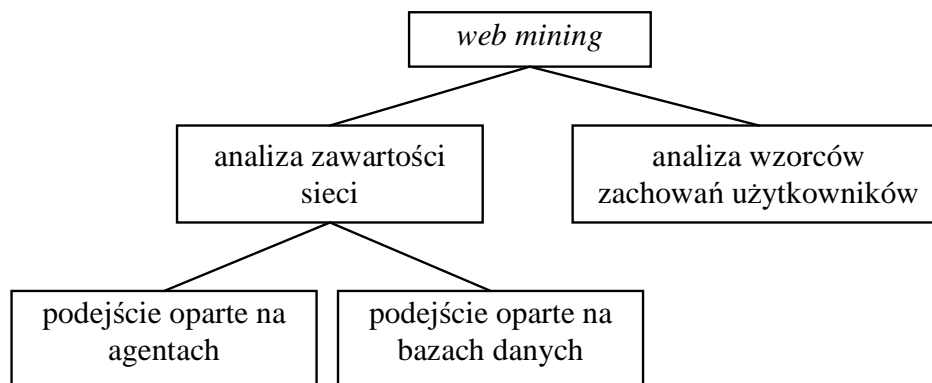


WEB USAGE MINING, CZYLI JAK SPRZEDAĆ SUKIENKĘ CIĄŻOWĄ W INTERNECIE

Mariusz Łapczyński, Uniwersytet Ekonomiczny w Krakowie, Katedra Analizy Rynku i Badań Marketingowych

Wprowadzenie do *web mining*

Duża popularność Internetu jako źródła wiedzy i rozrywki, a także rosnące znaczenie handlu elektronicznego i usług dostępnych za pośrednictwem sieci przyczyniają się do bardzo dynamicznego przyrostu liczby gromadzonych informacji. Analiza tych danych pozwala przedsiębiorstwom oszacować wartość życiową klienta, maksymalizować przychody ze sprzedaży (np. za pomocą sprzedaży krzyżowej), oceniać skuteczność kampanii promocyjnych, optymalizować wygląd i funkcjonalność witryn, dostarczać internautom spersonalizowany przekaz (ofertę, reklamę), czy znaleźć najbardziej skuteczną logiczną strukturę witryny.



Rys. 1. Obszary analityczne *web mining* (źródło: R. Cooley, B. Mobasher, J. Srivastava, *Web Mining: Information and Pattern Discovery on the World Wide Web*, Proceedings of the 9th International Conference on Tools with Artificial Intelligence, IEEE Computer Society, 1997, s. 558).

Web mining może być zdefiniowany jako odkrywanie i analiza użytecznych informacji z Internetu². Polega to na automatycznym przeszukiwaniu zasobów informacji dostępnych

² R. Cooley, B. Mobasher, J. Srivastava, *Web Mining: Information and Pattern Discovery on the World Wide Web*, Proceedings of the 9th International Conference on Tools with Artificial Intelligence, IEEE Computer Society, 1997, s. 558.



on-line (*web content mining*) oraz na odkrywaniu wzorców zachowań użytkowników na podstawie danych dostępnych z serwerów sieciowych (*web usage mining*). Schemat obszarów analitycznych *web mining* przedstawiono na rys. 1. Z punktu widzenia badań marketingowych interesujący może być wprawdzie każdy z nich, jednak jeśli brać pod uwagę analityczny CRM, to przedmiotem zainteresowań badacza jest przede wszystkim zachowanie użytkowników podczas wizyty na stronie internetowej.

Badanie zachowań Internautów odnosi się do automatycznego odkrywania i analizowania wzorców strumienia „kliknięć” (*clickstream*) wraz z innymi zmiennymi gromadzonymi lub wygenerowanymi w czasie kontaktu internauty z zasobami sieciowymi na jednej lub kilku witrynach internetowych. Odkryte w trakcie budowy modelu wzorce są zwykle przedstawione w postaci zestawu stron, obiektów lub zasobów charakteryzujących się dużą częstotliwością dostępu przez homogeniczne pod względem potrzeb i zainteresowań grupy nabywców.

Procedura analityczna składa się z trzech etapów³: 1/ zebranie i wstępne przygotowanie danych, 2/ odkrywanie wzorców, 3/ analiza wzorców. W pierwszym kroku dane są oczyszczone i dzielone na podzbiory transakcji z uwzględnieniem aktywności internautów na stronie podczas każdej wizyty. Zbiór danych transakcyjnych zostaje tu czasami powiększony o dodatkowe zmienne odnoszące się do zawartości strony, jej struktury lub obiektów takich, jak katalogi produktów. W drugim kroku odkrywa się nieznanne wcześniej wzorce zachowań za pomocą narzędzi bazodanowych, narzędzi statystycznych i narzędzi *data mining*. Oprócz wyszukiwania wzorców behawioralnych przeprowadza się tutaj wstępną eksplorację danych w zakresie zasobów internetowych, sesji i użytkowników. W ostatnim kroku procedury, odkryte wzorce i statystyki są filtrowane, agregowane i wykorzystywane jako dane wejściowe do różnych aplikacji, m.in.: silników rekomendacyjnych, aplikacji wizualizacyjnych czy aplikacji generujących raporty.

