ZESZYTY NAUKOWE UNIWERSYTETU SZCZECIŃSKIEGO

NR 508 STUDIA INFORMATICA NR 22 2008

TOMASZ ZDZIEBKO

SKRYPTOWY SYSTEM MONITOROWANIA ZACHOWAŃ UŻYTKOWNIKÓW SERWISÓW WWW

Wprowadzenie

Internet rozwija się w zdumiewającym tempie w wielu aspektach. Ze wzrostem popularności, liczby użytkowników, serwisów internetowych, sklepów internetowych i biznesowych zastosowań potrzeba poznania i zrozumienia zachowań indywidualnych użytkowników zyskuje na znaczeniu. Jednym z czynników przemawiających za tą tendencją jest możliwość dostosowania przekazywanych treści do indywidualnych gustów i preferencji użytkowników – tak zwana personalizacja. Jej waga dla właścicieli serwisów WWW jest coraz wyższa. Zastosowanie systemów personalizacji może się przyczynić do zwiększenia tak satysfakcji klientów, jak i obrotów sklepów internetowych oraz do pozyskania długoterminowych wartościowych klientów.

Personalizacja definiowana jest jako proces prezentowania użytkownikowi prawidłowych informacji w odpowiednim momencie[1]. Aby było to możliwe, konieczne jest poznanie indywidualnych preferencji użytkowników. W literaturze zaproponowano wiele inteligentnych technik budowania profili użytkowników na podstawie informacji o stronach, które są interesujące dla danego użytkownika[2]. We wcześniejszych badaniach często wykorzystywano techniki jaw-

[1] M. Speretta, S. Gauch, *Personalizing Search Based on User Search Histories*, Thirteenth International Conference on Information and Knowledge Management (CIKM 2004), 2004.

[2] H. Kim, P.K. Chan, *Learning implicit user interest hierarchy for context in personalization*, International Conference on Intelligent User Interfaces, 2003; J. Goecks, J. Shavlik, *Learning users' interests by unobtrusively observing their normal behavior*, Proc. 5th International Conference on Intelligent user Interfaces, 2000; B. Mobasher, H. Dai, T. Luo, M. Nakagawa,

nego poznawania preferencji użytkowników poprzez ich odpytywanie. Korzystający z informacji byli proszeni o wypełnienie ankiet lub o odpowiedź na proste pytanie, na przykład „Czy podoba Ci się ten produkt?" z dwoma możliwymi odpowiedziami „tak/nie". Jawne metody poznawania preferencji wymagają jednak świadomej zgody użytkownika, a także uczynienia pewnego wysiłku. Stanowi to często barierę nie do pokonania dla wielu osób. Niejawne metody pozyskiwania informacji o użytkownikach serwisów WWW opierają się na obserwacji ich zachowań. Nie są one z reguły tak dokładne, jak jawne metody, ale nie wymagają wysiłku ze strony użytkownika[3]. Oczywiście w celu uszanowania prywatności użytkowników informacja o działaniach monitorujących powinna być zawarta w polityce prywatności serwisu.

Większość technik niejawnego[4] odkrywania preferencji internautów proponowanych w literaturze bazuje na danych uzyskanych na podstawie kliknięć w poszczególne linki. Dane te pochodzą z logów serwerów WWW lub są zebrane poprzez oprogramowanie logujące po stronie klienta. Mając te informacje, można oszacować jeden z głównych wskaźników zainteresowania, jakim jest czas spędzony na stronie[5]. Jednakże miara ta może być niedokładna, ze względu na fakt, iż w wielozadaniowych systemach operacyjnych użytkownicy mogą wykonywać inne czynności, podczas gdy przeglądarka zostaje otwarta[6]. Z tego powodu powinny być mierzone inne niejawne wskaźniki zainteresowania.

Effective Personalization Based on Association Rule Discovery from Web Usage Data, Web Information and Data Management, 2005.

