Paweł Weichbroth
Mieczysław Owoc
Uniwersytet Ekonomiczny w Katowicach

WARTOŚCIOWANIE WIEDZY O ŚCIEŻKACH NAWIGACJI UŻYTKOWNIKÓW PORTALI INTERNETOWYCH

Wprowadzenie

Fenomen popularności sieci internetowej wynika z anektowania rozwiniętych metod przekazu informacji, znanych z telewizji i prasy. Spośród dostępnych usług internetu największą popularnością cieszy się usługa WWW, czyli publiczne udostępnianie treści w postaci tzw. stron internetowych. Ciągły wzrost rozmiaru i wykorzystania sieci WWW wytworzył nowe metody projektowania i rozwoju portali internetowych. Bogactwo i rozmaitość udostępnianych zasobów oraz zróżnicowany poziom zainteresowania użytkowników z nich korzystających może prowadzić do osłabienia użyteczności prezentowanych obiektów w obrębie portalu.

Nawigacja pomiędzy zasobami nieodpowiadającymi oczekiwaniom rodzi poczucie niechęci i ostatecznie prowadzi do opuszczenia portalu. Problem ten może być rozwiązany poprzez implementację systemów rekomendacji i personalizacji. Komercyjne zastosowania takich systemów dotyczy m.in. marketingu elektronicznego (*e-marketing*) czy handlu elektronicznego (*e-commerce*)[1].

Autorzy przeprowadzili badania mające na celu eksplorację plików logów serwera WWW, których celem było odkrycie ścieżek nawigacji użytkowników. Do tego celu została zaimplementowana dedykowana aplikacja *Web Log Miner* (WLM) w obiektowym języku C# w architekturze .NET, dostępnej w systemach operacyjnych Microsoft. Pliki logów zostały udostępnione przez portal onet.pl – czwartej najpopularniejszej stronie polskiego internetu w styczniu 2013 r.[2].

---

[1] J.R. Wen, *Enhancing Web Search Through Query Log Mining* [in:] Encyclopedia of Data Warehousing and mining, ed. J. Wang, Idea Group Reference, Hershey 2006, s. 438-442.

[2] Wirtualne Media, *Google i cała reszta – 150 najpopularniejszych stron w polskim Internecie*, http://www.wirtualnemedia.pl/artykul/google-i-cala-reszta-150-najpopularniejszych-stron-w-polskim-internecie# [02.07.2013].

Analiza odkrytej wiedzy w procesie eksploracji plików logów wskazała na konieczność jej weryfikacji oraz oceny przed jej integracją z bazą wiedzy, przetwarzaną w procesie rekomendacji interfejsu użytkownika. W procesie oceny stwierdzono, iż wiedza wyprowadzona (wygenerowana) z tego typu danych była obarczona pewnymi nieprawidłościami, które można sprowadzić do czterech antywłaściwości, takich jak: nieadekwatność, niekompletność, niespójność oraz niepewność. Autorzy postawili sobie za cel opracowanie metody wartościowania wiedzy o ścieżkach nawigacji użytkowników, uzyskanej z eksploracji plików logów serwera WWW. W jej zakresie zaproponowano zbiór ograniczeń, miernik oceny oraz metody i techniki jej weryfikacji i oceny.

1. Pozyskiwanie wiedzy o ścieżkach nawigacji użytkowników

Rekomendacja może być rozpatrywana jako proces identyfikowania preferencji użytkownika i adaptacji serwisu w celu satysfakcjonowania potrzeb użytkownika na podstawie historii zachowania bieżącego użytkownika lub innych, którzy współdzielą podobne zainteresowania do tego użytkownika[3]. Wiedza na potrzeby rekomendacji oraz personalizacji może być pozyskana w sposób[4]: jawny, gdzie użytkownik dobrowolnie i świadomie przekazuje informacje oraz niejawny, gdzie akcje użytkownika są rejestrowane przy zastosowaniu środków (mechanizmów) niezależnych w swojej pracy od interakcji z nim.

Jednym z wielu niejawnych źródeł danych o aktywności użytkowników są pliki logów serwera WWW (*log file*). Są one zapisem wykonanych żądań do zasobów, pozwalające m.in. na diagnozowanie błędów[5], określenie obciążenia serwera poprzez pomiar liczby użytkowników w określonym interwale czasowym[6] oraz analizę użytkowania udostępnianych zasobów[7].