Proces przygotowania danych jest najbardziej czasochłonnym i obciążającym zasoby sprzętowe etapem całej procedury. Wstępne przetwarzanie oryginalnych danych z różnych źródeł czy transformacja do postaci „akceptowanej” przez oprogramowanie analityczne jest niezbędne do uzyskania efektu końcowego, jakim jest wdrożenie modelu w życie. Podczas przygotowywania danych dotyczących użytkowników można napotkać następujące problemy⁴:

- ◆ Pojedynczy adres IP/kilka sesji – dostawca usług internetowych zwykle posiada kilka serwerów pośredniczących (proxy), pojedynczy serwer może z kolei mieć kilku internautów nawiązujących połączenie z witryną w tym samym czasie.
- ◆ Kilka adresów IP/pojedyncza sesja – niektórzy dostawcy Internetu przydzielają losowo każde zapytanie klienta do jednego z kilku adresów IP; pojedyncza sesja może mieć zatem wiele różnych adresów IP.

³ B. Mobasher, *Web Usage Mining*, w: *Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data*, red. B. Liu, Springer Verlag, Berlin Heidelberg 2007, s. 449.

⁴ J. Srivastava, R. Cooley, M. Deshpande, P-N. Tan, *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*, SIGKDD Explorations, Volume 1, Issue 2, January 2000, s. 14.



- ◆ Wiele adresów IP/pojedynczy użytkownik – użytkownik, który łączy się z siecią z różnych komputerów ma inny adres IP w różnych sesjach.
- ◆ Kilka przeglądarek/pojedynczy użytkownik – internauta korzystający z różnych przeglądarek internetowych (np. IE, Mozilla Firefox, Opera) z jednego komputera może zostać zidentyfikowany jako kilku różnych użytkowników.

Na etapie odkrywania wzorców korzysta się z wielu metod i algorytmów wywodzących się ze statystyki, *data mining*, uczenia się maszyn i rozpoznawania wzorców. Wyróżnia się następujące metody analityczne⁵:

- ◆ Analiza statystyczna – można tutaj wykorzystać statystyki opisowe (częstości, średnią, medianę itp.) dla zmiennych odnoszących się do przeglądanych stron, czasu przeglądania strony czy liczby odwiedzanych stron w trakcie jednej sesji. Pomimo że analiza tego typu jest dosyć powierzchowna, to jednak wykorzystuje się ją do usprawnienia działania systemu, poprawy jego bezpieczeństwa, modyfikacji wyglądu strony czy wsparcia decyzji marketingowych.
- ◆ Reguły asocjacyjne (*association rules*) – sprawdza się tutaj, które strony są odwiedzane podczas jednej sesji, ustalając wcześniej wartość wsparcia dla reguły. Strony te nie muszą być powiązane za pomocą odnośników (hiperłączy), zaś wyniki analizy mogą być użyte do zmiany struktury witryny (zob. str. 74).
- ◆ Grupowanie (*clustering*) – w przypadku eksploracji stron internetowych istnieją dwa rodzaje grupowania związane osobno z użytkownikami i osobno z przeglądаныmi stronami. W pierwszym wypadku dąży się do utworzenia skupisk użytkowników o podobnych wzorcach zachowań. Po włączeniu do analizy zmiennych demograficznych można przeprowadzić segmentację na potrzeby handlu elektronicznego lub personalizować zawartość stron przeglądanych przez użytkowników z poszczególnych skupisk. Z drugiej strony, grupowanie stron pozwala odkryć skupiska mające powiązaną zawartość. Informacja ta może być następnie użyta do dynamicznego przedstawienia internautom odpowiednich hiperłączy odnoszących się do ich zapytań lub historii poszukiwanych przez nich danych.
- ◆ Klasyfikacja/dyskryminacja (*classification*) – polega na przyporządkowaniu obserwacji do zdefiniowanych wcześniej klas i na znalezieniu profilu internautów należących do każdej z nich. Po dokonaniu selekcji zmiennych niezależnych należy wybrać któreś z narzędzi do budowy modeli wzorcowych (ukierunkowany *data mining*), np. drzewa klasyfikacyjne, naiwne klasyfikatory Bayesa, metodę najbliższego sąsiedztwa, metodę wektorów nośnych itp. Przykładowy profil mógłby brzmieć następująco: 35% klientów z działu „książki historyczne” to osoby w wieku 46-55 lat mieszkające w miastach o liczbie mieszkańców przekraczającej 100 tys. osób.
- ◆ Wzorce sekwencji (*sequential patterns*) – badacz poszukuje wzorców odwiedzin strony, gdzie każda wizyta oznacza osobną sesję. Przykładowo: poszukuje się schematów/reguł zakupów realizowanych podczas kolejnych wizyt w sklepie internetowym.

