INFORMATYKA EKONOMICZNA BUSINESS INFORMATICS 22 • 2011

ISSN 1507-3858

Łukasz Mikulski, Paweł Weichbroth
Uniwersytet Mikołaja Kopernika, Memex

USUWANIE ARTEFAKTÓW W WYKRYWANIU WZORCÓW UŻYTKOWANIA STRON WWW

Streszczenie: Aktywność użytkowników korzystających z zasobów portali internetowych zapisywana jest w plikach logów serwera WWW. W celu odkrycia i analizy wzorców zachowań takich użytkowników dane z surowych plików loga należy poddać przetwarzaniu wstępnemu. W niniejszym artykule przeprowadzono badanie złożone z dwóch etapów. W pierwszym z nich znajdowane są wszystkie częste zbiory przy arbitralnie założonym współczynniku wsparcia oraz reguły asocjacyjne, przy arbitralnie założonym współczynniku ufności. W drugim etapie podawane są analizie uzyskane wyniki – częste zbiory i wygenerowane na ich podstawie reguły asocjacyjne. Na podstawie tej analizy usuwane są zakwalifikowane jako szum subiektywnie wybrane zbiory, które pochodzą spoza zakresu danych objętego badaniem lub dominują nad pozostałymi elementami serwisu. Jeżeli zachodzi taka potrzeba, czynności związane z odszumianiem danych można iterować. Na podstawie tak przetworzonego pliku loga serwera WWW ostatecznie znajdowane są często zbiory. Na ich podstawie ekstrahowane są z kolei reguły asocjacyjne. Według tak przyjętego założenia, to one mogą odzwierciedlać istotne ścieżki nawigacji, które odpowiednio zagregowane posłużą do wyodrębnienia wzorców użytkowania portalu.

Słowa kluczowe: analiza użytkowania zasobów WWW, odkrywanie wiedzy z baz danych, drążenie danych.

1. Wstęp

W ostatniej dekadzie zauważalny jest wysoki stopień rozwoju przemysłu telekomunikacyjnego oraz informatycznego. Dobrym przykładem jest wysoka dostępność łączy internetowych i popularność komputerów osobistych. Internet, będący ogólnie dostępną siecią komputerową, jest z powodzeniem wykorzystywany jako kanał sprzedaży w handlu hurtowym i detalicznym. Elektroniczną ladą tradycyjnego sklepu są serwisy witryn internetowych, udostępniane w ramach usługi WWW.

Usługa WWW jest typową usługą w architekturze klient-serwer. Serwisy WWW są przechowywane i udostępniane po stronie serwera, użytkownicy zaś uzyskują do nich dostęp, korzystając z przeglądarki internetowej. Witryna interne-

towa składa się ze stron, do których domyślnie użytkownik otrzymuje anonimowy dostęp za pomocą protokołu http [Berners-Lee i in. 1995; Fielding i in. 1997; Fielding i in. 1999]. Aktywność użytkowników witryny WWW zapisywana jest po stronie serwera w tzw. plikach logów.

W literaturze przedmiotu eksploracji danych ekstrakcja wiedzy z plików loga serwera WWW [Mikulski, Weichbroth 2009; Weichbroth 2009a] nazywana jest analizą użytkowania sieci WWW (*web usage mining*) [Cooley i in. 1997; Mobasher i in. 1996]. W przeprowadzonym badaniu portalu Onet.pl wykorzystano dobrze znany i opisany algorytm Apriori [Agrawal i in. 1993; Agrawal, Srikant 1994; Weichbroth 2009b], rozszerzony w stosunku do pierwotnego rozwiązania o możliwość jednoczesnego generowania reguł asocjacyjnych podczas wyszukiwania „częstych" zbiorów [Mikulski, Weichbroth 2009]. Wyniki uzyskane w badaniu pierwotnym i ich interpretacja wskazały na potrzebę odszumienia danych. Niniejsza praca przedstawia wyniki kolejnego etapu badania plików logów serwera WWW, nazwanego badaniem wtórnym. Jednym z głównych założeń badania wtórnego jest analiza danych po usunięciu z nich znalezionych w badaniu pierwotnym artefaktów. Określona w ten sposób redukcja „szumu informacyjnego" pozwoli na znalezienie wzorców zachowań użytkowników witryny internetowej, dotąd niespełniających z góry założonego poziomu wsparcia. Odkryte profile użytkowników, zapisane w postaci reguł asocjacyjnych, mogą być z sukcesem wykorzystane do personalizacji treści, reklamy internetowej bądź handlu elektronicznego. Analiza użytkowania witryn internetowych jest obecnie zagadnieniem bardzo popularnym i przedmiotem intensywnych badań [Hatonen i in. 2003; Ivancsy, Vajk 2006; Kosala, Blockel 2000].

2. Matematyczna definicja problemu

Reguły asocjacyjne są zwykle wykorzystywane do ekstrakcji wiedzy i zależności ukrytych w danych. Ekstrakcja reguł asocjacyjnych wymaga z definicji znalezienia wszystkich „częstych" zbiorów. Odkrywanie reguł asocjacyjnych polega na znalezieniu grupy obiektów, które występują razem w określonym kontekście. Zadanie to jest realizowane poprzez wykorzystanie algorytmów analizy związków (*association rules analysis*), takich jak algorytm Apriori [Agrawal i in. 1993; Agrawal, Srikant 1994].

