ODKRYWANIE WZORCÓW ZACHOWAŃ UŻYTKOWNIKÓW WWW

Marek Wojciechowski

Instytut Informatyki
Politechnika Poznańska
ul. Piotrowo 3a, 60-965 Poznań
Marek.Wojciechowski@cs.put.poznan.pl

Streszczenie

Niniejszy artykuł poświęcony jest odkrywaniu wzorców zachowań użytkowników WWW poprzez zastosowanie technik eksploracji danych (*ang. data mining*) do analizy logu serwera WWW. W artykule przedstawiono podstawowe techniki eksploracji danych wraz z przykładami ich zastosowania w stosunku do logu serwera WWW. Szczególny nacisk położony został na praktyczne zastosowania odkrytej wiedzy oraz problemy specyficzne dla analizy logu serwera WWW, które nie występują w przypadku innych źródeł danych.

1. Wprowadzenie

Eksploracja danych jest stosunkowo nową, dynamicznie rozwijającą się, dziedziną informatyki obejmującą problematykę zautomatyzowanego analizowania bardzo dużych wolumenów danych w celu znalezienia nietrywialnych, wcześniej nieznanych, potencjalnie interesujących zależności. Techniki eksploracji danych znajdują zastosowanie wszędzie tam, gdzie dostępne są olbrzymie ilości danych. Jako przykłady zastosowań można wymienić analizę koszyka zakupów na podstawie danych o transakcjach w supermarkecie, rozpoznawanie trendów na rynkach finansowych, analizę medycznych baz danych, klasyfikację klientów firm ubezpieczeniowych, itd. Analiza informacji zawartych w logu serwera WWW jest tylko jednym z obszarów zastosowań eksploracji danych. Posiada ona jednak swoją specyfikę i wymaga rozwiązywania problemów, które nie występują w przypadku innych dziedzin zastosowań. Z tego względu dziedzina ta stała się ostatnio przedmiotem wielu prac badawczych. Określana jest często jako *Web mining* lub jako *Web usage mining* [6], gdyż termin *Web mining* używany jest również w stosunku do technik inteligentnego wyszukiwania informacji w Internecie traktowanym jako olbrzymia baza danych.

Specyfika eksploracji danych zawartych w logu serwera WWW wynika głównie z charakteru tych danych i sposobu ich gromadzenia. Każdy zapis w logu zawiera między innymi adres IP komputera, z którego nastąpiło odwołanie, identyfikator użytkownika, nazwę pliku, do którego nastąpiło odwołanie i etykietę czasową. Techniki eksploracji danych wykrywają często powtarzające się wzorce w ramach transakcji lub sekwencji transakcji użytkowników. W przypadku żądań kierowanych do serwera WWW pojęcie transakcji ma charakter rozmyty. Ponadto, informacje w logu mogą być niepełne lub zafałszowane

w wyniku działalności serwerów proxy i wykorzystywania przez przeglądarki pamięci podręcznej (*ang. cache*).

W analizie logu znajdują zastosowanie takie techniki eksploracji danych jak odkrywanie częstych ścieżek nawigacji oraz techniki ogólnego przeznaczenia, czyli: odkrywanie reguł asocjacyjnych i wzorców sekwencji oraz klasyfikacja i grupowanie. Jednak aby możliwe było zastosowanie algorytmów eksploracji danych, konieczne jest wstępne przetworzenie informacji zawartych w logu serwera WWW poprzez odfiltrowanie nieistotnych wpisów oraz identyfikację transakcji poszczególnych użytkowników. Ponadto, należy zapewnić aby zapisy w logu w jak największym stopniu odzwierciedlały faktyczne odwołania użytkowników do dokumentów znajdujących się na serwerze, a także stosować mechanizmy pozwalające na stwierdzenie, które odwołania są odwołaniami jednego użytkownika, co nie zawsze jest zadaniem trywialnym.

2. Identyfikacja użytkowników i ich odwołań do serwera WWW

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 logi zgodne z formatem znanym pod nazwą *Common Logfile Format* [8]. Nie jest to jednak w pełni obowiązujący standard, gdyż na przykład niektóre serwery pamiętają również pewne dodatkowe informacje. Ponadto, format w jakim informacje są zapisane może być różny. *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*.

Z punktu widzenia eksploracji danych 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. Eksploracja danych polega na znajdowaniu często powtarzających się wzorców w sekwencjach dostępów użytkowników do serwera WWW lub grupowaniu użytkowników wykazujących podobne zachowanie. Z tego powodu obowiązkowym 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.

Informacje zawarte w logu mogą być nie tylko niepełne, ale również zafałszowane ze względu na wykorzystywanie serwerów proxy i podręcznej pamięci przeglądarek (*ang. cache*) [12]. 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 [11] 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 [3], 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 i na jej bazie dokonywać analiz.