⁵ J. Srivastava, R. Cooley, M. Deshpande, P-N. Tan, *op. cit.*, s. 16 i następane.



Specjaliści ds. marketingu są dzięki temu w stanie przewidzieć kolejne zakupy i oddziaływać na grupę docelową za pomocą odpowiednio przygotowanego komunikatu reklamowego.

- ◆ Modelowanie zależności (*dependency modeling*) – ma na celu poszukiwanie związków między zmiennymi w obrębie witryny. Przykładowy model może zawierać zmienne niezależne odnoszące się do działań, jakie użytkownik podejmuje podczas wizyty w sklepie internetowym oraz zmienną zależną odnoszącą się do kategorii/marki/ceny nabywanego produktu. Do analizy używa się ukrytych modeli Markova albo sieci bayesowskich⁶. Informacje uzyskane w ten sposób są pomocne w formułowaniu strategii sprzedażowych lub usprawnieniu struktury witryny w celu łatwiejszej nawigacji przez użytkowników.

Analiza wzorców zachowań użytkowników

Wyróżnia się pięć głównych obszarów aplikacyjnych modeli *web mining*⁷:

1. personalizację,
2. usprawnienie systemu,
3. modyfikację witryny,
4. analitykę biznesową (*business intelligence*),⁸
5. charakterystykę używania strony.

W odniesieniu do handlu elektronicznego personalizacja przekazu to kluczowy obszar marketingu zindywidualizowanego. Dynamiczne tworzenie treści dla użytkowników w oparciu o ich profil i dotychczasową aktywność na stronie jest pomocne dla realizacji sprzedaży krzyżowej (*cross-selling*) i sprzedaży uzupełniającej (*up-selling*). Aplikacje personalizujące przekaz analizują aktywność internauty i wskazują na potencjalnie interesujące go odnośniki. Rekomendowane są strony odwiedzane przez użytkowników należących do tego samego segmentu.

Usprawnienie systemu opiera się na znajomości ruchu internautów w obrębie witryny. Bywa wykorzystywane w poprawie optymalizacji dostępu do strony (*web caching*), transmisji i dystrybucji danych oraz równoważenia obciążenia połączeń sieciowych (*load balancing*). Innym obszarem jest wykrywanie osób, które bez uprawnienia próbują połączyć się z siecią (intruzów), wykrywanie oszustw oraz prób włamań do systemu.

⁶ Sieci bayesowskie (*Bayesian Belief Networks*) bywają również nazywane sieciami przekonań Bayesa lub sieciami przekonań (tłumaczenie za D.T. Larose, *Metody i modele eksploracji danych*, PWN, Warszawa 2008, s. 241).

⁷ J. Srivastava, R. Cooley, M. Deshpande, P-N. Tan, *op. cit.*, s. 17 i następne.

⁸ Termin *business intelligence* bywa również tłumaczony jako „wywiad gospodarczy”, „biały wywiad” lub „inteligencja biznesowa”. W kontekście budowy modeli *data mining* lepiej pozostać przy podanym tu przekładzie – „analityka biznesowa”.



Modyfikacja strony ma na celu wzrost atrakcyjności witryny zarówno pod względem zawartości, jak i struktury. Jest stosowana np. przy tworzeniu katalogów w handlu elektronicznym. Zmiana projektu strony odbywa się automatycznie w oparciu o znajomość wzorców zachowań internautów rozpoznanych dzięki rejestrowi zdarzeń (plikom dziennika). Do łączenia stron wykorzystuje się omówione wcześniej grupowanie.

Analityka gospodarcza dostarcza informacji o tym, w jaki sposób klienci korzystają z witryny. Wiedza ta jest wykorzystywana przez specjalistów ds. marketingu w trzech obszarach: pozyskiwania klientów, sprzedaży krzyżowej i analizie migracji klientów (w tym w badaniu lojalności). Specjalistyczne oprogramowanie analizuje dane o sprzedaży, oblicza wskaźniki CTR⁹ (liczba kliknięć w link do liczby wyświetleń reklamy) i dostarcza prostych statystyk dotyczących odwiedzin witryny.

Charakterystyka użytkowania strony (*usage characterization*) to wprawdzie obszar, w którym nie wykorzystuje się narzędzi *data mining*, jednak w dużym stopniu pokrywa się on z eksploracją zachowań internautów (*web usage mining*). Analizuje się tutaj szczegółowo obsługę przeglądarki przez użytkowników oraz strategię poruszania się po poszczególnych witrynach. Programy przeznaczone do tego typu zadań rejestrują przykładowo użycie przycisków „w przód/w tył”, zapisywanie plików przez internautę lub dodawanie adresu strony do zakładek (do folderu „Ulubione”).