Reguła asocjacyjna dostarcza informacji w formie stwierdzenia „jeżeli – to" (*if – then*), które dzieli się na dwie części: poprzednika „jeżeli" (*antecedent*) oraz następnika (*consequent*) – „to". W przeciwieństwie do reguły logicznej, posiada cechy probabilistyczne. Reguła jest policzalna w postaci dwóch mierników, które wyrażają jej stopień niepewności.

Pierwszym miernikiem jest wsparcie reguły (*support*), który jest liczbą transakcji, zawierających jednocześnie poprzednik i następnik. W pracy wykorzystano

także współczynnik wsparcia (*support ratio*). Jeżeli $N_{A\to B}$ będzie liczbą sekwencji w postaci $A\to B$ oraz N to ogólna liczba transakcji, to współczynnik wsparcia jest ilorazem liczby sekwencji do liczby transakcji ogółem:

$$\text{support ratio}\{\text{A}\to\text{B}\}=\frac{N_{A\to B}}{N}.$$

Drugim miernikiem jest współczynnik ufności (*confidence ratio*), będący stosunkiem liczby transakcji zawierających wszystkie transakcje poprzednika i następnika do liczby transakcji zawierających poprzednika. Jeżeli $A\to B$ będzie regułą typu jeżeli A to B, to współczynnik ufności jest ilorazem wsparcia reguły $A\to B$ do wsparcia dla zmiennej A:

$$\text{confidence}\{\text{A}\to\text{B}\}=\frac{N_{A\to B}}{N_A}=\frac{\frac{N_{A\to B}}{N}}{\frac{N_A}{N}}=\frac{\text{support}\{\text{A}\to\text{B}\}}{\text{support}\{\text{A}\}}.$$

Dane wejściowe dla opisywanego problemu stanowi ciąg $(P'_i)_{i=1..M'}$ podzbiorów pewnego uniwersum U'. Zbiorom P'_i odpowiadają pojedyncze sesje użytkowników portalu internetowego, a elementy tych zbiorów, jak i całego uniwersum U', to pojedyncze żądania konkretnych podstron logowane przez serwer. Wszystkie dane pochodzą z tego samego pliku loga.

W czasie przetwarzania wstępnego następuje ograniczenie uniwersum. Odbywa się to przez usunięcie z niego niepotrzebnych elementów spełniających określone warunki, które mogą być definiowane na każdym z trzech poziomów organizacji portalu. Ostatecznie prowadzi to do określenia nowego uniwersum $U\subseteq U'$. Aplikując zmiany w uniwersum, uzyskujemy nowy ciąg $(P_i)_{i=1..M'}=(P'_i)_{i=1..M'}\cap U$. Operacja ta może spowodować, że pewne elementy ciągu P staną się zbiorami pustymi. Elementy takie zostają usunięte, elementy ciągu P przenumerowane, a ich liczba ustalona na M. Ostatecznie, dane wejściowe dla problemu stanowi ciąg $(P_i)_{i=1..M}$ podzbiorów pewnego uniwersum U.

W badaniu interesować nas będą tylko takie zbiory $A\subseteq U$, które jako podzbiory zbiorów P_i występują w danych wejściowych wystarczająco „często". Formalnie częstość tę będziemy opisywać za pomocą wartości współczynnika wsparcia *support ratio(A)*$=|\{i\ ;\ A\subseteq P_i\}|\ /\ |U|$, a zbiory, których współczynnik wsparcia przekroczy arbitralnie ustalony poziom *minsupp*, nazywać będziemy zbiorami częstymi.

Przez współczynnik zaufania dla pary rozłącznych zbiorów *(A,B)*, co czytać można *B* pod warunkiem że *A*, rozumieć będziemy wartość *confidence ratio(A,B)=support ratio(*$A\cup B$*)/support ratio(A)*. Jeśli współczynnik zaufania przekroczy arbitralnie ustalony poziom *minconf*, to parę *(A,B)* nazwiemy regułą asocjacyjną i oznaczać będziemy $A\to B$; A nazywać będziemy poprzednikiem reguły aso-

cjacyjnej, a B – następnikiem. Warto zauważyć, że interesować nas będą wyłącznie istotne reguły asocjacyjne, czyli takie reguły $A \rightarrow B$, że zbiór $A \cup B$ jest zbiorem „częstym".

W związku z decydującym wpływem częstości zbioru $A \cup B$ na istotność reguły asocjacyjnej $A \rightarrow B$ mówimy, że istotna reguła asocjacyjna $A \rightarrow B$ jest związana ze zbiorem „częstym" $A \cup B$. Związek ten został wykorzystany w trakcie projektowania struktury danych odpowiedzialnej za przechowywanie informacji o pojedynczym zbiorze „częstym" oraz przy procesie generowania reguł asocjacyjnych w trakcie wyznaczania zbiorów „częstych". Stanowi to istotne rozszerzenie algorytmu Apriori.