[3] A. Watson, M.A. Sasse, *Measuring perceived quality of speech and video in multimedia conferencing applications*, Proc. ACM Multimedia Conference, 1998.

[4] W angielskim stosuje się termin *implicit*, co w tym kontekście oznacza niejawny lub domniemany.

[5] L.A. Granka, T. Joachims, G. Gay, *Eye-tracking analysis of user behavior in WWW search*, Proc. 27th annual international conference on Research and development in information retrieval, 2004; M. Claypool, P. Le, M. Wased, D. Brown, *Implicit interest indicators*, In Proc. 6th international conference on Intelligent User Interfaces, 2001.

[6] K. Jung, *Modeling web user interest with implicit indicators*, Master Thesis, Florida Institute of Technology, 2001.

1. Badania zachowań użytkowników serwisów internetowych w literaturze

Monitorowanie zachowań użytkowników serwisów WWW zostało ostatnio szczegółowo opisane przez Velaythana i Yamadę[7]. Opracowali oni system GINIS wykorzystujący zmodyfikowaną wersję przeglądarki Internet Explorer, zawierającą rozszerzenia pozwalające na drobiazgowe rejestrowanie akcji dokonywanych przez użytkowników. System pozwala na rejestrowanie ponad 70 różnych akcji nawigacyjnych, które były klasyfikowane jako 40 różnych zachowań. Został przetestowany na grupie dziesięciu ochotników, którzy średnio przez 22 dni korzystali z tej przeglądarki do codziennego przeglądania sieci. W trakcie przeglądania zasobów WWW ich zachowanie było rejestrowane, a przy każdorazowym opuszczeniu strony byli proszeni o odpowiedź, czy dana strona ich zainteresowała. Rozwiązanie to wymagało świadomej zgody użytkownika na ujawnienie swoich preferencji odnośnie do poszczególnych stron WWW oraz wysiłku z jego strony. Po przeprowadzeniu analizy zebranych danych autorzy badania odkryli, iż przewijanie strony, korzystanie z klawiatury, wprowadzanie danych do formularzy, korzystanie z linków i wyszukiwanie tekstu stanowią najczęstsze akcje wykonywane przez internautów. Badanie to ukazało również, że czas spędzony na stronie nie jest najważniejszym wskaźnikiem zainteresowania użytkowników. Na podstawie tych badań stwierdzono, że inne akcje, takie jak na przykład scrollowanie, korzystanie z formularzy, wyszukiwanie tekstu, kopiowanie tekstu, powinny być brane pod uwagę przy ocenianiu zainteresowania użytkowników. Velayathan i Yamada odkryli również, że poszczególnych użytkowników cechowały różne wzorce przeglądania sieci, co oznaczało w praktyce, że niektóre wskaźniki oceny zainteresowania nie sprawdzały się jednakowo w odniesieniu do wszystkich uczestników badania.

Jedno z wcześniejszych badań zostało zaprezentowane przez Weinreicha i innych w 2006 roku[8]. Badanie to polegało na obserwacji 25 użytkowników średnio przez 25 dni. Skupiono się na poznaniu sposobów korzystania z przeglądarki do nawigacji po różnych serwisach internetowych. Wyniki badania wskazują, iż użytkownicy bardzo często pobieżnie przeglądają strony internetowe w celu określenia, czy dana treść jest interesująca. Jeśli nie, podejmują

---

[7] G. Velayathan, S. Yamada, *Can We Find Common Rules of Browsing Behavior?*, 6th International World Wide Web Conference, 2007.

[8] H. Weinreich, H. Obendort, E. Herder, M. Mayer, *Exploring Three Aspects of Web Navigation*, WWW Conference2006, ACM Press, 2006.

akcję nawigacyjną i przemieszczają się na inną stronę. W badaniu zauważono również silną tendencję do równoległego przeglądania różnych stron w kliku oknach lub zakładkach przeglądarek.