Autorzy na potrzeby reprezentacji wiedzy zbudowali ontologie w odniesieniu do każdego typu wiedzy, przy pomocy matematycznych formalizmów. Ontologie posłużyły także jako wzorce do implementacji określonych struktur da-

---

[3] G. Xu, Y. Zhang, X. Zhou, *Discovering Task-Oriented Usage Pattern for Web Recommendation* [in:] Proceedings of the 17th Australasian Database Conference – Volume 49, Australian Computer Society 2006, s. 167-174.

[4] T. Staś, *Wykorzystanie algorytmów mrowiskowych w procesie doskonalenia portali korporacyjnych*, Wydawnictwo Akademii Ekonomicznej, Katowice 2008 (praca doktorska).

[5] Ibid.

[6] The Apache Software Foundation, *Apache HTTP Server Version 2.2. Log Files*, http://httpd.apache.org/docs/current/logs.html#accesslog [2012.01.23].

[7] J. Pei, J. Han, B. Mortazavi-Asl, H. Zhu, *Mining* Access *Patterns Efficiently from Web Logs*, Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications, Springer-Verlag 2000, s. 396-407.

nych w rozwijanej przez autorów aplikacji *Web Log Miner*. Do reprezentacji wiedzy o ścieżkach nawigacji użytkowników zdefiniowano ontologie sekwencji i transakcji, sekwencyjnej reguły asocjacyjnej i reguły transakcji oraz macierzy transakcji.

Należy podkreślić, iż w procesie rekomendacji wystarczy, aby maszyna wnioskująca utylizowała wiedzę wyłącznie w formie sekwencji. Z drugiej strony, pozostałe typy wiedzy zostały wykorzystane w procesach weryfikacji i oceny wiedzy. Stanowią niejako uzupełnienie sekwencji o czynnik czasu, pozwalając tym samym na ocenę wiarygodności oraz oszacowanie stopnia zainteresowania użytkowników treścią udostępnianą w postaci stron internetowych.

2. Metodyczne podstawy wartościowania generowanych baz wiedzy

W niniejszej pracy przyjęto, iż wartościowanie jest procesem, który pozwala na określenie zgodności bazy wiedzy ze specyfikacjami sformułowanymi dla danej aplikacji[8]. Składa się na niego zbiór czynności, których realizacja pozwoli wyznaczyć wartości bazy wiedzy. Zbiór ten, w nawiązaniu do dychotomicznego ujęcia omawianego terminu, może być podzielony na dwa odrębne podzbiory czynności tj.: weryfikację i ocenę. Można to zapisać w formie równania, danego wzorem (1):

$$\textit{wartościowanie} = \textit{weryfikacja} \cup \textit{ocena} \tag{1}$$

W odniesieniu do bazy wiedzy „weryfikację" należy utożsamiać z procedurą (zbiorem czynności) analizy zapisanej tam wiedzy, która pozwoli na jednoznaczne stwierdzenie, która z jej formalnych specyfikacji została spełniona. Przez jej „ocenę" należy z kolei rozumieć procedurę przetwarzania (pośrednią lub bezpośrednią) bazy wiedzy, która pozwoli na umyślne stwierdzenie, która z pseudoformalnych specyfikacji została spełniona.

Wstępną listę kryteriów weryfikacji i oceny wiedzy sformułował Owoc[9], który ją szczegółowo opisał i ujął w następującej kolejności[10]:

---

[8] M. Owoc, *Wartościowanie wiedzy w inteligentnych systemach wspomagających zarządzanie*, Wydawnictwo Akademii Ekonomicznej, Wrocław 2004.

[9] M. Owoc, *Kryteria wartościowania wiedzy*, Wydawnictwo Akademii Ekonomicznej, Wrocław 1994, Idem, *Measuring Aspects of Knowledge Validation* [in:] materiały konferencyjne Rzeczka 1998, red. A. Baborski, Wydawnictwo Akademii Ekonomicznej, Wrocław 1998.

[10] M. Owoc, *Wartościowanie*…, op. cit.

