Automatyczna personalizacja serwerów WWW z wykorzystaniem metod eksploracji danych

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

Abstrakt

Niniejszy artykuł poświęcony jest zagadnieniom automatycznej personalizacji serwisów WWW w oparciu o tzw. adaptatywne serwery WWW. Personalizacja serwisów WWW polega na wykorzystywaniu znanych profili preferencji do dynamicznego dostosowywania zawartości serwisu do potrzeb poszczególnych użytkowników. Adaptatywne serwery WWW automatycznie odkrywają typowe schematy zachowań użytkowników analizując informacje o użytkowaniu serwisu zawarte w logu serwera technikami eksploracji danych. W artykule przedstawiono ogólną ideę adaptatywnych serwerów WWW oraz szczegółowe propozycje ich implementacji. Artykuł poświęca również dużo miejsca technikom gromadzenia wiarygodnych informacji o użytkowaniu serwisów WWW oraz ich wstępnego przetwarzania do formatu odpowiedniego dla technik eksploracji danych.

1. Wprowadzenie

Personalizacja serwisów WWW (ang. *Web personalization*) polega na wykorzystywaniu odkrytych profili preferencji do dynamicznego dostosowywania zawartości serwisu do potrzeb poszczególnych użytkowników [MCS99]. W chwili obecnej, personalizacja nabiera coraz większego znaczenia, ze względu na stale rosnącą liczbę konkurujących ze sobą serwisów WWW. Podstawową bronią w walce o klienta, zarówno w kontekście serwisów informacyjnych jak i sklepów internetowych, jest dostosowanie przedstawianej zawartości do jego potrzeb, oczekiwań i zainteresowań. Ze względu na oczywisty fakt, że różni użytkownicy mogą mieć różne oczekiwania i upodobania, serwisy przedstawiające tę samą zawartość wszystkim odwiedzającym, mogą tracić popularność na rzecz serwisów posiadających zdolność adaptacji do różnych i zmiennych w czasie preferencji użytkowników.

Narzucającym się rozwiązaniem, mogącym zapewnić użytkownikom dostosowany do ich potrzeb obraz danej witryny, jest umożliwienie im podjęcia decyzji jakie informacje i w jakiej formie mają im być przedstawiane. Rozwiązanie to jest stosowane w praktyce przez wiele serwisów i portali, np. *Yahoo!*. Jest ono określane raczej jako „możliwość dostosowywania" (ang. *customization*), a nie personalizacja, gdyż nie ma charakteru automatycznego i prawie całkowicie polega na użytkownikach. W przypadku wspomnianego serwisu *Yahoo!*, użytkownicy mają możliwość złożenia z setek dostępnych modułów swojej strony *My Yahoo!* [MPR00]. Dostępne moduły prezentują aktualne informacje o pogodzie, sporcie, kursach akcji, itp. Niestety praktyka pokazuje, że przeważająca większość użytkowników nie korzysta w ogóle z możliwości dostosowywania zawartości serwisu własnymi siłami. Z pewnością dzieje się tak nie tylko dlatego, że domyślne strony są tak skonstruowane aby odpowiadały preferencjom jak największej grupy użytkowników. Bardzo prawdopodobne jest, że dla większości użytkowników dostosowywanie zawartości odwiedzanych witryn jest zbyt czasochłonne, a nawet po prostu zbyt skomplikowane.

Celem automatycznej personalizacji serwisów WWW jest przerzucenie odpowiedzialności za dostosowywanie prezentowanej zawartości z użytkownika na serwer. W tym wypadku serwer stara się dopasować charakterystykę użytkownika do jednego ze znanych profili i następnie prezentuje użytkownikowi zawartość dostosowaną do dopasowanego profilu preferencji. Kluczowym elementem jest pozyskanie charakterystyk użytkownika takich jak wiek, płeć, zawód, adres zamieszkania, zainteresowania, itd. Najprostszym sposobem uzyskania takich informacji jest wymaganie od użytkownika jawnego ich podania poprzez wypełnienie formularza rejestracyjnego czy ankiety. Serwer po dopasowaniu użytkownika do jednego z profili preferencji dostosowuje zawartość przedstawianą użytkownikowi zgodnie z regułami związanymi z danym profilem. Reguły te mogą być ustalone a priori przez ekspertów (być może posługujących się narzędziami do analizy danych) lub modyfikowane dynamicznie poprzez obserwację zachowań użytkowników przypisanych do danego profilu. Drugie z wymienionych rozwiązań jest bardziej elastyczne – może uwzględnić np. rozbudowę witryny. Ponadto, jest rozwiązaniem bardziej praktycznym gdy zbiór produktów, dokumentów oferowanych użytkownikom w ramach danego serwisu jest bardzo liczny [G00]. Rozwiązanie to polega na proponowaniu użytkownikom tego, co zainteresowało innych użytkowników o podobnych charakterystykach (ang. *collaborative filtering*).

