A C T A U N I V E R S I T A T I S N I C O L A I C O P E R N I C I
EKONOMIA XXXIX – NAUKI HUMANISTYCZNO-SPOŁECZNE – ZESZYT 389 – TORUŃ 2009

Uniwersytet Mikołaja Kopernika w Toruniu
Katedra Ekonometrii i Statystyki

Michał Kukliński, Małgorzata Śniegocka-Łusiewicz

MIARY ASOCJACJI W ANALIZIE KOSZYKOWEJ – PRZYKŁAD EMPIRYCZNY

Z a r y s t r e ś c i. Poniższy artykuł jest próbą przedstawienia możliwości wykorzystania analizy koszykowej w badaniu zjawiska asocjacji na przykładzie danych transakcyjnych pochodzących z hurtowni spożywczej. Zaprezentowane zostały trzy metody analizy: uogólniona metoda indukcji reguł, *a priori* oraz CARMA. Zastosowanie wspomnianych reguł i analiza wyników pozwoli na uzyskanie niezbędnych informacji dla analiz marketingowych oraz daje podstawy do zastosowania w skutecznych strategiach marketingowych.

S ł o w a k l u c z o w e: analiza koszykowa, reguły asocjacji, GRI, *a priori*, CARMA.

1. WSTĘP

Uporządkowany zbiór danych pozwala na poznanie praw i reguł nim rządzących. Zasady te nazywamy regułami asocjacji. Odpowiednio wykonane badanie pod kątem wyszukiwania tychże reguł pozwala wyciągnąć wnioski przydatne dla badacza. Celem artykułu jest przedstawienie podstawowych wskaźników zjawiska asocjacji w analizie koszykowej na podstawie przykładu empirycznego.

Analizą koszykową, w sposób uproszczony, możemy nazwać analizę zawartości koszyka klienta. Przychodząc do sklepu (np. supermarketu, sklepu internetowego, punktu usługowego) klient ma do wyboru wiele różnego typu produktów. Do swojego „koszyka" może włożyć to, co chce, w jakiej kolejności chce i ile chce. Analiza koszykowa polega na rozpoznaniu reguł, którymi kierują się klienci przy zapełnianiu „koszyka", zwyczajów danego klienta, prawidłowości w korzystaniu z usług danego typu, badaniu, jakie produkty kupowane są razem lub w określonej sekwencji.

Obliczenia reguł asocjacyjnych zostały przeprowadzone na danych empirycznych pochodzących z hurtowni spożywczej obejmujące dwa pełne lata

sprzedaży. Hurtownia danych zawiera 238796 dokumentów sprzedaży, a w swoim asortymencie posiada 118 grup towarowych.

Pierwotna hurtownia danych zawierała 6.087.393 rekordów, w których zostały zawarte informacje dotyczące sprzedaży, tj. numeru dokumentu sprzedaży, daty sprzedaży, kodu produktu (zawierający grupę asortymentową), nazwę produktu, ceny, ilości, wysokości udzielonego rabatu oraz dane nabywcy.

Na potrzeby badania reguł asocjacyjnych pominięto część informacji, które opisywały sprzedaż. W celu uzyskania przejrzystego wyniku badania skupiono się na grupach towarowych produktów, a nie na poszczególnych produktach. W wyniku powyższego założenia do analizy wykorzystane zostały 2 pola hurtowni danych: numer dokumentu sprzedaży oraz numer grupy asortymentowej zawartej w każdym kodzie produktu. W związku z tym, że występowanie kilku produktów zawierających się w jednej grupie asortymentowej w ramach jednego dokumentu sprzedaży powodowało powtórzenia, należało dokonać grupowania. Grupowanie miało na celu doprowadzenie danych do postaci, w której dana grupa asortymentowa występowała na każdym dokumencie sprzedaży najwyżej jeden raz. W wyniku grupowania hurtownia danych zmniejszyła się do 2.373.090 rekordów.