Reguły sekwencyjne

Analiza reguł sekwencyjnych może odnosić się do sekwencji odwiedzin witryny w danym okresie czasu podczas kolejnych wizyt albo do sekwencji wykonanych operacji w trakcie jednej wizyty na stronie. W pierwszym wypadku sytuacja jest dosyć trudna, ponieważ konieczne jest posiadanie zmiennej identyfikującej internautę. Jest to możliwe wówczas, gdy użytkownik loguje się podczas każdej wizyty w sklepie (na stronie). Sama rejestracja zakupów bywa jednak niewystarczająca, bowiem nabywca przed dokonaniem transakcji dokonuje porównania ofert konkurencyjnych i zdarza się, że kilkakrotnie odwiedza witrynę sklepu przed zakupem, nie logując się na stronie. Znacznie łatwiej zidentyfikować użytkownika podczas jednej wizyty, ponieważ plik rejestru (log) ma zazwyczaj numer identyfikacyjny sesji. Przykładowe wyniki analiz zamieszczone poniżej dotyczą sekwencji zdarzeń mających miejsce podczas jednej sesji na stronie internetowej sklepu oferującego odzież ciążową.

Badacz ma możliwość sprawdzenia, z ilu elementów składają się najdłuższe reguły (tabela 1.) przy określonym z góry poziomie współczynnika *wsparcie* (*support*), tu: 1%, i współczynnika *zaufanie* (*confidence*), tu: 10%. Najdłuższe z nich mają po 6 elementów, przy czym częstość ich występowania wśród wszystkich sesji w badanym okresie nie przekracza 6,57%, co daje *de facto* ponad 1500 wzorców zachowań.

⁹ CTR – *click trough rate*.



Tabela. 1. Przykładowe najdłuższe sekwencje.

Min: wsparcie= 1,0%, zaufanie = 10,0%			
Popularne sekwencje	Liczba zestawów	Liczność	Wsparcie%
(Spodnie), (Spodnie), (Spodnie), (Spodnie), (Spodnie), (Spodnie)	6,000000	1578,000	6,567885
(Spodnice), (Spodnice), (Spodnice), (Spodnice), (Spodnice), (Spodnice)	6,000000	896,000	3,729293
(Promocje), (Promocje), (Promocje), (Promocje), (Promocje), (Promocje)	6,000000	607,000	2,526430
(Bluzki), (Bluzki), (Bluzki), (Bluzki), (Bluzki), (Bluzki)	6,000000	526,000	2,189295
(Spodnie), (Spodnie), (Spodnie), (Spodnie), (Spodnie), (Spodnice)	5,000000	428,000	1,781403
(Spodnie), (Spodnie), (Spodnie), (Spodnie), (Bluzki)	5,000000	387,000	1,610755
(Spodnie), (Bluzki), (Promocje)	3,000000	362,000	1,506701

Innymi słowy, podczas 1578 wizyt w sklepie internetowym użytkownicy sześć razy wybrali spodnie (różne kroje/fasony/kolory/marki), podczas 896 wizyt sześć razy wybrali spódnice itd. Najczęściej występujące długie sekwencje zawierają jednorodne elementy (spódnice, spodnie, promocje, bluzki), co wskazuje na to, że klienci, którzy długo przeglądają zasoby witryny koncentrują uwagę na konkretnym rodzaju odzieży.

Współczynnik wsparcia (*support*) informuje o tym, jak często dana sekwencja pojawia się w zbiorze wszystkich reguł w danym okresie czasu. Przykładowe najczęściej występujące reguły zamieszczono w tabeli 2. W zbiorze wszystkich analizowanych wizyt na stronie 32,4% odnosiło się do dwukrotnego wyboru spodni¹⁰, 24,7% do dwukrotnego wyboru spódnic lub sukienek, 21,4% do trzykrotnego wyboru spodni itd.

Tabela. 2. Najczęściej występujące reguły sekwencyjne.

Min: wsparcie= 1,0%, zaufanie = 10,0%				
Maks. liczność zestawu = 10				
Poprzednik	==>	Następnik	Wsparcie%	Zaufanie(%)
(Spodnie)	==>	(Spodnie)	32,39824	64,98581
(Spodnice)	==>	(Spodnice)	24,67327	60,48362
(Spodnie), (Spodnie)	==>	(Spodnie)	21,35187	65,90442
(Spodnie)	==>	Spodnie), (Spodnie)	21,35187	42,82852
(Bluzki)	==>	(Bluzki)	19,72030	55,70186
(Promocje)	==>	(Promocje)	16,87339	57,98055

Druga popularna miara jakości reguły sekwencyjnej to współczynnik zaufania (*confidence*), który informuje o tym, jak często po elementach poprzednika występują elementy następnika. W tabeli 3 znajdują się przykładowe reguły o najwyższych wartościach zaufania. 67,76% użytkowników, którzy 3-krotnie wybrali kategorię „spodnie”, ponownie wybierze ten produkt, 67,58% użytkowników, którzy 5-krotnie wskazali kategorię „spodnie”, zrobi to po raz kolejny itd. Uogólniając, miara *confidence* mówi o tym, jakie jest prawdopodobieństwo pojawienia się określonego zdarzenia po znanej sekwencji innych zdarzeń.