3. Algorytm Apriori

W prezentowanej pracy reguły asocjacyjne zostały użyte do analizy użytkowania portalu internetowego Onet.pl. Wyniki pierwszych badań zostały przedstawione na konferencji „Sejmik Młodych Informatyków" w Międzyzdrojach w 2009 roku. Do celów niniejszej pracy napisano program RuleMiner, gdzie został zaimplementowany zmodyfikowany algorytm Apriori (opisany powyżej). Jego modyfikacja polegała na generowaniu reguł asocjacji podczas wyszukiwania „częstych" zbiorów. Dodatkowo, do prezentacji wzorców (profili) użytkowników wykorzystano „wzbogaconą" analizę połączeń [Weichbroth 2010]. Wyniki otrzymane z przeprowadzonego badania pokazały konieczność przetwarzania wstępnego. Niezbędny okazał się „czynnik ludzki", który w subiektywny i arbitralny sposób wykluczy znalezione „częste" zbiory i tym samym reguły asocjacyjne, których interpretacja jest z góry znana, a liczba na tyle duża, że przesłania pozostałe zależności. Na przykład wygenerowane reguły asocjacji, które pokazują ścieżki nawigacji użytkowników w ramach serwisu poczty elektronicznej, są oczywiste i nie wnoszą żadnej nowej wiedzy.

Na uwagę zasługują dwa istotne ograniczenia algorytmów grupowania i odkrywania reguł asocjacyjnych, które są istotne również dla innych metod eksploracji danych: problem interpretacji i problem złożoności obliczeniowej.

Reguły otrzymywane w wyniku działania algorytmu odkrywania reguł asocjacyjnych na ogół w małej części stanowią wiedzę, która wcześniej nie była znana i jednocześnie jest użyteczna biznesowo. W większości wypadków otrzymywane reguły są trywialne i potwierdzają już posiadaną wiedzę. Osobną grupę stanowią reguły, których interpretacja biznesowa jest problematyczna i zwykle są pomijane jako niewyjaśnione i incydentalne. Analogiczny problem występuje w wypadku interpretacji reguł klasyfikacyjnych.

Złożoność analizy rośnie wykładniczo, co implikuje zwykle nieakceptowane czasy generowania reguł wynikowych. Problem eksplozji kombinatorycznej jest charakterystyczny dla większości algorytmów odkrywania wiedzy i stanowi jedno z podstawowych ograniczeń tych podejść.

4. Program RuleMiner

W programie RuleMiner wykorzystaliśmy dobrze znany i rozwijany algorytm Apriori autorstwa R. Agrawala i R. Srikanta [1994], służący do znalezienia „częstych" zbiorów. Danymi dla tego algorytmu jest ciąg podzbiorów (P_i) pewnego uniwersum U; każdy z podzbiorów reprezentowany jest przez ciąg par składających się z nazwy podzbioru (identyfikatora podzbioru) oraz elementu (identyfikatora elementu) należącego do tego podzbioru. Dodatkowo przyjętym założeniem jest posortowanie danych ze względu na pierwszą współrzędną. Powoduje to, że pary opisujące jeden podzbiór pojawiają się w danych wejściowych jako zwarty fragment.

Algorytm wyznacza kolejno rodziny „częstych" podzbiorów. Satysfakcjonujący poziom liczebności wyznacza przyjęty arbitralnie współczynnik wsparcia *minsupp*. W efekcie, do dalszego przetwarzania zakwalifikowane są tylko te podzbiory, których bezwzględna częstość występowania w danych wyjściowych jest nie mniejsza od poziomu współczynnika wsparcia. W kroku przygotowawczym algorytmu wyznaczana jest rodzina wszystkich 1-elementowych zbiorów „częstych". Kolejne kroki, wykonywane dopóty, dopóki powstają nowe rodziny zbiorów „częstych", składają się z trzech faz. W k-tym kroku ekstrahowane są wszystkie k-elementowe zbiory „częste" i związane z nimi reguły asocjacyjne.

W pierwszej fazie k-tego kroku algorytm wylicza, wykorzystując wyznaczoną w poprzednim kroku rodzinę zbiorów „częstych", nową rodzinę zbiorów kandydujących. Są to zbiory wielkości k, dla których przynajmniej dwa podzbiory wielkości k–1 są zbiorami „częstymi".

Drugą fazą jest dwustopniowa weryfikacja zbiorów kandydujących. Pierwszym kryterium pozwalającym odrzucić zbiór kandydujący wielkości k jest zbyt mała częstość któregokolwiek z jego podzbiorów wielkości k–1. Wykorzystujemy do tego ponownie zachowaną w poprzednim kroku rodzinę zbiorów „częstych" wielkości k–1. Ostateczna weryfikacja zbioru kandydującego odbywa się przez ponowne odwołanie do danych wejściowych i wyliczenie faktycznego wsparcia dla tego zbioru.

Ostatnia, trzecia faza iteracji algorytmu to zamiana statusu zweryfikowanej rodziny zbiorów kandydujących na status rodziny zbiorów „częstych" z jednoczesną weryfikacją niepustości tej rodziny i przygotowaniem do rozpoczęcia kolejnego kroku.