Kolejny wkład do poznania zachowań użytkowników serwisów internetowych wniosło badanie przeprowadzone przez Kima i Chana, którzy stworzyli przeglądarką będącą w stanie rejestrować zachowania użytkowników stron internetowych[9]. Jedenastu różnych użytkowników zostało poproszonych o spędzenie dwóch godzin na surfowanie po sieci. Kim i Chan odkryli, że wykorzystując wskaźniki *Complete*, *Active*, *LookAtIt*, *MousMove*, byli w stanie określić zachowanie ośmiu uczestników badania. Zaobserwowali także, że wskaźnik *MousMove*, a następnie *MouseClicks* mogą być najbardziej praktyczne w przy badaniu zainteresowania internautów.

Wyniki jednego z najwcześniej przeprowadzonych badań na dużą skalę zostały zawarte w pracy Catledge'a i Pitkowa[10]. Badanie zostało przeprowadzone na grupie 107 użytkowników i trwało 21 dni. Uczestnicy korzystali z przeglądarki XMosaic, do której dodano kod rejestrujący zdarzenia. Dane pozyskane z tego badania pozwoliły na klasyfikację sposobów nawigacji użytkowników na trzy różnorodne grupy. Jednym z często występujących wzorców nawigacji wśród wszystkich użytkowników było przemieszczanie w głąb struktury serwisu ze strony startowej, po czym następował powrót na stronę startową, skąd następowało kolejne przemieszczanie w głąb serwisu tym razem wzdłuż innej ścieżki.

Web browsing behaviour monitor (WBBM)

W pracy proponuje się system niejawnego monitorowania zachowań użytkowników serwisów WWW – WBBM, który bazuje na skryptach w technologii *JavaScript*. Rozwiązanie to ma niewątpliwie jedną zaletę – nie wymaga instalacji dodatkowych aplikacji po stronie użytkownika. Dzięki temu nie jest wymagana zgoda użytkownika na instalowanie dodatkowych programów. Oczywiście takie rozwiązanie może budzić obawy o naruszenie zasad prywatności. Aby

[9] H. Kim, P. Chan, *Implicit Indicators for Interesting Web Pages*. Proc. Intl. Conf. on Web Information Systems and Technologies, 2005, pp. 270–277.

[10] L.D. Catledge, J.E. Pitkow, *Characterizing Browsing Strategies in the World-Wide Web*, Proceedings of the Third International World-Wide Web Conference on Technology, Tools and Applications, 1995.

uchronić się przed takimi zarzutami, autor zdecydował się na monitorowanie zdarzeń w sposób anonimowy. Żadne monitorowane zdarzenia nie były powiązane z konkretnym komputerem, a tym bardziej jego użytkownikiem.

System WBBM wykorzystuje język *JavaScript*, Obiektowy Model Dokumentu – *DOM* i mechanizm *AJAX* – Asynchroniczny *JavaScript* i *XML*. Wymagania te spełnione są przez większość obecnie używanych przeglądarek internetowych. System był testowany na platformie Windows w następujących przeglądarkach: Opera 8.5, 9.21, Mozilla FireFox 1.5, 2.0, Microsoft Internet Explorer 6.0, 7.0. System WBBM powinien pracować prawidłowo również na innych systemach operacyjnych i innych przeglądarkach. Instalacja system WBBM polega jedynie na dołączeniu do każdej monitorowanej strony linku do skryptu znajdującego się w osobnym pliku oraz na umieszczeniu na serwerze skryptów w języku *PHP*, które zbierają dane przesyłane z monitorowanych przeglądarek.

System WBBM monitoruje wybrane zdarzenia występujące po stronie przeglądarki. Razem z informacją o typie zdarzenia zapisywane są dodatkowe informacje, takie jak na przykład czas zdarzenia, źródłowy element HTML zdarzenia. W celu ograniczenia połączeń nawiązywanych przez monitorowaną przeglądarkę z serwerem zbierającym dane, wszystkie zdarzenia są paczkowane. Paczka jest wysyłana do serwera w momencie, gdy osiągnie określony rozmiar lub gdy wystąpi zdarzenie *Unload* oznaczające opuszczenie bieżącej strony przez użytkownika. Komunikacja pomiędzy monitorowaną przeglądarką a serwerem zbierającym dane odbywa się przy użyciu mechanizmu *AJAX*. W celu identyfikacji zdarzeń na konkretnej maszynie wykorzystywany jest mechanizm *Cookies*.

