



## PRZYKŁAD BADANIA WZORCÓW ZACHOWAŃ KLIENTÓW ZA POMOCĄ ANALIZY KOSZYKOWEJ

*Agnieszka Pasztyła, StatSoft Polska Sp. z o.o.;  
Akademia Ekonomiczna w Krakowie, Katedra Statystyki*

### Cel analizy koszykowej

Analiza koszykowa służy do znajdowania w dużym zestawie danych ukrytych zależności w postaci prostych reguł. Pierwotnym zastosowaniem analizy koszykowej była analiza danych transakcyjnych pochodzących z supermarketów. Problemem, który postawili kierownicy hal przed analitykami, było znalezienie prawidłowości, które z dużym prawdopodobieństwem opisują zależności między kupowanymi produktami. Taka wiedza pozwoliłaby pracownikom m.in. tak rozmieścić produkty w sklepie, aby uzyskać największe wyniki sprzedaży, lub zaplanować promocje, nie zmniejszając w sposób nieprzewidziany potencjalnego zysku. W odpowiedzi do analizy danych historycznych zgromadzonych w bazach transakcyjnych zastosowano reguły asocjacyjne wsparte odpowiednio szybkim algorytmem przeszukiwania bazy. W efekcie analiza koszykowa pozwoliła odpowiedzieć na pytania o następującej konstrukcji:

- ◆ jakie produkty kupowane są najczęściej razem,
- ◆ jakie jest prawdopodobieństwo, że klienci, którzy kupili produkt A, kupią również produkt B,
- ◆ co kupują klienci, którzy uczestniczą w programach lojalnościowych,
- ◆ co wybrali klienci, którzy skorzystali z promocji?

Część z otrzymanych wyników potwierdza zwykle zależności, które znają pracownicy, jednak celem analizy koszykowej jest znalezienie „ukrytych” reguł, które nie są oczywiste i wzbogacają wiedzę specjalistów. Choć niewątpliwie potwierdzenie na podstawie danych historycznych hipotez, dotyczących prawidłowości obserwowanych w danej branży i formułowanych przez pracowników na bazie doświadczenia, stanowi również istotną wartość.

Analiza koszykowa ma oczywiście szersze zastosowanie niż badanie koszyków klientów hipermarketów. Mianowicie gdy przestajemy ograniczać się do postrzegania produktów w sensie fizycznym, charakterystycznego dla sklepów, np. kawy, ciastek, cukru, okazuje się, że nie ma ograniczeń co do przedmiotu analizy koszykowej. W szczególności możemy rozszerzyć obszar zainteresowań badawczych o usługi.



Przykładem może być np.:

- ♦ analiza usług pod kątem zastosowania metod zwiększania sprzedaży (*up-selling*, *cross-selling*),
- ♦ optymalizacja pakietów usług i opłat lub taryf w sektorze finansowym, telekomunikacyjnym i innych,
- ♦ planowanie kampanii promocyjnych na podstawie uzyskanych wyników,
- ♦ weryfikacja efektywności i skuteczności kampanii marketingowych poprzez porównanie wyników analiz z kilku okresów.

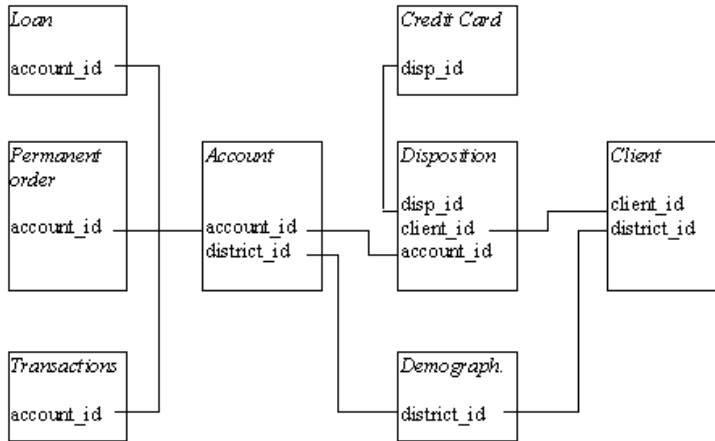
Rozszerzając obszar zastosowań reguł asocjacyjnych należy również zwrócić uwagę na dane transakcyjne dotyczące działań klientów nie tylko w aspekcie zysków lub strat z ewentualnego zakupu. Dane, w szczególności takie, które pozwalają na śledzenie zachowań w czasie, pozwalają na przeprowadzenie analiz o większym zakresie merytorycznym. Przykładem może być wykrywanie wszelkich nieprawidłowości. Analiza koszykowa i sekwencji umożliwia poznawanie np. mechanizmów nadużyć w bankowości, gdzie z danych transakcyjnych wyodrębniane są reguły opisujące pranie brudnych pieniędzy. Znalezione reguły opisują w sposób ilościowy kolejne etapy mechanizmu i mogą być wykorzystane w celu zapobiegania podobnym zjawiskom. Analiza koszykowa, sekwencji i połączeń może stanowić również cenne rozszerzenie narzędzi raportujących, wzbogacając standardowy zakres wyników o nowe spojrzenie na dane historyczne. Ogólnie rzecz biorąc, wyniki uzyskane w postaci reguł asocjacyjnych, czyli wyodrębnionych prawidłowości i odpowiadających im prawdopodobieństw pomagają w poznaniu prawidłowości działań klientów i zdobyciu przewagi konkurencyjnej przedsiębiorstwa. Mogą również wpłynąć na poziom zabezpieczeń przed nadużyciami.