- **adekwatność** (*adequacy*) – odpowiada takim atrybutom, jak: precyzyjność, odpowiedniość do wiedzy dziedzinowej, zgodność z aktualnymi zdarzeniami lub/ i wiedzą eksperta;
- **kompletność** (*completeness*) – oznacza wyczerpanie wszystkich możliwych przypadków użycia bazy wiedzy[11] (pojęciem przeciwstawnym jest niekompletność); wiedzę można określić jako kompletną kiedy posiada wszystkie elementy konieczne do generowania rozwiązań, jednak nie świadczy o poprawności dochodzenia do konkluzji[12]; w innym ujęciu dotyczy sytuacji pokrycia wszystkich możliwych kombinacji zmiennych przez odpowiedzi ze strony systemu w zakresie wyodrębnionej dziedziny wiedzy[13];
- **spójność** (*consistency*) – dotyczy takiego stanu bazy wiedzy, w której nie są przechowywane takie fakty, które – dla określonych więzów spójności (formuł przedstawiających strukturę wiedzy (*consistency constraint*)) – uniemożliwiałyby realizację celów systemu;
- **wiarygodność** (*reliability*) – została określona jako prawdopobieństwo osiągnięcia celu systemu pod warunkiem wykorzystania konkretnych segmentów bazy wiedzy[14];
- **efektywność** (*effectiveness*) – została określona jako relacja uzyskiwanych z bazy wiedzy zysków do poniesionych kosztów związanych z jej sporządzeniem.

Spójność i kompletność to kryteria realizowane w procesie weryfikacji, zaś wiarygodność, adekwatność oraz efektywność to kryteria przeprowadzone w procesie oceny wiedzy[15].

Wartościowanie bazy wiedzy jest realizowane przy pomocy określonych technik i metod. Pod pojęciem techniki wartościowania należy rozumieć konkretny algorytm, przypisany do określonego kryterium weryfikacji lub oceny bazy wiedzy. Metodę wartościowania będziemy z kolei utożsamiać z co najmniej jedną techniką wartościowania, możliwą do zastosowania w odniesieniu do wybranej grupy kryteriów. W kontekście niniejszego rozdziału metoda będzie tożsama z określonym przedmiotem oraz podmiotem. Przedmiot dotyczy

[11] A. Ligęza, *Logical Foundations for Knowledge–Based Systems*. *Knowledge Representation, Reasoning and Theoretical Properties*, Zeszyty Naukowe AGH. Automatyka, Vol. 63 (1529), Wydawnictwo AGH, Kraków 1993.

[12] M. Owoc, *Wartościowanie*…, op. cit.

[13] L.J. Morell, *Use of Metaknowledge in the Verification of Knowledge-Based Systems*, Proceedings of the 1st International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, Vol. 2, ACM, Tullahoma, Tennessee (USA) 1988, s. 847-857; M. Suwa, A.C. Scott, E.H. Shortliffe, *An Approach to Verifying Completeness and Consistency in a Rule-Based Expert System*, „AI Magazine" 1982, No. 3, s. 16-21.

[14] M. Owoc, *Wartościowanie*…, op. cit.

[15] M. Owoc, M. Ochmańska, T. Gladysz, *On Principles of Knowledge Validation* [w:] *Validation and Verification of Knowledge Based Systems*, eds. A. Vermesan, F. Coenen, Kluwer Academic Publishers, Dordrecht 2010, s. 25-35.

dedykowanego narzędzia – wykonywalnej aplikacji (programu komputerowego), będącej skomplilowanym kodem źródłowym, której poszczególne segmenty reprezentują algorytmy, niezbędne do realizacji procesów odkrywania i wartościowania wiedzy. Podmiot odnosi się z kolei do eksperta (człowieka), który dysponując niezbędną wiedzą i doświadczeniem, parametryzuje i nadzoruje procesy odkrywania i wartościowania wiedzy.

3. Hybrydowa metoda wartościowania wiedzy

Poniżej zostaną przedstawione formalne definicje, niezbędne w odniesieniu do kryteriów oraz mierników weryfikacji i oceny wiedzy generowanej z pliku logów serwera WWW.