W zmodyfikowanej wersji algorytmu, zaimplementowanej w programie RuleMiner, dodano do drugiej fazy kroku iteracyjnego generowanie reguł asocjacyjnych związanych ze zbiorami „częstymi" znalezionymi w tym kroku. W trakcie weryfikowania pierwszego stopnia k-elementowego zbioru kandydującego przepisywane są odpowiednio przetworzone reguły asocjacyjne wyznaczone w poprzednim kroku dla jego podzbiorów wielkości k–1. Uzupełnione są one o kilka potencjalnie występujących nowych reguł. Po weryfikacji drugiego stopnia, czyli wyliczeniu faktycznego wsparcia dla zbioru C, odbywa się sprawdzenie kandydujących

reguł asocjacyjnych. Odrzucone zostają te, dla których współczynnik zaufania jest zbyt niski. Szczegółowy opis programu RuleMiner znajduje się w pracy [Mikulski, Weichbroth 2009].

5. Przetwarzanie wstępne

Przygotowanie danych dla programu RuleMiner zostało przeprowadzone z wykorzystaniem standardowych narzędzi dostępnych w systemie Linux. Były to język skryptowy *awk* oraz programy *sort* i *uniq* służące do sortowania wierszy pliku tekstowego oraz usuwania z niego występujących w bezpośrednim następstwie kopii danego wiersza.

Dane otrzymane z portalu Onet.pl posiadają odpowiednio pola: (1) czas sesji, (2) identyfikator sesji, (3) identyfikator użytkownika, (4) identyfikator serwisu, (5) identyfikator podserwisu, (6) identyfikator pliku html w ramach serwisu (tab. 1).

Tabela 1. Wycinek pliku loga serwera WWW

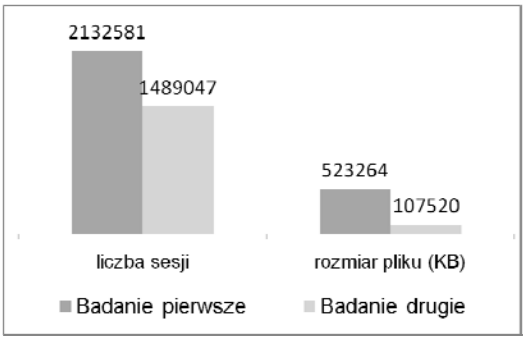
(1)	(2)	(3)	(4)	(5)	(6)
140229	8654368670	216633042432611969	2	5723	724
140229	8654368670	216633042432611969	2	5723	724
165434	8654372778	222771622445704681	124	5086	8052688
165100	8654373234	231008562444696641	26	158730	10076856
173226	8654372778	222771622445704681	124	5086	8151598

Źródło: Onet.pl.

Dane te uzupełnione są o pliki tekstowe zawierające słowniki tłumaczeń występujących w pliku loga identyfikatorów na ścieżki do plików html.

W każdym etapie badania, za pomocą programu napisanego w języku *awk*, z występujących w pliku źródłowym sześciu zmiennych wybierane były dwie. Pierwsza z nich, zmienna (2) – identyfikator sesji, służyła do ustalenia elementów ciągu $(P_i)_{i=1..M}$. Każdej sesji odpowiadał jeden zbiór będący elementem tego ciągu. Druga zmienna służyła do zidentyfikowania elementów uniwersum U, wchodzących w skład danego zbioru P_i. W badaniu wykorzystano zmienną (6) – identyfikator pliku html. W podobny sposób badanie można przeprowadzić, koncentrując się na zmiennych (4) – identyfikator serwisu lub (5) – identyfikator podserwisu.

Jednocześnie z wyborem zmiennych realizowane było ograniczanie uniwersum U. Zgodnie z wnioskami wyciągniętymi po pierwszym badaniu (zrealizowanym na pełnych danych, na poziomie plików html) ograniczenia zostały dokonane na poziomie pojedynczych stron (strona główna serwisu) oraz na poziomie serwisów (serwisy *poczta* oraz *sympatia*). Wiersze zawierające opisy żądań spełniających jeden z wymienionych warunków nie były przetwarzane. W sytuacji gdy wszystkie wiersze pliku źródłowego wchodzące w skład danej sesji spełniały warunki ograniczenia,



Rys. 1. Porównanie liczby sesji do rozmiaru plika loga serwera WWW

Źródło: opracowanie własne.

sesja taka była usuwana z danych wejściowych. Określone w ten sposób warunki redukcji spowodowały zmniejszenie liczby sesji z wyjściowych 2132581 do 1489047, rozmiar pliku zaś zmniejszył się z 511 MB do 105 MB. Natężenie szumu na poziomie 30,18% uzasadnia podejmowane działania, a z punktu widzenia celów badania jego usunięcie daje nadzieję na uzyskanie wiarygodniejszej wiedzy.

6. Wyniki badania

Wyniki badania pierwotnego pliku loga serwera WWW zasugerowały potrzebę przetwarzania wstępnego. Większość uzyskanych w postaci reguł asocjacyjnych informacji okazała się nie wnosić istotnej wiedzy na temat ścieżek nawigacji użytkowników. Dało się wyodrębnić trzy zbiory [{www},{sympatia},{poczta}], zakwalifikowane jako artefakty, które zostały usunięte ze zbioru danych. Tak przetworzony plik loga został ponownie poddany analizie.