Niewątpliwa wada, jaką ma WBBM, wynika z natury języka *JavaScript*, który nie pozwala na monitorowanie wszystkich interakcji użytkownika z przeglądarką. Przykładem jest korzystanie z menu przeglądarki lub poszczególnych funkcji wybieranych z menu kontekstowego. Z tego powodu system nie jest w stanie monitorować wszystkich zachowań i dostarczyć w pełni wiarygodnych informacji. Jednakże rozwiązanie pozwala na monitorowanie sporej liczby aktywności użytkowników, warte jest więc rozwijania, gdyż brak konieczności instalowania jakiekolwiek oprogramowania po stronie klienta czyni je łatwym do praktycznego zastosowania, a to może, z kolei, ulepszyć dotychczas stosowane rozwiązania.

System WBBM pozwala na monitorowanie zdarzeń wygenerowanych przez akcje wykonywane przy użyciu myszy lub klawiatury. Pierwszy problem,

jaki napotkano przy konstruowaniu systemu, wynikał z braku pełnej kompatybilności pomiędzy implementacjami standardów *DOM* i *JavaScript* w różnych przeglądarkach. Aby uniknąć tworzenia różnych wersji systemu dla różnych przeglądarek, problem ten udało się rozwiązać, wykorzystując głównie mechanizm detekcji obiektów[11]. Kolejny problem, którego jeszcze nie udało się rozwiązać, stanowi prawidłowa detekcja klawiszy naciskanych przez użytkowników[12]. Niestety, ze względu na różnice między przeglądarkami, różne języki i układy klawiatury na różnych systemach, jedynie klawisze CTRL, SHIFT i ALT mogły być odczytane prawidłowo.

Po przeanalizowaniu uprzednich badań w zakresie śledzenia zachowań użytkowników serwisów internetowych wybrano 14 zdarzeń, które w zamierzeniu miał monitorować system WBBM.

2. Wyniki badania

System WBBM został zainstalowany w serwisie konferencja.org, który stanowi największy polski katalog konferencji. W skład serwisu wchodzi wyszukiwarka oraz prosty katalog tematyczny danej konferencji. Każda pozycja znajdująca się w katalogu zawiera krótki opis konferencji, wraz z podstawowymi informacjami na temat terminu zgłoszeń, opłaty itp. Badanie przeprowadzono w okresie od 1 czerwca do 20 sierpnia 2007 roku. Całkowita liczba zarejestrowanych zdarzeń wyniosła ponad 26 tys., a całkowita liczba unikatowych użytkowników (identyfikowanych na podstawie cookies) wyniosła 770. Dane te zostały poddane wstępnemu przetwarzaniu, które miało na celu wyeliminowanie ruchu wygenerowanego przez roboty i błędnych zdarzeń spowodowanych przypadkowym przyblokowaniem klawiszy. Po ich usunięciu odnotowano ponad 23 tys. zdarzeń. Udział procentowy najczęściej występujących zdarzeń przedstawiono w tabeli 1. Uzyskane wyniki w znacznym stopniu pokrywają się z badaniami przedstawionymi w pracy Velayathana i Yamady[13]. Najczęściej występujące zdarzenia to: przewijanie strony, korzystanie z klawiatury, wypełnianie formularzy i korzystanie z linków nawigacyjnych – stanowią one w su-

[11] A. Watson, M.A. Sasse, *Measuring perceived quality of speech and video in multimedia conferencing applications*, Proc. ACM Multimedia Conference, 1998.