Pozyskiwanie informacji o preferencjach użytkowników w oparciu o formularze rejestracyjne i ankiety ma jednak kilka istotnych wad. Po pierwsze, ta forma pozyskiwania wiedzy cechuje się dużą subiektywnością i spotyka się z niechęcią użytkowników, „zmuszanych" do wypełniania dodatkowych formularzy. Ponadto, tak budowane profile użytkowników posiadają charakter statyczny i z upływem czasu ulegają degradacji. Rozwiązania bazujące na jednym statycznym profilu preferencji użytkownika nie biorą również pod uwagę faktu, że ten sam użytkownik może w różnych momentach czasowych poszukiwać różnych 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.

W związku z powyższym, 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* (ang. *adaptive web sites*), opisujący serwery WWW, które automatycznie ulepszają swoją zawartość i organizację na podstawie obserwacji ścieżek dostępów użytkowników. Idea wykorzystania koncepcji adaptatywnych serwerów do personalizacji serwisu polega na analizie plików logu serwera, wyławianiu z nich statystycznych korelacji pomiędzy pobieranymi dokumentami lub pracującymi użytkownikami, a następnie wykorzystywaniu znalezionych korelacji do modyfikacji struktury dokumentów WWW, wysyłanych użytkownikom. Do analizy logu, adaptatywne serwery WWW wykorzystują techniki eksploracji danych (ang. *data mining*). Eksploracja danych polega na automatycznym odkrywaniu nietrywialnych, interesujących, wcześniej nieznanych wzorców i zależności w dużych wolumenach danych. Eksploracja logów serwerów WWW (ang. *Web usage mining* [CMS97b]) jest jednym z podstawowych obszarów zastosowań technik eksploracji danych, posiadającym swoją specyfikę. W przypadku eksploracji danych o zachowaniach użytkowników WWW, szczególne znaczenie mają techniki gromadzenia danych źródłowych gwarantujące ich wiarygodność

oraz algorytmy wstępnej transformacji i „czyszczenia” zgromadzonych danych źródłowych.

Niniejszy artykuł opisuje stan nauki i technologii w zakresie personalizacji serwisów WWW w oparciu o adaptatywne serwery WWW, wykorzystujące techniki eksploracji danych. Artykuł przedstawia ogólną ideę adaptatywnych serwerów WWW, prezentując różne propozycje ich implementacji, które pojawiły się dotychczas w literaturze.

2. Gromadzenie i wstępne przetwarzanie informacji źródłowych dla eksploracji danych

Podstawowym źródłem danych dla wszelkich analiz zachowania użytkowników serwisu WWW jest plik logu (dziennik) serwera WWW. Najczęściej jest on jedynym źródłem informacji, choć serwisy działające w oparciu o serwery aplikacji (ang. *application servers*) mogą gromadzić również informacje o działaniach użytkowników specyficzne dla danego serwisu na poziomie aplikacji. Podstawowe problemy występujące w przypadku zaawansowanej analizy logu serwera WWW (np. na potrzeby automatycznej adaptacji serwera), to zapewnienie jak największej wiarygodności gromadzonych danych oraz transformacja danych zawartych w logu do formatu odpowiedniego dla konkretnej techniki eksploracji danych.