3. Filtracja danych wejściowych i identyfikacja transakcji

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 log serwera WWW procesowi filtracji, w wyniku którego ignorowane są zapisy dotyczące plików nie będących głównymi dokumentami odpowiadającymi tzw. stronom WWW (*ang. Web page*).

Doprowadzenie danych źródłowych do postaci sekwencji dostępów poszczególnych użytkowników do stron w zasadzie może zakończyć proces wstępnej obróbki logu. Niektóre z metod eksploracji danych (odkrywanie reguł asocjacyjnych i wzorców sekwencji) posługują się jednak pojęciem transakcji i wymagają danych źródłowych w postaci zbioru transakcji lub zbioru sekwencji transakcji. Wspomniane metody zostały stworzone do odkrywania często powtarzających się wzorców zachowań klientów supermarketów czy firm zajmujących się sprzedażą wysyłkową, gdzie transakcja obejmuje po prostu zbiór produktów zakupionych podczas jednej wizyty lub w ramach jednego zamówienia. W przypadku operacji dostępu do serwera WWW pojęcie transakcji ma charakter rozmyty. Dwa skrajne podejścia polegają na traktowaniu każdego dostępu do strony jako oddzielnej transakcji albo wszystkich odwołań tego samego użytkownika jako jednej transakcji. Można również wyodrębniać transakcje odpowiadające sesjom 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.

W [5] zaproponowano metody identyfikacji transakcji oparte na podziale 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.

W oparciu o wspomniany powyżej podział odwołań do stron można wyodrębniać transakcje obejmujące tylko dostępy do stron zorientowane na zawartość lub też dowolne. Do wykrywania zależności między dostępami do różnych stron i grupowania użytkowników wykazujących podobne zainteresowania lepiej nadają się transakcje obejmujące tylko odwołania ze względu na zawartość. Transakcje zawierające dostępy obu typów, posiadające strukturę listy odwołań nawigacyjnych zakończonej odwołaniem do strony zorientowanym na zawartość, znajdują z kolei zastosowanie przy odkrywaniu częstych ścieżek nawigacji w obrębie witryn internetowych.

4. Eksploracja danych zawartych w logu serwera WWW

Techniki eksploracji danych stosowane w analizie logu serwera WWW obejmują odkrywanie częstych ścieżek nawigacji, odkrywanie reguł asocjacyjnych i wzorców sekwencji, klasyfikację i grupowanie. Pierwsza z wymienionych technik została zaproponowana w [4] specjalnie z myślą o analizie nawigacji użytkowników w środowiskach, gdzie informacja dostarczana jest w postaci wielu dokumentów powiązanych ze sobą siecią wzajemnych odwołań. Pozostałe są metodami ogólnego przeznaczenia, które zostały zaadaptowane do odkrywania wiedzy ukrytej w logu serwera WWW.

4. 1 Odkrywanie częstych ścieżek nawigacji

Punktem wyjścia do odkrywania częstych ścieżek nawigacji (*ang. path traversal patterns*) jest wyodrębnienie z logu transakcji mających postać tzw. maksymalnych odwołań w przód (*ang. maximal forward reference*). Transakcje tego typu są sekwencjami dostępów do stron realizowanymi jako odwołania do wcześniej nie odwiedzonych dokumentów. Każda sekwencja kończy się stroną, z której nastąpił powrót do poprzedniego dokumentu lub która kończy sesję użytkownika. Przyjmuje się, że tylko ostatnia ze stron tworzących sekwencję była odwiedzona ze względu na jej zawartość. Pozostałe są traktowane jako strony odwiedzone w ramach nawigacji do interesującego użytkownika dokumentu. Załóżmy, że log zawiera informacje o odwołaniach użytkownika do stron w następującej kolejności: {*A, B, C, B, D, E, D, F*}. Taką sekwencję można rozbić na trzy transakcje mające postać maksymalnych odwołań w przód: {*A, B, C*}, {*A, B, D, E*} i {*A, B, D, F*}, w których za strony, do których dostęp miał miejsce ze względu na ich zawartość, uważane są: *C*, *E* i *F*.

Po dokonaniu transformacji zawartości logu do postaci zbioru sekwencji będących maksymalnymi odwołaniami w przód, odkrycie częstych ścieżek nawigacji sprowadza się do znalezienia wszystkich podsekwencji występujących w odpowiednio dużej liczbie maksymalnych odwołań w przód. Zbiór odkrytych ścieżek można ograniczyć do tzw. maksymalnych częstych ścieżek, czyli takich które nie są fragmentami innych częstych ścieżek.