Aplikacja RuleMiner pozwala na wyekstrahowanie „częstych" zbiorów i wygenerowanie na ich podstawie reguł asocjacyjnych z pliku loga serwera WWW. Zostały przyjęte dwa kryteria selekcji wyników, tj. współczynnik wsparcia oraz współczynnik zaufania.

Tabela 2. Liczba znalezionych „częstych" zbiorów

Liczba częstych zbiorów	Współczynnik wsparcia			Liczba częstych zbiorów	Współczynnik wsparcia		
	0,01	0,005	0,001		0,01	0,005	0,001
Jednoelementowych	98	98	284	Czteroelementowych	1	54	1080
Dwuelementowych	74	74	1124	Pięcioelementowych	0	0	326
Trzyelementowych	29	262	1663	Sześcioelementowych	0	0	38
Suma					202	488	4515

Źródło: opracowanie własne.

Dla porównania, na poziomie wsparcia 0,01 w pierwszym badaniu zostało znalezionych 59 jednoelementowych zbiorów, spośród których 18 zakwalifikowanych zostało jako artefakty. Ich usunięcie z pliku loga serwera WWW pozwoliło na wyeliminowanie temporalnych ścieżek nawigacji użytkowników, które zostały zaobserwowane na podstawie „bieżących" wiadomości na stronie głównej portalu. Ponadto wykryte zostały wewnętrzne zależności między stronami serwisów: *poczta* i *sympatia*, które wydają się mało użyteczne.

Usunięcie z pliku logów serwera małej liczby artefaktów spowodowało znaczny spadek liczby reguł asocjacyjnych. Prezentowane przez nie zależności dotyczą statycznych stron w schemacie portalu. Treść tych stron jest na bieżąco aktualizowana, nazwy plików zaś pozostają niezmienne. Warto zauważyć, że taka analiza dotyczy trwałych ścieżek nawigacji użytkowników – powtarzalnych i niezależnych od dodawanych na bieżąco stron. Na statyczny szkielet portalu składają się między innymi zbiory [[info]/kraj/item.html], [[biznes]/pap.html] czy [[biznes]/gielda.html], których treść jest okresowo aktualizowana.

Tabela 3. Liczba otrzymanych reguł asocjacyjnych w badaniu wtórnym dla współczynnika wsparcia 0,01

Liczba reguł asocjacyjnych	Współczynnik zaufania			Liczba reguł asocjacyjnych	Współczynnik zaufania		
	0,9	0,8	0,7		0,9	0,8	0,7
Dwuelementowych	2	3	5	Pięcioelementowych	0	0	0
Trzyelementowych	1	1	15	Sześcioelementowych	0	0	0
Czteroelementowych	0	1	1	Suma	3	5	21

Źródło: opracowanie własne.

Tabela 4. Liczba otrzymanych reguł asocjacyjnych w badaniu pierwotnym dla współczynnika wsparcia 0,01

Liczba reguł asocjacyjnych	Współczynnik zaufania			Liczba reguł asocjacyjnych	Współczynnik zaufania		
	0,9	0,8	0,7		0,9	0,8	0,7
Dwuelementowych	52	72	87	Pięcioelementowych	156	293	411
Trzyelementowych	148	225	308	Sześcioelementowych	42	88	119
Czteroelementowych	216	370	518	Suma	614	1048	1443

Źródło: opracowanie własne na podstawie [Mikulski, Weichbroth 2009].

Wyniki otrzymane w badaniu pierwotnym w 95% [Mikulski, Weichbroth 2009] wykazały zależności pomiędzy stroną główną i podstronami bezpośrednio z niej osiągalnymi oraz zależności wewnątrz autonomicznych serwisów *poczta* i *sympatia*. W porównaniu z badaniem wtórnym wydawać się może, że odkryta wiedza na temat użytkowania portalu jest bogatsza. Biorąc jednak pod uwagę zawartość znalezionych „częstych" zbiorów pierwszego typu, należy stwierdzić, że jest to przeważnie wiedza temporalna. Zależności te dotyczą wyłącznie „bieżących"

wiadomości umieszczonych na stronie głównej portalu. Analiza takiego ruchu wydaje się zbędna i niczego nie wnosi do analizy profili użytkowników portalu. Z drugiej jednak strony, dostarcza informacji na temat częstości, z jaką użytkownicy dostają się na daną podstronę, wykorzystując „głębokie" linki [Wassom 1998].

W przeprowadzonym badaniu udało się wyodrębnić trzy profile użytkowania (w nawiasie podano liczbę porządkową reguły asocjacyjnej):

– informacyjny (1,2),
– biznesowy (3,4,6),
– sportowy (6,7).