## Dane wykorzystane w analizie

W artykule zostanie wykorzystana baza danych zawierająca dane demograficzne oraz szczegóły transakcji bankowych przeprowadzonych przez klientów jednego z czeskich banków detalicznych w latach 1996–1999. Baza została przygotowana i udostępniona przez organizatorów konferencji PKDD 1999<sup>1</sup>, która odbywała się w Pradze. Zawiera ona dwa miliony transakcji, takich jak: przelewy, zlecenia stałe, płatności kartami, a także informacje o tym, kiedy zostało założone konto, cechy właściciela konta: data urodzenia, płeć, miejsce zamieszkania, czy właściciel zaciągnął kredyt i czy ma problemy z jego spłaceniem, czy ma kartę kredytową i od kiedy itp.

---

<sup>1</sup> Szczegółowy opis tabel i struktury bazy znajduje się pod adresem: <http://lisp.vse.cz/pkdd99/>.



Rys. 1. Struktura bazy klientów banku i ich transakcji (źródło: strona internetowa konferencji PKDD 1999)

Na podstawie tego typu danych możemy uzyskać bardzo wiele cennych informacji o klientach banku. Typowym przykładem może być segmentacja klientów, której celem byłoby wyodrębnienie i scharakteryzowanie jednorodnych grup klientów, np. ze względu na rodzaj i kwoty najczęściej przeprowadzanych transakcji, wybrane produkty bankowe, cechy demograficzne itp. Celem niniejszego artykułu jest pokazanie możliwości pozyskiwania wiedzy o klientach za pomocą analizy koszykowej oraz wyszukiwania wzorców zachowań z wykorzystaniem analizy koszykowej oraz sekwencji. Innymi słowy, chcemy opisać klientów banku za pomocą prawidłowości (reguł), które dotyczą przeprowadzanych operacji za pomocą rachunku bankowego oraz ich cech demograficznych. W zależności od tego, jakiego typu dane wykorzystamy do analizy – chodzi tu głównie o dane zagregowane i surowe oraz jakościowe i ilościowe – efektem analizy koszykowej może być zarówno szczegółowa charakterystyka klientów, będąca przepustką do efektywnych działań marketingowych zorientowanych na klienta (CRM), jak i wyszukiwanie różnego typu nadużyć (ang. *fraud detection*).

Warunkiem przeprowadzenia wymienionych analiz jest połączenie danych transakcyjnych i danych demograficznych, często jednak wykorzystywane są również dodatkowe źródła danych. Jak się okazuje, nie jest to zadanie proste. W analizowanej bazie, dane transakcyjne gromadzone są w tabelach *Transactions*, *Permanent orders* i *Loans*, które przypisane są do rachunków (tabela *Account*). Natomiast dane o klientach banku znajdują się w tabelach *Client* i *Demographic\_data*. Przy pierwszym zetknięciu się z bazą można odnieść wrażenie, że osiǳ jest tabela *Account* zawierająca informacje o rachunkach klientów. Takie podejście jest zgodne z podejściem pracowników banku, dla których, zarówno z punktu widzenia rachunkowości, jak i realizowanych operacji, jest to podstawowe spojrzenie na dane. Po szczegółowym zapoznaniu się ze strukturą bazy okazuje się, że równorzędną rolę odgrywa tabela *Disposition*, która łączy identyfikator rachunku z identyfikatorem klienta banku i w dodatkowym polu informuje jednocześnie o tym, czy