¹⁰ Każda wizyta na stronie była identyfikowana za pomocą unikatowego numeru sesji. Dwukrotne „kliknięcie” w kategorię „spodnie” nie oznacza, że internauta wykonał tylko te dwa działania, ale że oba znalazły się w zbiorze wszystkich wykonanych przez niego działań podczas jednej sesji.

Tabela 3. „Najsilniejsze” reguły sekwencyjne.¹¹

Min: wsparcie= 1,0%, zaufanie = 10,0%				
Maks. liczność zestawu = 10				
Poprzednik	==>	Następnik	Wsparcie%	Zaufanie(%)
(Spodnie), (Spodnie), (Spodnie)	==>	(Spodnie)	14,46766	67,75828
(Spodnie), (Spodnie), (Spodnie), (Spodnie), (Spodnie)	==>	(Spodnie)	6,56788	67,58030
(Spodnie), (Spodnie), (Spodnie), (Spodnie)	==>	(Spodnie)	9,71864	67,17491
(Spodnie), (Spodnie)	==>	(Spodnie)	21,35187	65,90442
(Spodnie)	==>	(Spodnie)	32,39824	64,98581
(Spodnice), (Spodnice), (Spodnice), (Spodnice)	==>	(Spodnice)	6,18080	64,36931

W trakcie analizy badacz ma możliwość wyboru konkretnego elementu poprzednika lub następnika. W tabeli 4 znajdują się reguły, których poprzednikiem jest kategoria „bluzki”. Okazuje się, że najczęściej po wyborze tego produktu internauci ponownie wybierają inny model bluzki (56%), inne dwa modele bluzek (33%), inne trzy modele bluzek (19%), produkt z działu „promocje” (17%), spodnie (16%), spódnice (12%) lub inne cztery wzory bluzek (11%).

Tabela 4. Wybrane reguły sekwencyjne ze zdefiniowanym elementem poprzednika (tu: bluzki).

Min: wsparcie= 1,0%, zaufanie = 10,0%				
Maks. liczność zestawu = 10				
Poprzednik	==>	Następnik	Wsparcie%	Zaufanie(%)
(Bluzki)	==>	(Bluzki)	19,72030	55,70186
(Bluzki)	==>	(Bluzki), (Bluzki)	11,72480	33,11780
(Bluzki)	==>	(Bluzki), (Bluzki), (Bluzki)	6,85507	19,36280
(Bluzki)	==>	(Promocje)	5,86448	16,56478
(Bluzki)	==>	(Spodnie)	5,74378	16,22384
(Bluzki)	==>	(Spodnice)	4,33281	12,23842
(Bluzki)	==>	(Bluzki), (Bluzki), (Bluzki), (Bluzki)	3,85000	10,87468

Oprócz zmiennych odnoszących się do rodzaju wybieranej odzieży warto do zbioru obserwacji wprowadzić dodatkowe zmienne charakteryzujące poszczególne produkty, np. kolor odzieży, rozmieszczenie fotografii produktu na stronie, długość nogawki/rękawa, sposób wykończenia itp. Z danych zamieszczonych w tabeli 5 wynika, że klienci podczas poszukiwania produktów są zainteresowani konkretnym kolorem ubrania. Najpopularniejsze wzorce zawierają ten sam kolor w poprzedniku i następniku reguły, chociaż można zauważyć łączenie koloru niebieskiego (zapewne džinsów) z innymi kolorami, np. czarnym, brązowym.