2.1 Struktura pliku logu

Jak już wspomniano wcześniej, informacje o dostępach do serwera WWW zapisywane są w pliku 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]. 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* jest wypełnione gdy serwer przeprowadza autoryzację użytkowników przy dostępie do danego zasobu i zawiera w takim przypadku nazwę użytkownika podaną przy autoryzacji. 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 1.

```
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. 1. Przykładowy plik logu serwera WWW.

2.2 Czyszczenie plików logu

Z punktu widzenia eksploracji danych, mającej na celu odkrycie typowych schematów zachowań użytkowników WWW, dane źródłowe powinny mieć postać zbiorów lub sekwencji stron WWW (ang. *Web page*) pobieranych z serwera przez poszczególnych użytkowników. Zapisy w logu dotyczą jednak pojedynczych plików, a nie stron traktowanych jako obiekty złożone. W przypadku dostępu do strony zawierającej np. obrazy, dźwięki lub filmy, w logu znajdzie się zapis dotyczący głównego dokumentu HTML (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 stronom WWW.

Należy tu zwrócić uwagę, że filtracja logu może nie być zadaniem trywialnym, gdy poszczególne elementy stron WWW są generowane na serwerze dynamicznie przez programy CGI lub serwlety. Na szczęście najczęściej dynamicznie generowane są dokumenty HTML, a zagnieżdżone w nich obiekty multimedialne pobierane są z systemu plików. Wtedy można założyć, że odwołania do programów czy serwletów są odwołaniami do stron WWW, tak jak odwołania do statycznych dokumentów HTML. Tak samo traktowane są dokumenty tworzone w oparciu o różnego rodzaju technologie dynamicznej generacji stron, takie jak np. ASP, PHP, JSP, posiadające swoje specyficzne rozszerzenia plików, oraz dokumenty XML.

2.3 Identyfikacja dostępów poszczególnych użytkowników

Z punktu widzenia analizy zachowań użytkowników 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. W oparciu o te informacje plik logu jest transformowany do postaci zbiorów lub sekwencji dostępów poszczególnych użytkowników do stron WWW. Operacja ta polega na wyodrębnieniu z logu informacji o dostępach poszczególnych użytkowników. Jest to realizowane 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. 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].

Typowy profil zachowania użytkowników może być rozumiany jako grupa stron lub grupa ścieżek nawigacyjnych, często powtarzająca się w poszczególnych sesjach. W przypadku śledzenia ścieżek nawigacji istotne są informacje o wszystkich stronach, do których odwoływał się dany użytkownik z uwzględnieniem kolejności odwołań. Alternatywą jest ograniczenie się tylko do tych stron, których treść zainteresowała użytkownika (strony służące jedynie jako ścieżka dostępu do szukanego dokumentu nie są uwzględniane). W [CMS97a] 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.

2.4 Wiarygodność informacji zawartych w logu serwera WWW

Dokonując jakichkolwiek analiz plików logu serwera WWW należy zdawać sobie sprawę z niedoskonałości mechanizmu odnotowywania przez serwer odwołań użytkowników do dokumentów. 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.

3. Automatyczna adaptacja serwera WWW w oparciu o wyniki eksploracji danych

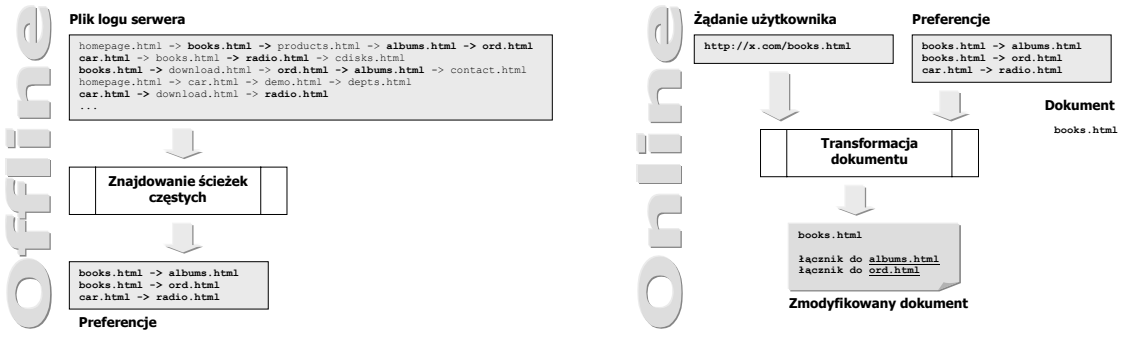
Proces adaptacji serwera WWW przebiega w dwóch fazach:

1. Offline: wykorzystanie pliku logu serwera do odkrycia typowych profili zachowań użytkowników reprezentowanych przez zbiory lub sekwencje stron WWW. Faza ta realizowana jest asynchronicznie względem połączeń użytkowników, np. w odstępach tygodniowych lub miesięcznych. W fazie tej stosowane są różne techniki eksploracji danych, po uprzedniej transformacji i oczyszczeniu pliku logu. Podstawowe znaczenie ma grupowanie (ang. *clustering*) [H75], polegające na podziale zbioru obiektów na grupy w taki sposób, aby obiekty wewnątrz każdej z grup były maksymalnie podobne do siebie, a jednocześnie możliwie jak najbardziej różniące się od obiektów przydzielonych do innych grup. W kontekście adaptatywnych serwerów WWW, grupowanymi obiektami są sekwencje lub zbiory stron reprezentujące poszczególne sesje

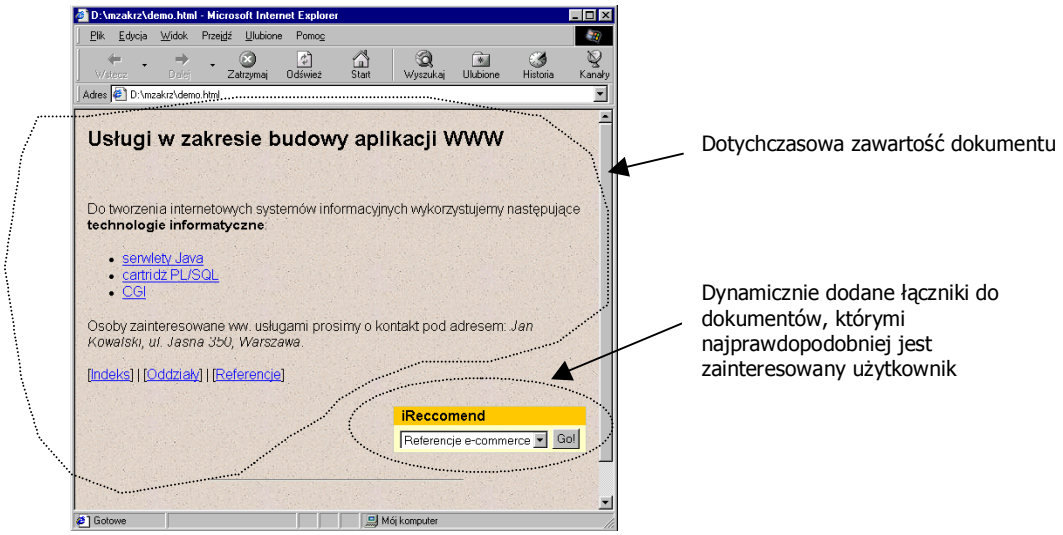
użytkowników. Do grupowania wybierane są takie algorytmy, które oprócz podziału na grupy dostarczają opis poszczególnych grup, w tym wypadku w postaci zbioru stron WWW lub ścieżek nawigacyjnych typowych dla danej grupy. Proces grupowania może być poprzedzony fazą odkrywania asocjacji [AIS93] lub wzorców sekwencyjnych [AS95], jeśli stosowany algorytm grupowania tego wymaga.

2. Online: wykorzystywanie znalezionych grup stron lub 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. 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 odkrytych profili zachowań i wybierane są te profile, które wykazują się największym dopasowaniem. Zbiór łączników do dokumentów opisujących dopasowane profile staje się dodatkowym elementem wizualnym, który dynamicznie jest dołączany do żądanego dokumentu [YJG+96].

Przedstawmy przykład prostej adaptacji serwera WWW, zilustrowany na rysunku 2. 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 3).



Rys. 2. Przykładowy proces adaptacji serwera WWW



Rys. 3. Przykład dokumentu WWW wzbogaconego o dynamicznie wygenerowane rekomendacje

W ramach ogólnej koncepcji adaptatywnych serwerów WWW zaproponowano w literaturze kilka sposobów ich implementacji, różniących się stosowanymi algorytmami grupowania, formą reprezentacji odkrytych profili zachowań użytkowników oraz sposobem wykorzystania odkrytych schematów do dynamicznej modyfikacji dokumentów.

W [YJG+96] zaproponowano podejście bazujące na wektorowej reprezentacji sesji użytkowników. Wymiary przestrzeni odpowiadają w tym przypadku stronom WWW serwisu, wartość konkretnej współrzędnej wektora zależy od czasu spędzonego przez użytkownika na stronie odpowiadającej tej współrzędnej (lub liczby odwiedzeń strony w ramach sesji). Sesje użytkowników są grupowane algorytmem lidera (ang. *leader algorithm*) [H75]. Średnie wektory dla poszczególnych grup sesji reprezentują typowe schematy zachowań. W trakcie normalnej eksploatacji serwisu, dla każdego użytkownika tworzony i na bieżąco uaktualniany jest wektor odpowiadający jego dotychczasowemu zachowaniu. Wektor ten jest dopasowywany do odkrytych schematów. Dokument wysyłany użytkownikowi jest uzupełniany o łączniki do stron, których użytkownik jeszcze nie odwiedził, a które występowały w schematach zachowań najlepiej pasujących do dotychczasowego zachowania użytkownika.