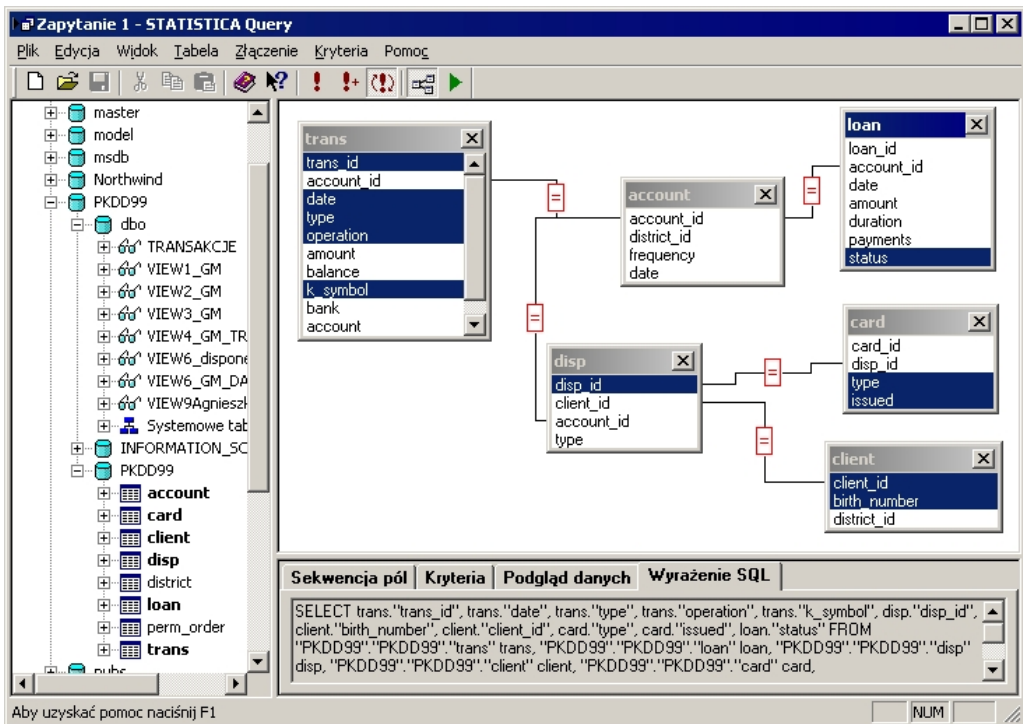


jest on właścicielem konta czy użytkownikiem, i w ten sposób definiuje uprawnienia klienta. Tabela ta jest niezwykle ważna, gdy realizowane transakcje chcemy połączyć z konkretną osobą, którą możemy opisać przez: wiek, płeć, wysokość dochodów itp. Natomiast nie byłaby istotna, gdyby interesowały nas wyłącznie podsumowania dotyczące operacji bankowych wybranego typu. Warto zauważyć, że baza ta jest tak zaprojektowana, aby szybko i efektywnie wykonywać zapytania właśnie tego ostatniego rodzaju. Dużo trudniej natomiast uzyskać dane będące podstawą analiz, w których wykorzystywane są zarówno dane transakcyjne oraz demograficzne.

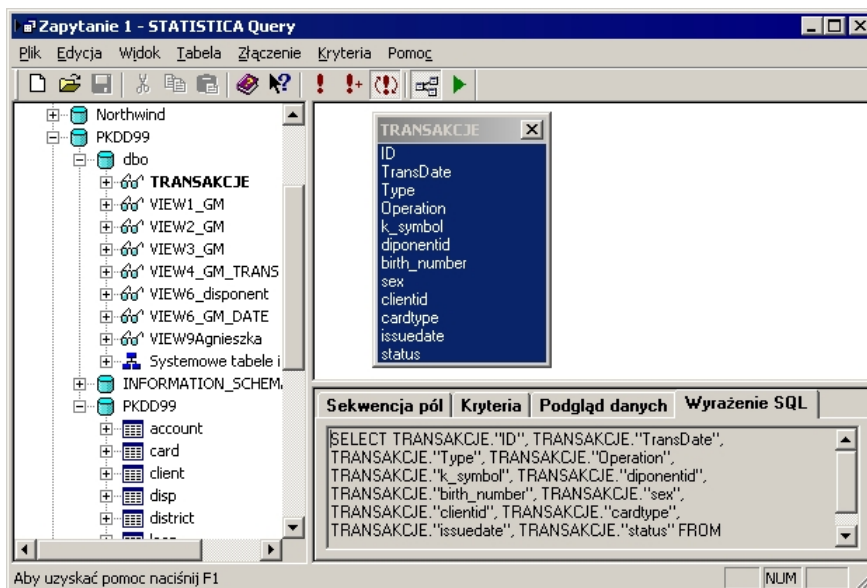
Na potrzeby niniejszego artykułu wykorzystano następujące tabele i pola:

- ♦ tabela *Transactions*: id\_transakcji, data, typ\_transakcji (wpłata/wypłata), operacja (np. wypłata kartą, przelew na konto; 5 wartości), k\_symbol (szczegółowy opis transakcji – 7 wartości);
- ♦ tabela *Client*: id\_klienta, birth\_number (zakodowana data urodzin i płeć);
- ♦ tabela *Card*: typ\_karty (junior, classic, gold), data\_wydania;
- ♦ tabela *Loan*: status (4 wartości określające, czy kredyt jest lub był spłacany w terminie).

Korzystając ze *STATISTICA Query*, narzędzia do tworzenia zapytań w trybie graficznym lub tekstowym, utworzono perspektywę na podstawie wybranych pól (rys. 2 i 3).



Rys. 2. Środowisko *STATISTICA Query* - budowa perspektywy



Rys. 3. Perspektywa z wybranymi polami z bazy danych

W ten sposób powstał plik płaski (rys. 4), który mógł być przeniesiony bezpośrednio do arkusza za pomocą prostego zapytania. W trakcie przygotowywania danych przyjęto ograniczenie, że analizowani będą klienci, którzy mają uprawnienia do zaciągania kredytu i podejmowania decyzji o wydaniu karty płatniczej. Innymi słowy, wybrano takie transakcje, dla których `id_dysponenta` było równe `id_klienta`. Okazało się jednak, że po wykonaniu zapytania ok. 20% transakcji zostało dwa razy umieszczonych w arkuszu. Były to transakcje realizowane z rachunków, gdzie oboje małżonkowie mieli równe uprawnienia w prowadzeniu konta. Ponieważ nie dysponujemy informacją o tym, kto daną operację wykonał, transakcje były uwzględniane dwa razy, dla każdego dysponenta. Tak przygotowany zbiór prowadził do błędów w analizie. Zgodnie ze strukturą bazy, tylko jedna osoba mogła być właścicielem konta (pole *Disponent.type*), stąd w następnym zapytaniu uwzględniono tylko właścicieli konta.