W niniejszej publikacji przedstawione zostało każdorazowo 15 wyników poszczególnych metod analizy asocjacji: metody uogólnionej indukcji reguł (GRI), metody *a priori* i metody CARMA.[1]

2. METODA UOGÓLNIONEJ INDUKCJI REGUŁ (GRI)

GRI – uogólniona indukcja reguł – wskazuje występowanie asocjacji danych. Przykładowo: klienci kupujący golarki i piankę do golenia prawdopodobnie kupią również płyn po goleniu. GRI wskazuje zasady o najwyższej zawartości informacji oparte na indeksie biorącym pod uwagę zarówno uogólnienie oraz dokładność. GRI może wykorzystywać dane ilościowe oraz jakościowe, ale cel musi być jakościowy. W przypadku algorytmu *a priori* wszystkie dane muszą mieć charakter jakościowy, natomiast dane ilościowe można dyskretyzować, lub zastosować wspomniany algorytm GRI. GRI stosuje podejście teorii informacji, aby określić czy kandydująca reguła jest interesująca. „GRI wykorzystuje *J-miarę*.

$$J = p(A)\left[p(A \mid B) \ln \frac{p(B \mid A)}{p(B)} + \left[1 - p(B \mid A)\right] \ln \frac{1 - p(B \mid A)}{1 - p(B)} \right],$$

gdzie: $p(A)$ – reprezentuje prawdopodobieństwo lub ufność obserwowanej wartości *A*. Jest to miara zakresu poprzednika; $p(B)$ - reprezentuje prawdopo-

[1] Opracowanie na podstawie przywołanej literatury.

dobieństwo lub ufność obserwowanej wartości *B*, jest to miara zakresu następnika; $p(B \mid A)$ – reprezentuje prawdopodobieństwo warunkowe lub późniejszą ufność zmiennej *B* dla danego *A*, które następuje.

Jest to miara prawdopodobieństwa zaobserwowania wartości *B* pod warunkiem, że występuje *A*. Zatem, $p(B \mid A)$ reprezentuje uaktualnione prawdopodobieństwo obserwowania wartości B po zyskaniu dodatkowej wiedzy o wartości *A*. W terminologii reguł asocjacyjnych $p(B \mid A)$ jest mierzone bezpośrednio, jako ufność reguły. (...) Dla reguł z więcej niż jednym poprzednikiem, $p(A)$ jest obliczane, jako prawdopodobieństwo koniunkcji wartości zmiennych w poprzedniku" (Larose, 2006).

W przypadku tego algorytmu badacz określa maksymalną liczbę reguł, które chce osiągnąć, ponieważ znajdowanie kolejnych reguł polega na obliczaniu *J-miary* dla kolejnych przypadków i porównywaniu wartości z najniższą dostępną wartością z tabeli. Jeżeli nowa wartość jest większa, to zostaje nadpisana na poprzednią. *J-miara* osiąga najwyższe wartości w przypadku, gdy $p(B)$ lub $p(B \mid A)$ przyjmuje wartości skrajne, tzn. prawdopodobieństwo wynosi prawie 1 lub prawie 0. Sytuacje, gdy poziom ufności, czyli $p(B \mid A)$, jest minimalny również są przydatne dla badacza, ponieważ umożliwiają utworzenie reguły przeciwnej (jeżeli *A* to NIE *B*).