Niech $P = \{p_1, p_2, \ldots, p_m\}$ będzie skończonym zbiorem stron internetowych. Sekwencja s to ciąg stron $<s_1\ s_2\ \ldots\ s_k>$ składający się z k elementów zbioru P, niekoniecznie różnych, dla którego jest spełniony warunek $s_{k-1} \neq s_k$. Długością sekwencji (ciągu) $<s_1\ s_2\ \ldots\ s_k>$ nazywamy liczbę k jego elementów (stron).

Dany jest zbiór danych D złożony z sesji (użytkowników), którego elementami są uporządkowane pary postaci $d = <d.id,\ d.seq>$, gdzie $d.id$ oznacza unikatowy identyfikator sesji (użytkownika), zaś $d.seq$ jest opisującą go sekwencją (ciągiem) $s^i = <p_1^i\ p_2^i\ \ldots\ p_{n_i}^i>$ elementów zbioru P, gdzie $i = d.id$, a n_i jest długością sesji (sekwencji) s^i. Dwie różne sesje mogą posiadać jednakowe sekwencje, czyli mogą istnieć takie $d_1, d_2 \in D$, że $d_1.seq \neq d_2.seq$, jednak zachodzi $d_1.id = d_2.id$.

Wsparcie sekwencji s to tzw. nośnik ciągu s w danym zbiorze sesji D – równy liczbie sesji, które zawierają s i są oznaczone jako $supp_D(s)$:

$$supp_D(s) = |\{d.id: d \in D \text{ oraz } s \sqsubseteq d.seq\}|.$$

W dalszej części pracy, tam gdzie nie będzie to prowadzić do niejasności, indeks dolny „D” będzie pomijany.

W danym zbiorze danych D sekwencja (ciąg) s będzie określana jako częsta, jeśli dla ustalonego minimalnego wsparcia (nośnika) min_supp, jest spełniona nierówność:

$$supp(s) \geq min\_supp.$$

Zbiór sekwencji częstych długości k oznaczymy przez LS_k, a zbiór wszystkich sekwencji częstych przez $\boldsymbol{LS}$, gdzie N jest maksymalną długością sesji w D. Poszczególne sekwencje o k-tej długości należące do zbioru LS_k oznaczymy przez $\{<ls_{1k}>, <ls_{2k}>, \ldots, <ls_{jk}>; j \in \{1, 2, \ldots, n\}; k \in \{1, 2, \ldots, n\}\}$.

Użytkownik (ekspert) procesu eksploracji pliku logów serwera ma do wyboru w obrębie aplikacji WLM dziewięć ograniczeń, tj.:

1. ***Data*** (*date*) – pozwala wyodrębnić ze zbioru danych *D* wyłącznie te sesje, których data rozpoczęcia spełnia zdefiniowane warunki.
2. ***Wyklucz element*** (*exclude item*) – pozwala usunąć ze zbioru danych *D* wyłącznie te sesje, które nie rozpoczynają się, nie kończą lub nie zawierają wskazanego elementu.
3. ***Zawiera element*** (*include item*) – umożliwia pozostawienie w zbiorze danych *D* wyłącznie te sesje, które rozpoczynają się, kończą lub zawierają wskazany element.
4. ***Długość sesji*** (*session lenght*) – umożliwia pozostawienie w zbiorze danych *D* wyłącznie tych sesji, których długość (łączna liczba żądanych stron przez użytkownika) spełnia zdefiniowane warunki.
5. ***Całkowity czas trwania sesji*** (*session duration*) – umożliwia pozostawienie w zbiorze danych *D* wyłącznie tych sesji, których łączny czas przeglądania (z wyjątkiem ostatniej strony) spełnia zdefiniowane warunki.
6. ***Czas*** (*time*) – umożliwia pozostawienie w zbiorze danych D wyłącznie tych sesji, których czas rozpoczęcia spełnia zdefiniowane warunki.
7. ***Pozycja*** (*position*) – pozwala na „przycięcie" sesji użytkowników zbioru danych *D* o liczbę żądań równą zdefiniowanemu warunkowi.
8. ***Wsparcie*** (*support*) – to miara atrakcyjności sekwencji przedstawiająca liczbę jej wystąpień w zbiorze danych *D*.
9. ***Zaufanie*** (*confidence*) – to miara atrakcyjności sekwencyjnych reguł asocjacyjnych, przedstawiająca prawdopodobieństwo warunkowe zajścia następnika pod warunkiem wystąpienia poprzednika. Działa na poziomie algorytmu eksploracji danych, który usuwa reguły, niespełniające zdefiniowanego progu odcięcia.