1	2	3	4	5	6	7	8	9	10
ID	TransDate	Type	Operation	k_symbol	disponent_id	birth_number	sex	client_id	cardtype
188 102	11/07/98	CREDIT	COLLECTION FROM ANOTHER BAN	PENSION	763	8/24/27	F	763	
188 103	12/07/98	CREDIT	COLLECTION FROM ANOTHER BAN	PENSION	763	8/24/27	F	763	
188 112	09/09/97	WITHDRAW.	REMITTANCE TO ANOTHER BANK	HOUSEHOLD	763	8/24/27	F	763	
188 113	10/09/97	WITHDRAW.	REMITTANCE TO ANOTHER BANK	HOUSEHOLD	763	8/24/27	F	763	
188 114	11/09/97	WITHDRAW.	REMITTANCE TO ANOTHER BANK	HOUSEHOLD	763	8/24/27	F	763	
188 115	12/09/97	WITHDRAW.	REMITTANCE TO ANOTHER BANK	HOUSEHOLD	763	8/24/27	F	763	

Rys. 4. Dane wyjściowe otrzymane z zapytania do bazy – fragment arkusza

Arkusz z surowymi danymi uzyskanymi bezpośrednio z bazy zawierał dwanaście zmiennych definiujących transakcję, które zostały opisane powyżej, przy czym trzy zmienne były



identyfikatorami. W kolejnym etapie, korzystając z modułu transformacje danych programu *STATISTICA*, utworzono trzy dodatkowe zmienne:

- ♦ wiek klienta (w latach) w dniu realizacji transakcji,
- ♦ informacja, czy klient korzystał lub korzysta z kredytu (kredyt\_tak/kredyt\_nie),
- ♦ informacja, czy klient korzysta z karty kredytowej (karta\_tak/karta\_nie).

Pierwsza zmienna (wiek klienta przy realizacji transakcji) rozszerza zakres dostępnych informacji o klientach banku o istotną charakterystykę, natomiast pozostałe dwie dodane zmienne agregują rozproszoną informację. W pierwszym kroku dalszej analizy będzie dla nas istotne, czym charakteryzują się klienci, którzy korzystają z kart płatniczych i kredytów, a dopiero w następnych etapach będziemy zawiązać analizę do wybranych grup klientów, np. posiadaczy kart typu *gold*.

Przed rozpoczęciem omawiania wyników analizy warto przypomnieć podstawowe pojęcia związane z analizą koszykową.

## Reguły asocjacyjne – ujęcie teoretyczne

Wynikiem analizy koszykowej są reguły asocjacji postaci: JEŻELI [poprzednik] TO [następnik] zapisywane za pomocą warunków: [warunki poprzednika] => [warunki następnika]. Na przykład: [kawa, śmietanka] => [ciastka]. Przykładowa reguła dotyczy osób, które kupując kawę i śmietankę, kupią również ciastka. Przecinki stosowane w zapisie warunków *poprzednika (body)* lub *następnika (head)* odpowiadają spójnikowi „i”. Jeżeli transakcja (rekord), czyli pojedynczy przypadek ze zbioru danych, „pasuje” do reguły, czyli spełnia wszystkie warunki poprzednika i następnika, mówimy, że reguła zawiera tę transakcję, lub że *transakcja wspiera regułę asocjacji*. Zazwyczaj pożądanym jest znalezienie tych reguł, w których *następnik* wyraża się jednym warunkiem (lub niewielką ich liczbą). W literaturze anglojęzycznej stosowany jest termin *itemset* lub *k-itemset* (rzadziej *item set*) – nie oznacza on zbioru przypadków w próbie uczącej, lecz zbiór elementów reguły lub transakcji. Przykładowo: zestaw {kawa, śmietanka} to *2-itemset* (zestaw dwuelementowy) i w tym przypadku opisuje on zawartość koszyka zakupowego.

Wśród reguł asocjacyjnych wyróżniamy takie, które oparte są na zmiennych jakościowych lub ilościowych, które dotyczą jednego lub więcej wymiarów (cech, atrybutów) danych, oraz reguły, które dotyczą jednego lub więcej poziomów agregacji zmiennych (chodzi tu na przykład o kategorie produktów, branże).

### Reguły oparte na zmiennych jakościowych i ilościowych

W zależności od postaci danych wejściowych możemy mieć do czynienia z różnymi zmiennymi odzwierciedlającymi te same dane. Najczęściej dla pojedynczej transakcji dysponujemy zbiorem wartości opisujących tę transakcję. Możemy również zawartość koszyka przedstawić za pomocą tabeli, w której kolejne zmienne oznaczają towary, które można nabyć. Zmienne te będą miały charakter binarny, na przykład osoba kupiła chleb



(1 lub tak) lub nie kupiła (0 lub nie). Tego typu zmienne określamy mianem jakościowych. Oczywiście mogą one przyjmować więcej niż dwie wartości – na przykład dla wielkości opakowania lub wersji produktu.

Zmienne jakościowe mają w analizie koszykowej dwa zastosowania. Przede wszystkim mogą dotyczyć cech (atrybutów) niemierzalnych, które chcemy uwzględnić w regułach, np. płci klienta, wykształcenia, nazwy nabytego produktu, koloru itp. Zmienne jakościowe mogą też być utworzone sztucznie, w celu zakodowania informacji o tym, czy klient nabył dany produkt (zmienna binarna), w sytuacji, gdy nie są nam potrzebne ilości kupionych towarów.

Zmienne ilościowe opisują ilość lub wartość towaru (usługi) i są wykorzystywane wtedy, gdy chcemy w sposób bardziej precyzyjny sformułować zależności dotyczące atrybutów.