Tabela 5. Wybrane reguły asocjacyjne składające się na wyodrębnione profile użytkowników na minimalnym poziomie wsparcia 0,01 oraz zaufania 0,5

Lp.	Reguła asocjacyjna	Wsparcie	Zaufanie
1	[[info]/forum/czytaj/item-swiat.html]→[[info]/swiat/item.html]	0,026	0,983
2	[[info]/forum/czytaj/item-kraj.html]→ [[info]/kraj/item.html]	0,015	0,961
3	[[biznes]/pap.html [biznes]/gielda.html]→ [[biznes]/gielda/wiadomosci.html]	0,015	0,808
4	[[biznes]/pap.html [info]/swiat/item.html]→ [[biznes]/gielda/wiadomosci.html]	0,013	0,826
5	[[biznes]/pap.html[biznes]/wiadomosci.html] → [[biznes]/gielda/wiadomosci.html]	0,02	0,791
6	[[sport]/pilka_nozna/puchar_uefa/wiadomosci.html [sport]/formula_1/wiadomosci.html]→ [[sport]/pilka_nozna/ekstraklasa/wiadomosci.html]	0,024	0,503
7	[[sport]/pilka_nozna/puchar_uefa/wiadomosci.html [sport]/pilka_nozna/ekstraklasa/wiadomosci.html]→ [[sport]/formula_1/wiadomosci.html]	0,023	0,542

Źródło: opracowanie własne.

Podobne profile wyekstrahowane zostały już w czasie badania pierwotnego. Składające się na te profile reguły asocjacyjne były wówczas mało widoczne i ukryte wśród mało istotnych informacji, zakwalifikowanych jako opisany wyżej szum. Pokazuje to, że wyniki badania wtórnego nie są sprzeczne, ale wręcz potwierdzają wcześniejsze wyniki. Uzyskano w ten sposób szerszą wiedzę na temat ścieżek nawigacji użytkowników w ramach portalu, która może być z powodzeniem wykorzystana do pozycjonowania treści w ramach określonych profili. Jak wynika z tab. 5, profile te odzwierciedlają sztywną strukturę portalu.

W pracy [Markov, Larose 2007] dokonano dwustopniowego podziału stron internetowych na strony nawigacyjne oraz strony informacyjne. W kontekście niniejszego badania strony *www* i *poczta* to strony nawigacyjne, a *sympatia* to strona informacyjna. Ogólnie można stwierdzić, że odszumienie polega na usunięciu z danych stron nawigacyjnych oraz mało istotnych stron informacyjnych. Poprzez ich wykluczenie analizowany zbiór danych posiada mniejszą wariancję. Ze względu na częstość

oraz treść aktualizacji portalu niezbędne wydaje się poszerzenie schematu analizy użytkowania zasobów internetowych o opisany w pracy proces odszumiania.

7. Podsumowanie

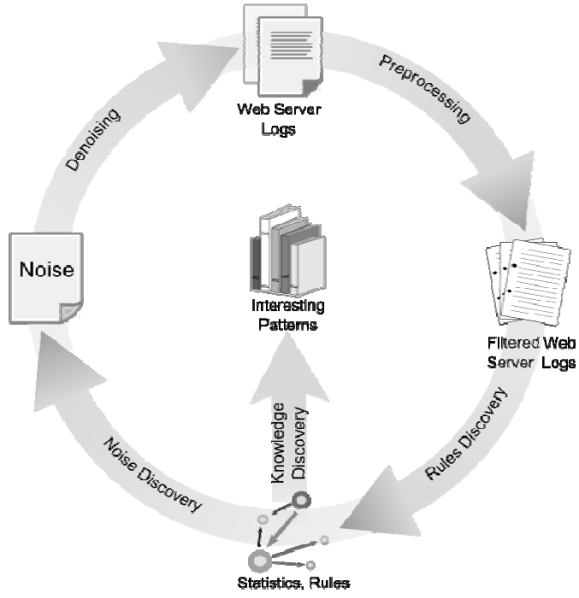
W pracy [Yan i in. 1996] zaprezentowano ogólną koncepcję systemu, która implementuje połączenie podejścia clusteringu i dynamicznie pozycjonowanych odsyłaczy. Na system składają się trzy główne komponenty: serwer WWW, moduł analizy logów typu offline i moduł dynamicznie pozycjonowanych odsyłaczy typu online. Moduł offline wykonuje przetwarzania wstępne okresowo (np. tygodniowo) plików loga serwera WWW, dokonując ekstrakcji wiedzy. Na podstawie zapisanych rekordów tworzone są klastry (kategorie) podobnych sesji użytkowników. *WebWatcher* proponuje podejście oparte na uczeniu maszynowym. Ma to na celu dostarczenie wskazówek nawigacji poprzez wykorzystanie opinii użytkowników końcowych. *Letizia* prezentuje podejście spersonalizowane. Zapisuje ona otwierane przez użytkownika strony, wywołane odsyłacze oraz wpisane w wyszukiwarce słowa kluczowe.

W pracy [Srivastava i in. 2000] zaproponowano schemat analizy użytkowania sieci WWW, który wyodrębnia cztery etapy. Występujący tam etap przetwarzania wstępnego różni się od przedstawionego w tej pracy. W literaturze przedmiotu analizy użytkowania sieci WWW działanie polegające na usunięciu subiektywnie wybranych zbiorów i ponownej analizie pliku loga serwera WWW nie jest utożsamiane z procesem przetwarzania wstępnego. Klasyczne podejście kwalifikowałoby do niego tylko proces projekcji danych z pliku loga na dwie zmienne. W wykonanym badaniu opisane odszumianie danych wejściowych było możliwe tylko przed projekcją i wymagało jej powtórzenia. Skłania to do uznania procesu odszumienia danych za część przetwarzania wstępnego.