W perspektywie procesu pozyskiwania i modelowania wiedzy o ścieżkach nawigacji użytkowników, użytkownik procesu eksploracji danych w zakresie jej weryfikacji i oceny ma do dyspozycji dziewięć ograniczeń. Definiowanie ww. ograniczeń ma dwojaki charakter. Z jednej strony, pierwszych siedem ograniczeń ma bezpośrednie zastosowanie w odniesieniu do zbioru danych *D*. W praktyce zasadniczo ma to na celu redukcję rozmiaru zbioru danych *D*, tuż przed rozpoczęciem procesu ich eksploracji, co pozwala na skrócenie czasu jego trwania. Z drugiej strony, wielowymiarowa redukcja rozmiaru zbioru danych *D* umożliwia odkrycie takiej wiedzy, która przy zastosowaniu typowych ograniczeń, tj. wsparcie i zaufanie, byłaby zbyt czasochłonna lub nawet niemożliwa do realizacji. Innymi słowy, racjonalna i zasadna manipulacja powyższym zbiorem ograniczeń ewidentnie usprawnia proces odkrywania wiedzy.

Procedura weryfikacji wygenerowanych struktur wiedzy została zaimplementowana w aplikacji WLM. Jest domyślnie uruchamiana po procedurze eksploracji danych i dotyczy wyłącznie kryterium spójności. Przez spójność w kon-

tekście wiedzy generowanej z pliku logów serwera należy rozumieć poprawne mapowanie (przyporządkowanie) zmiennej typu arytmetycznego do zmiennej typu łańcuchowego. Niech zbiór $Dict = \{p_1 = str_1, p_2 = str_2, \ldots, p_m = str_l\}$ oznacza zbiór słownika stron internetowych, gdzie poszczególnym stronom ze zbioru P zostały przyporządkowane unikatowe łańcuchy tekstowe.

Współczynnik spójności sekwencji *SCR* (*Sequence Consistency Ratio*) przedstawia stosunek sekwencji, dla których mapowanie zostało wykonane poprawnie do wszystkich wygenerowanych sekwencji (dla zdefiniowanych przez użytkownika ograniczeń) i został wyrażony wzorem (1):

$$SCR = \bigcup_{kj} \{ls_{kj}; \forall_i (ls[i]_{kj}) \in Dict\} \div |\boldsymbol{LS}|. \quad (1)$$

Pomnożony przez 100 pokazuje jaki odsetek sekwencji jest spójny, tj. możliwy w całości do interpretacji przez właściciela procesu.

Niech $\boldsymbol{SR} = \{SR_2, SR_3, \ldots, SR_n\}$ oznacza rodzinę zbiorów sekwencyjnych reguł asocjacyjnych, która jest sumą zbiorów reguł o długości co najmniej równej dwa, co można zapisać $\boldsymbol{SR} = \bigcup_{k=2} SR_k$. Poszczególne sekwencyjne reguły asocjacyjne o *k*-tej długości należące do zbioru SR_k oznaczymy przez $\{<sr_{1k}>, <sr_{2k}>, \ldots, <sr_{jk}>; j \in \{1, 2, \ldots, n\}; k \in \{2, 3, \ldots, n\}\}$. Analogicznie do współczynnika SCR został określony współczynnik spójności sekwencyjnych reguł asocjacyjnych RCR (*sequential association Rule Consistency Ratio*), który przedstawia stosunek sekwencyjnych reguł asocjacyjnych, dla których mapowanie zostało wykonane poprawnie do wszystkich wygenerowanych reguł (dla zdefiniowanych przez użytkownika ograniczeń) i jest dany wzorem (2):

$$RCR = \bigcup_{kj} \{sr_{kj}; \forall_i (sr[i]_{kj}) \in Dict\} \div |\boldsymbol{SR}.| \quad (2)$$

Pomnożony przez 100 pokazuje jaki odsetek sekwencyjnych reguł asocjacyjnych jest spójny, tj. możliwy w całości do interpretacji przez właściciela procesu.