Oczywiście wyszukiwanie reguł dla wszystkich realizacji zmiennej ilościowej jest nieefektywne. W szczególności gdy zmienna ma duży obszar zmienności (np. masa produktu mierzona z dokładnością do trzech miejsc po przecinku), możemy mieć do czynienia z bardzo dużą liczbą jej realizacji i w efekcie z przytłaczającą liczbą potencjalnych reguł asocjacyjnych. Dlatego obszar zmienności zmiennej dzielony jest na przedziały i w oparciu o nie tworzone są reguły asocjacyjne. W celu utworzenia optymalnych przedziałów wartości zmiennych ilościowych, potrzebnych do uzyskania reguł asocjacyjnych, które najlepiej będą odzwierciedlać zależności między kupowanymi produktami i czynnikami zewnętrznymi, stosowane są metody klasyfikacyjne, takie jak analiza skupień i inne. Klasyfikacja koszykowa, zapoczątkowana w latach dziewięćdziesiątych m.in. przez Agrawala, Srikanta i Imielińskiego [1], dotyczy jedynie zmiennych jakościowych.

### ***Liczba wymiarów uwzględniana w regułach asocjacyjnych***

Przez wymiar reguły asocjacyjnej rozumiemy liczbę uwzględnionych zmiennych (atrybutów). Reguła, która dotyczy jednej cechy jest określana jako jednowymiarowa. Przykładem reguły jednowymiarowej może być reguła postaci: [komputer, drukarka] => [oprogramowanie]. Przykładowa reguła opisuje prawidłowość, że klienci, którzy kupują komputer i drukarkę, z obliczonym prawdopodobieństwem kupują również oprogramowanie. Zarówno warunki poprzednika, jak i następnika, dotyczą tego samego atrybutu, czyli tego, co zawiera transakcja, lub inaczej, co kupił klient.

Warunki reguły wielowymiarowej uwzględniają więcej niż jeden atrybut. Reguła wielowymiarowa może określać zależności między na przykład: wiekiem klienta, godziną zakupów i tym, co klient nabył: [wiek\_klienta  $\in$  <14; 19>, godzina\_zakupów  $\in$  <10; 15>] => [produkt = napoje, słodycze]. Na podstawie powyższej reguły dowiadujemy się, że klienci w wieku od 14 do 19 lat, którzy przychodzą do sklepu w godzinach od 10 do 15, z określonym prawdopodobieństwem kupują napoje i słodycze.

### ***Reguły dotyczące danych zagregowanych i surowych***

Jak już wspomniano, analiza koszykowa może być prowadzona na dowolnym poziomie ogólności. To znaczy, że reguły asocjacyjne mogą dotyczyć zależności między



kupowanymi produktami, kategoriami produktów lub między branżami. Ogólnie w celu wyszukiwania prawidłowości w zbiorze danych powinniśmy szukać najpierw silnych reguł, czyli takich prawidłowości, które występują z dużym prawdopodobieństwem w odniesieniu do danych zagregowanych. Przykładowo w pierwszej kolejności badamy, jak często kupowane są razem pieczywo i nabiał, i jeśli okaże się, że jest to silna reguła, to warto zbadać szczegółowe zależności w tych kategoriach produktów, na przykład pomiędzy niskotłuszczowym mlekiem i jogurtami a ciemnym pieczywem itp.

### **Miary jakości uzyskanych reguł**

Wybierając postać danych wejściowych musimy pamiętać, że potencjalnych reguł jest bardzo dużo. Na przykład w przypadku trzech zmiennych dychotomicznych (odpowiedzi typu TAK-NIE na trzy pytania) możemy otrzymać maksymalnie 650 reguł. Jest to liczba wariacji bez powtórzeń dla trzech zmiennych i dwóch możliwych wartości i nie ma w niej bezużytecznych powtórzeń typu [ODP1=NIE => ODP1=NIE], czyli na przykład jeśli klient nie kupił produktu A, to nie kupił również produktu B. Dla supermarketu, gdzie mamy do czynienia z tysiącami produktów w ofercie, trudno wyrazić słowami liczbę możliwych reguł, ponieważ wzrost tej liczby jest typu  $n!$  Interesujące są oczywiście tylko te reguły, które występują często w danych historycznych (próbie uczącej), czyli opisują często występujące zachowania, a nie są jedynie pustymi sformułowaniami. Dlatego, aby wyodrębnić te reguły, które niosą dla nas istotną informację, wykorzystujemy trzy parametry służące do oceny „ważności” reguł. Są to:

- ♦ wsparcie reguły (*support*) – odsetek transakcji w danych historycznych, które zawierają wybraną regułę, jest to prawdopodobieństwo kupienia danego produktu przez losowo wybranego klienta,
- ♦ pewność reguły (*confidence*) – odsetek transakcji zawierających analizowaną regułę w zbiorze tych, które zawierają poprzednik (dla reguły  $A \Rightarrow B$  odpowiada to prawdopodobieństwu warunkowemu  $P(B|A)$ ), jest to prawdopodobieństwo, że losowo wybrany klient, który nabył produkt A, kupi również produkt B,
- ♦ korelacja (*correlation*) – jest to miara, która określa kierunek skorelowania produktów, inaczej – określa, w jaki sposób fakt, że klient wybrał produkt A, zwiększa lub zmniejsza prawdopodobieństwo, że wybierze on również produkt B,
- ♦ przyrost (*lift*) – jest to modyfikacja korelacji reguły; informuje o tym, jaki jest wpływ produktu A na sprzedaż produktu B (lub występowanie zjawiska B).