Znajomość częstych ścieżek nawigacji może być przydatna przy ocenie schematu powiązań między stronami w ramach witryny internetowej. Może posłużyć do usprawnienia sieci połączeń lub wskazać bardziej odpowiednie miejsca do umieszczania reklam.

4. 2 Odkrywanie reguł asocjacyjnych

Reguły asocjacyjne (*ang. association rules*) zostały zaproponowane w [1] z myślą o analizie koszyka zakupów (*ang. market basket analysis*). Mają one postać implikacji $X \rightarrow Y$, gdzie X i Y są zbiorami elementów. Przykładem reguły asocjacyjnej w kontekście analizy koszyka zakupów jest reguła mówiąca, że 70% klientów kupujących *chleb* kupuje w ramach tej samej transakcji *masło*.

Odkrywanie reguł asocjacyjnych w logu serwera WWW może posłużyć do uzyskania informacji o zbiorach stron, do których użytkownicy mają tendencję odwoływać się w ramach pojedynczej sesji. Dlatego najczęściej przed odkrywaniem reguł asocjacyjnych dane zawarte w logu są transformowane do postaci zbioru transakcji obejmujących dostępy do stron w ramach jednej sesji. Przykładem reguły odkrytej na podstawie analizy logu może być reguła mówiąca, że 40% użytkowników, którzy odwiedzili stronę */products/prod1.htm*, odwiedziło w tej samej sesji stronę */products/prod2.htm*. W związku z tym, że takie reguły wskazują, iż użytkownicy zainteresowani stroną *A*, często wykazują również zainteresowanie stroną *B*,

mogą one być wykorzystane do lepszej organizacji połączeń między stronami w ramach witryny. Ponadto, w przypadku gdy odkryte asocjacje dotyczą stron będących formatkami służącymi do składania zamówień, mogą one znaleźć takie samo zastosowanie jak reguły odkrywane w analizie koszyka zakupów, czyli np. wspomagać planowanie strategii marketingowych.

4. 3 Odkrywanie wzorców sekwencji

Problem odkrywania wzorców sekwencji (*ang. sequential patterns*) [2][13] dotyczył z założenia często powtarzających się wzorców w zachowaniach klientów reprezentowanych przez sekwencje ich transakcji. Wzorzec sekwencji może np. nieść informację o tym, że 20% klientów firmy zajmującej się wysyłkową sprzedażą książek najpierw zamówiło „*Ogniem i mieczem*", potem „*Potop*", a następnie „*Pana Wołodyjowskiego*". Bardzo podobnym problemem jest odkrywanie częstych epizodów (*ang. episodes*) w sekwencjach zdarzeń [9][10], gdzie zdarzeniami mogą być np. awarie w sieciach telekomunikacyjnych, ale także odwołania do konkretnych stron przez użytkowników WWW.

Generalnie, w przypadku analizy logu serwera WWW odkrywanie wzorców w sekwencjach odwołań mających miejsce w różnych sesjach użytkownika jest utrudnione, gdyż jak wspomniano wcześniej identyfikacja użytkowników w czasie wykraczającym poza pojedynczą sesję jest zadaniem bardzo trudnym. Przykładem wzorca sekwencji w odwołaniach do serwera WWW jest informacja, że 20% użytkowników najpierw odwiedziło stronę */products/prod1.htm*, a później (w którejś z kolejnych sesji) stronę */products/prod2.htm*. Zastosowanie wzorców sekwencji w analizie logu serwera WWW wiąże się ściśle z elektronicznym handlem. Znajomość tego typu wzorców może być pomocna przy planowaniu strategii marketingowych.

4. 4 Grupowanie i klasyfikacja użytkowników

Klasyfikacja (*ang. classification*) [14] i grupowanie (*ang. clustering*) [7] są problemami znanymi od dawna w kontekście uczenia maszynowego (ang. *machine learning*) i sztucznej inteligencji. Eksploracja danych dostarcza nowe algorytmy, lepiej radzące sobie z dużą ilością danych wejściowych, z dużą liczbą atrybutów opisujących klasyfikowane lub grupowane obiekty oraz niekiedy brakiem naturalnych miar podobieństwa między obiektami.

Klasyfikacja polega na znalezieniu opisów dla poszczególnych klas, na które został podzielony zbiór klasyfikowanych obiektów. Opisy te mogą mieć np. postać reguł pozwalających ocenić do jakiej klasy dany obiekt należy. Celem klasyfikacji użytkowników WWW na podstawie ich zachowania może być określanie profilu klientów zainteresowanych konkretnym zbiorem dokumentów. Klasyfikacja oparta na danych zawartych w logu serwera WWW może prowadzić do odkrycia takich zależności jak np. reguła mówiąca, że użytkownicy mieszkający w północnych stanach USA często wykazują zainteresowanie stroną */products/sports/winter.htm*.