Drugim kryterium weryfikacji wiedzy jest kompletność. Jak zauważa Owoc[16] w procesie weryfikacji „(...) istotna jest znajomość logiki generowania ekspertyz". Przymiotnik „częsty" odzwierciedla kryterium procesu indukcji konstrukcji nowych struktur danych, określanych mianem wiedzy, która pozwala na wnioskowanie z prawdziwości przesłanek w odniesieniu do zaistniałych w ich rezultacie następstw. Podstawowym „kawałkiem" tego typu wiedzy w sztucznej inteligencji jest reguła, która implikuje powyższy typ wnioskowania, zarówno w swojej

[16] M. Owoc, *Wartościowanie*…, op. cit.

trywialnej formie zapisu: $\alpha \rightarrow \beta$, jak również domyślnej interpretacji: „jeżeli α, to β". „Częsta" reguła to taka, która wystąpiła co najmniej tyle razy w zbiorze przypadków na ile określił to ekspert lub użytkownik procesu. Oznacza to, iż pojęcie „częstej" reguły jest w pewnym stopniu subiektywne. W skrajnych przypadkach kryterium to może być określone na poziomie jednego przypadku (wsparcie równe 1) lub na poziomie równym mocy zbioru przypadków (wsparcie równe liczbie przypadków).

W odniesieniu do systemu rekomendacji, pożądane jest generowanie rekomendacji w obrębie każdego żądania użytkownika. Innymi słowy, jest to sytuacja, w której maszyna wnioskująca, utylizująca bazę wiedzy, będzie w stanie wyznaczyć rekomendację dla każdej sekwencji żądań. Zdaniem autorów jest to zadanie możliwe do realizacji, jednak przy z góry przyjętych założeniach, upraszczających metodę generowania rekomendacji. Na przykład zakładając stałą liczbę obiektów rekomendacji, procedura weryfikacji kompletności bazy wiedzy polegałaby na „odpytaniu" systemu z każdego obiektu i przeglądu udzielonych „odpowiedzi". Z drugiej strony, kompletność bazy wiedzy w kontekście jej rozmiaru jest krytycznie uzależniona od zdefiniowanego wsparcia i zaufania. Próg odcięcia, określany również jako minimalny poziom wsparcia dla sekwencji oraz minimalny poziom zaufania dla reguł, zdefiniowany na zbyt wysokim poziomie, negatywnie wpłynie zarówno na liczbę, jak i na długość wygenerowanych częstych sekwencji (reguł). Mała liczba krótkich sekwencji (reguł) może oznaczać brak możliwości wyznaczania rekomendacji dla użytkowników portalu.

Jak już zasygnalizowano powyżej, wynikiem procesu eksploracji pliku logów serwera, udostępniającego zasoby w postaci stron internetowych, jest rodzina zbiorów częstych sekwencji, reprezentujących częste ścieżki nawigacji jego użytkowników. Biorąc pod uwagę określone w punkcie piątym kryteria wartościowania wiedzy, w pierwszej kolejności ekspert dokonuje oceny adekwatności uzyskanej wiedzy. Podobnie jak przypadku minimalnego wsparcia, ekspert subiektywnie określa próg adekwatności, będący odbiciem posiadanego doświadczenia i wiedzy w zakresie dziedziny problemu.

Niech *adq* oznacza ustalony przez eksperta próg adekwatności, wyrażający maksymalną długość (liczbę elementów) sekwencji częstych. Współczynnik adekwatności SAR (*Sequence Adequacy Ratio*) został określony jako iloraz sumy mocy zbiorów sekwencji częstych o długości co najwyżej równej *adq* do mocy rodziny zbiorów częstych sekwencji, zapisany wzorem (3).

$$SAR = \sum_{k=1}^{adq} |LS_k| \div |\boldsymbol{LS}.| \tag{3}$$

Pomnożony przez 100 pokazuje jaki odsetek wygenerowanych sekwencji jest adekwatny, tj. zgodny z wiedzą eksperta. Z drugiej strony, dla przyjętych założeń w procesie rekomendacji, co było już sygnalizowane przy omawianiu kryterium kompletności, adekwatność może być intepretowana jako odsetek sekwencji: (a) możliwych do zastosowania w procesie rekomendacji lub (b) odpowiadających przesłankom integracji w struktury istniejącej bazy wiedzy.