Tabela 1. Wyniki asocjacji z wykorzystaniem metody GRI

Następnik	Poprzednik	Identyfikator	Ilość	Wsparcie %	Poziom ufności %	Wsparcie	Dźwignia	Wdrażalność
sery, serki	• przetwory warzywne	12	31 411	13,15	86,82	11,42	2,341	1,733
jogurty	• sery, serki • desery	3	31 805	13,32	86,49	11,52	3,183	1,8
jogurty	• desery	2	37 228	15,59	82,98	12,937	3,054	2,653
sery, serki	• jogurty	1	64 882	27,17	81,13	22,044	2,188	5,127
cukierki, pastylki,	• czekolady • ciastka	73	25 963	10,87	74,67	8,119	2,09	2,753
cukierki, pastylki,	• bombonierki	55	31 366	13,14	74,26	9,754	2,078	3,382
jogurty	• mleka, maślanki, kefiry	4	43 376	18,16	73,73	13,393	2,714	4,771
ciastka	• zupy w proszku • wafelki	81	26 017	10,9	73,51	8,009	2,043	2,887
ciastka	• kawy mielone • wafelki	64	30 821	12,91	73,39	9,472	2,039	3,435

Ciąg dalszy tabeli 1

Następnik	Poprzednik	Identyfikator	Ilość	Wsparcie %	Poziom ufności %	Wsparcie	Dźwignia	Wdrażalność
ciastka	• sery, serki • wafelki	63	31 648	13,25	72,96	9,67	2,027	3,583
przetwory warzywne	• dżemy, powidła, konfitury	40	32 678	13,68	72,95	9,983	2,241	3,7
ciastka	• cukierki, pastylki, draże	31	46 163	19,33	72,89	14,091	2,025	5,24
ciastka	• przetwory warzywne	65	31 868	13,35	72,66	9,697	2,019	3,65
ciastka	• czekolady • wafelki	89	24 891	10,42	72,56	7,563	2,016	2,859
wafelki	• batony • ciastka	50	28 274	11,84	72,44	8,577	2,308	3,263

Objaśnienia:

Następnik – w przypadku zależności, jeżeli *A* to *B*, jest to szukane *B*.

Poprzednik – w przypadku zależności, jeżeli *A* to *B*, jest to szukane *A*.

Identyfikator – pozwala zidentyfikować, które reguły zostały zastosowane do danej predykcji. Numer identyfikacyjny reguły pozwala też na późniejsze dołączenie dodatkowych informacji na temat reguł takich jak wdrażalność, informacja o produkcie, poprzedniki.

Ilość – ilość przypadków; pokazuje ile jest pojedynczych przypadków o unikatowym kluczu podstawowym, dla których miała miejsce poszukiwana reguła. Oznacza to liczbę wystąpień A w całym zbiorze danych.

Wsparcie % – Wsparciem w ujęciu procentowym (pokryciem procentowym) zbioru A nazywamy stosunek liczby transakcji wspierających A do liczby wszystkich transakcji. Innymi słowy jest to prawdopodobieństwo wystąpienia zdarzenia A.

Poziom ufności % – wskazuje on, w postaci procentowej, proporcje liczby rekordów z zarówno odpowiednim poprzednikiem oraz jego następnikiem do liczby rekordów z jedynie odpowiednim poprzednikiem. Oznacza to, iż poziom ufności jest to stosunek rekordów z A i B do wszystkich rekordów z A.

Wsparcie – pokrycie reguły w ujęciu procentowym, pokazuje udział rekordów z zarówno odpowiednim poprzednikiem jak i następnikiem w stosunku do ogólnej liczby rekordów. Jest to stosunek rekordów z A i B do wszystkich rekordów.

Dźwignia – pokazuje poziom ufności dla reguły przed prawdopodobieństwem wystąpienia następnika. Oznacza to, że jeżeli mamy zasadę, jeżeli A to B to ciężarem tej zasady będzie następująca wartość: stosunek poziomu ufności dla reguły: jeżeli A to B do wsparcia reguły, jeżeli B to A.