Rozwiązanie przedstawione w [MCS99] tworzy profile preferencji w oparciu o współwystępowanie dokumentów w historycznych sesjach użytkowników odnotowanych w logu. Najpierw odkrywane są często występujące asocjacje stron, następnie na bazie znalezionych asocjacji generowane są potencjalnie nakładające się na siebie grupy stron stanowiące profile zachowań. Grupowanie jest realizowane za pomocą algorytmu partycjonowania hipergrafu [HKM97], w którym wierzchołki odpowiadają stronom serwisu, a krawędzie odkrytym asocjacjom. Moduł online systemu automatycznej personalizacji działa na tej samej zasadzie co w implementacji opisanej w [YJG+96], z uwzględnieniem dwóch drobnych usprawnień. Po pierwsze, w trakcie dopasowywania bieżącego zachowania użytkownika do odkrytych wcześniej profili, uwzględniane są dostępy realizowane w obrębie pewnego ustalonego okna czasowego, tak aby na generowane dynamicznie łączniki z podpowiedziami wpływ miał tylko ostatni fragment

historii sesji. Po drugie, dołączane do wysyłanych użytkownikom dokumentów dynamicznie generowane łączniki są posortowane wg ich wagi (im dalej proponowany dokument znajduje się od bieżącego w topologii serwisu, tym większa jest mu przypisywana waga).

Propozycja opisana w [MWZ00] może być postrzegana jako rozszerzenie implementacji opisanej powyżej o uwzględnienie kolejności odwołań realizowanych w ramach poszczególnych sesji. Preferencje użytkowników są w tym wypadku reprezentowane przez zbiory podobnych najczęściej stosowanych ścieżek nawigacyjnych. W celu znalezienia preferencji, realizowany jest dwufazowy algorytm. W pierwszej fazie przeszukiwany jest log serwera WWW w celu znalezienia wszystkich najczęściej występujących ścieżek nawigacyjnych, mających postać wzorców sekwencyjnych. W fazie drugiej, znalezione ścieżki nawigacyjne są grupowane algorytmem o nazwie POPC, kierującym się współwystępowaniem ścieżek w historiach dostępów użytkowników (tzn. podobieństwo dwóch ścieżek wynika z tego, iż wielu użytkowników, którzy podążają jedną z nich, podąża również drugą).

4. Podsumowanie

W artykule przedstawiono aktualne podejścia do personalizacji serwisów WWW, ze szczególnym naciskiem na automatyczną personalizację w oparciu o adaptatywne serwery WWW. Opisana została ogólna koncepcja adaptatywnych serwerów WWW, ze wskazaniem obszarów, w których zastosowanie znajdują techniki eksploracji danych. Szczególny nacisk położony został na problematykę obróbki pliku logu serwera, stanowiącego podstawowe źródło danych do eksploracji.

Literatura

[AIS93] Agrawal R., Imielinski T., Swami A., “Mining association rules between sets of items in large databases”. Proc. of the ACM SIGMOD Conference on Management of Data, 1993.

[AS95] Agrawal R., Srikant R.: “Mining Sequential Patterns” Proc. of the 11th International Conference on Data Engineering, 1995.

[CMS97a] 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), 1997.

[CMS97b] Cooley R., Mobasher B., Srivastava J., “Web Mining: Information and Pattern Discovery on the World Wide Web”, Proc. of ICTAI’97, 1997.

[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.

[G00] Greening D.R., ”Data Mining on the Web”, Web Techniques, 2000.

[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, 1995.

[MCS99] Mobaser, B., Cooley, R., Srivastava, J., ”Creating Adaptive Web Sites Through Usage-Based Clustering of URLs”, Proc. of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX), 1999.

[MPR00] Manber U., Patel A., Robison J., “Experience with Personalization on Yahoo!”, Communications of the ACM, Vol. 43, No. 8, 2000.

[MWZ00] Morzy T., Wojciechowski M., Zakrzewicz M.: “Web Users Clustering”, Proc. of the ISCIS 2000 Conference, 2000.

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

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

[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), 1996.

[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.