Analogicznie został zdefiniowany współczynnik adekwatności sekwencyjnych reguł asocjacyjnych RAR (*sequential association Rule Adequacy Ratio*), jako iloraz sumy mocy zbiorów sekwencyjnych reguł asocjacyjnych o długości co najwyżej równej *adq* do mocy rodziny zbiorów sekwencyjnych reguł asocjacyjnych, dany wzorem (4):

$$RAR = \sum_{k=2}^{adq} |SR_k| \div |\boldsymbol{SR}|. \tag{4}$$

Pomnożony przez 100 pokazuje jaki odsetek wygenerowanych sekwencyjnych reguł asocjacyjnych jest adekwatny, tj. zgodny z wiedzą eksperta. Możliwa do przyjęcia jest alternatywna interpretacja adekwatności reguł, analogiczna do tej zaproponowanej w odniesieniu do adekwatności sekwencji.

Drugim kryterium oceny wiedzy jest wiarygodność. Główne czynniki wiarygodności są dość typowe[17], a mianowicie dotyczą źródeł wiedzy, zawartości wiedzy i zastosowanych metod reprezentacji wiedzy. W odniesieniu do analizy użytkowania zasobów internetowych, zdaniem autorów, wiarygodność wygenerowanej wiedzy odnosi się do wybranej metody (lub metod) rekonstrukcji sesji użytkowników oraz przyjętych ograniczeń. Ponadto, ocena wiarygodności wiedzy odnosi się do czasu jaki użytkownik poświęcił na przeglądanie określonych zasobów (stron internetowych). W tym celu są generowane dwa typy wiedzy: transakcje użytkowników oraz macierzy transakcji. Pierwszy typ wiedzy to ważona sekwencja, gdzie każda pojedyncza waga jest medianą czasu trwania, jaką użytkownik spędził na danej stronie (z wyjątkiem ostatniej). Rozwinięciem transakcji jest tzw. macierz transakcji, gdzie każdy pojedynczy wiersz przedstawia czas, jaki użytkownik spędził na danej stronie. Liczba wierszy macierzy jest równa liczbie wystąpień sekwencji (równa wsparciu sekwencji), zaś każdy wiersz dodatkowo posiada informację o dacie i czasie wystąpienia sesji użytkownika. Niskie czasy, pomimo tego, iż transakcja jest częsta, mogą świadczyć o małym zainteresowaniu prezentowaną treścią w obrębie danej strony.

[17] Ibid.

Trzecim i ostatnim kryterium oceny wiedzy jest efektywność. Autorzy nie posiadając dostępu do danych finansowych, nie mieli możliwości jej oszacowania. Ponadto, biorąc pod uwagę ograniczoną objętość niniejszej pracy, rozważania teoretyczne umyślnie pominięto.

Podsumowanie

Realizacja procesu wartościowania wiedzy odbywa się przy pomocy dedykowanych technik, reprezentowanych w postaci zaimplementowanych algorytmów, uruchamianych na plikach tekstowych, przechowujących wygenerowaną wiedzę. Proces ten występuje przed jej integracją z bazą wiedzy, utylizowaną przez maszynę wnioskującą w procesie adaptacji interfejsu użytkownika portalu internetowego. Propozycja autorów w tym zakresie dotyczy czterech technik, sekwencyjnie uruchamianych na poszczególnych typach wiedzy, które obejmują:

1. **Inspekcję** (*inspection*), przeprowadzaną przez eksperta w celu przeglądu wygenerowanej wiedzy; jest to jedyna technika w pełni manualna i tym samym „obciążona" czynnikiem ludzkim, co oznacza wysoki poziom arbitralności i subiektywizmu.
2. **Identyfikację** (*identification*), przeprowadzaną przez eksperta w celu diagnozy stanu wiedzy i rozpoznania przyczyn ewentualnych anomalii wiedzy; jest to technika półautomatyczna, realizowana wespół z dedykowanym narzędziem (aplikacją), której funkcjonalność wspomaga ww. procesy; w przeciwieństwie do inspekcji posługuje się formalnie i jawnie określonymi kryteriami w stosunku do dziedziny problemu.
3. **Eliminację** (*elimination*), przeprowadzoną w celu usunięcia źródeł anomalii wiedzy przy wykorzystaniu dedykowanego narzędzia; jest to technika automatyczna (nienadzorowana) lub półautomatyczna (nadzorowana, tj. systematycznie obserwowana i zatwierdzana przez eksperta).
4. **Zastępstwo** (*substitution*), przeprowadzona w celu zastąpienia brakującej (niepełnej) wiedzy lub jej modyfikacji; jest to technika półautomatyczna; stanowi uzupełnienie procesu inspekcji i jest przeprowadzana „na żądanie" przez eksperta.