Żądania użytkowników do stron internetowych udostępnianych w ramach witryny internetowej zapisywane są w plikach loga serwera WWW. Biorąc pod uwagę format oraz ilość informacji zapisywanych w nich, pliki takie wymagają przetwarzania. Stopień filtrowania zależny jest od aplikacji, dla której pliki te stanowią dane wejściowe, oraz od zakresu podejmowanego badania. Aplikacja RuleMiner, której opis znajduje się w [Mikulski, Weichbroth 2009], przetwarza przygotowane pliki loga serwera WWW. Dla zadanych z góry parametrów: współczynnika wsparcia i współczynnika zaufania, program znajduje „częste" zbiory i na ich podstawie generuje reguły asocjacyjne oraz ich współczynniki wsparcia i zaufania. Ich analiza dostarcza informacji na temat zależności pomiędzy poszczególnymi zbiorami. Użytkownik, klasyfikując reguły w homogeniczne grupy, tworzy wzorce użytkowania. Wartość, jaką niesie ze sobą taka wiedza, jest kwestią subiektywną. Implikuje to podjęcie decyzji o odrzuceniu lub przyjęciu otrzymanych profili. Odrzucenie oznacza wskazanie, które zbiory zakłócają proces analizy i uznanie ich za tzw. szum. W kolejnym kroku zbiory takie są usuwane ze źródłowych plików loga

serwera WWW. Opisany powyżej proces można powtarzać – liczba i skala iteracji jest arbitralną decyzją użytkownika. Końcowym efektem analizy powinna być obiektywna wiedza na temat użytkowania zasobów witryny internetowej, użyteczna z punktu widzenia założonych celów badania.

Na rysunku 2 został pokazany schemat procesu analizy zasobów WWW, który w porównaniu z pracą [Srivastava i in. 2000] został rozszerzony o detekcję tzw. szumu.



Rys. 2. Schemat procesu analizy użytkowania zasobów internetowych

Źródło: opracowanie własne na podstawie [Srivastava i in. 2000].

Wnioski wysnute na podstawie przeprowadzonego badania można podzielić na dwie grupy. W pierwszej znajduje się potwierdzenie dla ogólnych własności metody reguł asocjacyjnych oraz szczegółowych obserwacji dotyczących algorytmu Agrawala i Srikanta i jego konkretnej implementacji w postaci programu RuleMiner. Dzięki zastosowanemu odcięciu na poziomie współczynnika wsparcia otrzymujemy tylko istotne z punktu widzenia badania dane, wraz ze wzrostem tego współczynnika wzrasta liczba oraz rozmiar znalezionych reguł asocjacyjnych. Odcięcie nieistotnych zbiorów częstych na poziomie współczynnika wsparcia realizowane było jako pierwsze, przez co jego poziom ma przy ustalonych danych wejściowych decydujący wpływ na czas działania aplikacji. Współczynnik zaufania, określający poziom, na którym dokonujemy drugiego odcięcia, i realizowany po odcięciu zbiorów częstych o niskim poziomie wsparcia, ma istotny wpływ na ostateczną liczbę reguł, ale zmiana tego współczynnika nie wpływa na czas działania

aplikacji. Liczba elementów składających się na reguły asocjacyjne określa siłę korelacji (im więcej elementów w poprzedniku reguły, tym korelacja jest słabsza; im więcej elementów w następniku reguły, tym korelacja jest silniejsza). Ma to, obok faktycznych wartości wsparcia i zaufania, decydujący wpływ na hierarchizowanie reguł w ostatecznej analizie wyników badania.

Druga grupa wniosków dotyczy przeprowadzonego badania, specyficznych danych, jakie zostały mu poddane, oraz zastosowania wielostopniowego procesu filtracji szumu. Badanie pierwotne ujawniło strukturę portalu. W szczególności wyodrębnione zostały dwa autonomiczne serwisy z niezwykle silnymi, statycznymi ścieżkami nawigacji, posiadające własne grupy odbiorców i wprowadzające szum utrudniający właściwą analizę. Ponadto hierarchiczna budowa samego serwisu informacyjnego znalazła odzwierciedlenie w wyekstrahowanych regułach asocjacyjnych. Dokonane odszumienie danych wejściowych, mimo stosunkowo niewielkiego zmniejszenia ich rozmiaru, pozwoliło znacznie obniżyć parametry w badaniu wtórnym oraz uzyskać dużo ostrzejsze i zgodne z badaniem pierwotnym wyniki. Reasumując, wielostopniowe badanie z usuwaniem zdiagnozowanego szumu informacyjnego wydaje się trafne w przypadku wykorzystania w analizie metody reguł asocjacyjnych. Ponadto sama metoda wydaje się dawać zadowalające rezultaty w wykrywaniu wzorców użytkowania stron WWW.

Literatura

Agrawal R., Imielinski T., Swami A. [1993], *Mining Association Rules between Sets of Items in Large Databases*, [w:] SIGMOD '93: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, ACM, Nowy Jork, s. 207–216.

Agrawal R., Srikant R. [1994], *Fast algorithms for mining association rules*. Proceedings of the Twentieth International Conference on Very Large Data Bases, Morgan Kaufmann, San Francisco, s. 487–499.