[12] Opera: http://my.opera.com/hallvors/blog/show.dml/217592, 12.07.2006.

[13] G. Velayathan, S. Yamada, *Can We Find Common Rules of Browsing Behavior?*, 6th International World Wide Web Conference, 2007.

mie 96,8% odnotowanych czynności wykonanych przez użytkowników. Nie udało się, co było dużym zaskoczeniem, zarejestrować czynności związanych z drukowaniem, dodawaniem do ulubionych czy wyszukiwaniem tekstu. Wynika to prawdopodobnie z tego, że użytkownicy wywołują te akcje, korzystając z menu przeglądarki lub menu kontekstowego zamiast skrótów klawiaturowych. Potwierdzeniem dla tej hipotezy może być fakt dość częstego korzystania z menu kontekstowego, który zanotowano w 2% przypadków.

Tabela 1

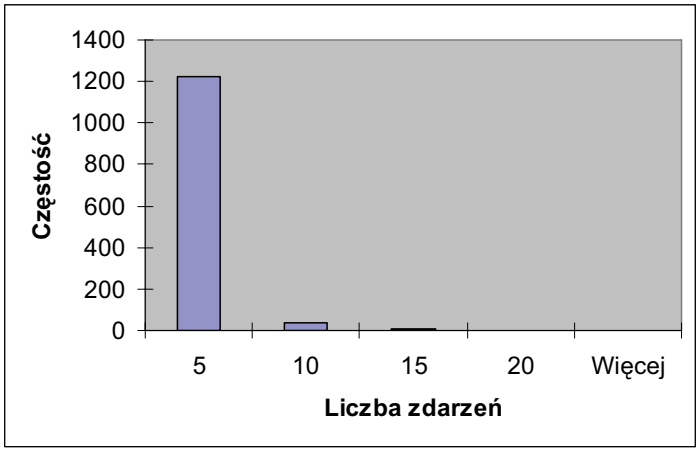
Zdarzenia i ich częstość występowania

Zachowanie	**Odsetek**
przewijanie strony (*scroll*)	35,9
korzystanie z klawiatury	32,8
wypełnianie formularzy	14,6
korzystanie z linków nawigacyjnych	13,5
korzystanie z menu kontekstowego	2,0
zmiana rozmiaru tekstu	0,6
kopiowanie	0,4
zapisywanie strony	0,1
zaznaczanie tekstu	0,1

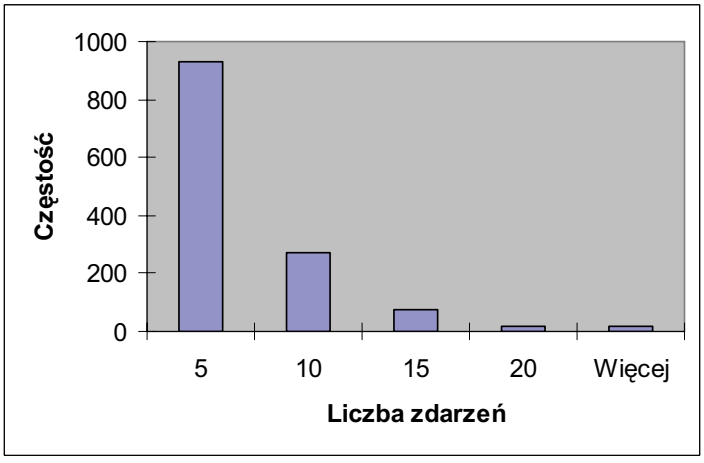
Źródło: opracowanie własne.