Celem grupowania jest podział zbioru obiektów na grupy w taki sposób, aby podobieństwo między obiektami, które znajdują się w tej samej grupie, było jak największe, a między obiektami z różnych grup jak najmniejsze. Grupowanie stosowane jest na przykład do wyodrębniania grup klientów posiadających podobne charakterystyki i zainteresowania w celu opracowania trafniejszych strategii marketingowych. Dotyczyć to może oczywiście zarówno firm zajmujących się tradycyjnymi metodami sprzedaży jak i firm prowadzących działalność za pośrednictwem Internetu.

Ciekawą propozycją wykorzystania grupowania w środowisku WWW jest dynamiczna zmiana połączeń między dokumentami zgodnie z przewidywanymi preferencjami użytkowników [15]. Punktem wyjścia dla dynamicznej konfiguracji witryny jest identyfikacja grup użytkowników wykazujących zainteresowanie tym samym zbiorem dokumentów. Każda z odkrytych grup odpowiada jakiemuś profilowi użytkownika i obejmuje zbiór stron. W sytuacji gdy użytkownik zachowuje się zgodnie z jednym z odkrytych profili, tj. odwoływał się do stron związanych z danym profilem, do żądanej przez niego strony dołączane są odnośniki do pozostałych stron należących do danego profilu.

5. Podsumowanie

Informacje zawarte w logu serwera WWW mogą być źródłem użytecznej wiedzy pozwalającej m. in. na lepsze zaprojektowanie witryn internetowych, trafniejsze kierowanie ogłoszeń i reklam do użytkowników, a także określanie strategii marketingowych w ramach elektronicznego biznesu. W odkrywaniu tej wiedzy znajdują zastosowanie techniki eksploracji danych takie jak odkrywanie reguł asocjacyjnych i wzorców sekwencji, klasyfikacja i grupowanie oraz odkrywanie częstych ścieżek nawigacji. Specyfika eksploracji danych zawartych w logu serwera WWW wynika głównie z faktu, że wymagana wstępna obróbka danych nie jest zadaniem trywialnym. Ponadto, aby zwiększyć wiarygodność odkrytej wiedzy wskazane jest stosowanie technik pozwalających na gromadzenie pełniejszej informacji o dostępach do serwera (dodatkowa autoryzacja i eliminacja wpływu pamięci podręcznej). Niestety techniki te wiążą się ze zmniejszeniem poziomu anonimowości użytkowników i wydłużeniem czasu odpowiedzi, co zmniejsza poparcie dla ich stosowania.

Bibliografia

[1] Agrawal R., Imielinski T., Swami A., “Mining Association Rules Between Sets of Items in Large Databases”, Proc. ACM SIGMOD Conference on Management of Data, Washington DC, USA, May 1993.

[2] Agrawal R., Srikant R., "Mining Sequential Patterns", Proc. of the 11th Int’l Conference on Data Engineering (ICDE), Taipei, Taiwan, March 1995.

[3] 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.

[4] Chen M.-S., Park J.S., Yu P.S., “Efficient Data Mining for Path Traversal Patterns”, IEEE Transactions on Knowledge and Data Engineering, Vol. 10, No. 2, March/April 1998.

[5] 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.

[6] Cooley R., Mobasher B., Srivastava J., “Web Mining: Information and Pattern Discovery on the World Wide Web”, Proc. of the 9th IEEE Int’l Conference on Tools with Artificial Intelligence (ICTAI), Newport Beach, California, November 1997.

[7] Hartigan J., “Clustering Algorithms”, John Wiley, 1975.

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

[9] Mannila H., Toivonen H., "Discovering generalized episodes using minimal occurrences", Proc. of the 2nd Int’l Conference on Knowledge Discovery and Data Mining (KDD), Portland, Oregon, August 1996.

[10] Mannila H., Toivonen H., Verkamo A.I., "Discovering frequent episodes in sequences", Proc. of the 1st Int'l Conference on Knowledge Discovery and Data Mining (KDD), Montreal, Canada, August 1995.
[11] 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.
[12] Pitkow J., "In search of reliable usage data on the www", Sixth Int'l World Wide Web Conference, Santa Clara, California, 1997.
[13] Srikant R., Agrawal R., "Mining Sequential Patterns: Generalizations and Performance Improvements", Proc. of the 5th Int'l Conference on Extending Database Technology (EDBT), Avignon, France, March 1996.
[14] Weiss S.M., Kulikowski C.A., "Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems", Morgan Kaufmann, San Mateo, CA, 1991.
[15] 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.