentu, historia jego sesji jest dopasowywana do istniejących grup ścieżek nawigacyjnych i wybierane są te grupy, które wykazują się największym dopasowaniem. Zbiór łączników do dokumentów opisanych w ścieżkach nawigacyjnych dopasowanych grup staje się dodatkowym elementem wizualnym, który dynamicznie jest dołączany do żądanego dokumentu [YJG+96].

3. Podsumowanie

W artykule przedstawiono architekturę systemu automatycznej personalizacji zawartości serwerów WWW, który umożliwia tworzenie środowisk WWW dopasowujących się do zachowań użytkowników. Obecnie w Instytucie Informatyki Politechniki Poznańskiej rozwijany jest moduł rozszerzający funkcjonalność Oracle Application Servera o tak rozumianą adaptatywność. Dzięki zastosowaniu zaprezentowanej filozofii, część odpowiedzialności za wygląd i strukturę zawartości serwera WWW jest przenoszona z projektantów na użytkowników.

Literatura

[CP95] Catledge L.D., Pitkow J.E., ”Characterizing Browsing Strategies in the World Wide Web”, Proc. of the 3rd Int’l World Wide Web Conference, 1995.

[CM99] Cooley, R., Mobaser, B., Srivastava, J., „Data preparation for mining World Wide Web browsing patterns”, Journal of Knowledge and Information Systems, 1, 1999.

[CMS97] Cooley R., Mobasher B., Srivastava J., “Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns”, Proc. of the 1997 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX), Newport Beach, California, November 1997.

[H75] Hartigan J., Clustering Algorithms, John Wiley, 1975.

[HKM97] Han, E-H, Karypis, G., Kumar, V., Mobasher, B., “Clustering based on association rule hypergraphs”, Proc. of SIGMOD’97 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD’97), May, 1997.

[L95] Luotonen A., “The common log file format”, http://www.w3.org/pub/WWW/, 1995.

[PPR96] Pirolli P., Pitkow J., Rao R., “Silk From a Sow’s Ear: Extracting Usable Structure from the World Wide Web”, Conference on Human Factors in Computing Systems (CHI 96), Vancouver, British Columbia, Canada, 1996.

[P97] Pitkow J., “In search of reliable usage data on the www”, Sixth Int’l World Wide Web Conference, Santa Clara, California, 1997.

[YJG+96] Yan T.W., Jacobsen M., Garcia-Molina H., Dayal U., “From User Access Patterns to Dynamic Hypertext Linking”, Proc. of the 5th Int’l World Wide Web Conference, 1996.

[PE97] Perkowitz, M., Etzioni, O., “Adaptive Web Sites: an AI challenge”, Proc. 15th Int. Joint Conf. AI, 1997.