Bezpośrednie zastosowanie powyższych technik uszlachetniania generowanych baz wiedzy pozwoliło na eliminację irrelewantnej wiedzy z punktu widzenia zastosowania systemu rekomendacji w procesie adaptacji interfejsu użytkownika.

Bazy wiedzy wykorzystywane w procesie rekomendacji interfejsu użytkowników portali internetowych ulegają okresowej fragmentacji w konsekwencji aktualizacji udostępnianych tam zasobów. Innymi słowy przechowywana tam

wiedza naturalnie „starzeje się" wraz z zawartością, której dotyczy. Proces wartościowania wiedzy należy zatem przeprowadzać w takich odstępach czasu, które korespondują ze zmianami udostępnianej zawartości w obrębie portalu.

Literatura

Ligęza A., *Logical Foundations for Knowledge–Based Systems*. *Knowledge Representation, Reasoning and Theoretical Properties*, Wydawnictwo AGH, Kraków 1993.

Morell L. J., *Use of Metaknowledge in the Verification of Knowledge-based Systems*, Proceedings of the 1st International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems 1988, Vol. 2, ACM, Tullahoma, Tennessee (USA) 1988.

Owoc M., *Kryteria wartościowania wiedzy*, Wydawnictwo Akademii Ekonomicznej, Wrocław 1994.

Owoc M., *Measuring Aspects of Knowledge Validation* [w:] materiały konferencyjne Rzeczka 1998, red. A. Baborski, Wydawnictwo Akademii Ekonomicznej, Wrocław 1998.

Owoc M., *Wartościowanie wiedzy w inteligentnych systemach wspomagających zarządzanie*, Wydawnictwo Akademii Ekonomicznej, Wrocław 2004.

Pei J., Han J., Mortazavi-Asl B., Zhu H., *Mining Access Patterns Efficiently from Web Logs*, Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications, Springer-Verlag 2000.

Staś T., *Wykorzystanie algorytmów mrowiskowych w procesie doskonalenia portali korporacyjnych*, Wydawnictwo Akademii Ekonomicznej, Katowice 2008 (praca doktorska).

Suwa M., Scott A.C., Shortliffe E.H., *An Approach to Verifying Completeness and Consistency in a Rule-Based Expert System*, „AI Magazine" 1982, No. 3.

The Apache Software Foundation, Apache HTTP Server Version 2.2. Log Files. http://httpd.apache.org/docs/current/logs.html#accesslog [23.01.2012].

Wen J.R., *Enhancing Web Search Through Query Log Mining* [w:] *Encyclopedia of Data Warehousing and mining*, ed. J. Wang, Idea Group Reference, Hershey 2006.

Wirtualne Media, Google i cała reszta – 150 najpopularniejszych stron w polskim Internecie, http://www.wirtualnemedia.pl/artykul/google-i-cala-reszta-150-najpopularniejszych-stron-w-polskim-internecie# [02.07.2013].

Xu G., Zhang Y., Zhou X., *Discovering Task-oriented Usage Pattern for Web Recommendation* [in:] Proceedings of the 17th Australasian Database Conference, Vol. 49, Australian Computer Society 2006.

EVALUATING KNOWLEDGE OF WEB PORTAL USERS' NAVIGATION PATHS

Summary

The aim of this article is presentation fundamentals of the proposed hybrid method knowledge validation concerning web user navigation patterns discovery. Four techniques of knowledge validation are employed in the described method: inspection, identification, elimination and substitution. In the implemented program algorithm WLM necessary constraints and indicators have been elaborated: sequence consistency ratio, sequence adequacy ratio and sequential association rule adequacy ratio.