### **Wykrywanie wzorców zachowań klientów – wyniki analizy**

Celem przedstawionej poniżej analizy koszykowej jest wyszukanie reguł, które z dużym prawdopodobieństwem charakteryzują klientów banku lub grupy klientów banku. W przykładzie wykorzystano 15 zmiennych opisujących transakcje i klienta, który ją zrealizował. Zmienne te zostały opisane w punkcie 2. Wśród nich znajduje się jedna zmienna ilościowa – wiek klienta w dniu zawarcia transakcji. Obliczenia zostały wykonane za pomocą

programu *Sequence, Association and Link Analysis* pakietu *STATISTICA*, który pozwala na przeprowadzenie analizy koszykowej i sekwencji z wykorzystaniem zmiennych jakościowych i ilościowych.

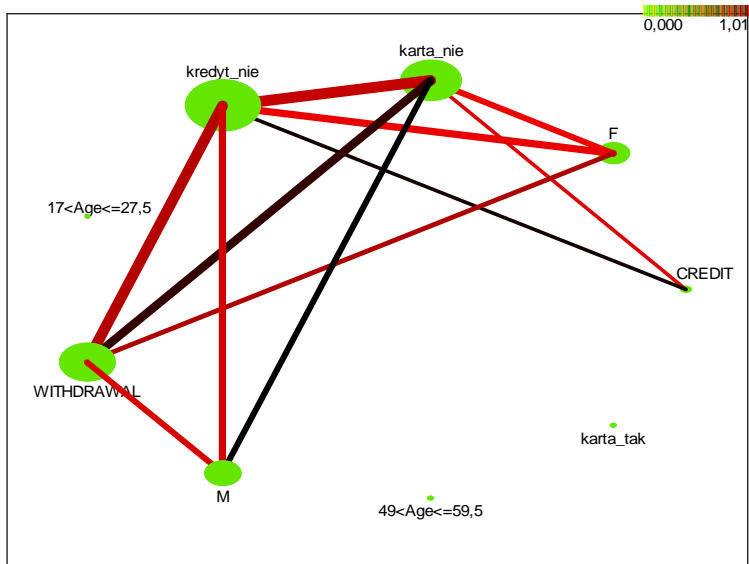
W pierwszym kroku wyodrębniono silne reguły, czyli takie, które informują o najczęściej występujących prawidłowościach i jednocześnie takie, które charakteryzują zbiorowość klientów w sposób ogólny, za pomocą zagregowanych cech. Jako kryteria wyszukiwania reguł przyjęto minimalne wsparcie reguły równe 20% i minimalną pewność reguły 10%. Oznacza to, że co najmniej 20% transakcji zostało wykonanych z rachunków klientów o wyróżnionych cechach (wsparcie). Rys. 5 przedstawia najsilniejsze reguły asocjacyjne, które zostały wyodrębnione przy tak zadanych ograniczeniach.

Summary of association rules (Transakcje_200_bezpowtorek)						
Min. support = 20,0%, Min. confidence = 10,0%						
Max. size of an itemset = 10						
	Body	==>	Head	Support(%)	Confidence(%)	Lift
79	kredyt_nie	==>	karta_nie	73,66	78,80	1,00
80	karta_nie	==>	kredyt_nie	73,66	93,86	1,00
89	WITHDRAWAL	==>	kredyt_nie	68,39	93,51	1,00
90	kredyt_nie	==>	WITHDRAWAL	68,39	73,17	1,00
71	WITHDRAWAL	==>	karta_nie	57,54	78,67	1,00
72	karta_nie	==>	WITHDRAWAL	57,54	73,32	1,00
73	WITHDRAWAL	==>	karta_nie, kredyt_nie	54,02	73,85	1,00
74	kredyt_nie	==>	karta_nie, WITHDRAWAL	54,02	57,79	1,00
75	kredyt_nie, WITHDRAWAL	==>	karta_nie	54,02	78,98	1,01
76	karta_nie	==>	kredyt_nie, WITHDRAWAL	54,02	68,83	1,01
77	karta_nie, WITHDRAWAL	==>	kredyt_nie	54,02	93,87	1,00
78	karta_nie, kredyt_nie	==>	WITHDRAWAL	54,02	73,33	1,00
81	M	==>	kredyt_nie	49,24	93,84	1,00
82	kredyt_nie	==>	M	49,24	52,68	1,00
33	kredyt_nie	==>	F	44,23	47,32	1,00
34	F	==>	kredyt_nie	44,23	93,07	1,00
43	M	==>	karta_nie	41,64	79,35	1,01
44	karta_nie	==>	M	41,64	53,05	1,01
51	M	==>	karta_nie, kredyt_nie	39,01	74,34	1,01
52	kredyt_nie	==>	karta_nie, M	39,01	41,73	1,00
53	kredyt_nie, M	==>	karta_nie	39,01	79,22	1,01
54	karta_nie	==>	kredyt_nie, M	39,01	49,70	1,01
55	karta_nie, M	==>	kredyt_nie	39,01	93,69	1,00
56	karta_nie, kredyt_nie	==>	M	39,01	52,96	1,01
91	M	==>	WITHDRAWAL	38,22	72,84	1,00
92	WITHDRAWAL	==>	M	38,22	52,26	1,00

Rys. 5. Wybrane reguły asocjacyjne wyodrębnione przy zadanym minimalnym wsparciu 20% i pewności 10%