Berners-Lee T., Fielding R., Frystyk H. [1995], *Hypertext Transfer Protocol - HTTP/1.0. Internet Draft*, http://www.w3.org/Protocols/HTTP/1.0/draft-ietf-http-spec.html (24.11.2009).

Cooley R., Mobasher B., Srivastava J. [1997], *Web mining: Information and pattern discovery on the world wide web*. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, Los Alamitos, s. 558–567.

Evfimievski A., Srikant R., Agrawal R., Gehrke J. [2002], *Privacy Preserving Mining of Association Rules*. Proceedings of the eighth ACM SIGKDD International Conference on Knowledge discovery and data mining, ACM, Nowy Jork, s. 217–228.

Fielding R., Gettys J., Mogul J., Frystyk H., Berners-Lee T. [1997], *Hypertext Transfer Protocol - HTTP/1.1. Internet Official Protocol Standards (RFC 2068)*, http://tools.ietf.org/html/rfc2068 (24.11.2009).

Fielding R., Gettys J., Mogul J., Frystyk H., Masinter L., Leach P., Berners-Lee T. [1999], *Hypertext Transfer Protocol - HTTP/1.1. Internet Official Protocol Standards (RFC 2616)*, http://www.w3.org/Protocols/ rfc2616/ rfc2616.html (24.11.2009).

Hatonen K., Boulicaut J.F., Klemettinen M., Miettinen M., Mason C. [2003], *Comprehensive Log Compression with frequent patterns*, DaWaK 2003, LNCS 2737, Springer-Verlag, Berlin, s. 360–370.

Ivancsy R., Vajk I. [2006], *Frequent pattern mining in web log data*, Acta Polytechnica Hungarica, vol. 3, no. 1, s. 77–90.

Kosala R., Blockel H. [2000], *Web mining research: A survey*, Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mininig SIGKDD: GKDD Explorations 1.

Markov Z., Larose D.T. [2007], *Data Mining the Web*: *Uncovering Patterns in Web Content Structure*, John Wiley & Sons, New York.

Mikulski Ł., Weichbroth P. [2009], *Discovering patterns of visits on the Internet web sites in the perspective of associative models*, "Polish Journal of Environmental Studies", vol. 18, no. 3B, Olsztyn, s. 267–271.

Mobasher B., Jain N., Han E.S., Srivastava J. [1996], *Web Mining: Pattern Discovery from World Wide Web Transactions*, Technical Report 96-050. University of Minnesota, Minnesota.

Scime A. [2005], *Web Mining: Applications and Techniques*, Idea Group Publishing, Hershey.

Spiliopoulou M., Faulstich L.C. [1998], *WUM: A Web Utilization Miner*, Proceedings of EDBT Workshop WebDB98, Springer Verlag, Berlin, s. 109–115.

Srivastava J., Cooley R., Deshpande M., Tan P.N. [2000], *Web usage mining: discovery and applications of usage patterns from web data*, ACM SIGKDD Explorations Newsletter, vol. 1, issue 2, New York.

Yan T.W., Jacobsen M., Garcia-Molina H., Dayal U. [1996], *From User Access Patterns to Dynamic Hypertext Linking*, Computer Networks and ISDN Systems, vol. 28, issue 7–11, s. 1007–1014.

Wassom B.D. [1998], Note: Copyright Implications of "Unconventional Linking" on the World Wide Web: Framing, Deep Linking and Inlinking, Law Review Case Western Reserve University.

Weichbroth P. [2009a], *Analiza zachowań użytkowników portalu onet.pl w ujęciu reguł asocjacyjnych*, [w:] *Inżynieria Wiedzy i Systemy Ekspertowe*, red. A. Grzech, K. Juszczyszyn, H. Kwaśnicka, N.T. Nguyen, Akademicka Oficyna Wydawnicza Exit, Warszawa, s. 81–88.

Weichbroth P. [2009b], *Odkrywanie reguł asocjacyjnych z transakcyjnych baz danych*, [w:] *Informatyka Ekonomiczna. Rynek usług informatycznych*, red. A. Nowicki, I. Chomiak-Orsa, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu 14, Wyd. Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław, s. 301–309.

Weichbroth P. [2010], *A framework of rule based expert system for market basket analysis*, [w:] *Advanced Information Technologies for Management – AITM2010*, red. J. Korczak, H. Dudycz, M. Dyczkowski, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław, s. 276–291.

DENOISING AS A METHOD OF DISCOVERING HIDDEN WEB USAGE PATTERNS

Summary: The activity of web portals' users is recorded in a WWW server log file. In order to reveal and analyse the web usage patterns, the data from unprocessed log files should be preprocessed. In this article the two-stage research was conducted. In the first one all frequent sets were found, with arbitrarily assumed support ratio, and association rules, with arbitrarily assumed confidence ratio. In the second stage the obtained results were analysed – frequent sets, and based on them generated association rules. Based on this analysis the subjectively chosen sets, classified as noise, are removed. Those are either outside the scope of research data or the ones which dominate other elements. If necessary the activities connected with data denoising can be iterated. Based on such a processed WWW server log file, finally frequent sets are selected. In turn, based on aforementioned, association rules are extracted. Those are the ones reflecting the relevant navigation paths which, while adequately aggregated, would be used to select the web usage patterns.

Keywords: web usage mining, data mining, knowledge discovery from databases.