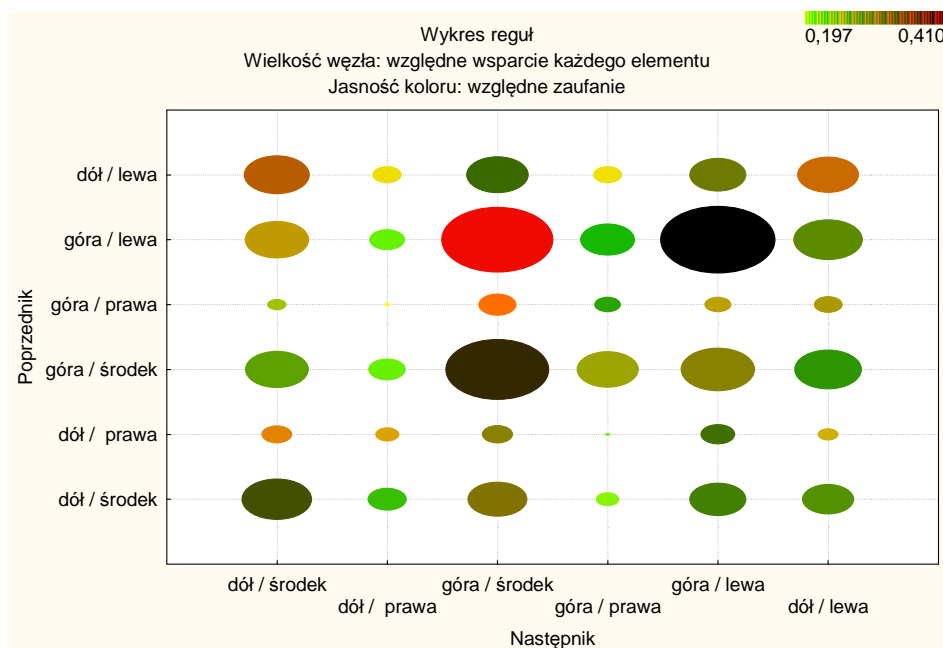
Z wykresu reguł (rys. 2) wynika ponadto, że internauci wybierają najczęściej produkty, których fotografia jest zlokalizowana u góry ekranu na środku lub u góry ekranu z prawej strony (wskazuje na to wielkość węzłów na rysunku).

¹¹ Ponieważ w literaturze przedmiotu współczynnik „confidence” bywa również określany mianem „strength”, więc synonimem pojęcia „reguła o dużym zaufaniu” może być termin „silna reguła”.

Tabela 5. Reguły sekwencyjne dla dodatkowych zmiennych opisujących produkty – kolor ubrania.

Min: wsparcie= 1,0%, zaufanie = 10,0%				
Maks. liczność zestawu = 10				
Poprzednik	==>	Następnik	Wsparcie%	Zaufanie(%)
(niebieski)	==>	(niebieski)	20,00333	48,36957
(czarny)	==>	(czarny)	16,54041	38,74050
(niebieski)	==>	(czarny)	12,82777	31,01852
(niebieski)	==>	(niebieski), (niebieski)	10,56772	25,55354
(niebieski), (niebieski)	==>	(niebieski)	10,56772	52,82980
(niebieski)	==>	(brązowy)	8,25356	19,95773
(brązowy)	==>	(czarny)	8,12453	29,84253
(czarny)	==>	(szary)	8,02048	18,78534
(czarny)	==>	(brązowy)	7,94140	18,60012

Co ważne, osoba, która „klika” zdjęcie w lewym górnym rogu, wybiera kolejny produkt położony w tym samym miejscu bądź u góry na środku ekranu (wskazuje na to kolor węzłów na rysunku¹²).



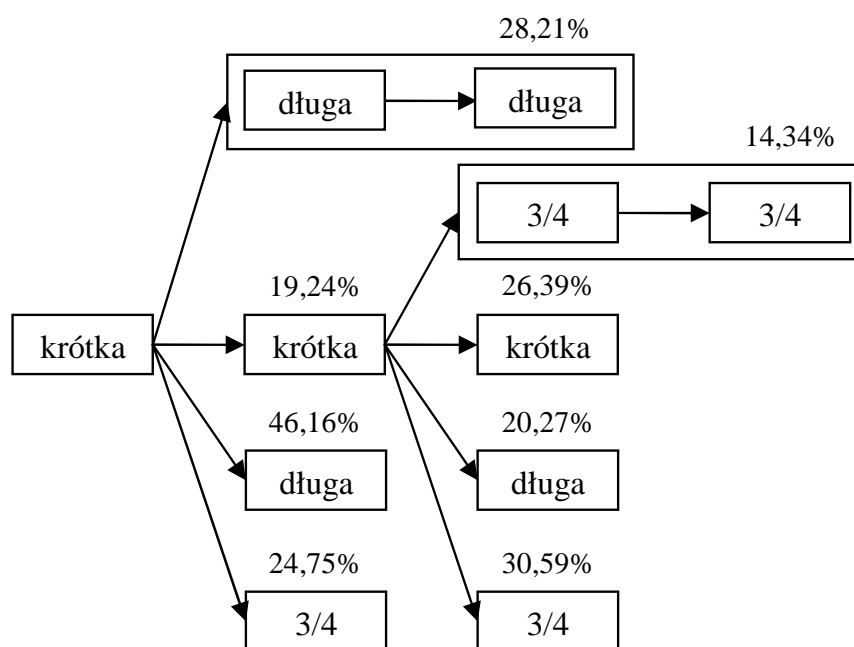
Rys. 2. Reguły sekwencyjne dla dodatkowych zmiennych opisujących produkty – położenie fotografii produktu na stronie.