Na podstawie uzyskanych wyników możemy stwierdzić, że ponad 73% transakcji jest realizowanych przez osoby, które dotychczas nie korzystały z kredytu i nie posiadają karty płatniczej. Przy czym jest większe prawdopodobieństwo, że jeśli transakcję zlecił klient, który nie ma karty, to nie brał on dotychczas kredytu (93,86%), niż odwrotnie (78,8%) (por. rys. 5). Najsilniejsze okazały się reguły, które dotyczą kart płatniczych i kredytów, a dokładnie niekorzystania z tych produktów, ponieważ w latach 1996–1999 w Czechach karty płatnicze i kredyty dla klientów indywidualnych wiązały się z wysokimi kosztami i nie cieszyły się dużym zainteresowaniem. Wyniki przedstawione w tabeli zostały zilustrowane na wykresie pajęczynowym (ang. *web graph*). Wielkość pola przy wartościach zmiennych

symbolizuje wsparcie, czyli jak wiele operacji bankowych charakteryzujących się daną własnością zostało wykonanych. Grubość linii łączącej dwa produkty oznacza wsparcie danej pary własności, natomiast intensywność koloru linii informuje o wartości korelacji.



Rys. 6. Reguły asocjacyjne przy zadanym minimalnym wsparciu 20% i pewności 10%

Zmniejszając wartości progowe wsparcia i pewności reguł oraz dodając do analizy zmienne, takie jak: rodzaj płatności, typ karty kredytowej, historię kredytową klienta, możemy otrzymywać prawidłowości, które coraz precyzyjniej będą charakteryzować klientów.

Celem drugiego etapu analizy było uzyskanie wspólnych cech klientów, którzy najczęściej płacą kartą. W tym celu wyodrębniono reguły, które w poprzedniku zawierały wartość „credit card withdrawal” (rys. 7).

Na podstawie otrzymanych wyników tę grupę klientów możemy scharakteryzować następująco:

- ◆ blisko 96% płatności kartami jest realizowanych przez osoby, które nie korzystały z kredytów (pewność reguły);
- ◆ najwięcej transakcji, ponad 68%, zostało zawartych za pomocą kart typu „classic”; drugą pod względem popularności jest karta „junior” (22,1%);
- ◆ kobiety nieznacznie częściej płacą kartą niż mężczyźni, odpowiednio 51,2% i 48,8% wszystkich transakcji za pomocą kart;
- ◆ grupa wiekowa, która najczęściej korzysta z kart płatniczych, to klienci w wieku od 49 do 60 lat (25,6% transakcji), od 27 do 38 lat (24,05% transakcji) i od 17 do 27 lat (23,6%).



Summary of association rules with selected items (Transakcje_200_bezpowtorek)						
Min. support = 0,1%, Min. confidence = 10,0%						
Max. size of an itemset = 10						
	Body	==>	Head	Support(%)	Confidence(%)	Lift
43	CREDIT CARD WITHDRAWAL	==>	kredyt_nie	1,03	95,98	1,03
85	CREDIT CARD WITHDRAWAL	==>	classic	0,74	68,88	4,40
33	CREDIT CARD WITHDRAWAL	==>	kredyt_nie, classic	0,69	65,04	4,63
15	CREDIT CARD WITHDRAWAL	==>	F	0,55	51,19	1,08
71	CREDIT CARD WITHDRAWAL	==>	M	0,52	48,81	0,93
20	CREDIT CARD WITHDRAWAL	==>	F, kredyt_nie	0,52	48,71	1,10
46	CREDIT CARD WITHDRAWAL	==>	kredyt_nie, M	0,51	47,26	0,96
5	CREDIT CARD WITHDRAWAL	==>	F, classic	0,39	36,03	4,51
14	CREDIT CARD WITHDRAWAL	==>	F, kredyt_nie, classic	0,36	33,74	4,81
67	CREDIT CARD WITHDRAWAL	==>	M, classic	0,35	32,85	4,30
41	CREDIT CARD WITHDRAWAL	==>	kredyt_nie, M, classic	0,33	31,31	4,45
79	CREDIT CARD WITHDRAWAL	==>	49<Age<=60	0,27	25,60	1,21
65	CREDIT CARD WITHDRAWAL	==>	27<Age<=38	0,26	24,10	1,27
47	CREDIT CARD WITHDRAWAL	==>	kredyt_nie, 27<Age<=38	0,26	24,05	1,39
59	CREDIT CARD WITHDRAWAL	==>	17<Age<=27	0,25	23,68	1,14
48	CREDIT CARD WITHDRAWAL	==>	kredyt_nie, 17<Age<=27	0,25	23,58	1,18
45	CREDIT CARD WITHDRAWAL	==>	kredyt_nie, 49<Age<=60	0,25	23,54	1,25
77	CREDIT CARD WITHDRAWAL	==>	49<Age<=60, classic	0,25	23,49	4,77
29	CREDIT CARD WITHDRAWAL	==>	kredyt_nie, junior	0,24	22,09	5,87
84	CREDIT CARD WITHDRAWAL	==>	junior	0,24	22,09	5,87
39	CREDIT CARD WITHDRAWAL	==>	kredyt_nie, 49<Age<=60, classic	0,23	21,43	5,15
63	CREDIT CARD WITHDRAWAL	==>	27<Age<=38, classic	0,22	20,59	5,11
42	CREDIT CARD WITHDRAWAL	==>	kredyt_nie, 27<Age<=38, classic	0,22	20,54	5,25

Rys. 7. Najsilniejsze reguły asocjacyjne dotyczące klientów korzystających z kart płatniczych wyodrębnione przy minimalnym wsparciu 0,1% i pewności 10%

Aby uzyskać pełny obraz klientów, którzy posiadają karty kredytowe, możemy sformułować problem inaczej i zapytać, co charakteryzuje wybraną grupę klientów, gdy weźmiemy pod uwagę wszystkie dokonywane przez nich transakcje bankowe. Aby uzyskać odpowiedź na tak postawione pytanie, należy wyszukać reguły, które w poprzedniku zawierają wartość „karta\_tak”. Wyniki zostały przedstawione na rys. 8.