W dalszej części analizy postanowiono porównać liczbę wykonywanych akcji na stronie w zależności od czasu korzystania z niej. Dla celów analizy postanowiono podzielić wizyty na cztery grupy: wizyty znajdujące się w następujących przedziałach czasu (podane w sekundach): (0, 10>; (10, 60>; (60, 600>; (600, +∞). Liczbę zdarzeń w trakcie wizyty na stronie podzielono na kolejne cztery grupy. Dla wizyt o czasie krótszym lub równym 10 sekund dominująca liczba zdarzeń mieści się w przedziale (0, 5> (rysunek 1). Ze wzrostem czasu wizyty rośnie liczba zarejestrowanych zdarzeń, aczkolwiek nadal nie jest ich dużo (rysunki 1–4). Dla wizyt o czasie powyżej 10 minut nie następuje jak wcześniej wzrost częstości wizyt o liczbie zdarzeń powyżej 5, jak to miało miejsce w wypadku zdarzeń o krótszym czasie trwania. W ponad połowie wizyt o czasie trwania dłuższym niż 10 minut odnotowano 5 lub mniej zdarzeń. Jest to zdecydowanie zbyt niska wartość ze względu na czas trwania wizyty. Jedną z przyczyn może być wykonywanie innych zadań, podczas gdy okno przeglą-

darki pozostaje otwarte. Rezultat ten potwierdza tezę sformułowaną w pracy Velayathana, Yamady[14], iż czas spędzony na stronie nie jest najlepszym wskaźnikiem zainteresowania nią.

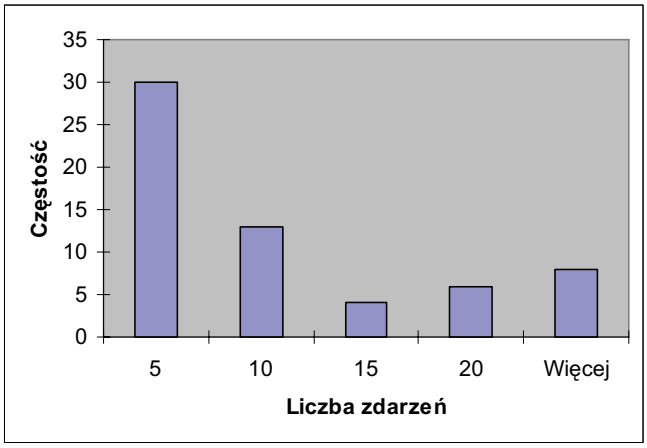


Rys. 1. Histogram częstości liczby zdarzeń dla wizyt o czasie <= 10 s
Źródło: opracowanie własne.

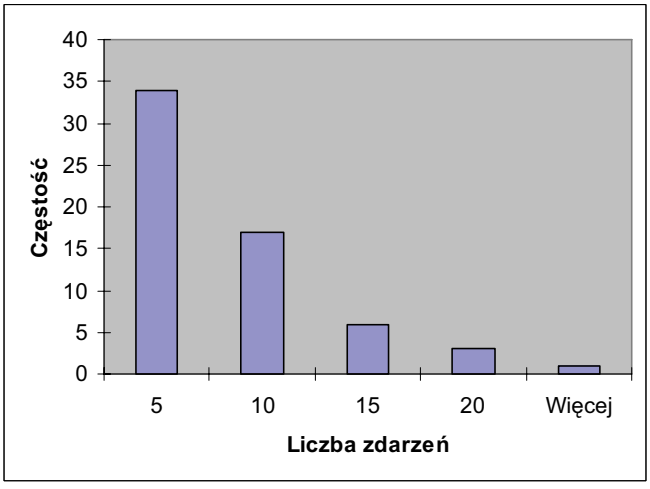


Rys. 2. Histogram częstości liczby zdarzeń dla wizyt o czasie (10; 60 s>
Źródło: opracowanie własne.

[14] Tamże.



Rys. 3. Histogram częstości liczby zdarzeń dla wizyt o czasie (60; 600 s>
Źródło: opracowanie własne.



Rys. 4. Histogram częstości liczby zdarzeń dla wizyt o czasie dłuższym niż 10 min
Źródło: opracowanie własne.