Korzystając z reguł sekwencyjnych, można dodatkowo prześledzić ścieżkę kliknięć i narysować schemat, według którego internauci poszukują produktu w obrębie witryny. Za przykład posłużą reguły sekwencyjne, których elementami są długość nogawki spodni (tabela 6.).

¹² Oryginalny kolorowy wykres jest łatwiejszy w interpretacji niż widoczny na wydruku czarno-biały.

Tabela 6. Reguły sekwencyjne dla dodatkowych zmiennych opisujących produkty – długość nogawki spodni.

Min: wsparcie= 0,5%, zaufanie = 10,0%				
Maks. liczność zestawu = 10				
Poprzednik	==>	Następnik	Wsparcie%	Zaufanie(%)
(krótka)	==>	(długa)	9,593334	46,15668
(krótka)	==>	(długa), (długa)	5,863018	28,20890
(krótka)	==>	(trzy-czwarte)	5,144473	24,75175
(krótka)	==>	(krótka)	3,997860	19,23501
(krótka)	==>	(długa), (długa), (długa)	3,088213	14,85840
(krótka)	==>	(trzy-czwarte), (trzy-czwarte)	2,354380	11,32769
(krótka), (krótka)	==>	(trzy-czwarte)	1,223055	30,59273
(krótka), (krótka)	==>	(krótka)	1,054885	26,38623
(krótka), (krótka)	==>	(długa)	0,810274	20,26769
(krótka), (krótka)	==>	(trzy-czwarte), (trzy-czwarte)	0,573307	14,34034



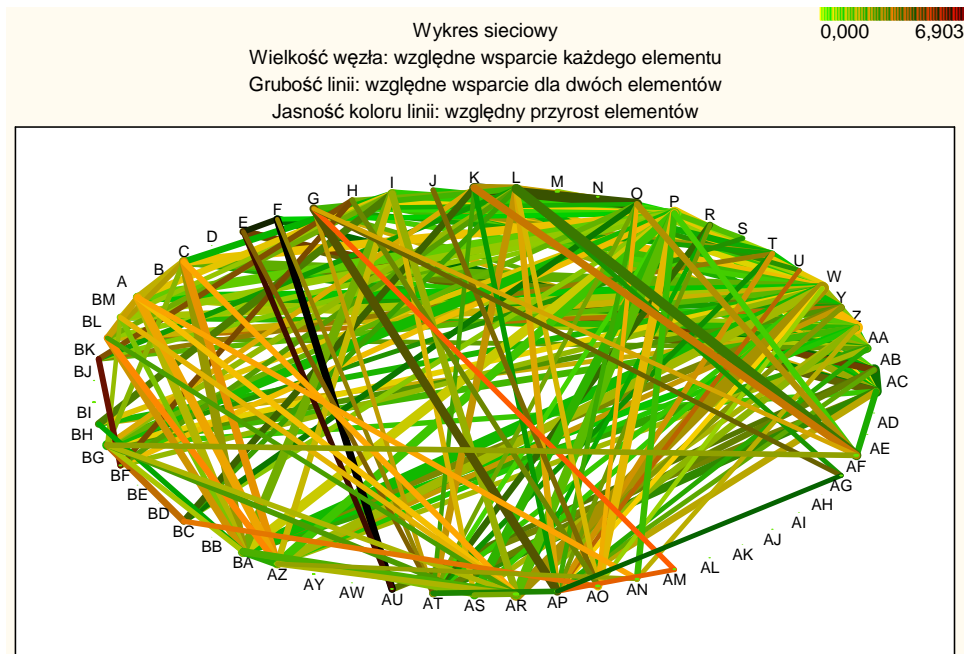
Rys. 3. Długość nogawki spodni – fragment ścieżki „kliknięć”.

Na rysunku 3 znajdują się wartości współczynnika zaufania, które można interpretować następująco:

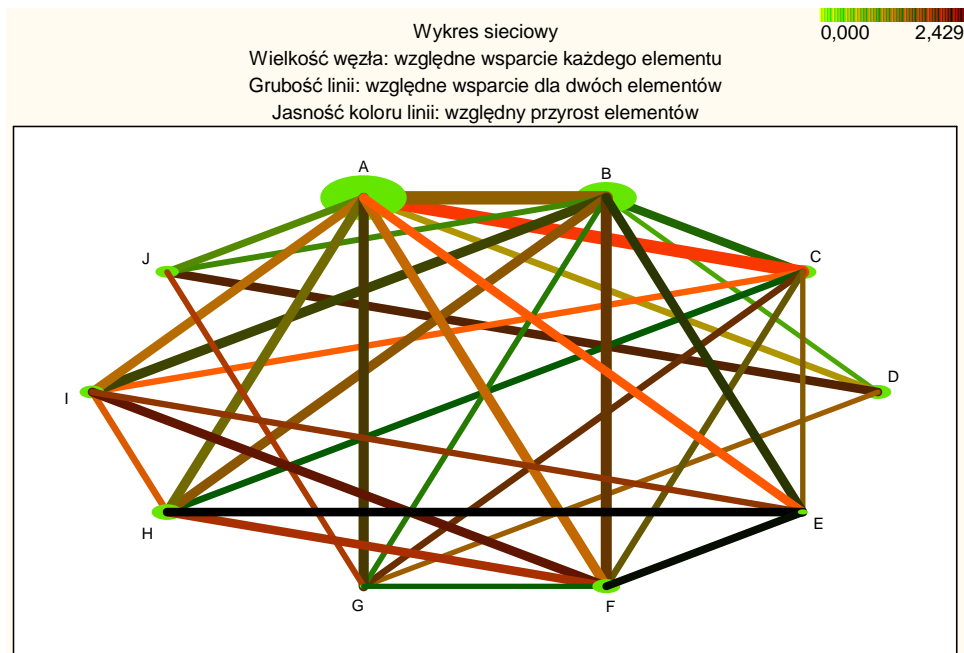
- ◆ 46,16% użytkowników, którzy wybrali krótkie spodnie, wybierze następnie długie spodnie (28,21% wybierze dwa wzory długich spodni, 24,75% wybierze spodnie z nogawką typu $\frac{3}{4}$, a 19,24% ponownie wybierze inny model krótkich spodni);
- ◆ 30,59% użytkowników, którzy wybrali dwa wzory krótkich spodni, wybierze spodnie z nogawką $\frac{3}{4}$ (26,39% po raz trzeci wybierze krótkie spodnie, 20,27% wybierze długie spodnie, a 14,34% wybierze dwa różne modele spodni z nogawką typu $\frac{3}{4}$).

Reguły asocjacyjne

W przypadku, gdy modele *web mining* dotyczą zakupów za pośrednictwem Internetu, można przeprowadzić analizę koszykową, wykorzystując w tym celu reguły asocjacyjne. Liczba wygenerowanych reguł jest w takich wypadkach dosyć duża – sięgająca nierzadko kilkudziesięciu tysięcy pozycji.



Rys. 4. Wykres sieciowy dla wszystkich wzorów odzieży.



Rys. 5. Wykres sieciowy dla wzorów odzieży o najwyższym współczynniku *support*.



Próba wizualizacji wszystkich reguł skojarzeniowych bywa kłopotliwa (rys. 4) i najlepiej ograniczyć wtedy liczbę produktów do tych o najwyższym współczynniku wsparcia (rys. 5.) albo sporządzać wykresy dla wybranych podzbiorów asortymentu¹³.

Drugie ograniczenie analizy koszykowej w przypadku sklepu internetowego z odzieżą odnosi się do opisu pozycji z oferty. Używanie nazw produktów nie jest skuteczne, ponieważ dotyczą one tylko bieżącej kolekcji, której wygląd i nazewnictwo w następnym sezonie będą całkiem inne. W tym wypadku konieczne jest wprowadzenie do zbioru obserwacji dodatkowych charakterystyk asortymentu.

Podsumowanie

Analiza zachowań internautów jest najbardziej atrakcyjnym, z punktu widzenia badań marketingowych, obszarem *web mining*. Eksploracja użytecznych reguł odnoszących się do rzeczywistych działań użytkowników podczas wizyty na stronie internetowej pozwala zwiększyć użyteczność i funkcjonalność witryn. Jeśli spojrzeć na analizę danych, to wykorzystywane tutaj narzędzia i metody są stosunkowo proste w użyciu, a wyniki łatwe do interpretacji. Bardzo duża liczba reguł powoduje, że niezbędna jest ścisła współpraca pomiędzy badaczem a menedżerem, co w literaturze przedmiotu jest określane terminem „proces filtrowania reguł przez ekspertów” (*expert filtering process*).

Literatura

1. Berry M.J.A., Linoff G.S., *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management*, Second Edition, John Wiley & Sons, 2004.
2. Cooley R., Mobasher B., Srivastava J., *Web Mining: Information and Pattern Discovery on the World Wide Web*, Proceedings of the 9th International Conference on Tools with Artificial Intelligence, IEEE Computer Society, 1997, s. 558-567.
3. Giudici P., *Applied Data Mining. Statistical Methods for Business and Industry*, John Wiley & Sons, 2003.
4. Liu B., *Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data*, Springer Verlag, Berlin Heidelberg 2007.
5. Mobasher B., *Web Usage Mining, w: Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data*, red. B. Liu, Springer Verlag, Berlin Heidelberg 2007, s. 449-483.
6. Srivastava J., Cooley R., Deshpande M., Tan P-N., *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*, SIGKDD Explorations, Volume 1, Issue 2, January 2000, s. 12-23.

¹³ Ze względu na poufność danych, nazwy produktów zastąpiono dużymi literami: A, B, C itd.