Summary of association rules with selected items (Transakcje_200_bezpowtorek)						
Min. support = 5,0%, Min. confidence = 10,0%						
Max. size of an itemset = 10						
	Body	==>	Head	Support(%)	Confidence(%)	Lift
6	karta_tak	==>	kredyt_nie	19,81	92,06	0,98
11	karta_tak	==>	WITHDRAWAL	15,60	72,49	0,99
8	karta_tak	==>	kredyt_nie, WITHDRAWAL	14,38	66,81	0,98
13	karta_tak	==>	M	10,83	50,35	0,96
2	karta_tak	==>	F	10,69	49,65	1,04
7	karta_tak	==>	kredyt_nie, M	10,23	47,54	0,97
4	karta_tak	==>	F, kredyt_nie	9,58	44,52	1,01
3	karta_tak	==>	F, WITHDRAWAL	7,88	36,62	1,05
12	karta_tak	==>	WITHDRAWAL, M	7,72	35,87	0,94
9	karta_tak	==>	kredyt_nie, WITHDRAWAL, M	7,28	33,82	0,94
5	karta_tak	==>	F, kredyt_nie, WITHDRAWAL	7,10	32,99	1,01
1	karta_tak	==>	CREDIT	5,33	24,75	0,97
14	karta_tak	==>	49<Age<=60	5,30	24,65	1,17
10	karta_tak	==>	17<Age<=27	5,01	23,27	1,12

Rys. 8. Reguły asocjacyjne dotyczące posiadaczy kart płatniczych wyodrębnione przy minimalnym wsparciu 10% i pewności 10%

Dodatkowe zapytanie potwierdziło dotychczasowe wyniki i ze względu na ograniczony zakres danych wniosło niewiele dodatkowych informacji o posiadaczach kart płatniczych.



Dowiadujemy się jedynie, że najczęściej wykonywaną przez nich operacją jest wypłata (*withdrawal*). Wypłaty środków stanowią 72,5% wszystkich transakcji, które wykonała analizowana grupa klientów. W celu uzyskania szczegółowych informacji o wypłacanych kwotach i częstotliwości wypłat należy uzupełnić zbiór danych o zmienne zawierające kwoty transakcji. Takie dane umożliwiłyby również przeprowadzenie analizy sekwencji i wyodrębnienie np. najczęściej powtarzanego schematu wypłaty środków w trakcie miesiąca.

## Podsumowanie

Podsumowując wyniki i etap przygotowania danych, należy zwrócić uwagę na kilka ważnych aspektów, charakterystycznych dla analiz przeprowadzanych na dużych zbiorach danych, otrzymanych w wyniku zapytań do baz danych.

Przede wszystkim należy podkreślić, że przed przystąpieniem do etapu przygotowywania danych, jak i do samej analizy, należy bardzo dokładnie zaprojektować analizę oraz precyzyjnie określić cel badania. W tym celu należy odpowiednio dużo czasu poświęcić na zapoznanie się z danymi, w szczególności nad ich strukturą oraz informacjami, jakie można uzyskać z uporządkowania i przekształceń. Kolejnym etapem jest także przygotowanie danych, aby uzyskać wyniki, które będą dotyczyć problemu sformułowanego w celu badania, a także, aby były to wyniki kompletne, tzn. dotyczące całego zbioru. Często zdarza się, że przy przekształcaniu danych do wybranej postaci tracimy część informacji o badanej zbiorowości lub występuje redundancja danych i uzyskane wyniki dotyczą danych niekompletnych, zniekształconych lub tylko podzbioru. Struktura danych pozyskanych z baz danych, które mają być poddane analizie statystycznej, wymaga wielu przekształceń. Jest to bardzo czasochłonny proces i często wymaga umiejętności z wielu dziedzin. Jeśli przekształcenia robione są bez żadnych założeń i bez wizji celu, któremu mają służyć, wówczas uzyskiwane wyniki są najczęściej nieprawidłowe.

## Literatura

1. R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, *Proceedings of ACM SIGMOD International Conference on Management of Data*, Washington DC, May 1993.
2. Hastie T., Tibshirani R., Friedman J., *The elements of statistical learning. Data mining, inference and prediction*, Springer Verlag, 2001.
3. Han, Jiawei, *Data mining: concepts and techniques*, Morgan Kaufman Publishers, 2001.
4. R. Kita, Analiza sposobu poruszania się użytkowników po portalu internetowym”, [w] *Data mining – metody i przykłady*, StatSoft Polska 2002 (artykuł dostępny na stronie [www.statsoft.pl/czytelnia/dm/wstepdm.html](http://www.statsoft.pl/czytelnia/dm/wstepdm.html)).
5. Westphal C., Blaxton T., *Data mining solutions. Methods and Tools for Solving Real-World Problems*, Wiley Computer Publishing, 1998.