Wnioski

W pracy przedstawiono system WBBM służący do niejawnego monitorowania zachowań użytkowników serwisów WWW. Jedną z zalet tego rozwiązania jest brak konieczności instalacji dodatkowego oprogramowania na komputerze użytkownika oraz brak konieczności podejmowania dodatkowych działań przez użytkowników, w celu ujawnienia swoich preferencji. Najpoważniejszą wadą przedstawionego rozwiązania jest brak możliwości pełnego monitorowania wszystkich zdarzeń wykonywanych w przeglądarce.

Uzyskane wyniki są zgodne z uprzednio prowadzonymi badaniami. Potwierdzają również ważny tezę, iż czas spędzony na stronie nie jest najważniejszym wyznacznikiem zainteresowania nią.

W przyszłości autor planuje rozszerzyć system monitorowania o dodatkowe wskaźniki, takie jak na przykład dystans, o jaki użytkownik przemieścił kursor na stronie, dystans, o jaki została przewinięta zawartość strony, oraz porzucone linki (to znaczy takie, które były rozważane albo zostały pominięte). Planuje również zaadaptować system WBBM do budowania profili na potrzeby systemów personalizacyjnych.

Literatura

Catledge L.D., Pitkow J.E., *Characterizing Browsing Strategies in the World-Wide Web*, Proceedings of the Third International World-Wide Web conference on Technology, tools and applications, 1995.

Claypool M., Le P., Wased M., and Brown, D., *Implicit interest indicators*, In Proc. 6th International Conference on Intelligent User Interfaces, 2001.

Goecks J., Shavlik J., *Learning users' interests by unobtrusively observing their normal behavior.*, In Proc. 5th International Conference on Intelligent User Interfaces, 2000.

Granka L.A., Joachims T., Gay, G., *Eye-tracking analysis of user behavior in WWW search.*, In Proc. 27th Annual International Conference on Research and Development in Information Retrieval, 2004.

Kim H., and Chan P.P., *Implicit Indicators for Interesting Web Pages*. Proc. Intl. Conf. on Web Information Systems and Technologies, pp. 270–277, 2005, pp. 270–277.

Jung K.: *Modeling web user interest with implicit indicators*, Master Thesis, Florida Institute of Technology, 2001.

Kim H., Chan P.K., *Learning implicit user interest hierarchy for context in personalization*., In International Conference on Intelligent User Interfaces, 2003.

Mobasher B., Dai H., Luo T., Nakagawa M., *Effective Personalization Based on Association Rule Discovery from Web Usage Data*, Web Information and Data Management, 2005.

Opera, http://my.opera.com/hallvors/blog/show.dml/217592, 12.07.2006 .

Peter-Paul Koch site, http://www.quirksmode.org/js/support.html, 07.2007.

Speretta M., Gauch S., *Personalizing Search Based on User Search Histories*, Thirteenth International Conference on Information and Knowledge Management (CIKM 2004), 2004.

Velayathan G., Yamada S., *Can We Find Common Rules of Browsing Behavior?*, 6th International World Wide Web Conference, 2007.

Watson A., Sasse M.A., *Measuring perceived quality of speech and video in multimedia conferencing applications*, In Proc. ACM Multimedia Conference, 1998.

Weinreich H., Obendort H., Herder E., Mayer M., *Exploring Three Aspects of Web Navigation*, WWW Conference 2006, ACM Press, 2006.

SCRIPT SYSTEM OF BEHAVIOURS MONITORING OF WWW-SERVICES USERS

Summary

A user's interest in web pages can be estimated explicitly by querying his preferences or implicitly by observing his behaviors. Implicit methods are by nature less accurate than explicit methods, but they don't require any effort from user. Most of the previous studies used click stream data to infer user's interest. Starting with analysis of previous studies in web browsing behaviours, this study presents WBBM system for evaluating such behaviours. For purpose of these study author developed script-based system for monitoring web users activities. These system unobtrusively monitor low level activities at client-side and sends them to server. Results of experimental evaluation WBBM system has been also presented.

Translated by Tomasz Zdziebko