Wdrażalność – jest procentową miarą wystąpienia danych spełniających warunki poprzednika, ale niespełniających warunków następnika. W odniesieniu do zakupu produktów oznacza to, jaki odsetek klientów posiada (lub zakupiło) poprzednika, ale nie zakupiło jeszcze następnika. Statystyka wyrażalności jest określona, jako (ilość rekordów spełniająca warunek poprzednika – ilość rekordów spełniająca regułę)/ilość rekordów; gdzie spełniająca warunek poprzednika oznaczają ilość rekordów, dla których poprzednik jest prawdziwy. – ilość rekordów spełniająca regułę oznacza ilość rekordów, dla których zarówno poprzednik jak i następnik są prawdziwe. Innymi słowy wdrażalność jest to iloraz różnicy liczby A i liczba rekordów z zarówno A i B oraz sumy wszystkich rekordów.

Źródło: obliczenia własne za pomocą programu SPSS Clementine 11.1.

Przy założonym: minimalnym pokryciu procentowym (wsparciu) na poziomie 10%, najniższym poziomie ufności wynoszącym 50% oraz przy liczbie poprzedników nieprzekraczającej trzech otrzymaliśmy 100 reguł asocjacyjnych, które zostały obliczone na podstawie 238800 istotnych transakcji. Maksymalne pokrycie procentowe reguły wyniosło 37,08%, najwyższy poziom ufności reguły osiągnął 86,82%, dźwignia reguł zawierała się między wartościami 1,62% a 3,48%.

W wyniku przeprowadzonego badania możemy zaobserwować wzajemny wpływ serów, serków i jogurtów na siebie, co obrazują wyniki z poziomem ufności przekraczającym 80%, w szczególności w przypadku wyniku o identyfikatorze reguły 12 z poziomem ufności 86,82%, wyniku z identyfikatorem reguły 3 oraz poziomem ufności 86,49%, wynik identyfikatorze reguły 1 z poziomem ufności 81,13%, ale również wynik z niższym poziomem ufności, jednak nadal dość wysokim – 73,73% o identyfikatorze reguły 4. Należy zwrócić uwagę na fakt, iż reguła posiadająca najwyższy poziom ufności (86,82%, identyfikator reguły 12) dotyczy zależności serów, serków od jogurtów (przy współtowarzyszeniu przetworów warzywnych) oraz reguła posiadająca największą ilość przypadków (identyfikator reguły 1 z poziomem ufności wynoszącym 81,13%) również dotyczy zależności serów, serków od jogurtów, co potwierdza wcześniejszy wniosek. Pozostałe reguły dają zróżnicowane wyniki, dlatego wskazane jest badanie również innymi metodami.

3. METODA Z WYKORZYSTANIEM ALGORYTMU *A PRIORI*

Algorytm *a priori* wydobywa zestaw reguł z danych wybierając reguły z najwyższą zawartością informacji. Metoda *a priori* oferuje 5 różnych metod selekcji reguł i wykorzystuje wyrafinowany model indeksowania do wydajnego przetwarzania dużych baz danych. W przypadku dużych problemów badawczych algorytm *a priori* jest szybszy niż GRI, nie ma on odgórnego limitu ilości reguł, które można uzyskać i może obsługiwać reguły aż do 32 założeń. Metoda *a priori* wymaga, aby wszystkie dane wejściowe i wyjściowe były jakościowe, ale oferuje lepszą wydajność, ponieważ został zoptymalizowany do tego rodzaju danych. Algorytm *a priori* wykorzystuje właściwość *a priori*, która mówi o tym, że jeżeli zbiór zdarzeń Z nie jest pusty, to dla dowolnego elementu A, gdzie $Z \cup A$ także nie będzie pusty. Oznacza to, że dodanie dowolnego artykułu do zbioru niepustego nie spowoduje, iż zbiór ten stanie się pustym. Kolejnym wnioskiem jest, iż żaden nadzbiór niepusty zbioru nie będzie pusty. Oznacza to, że poszukując zbiorów częstych algorytm najpierw przeanalizuje wszystkie jednoelementowe podzbiory i dopiero wśród tych częstych będzie szukał kandydatów na częste zbiory dwuelementowe i tak dalej. Posiadając już wszystkie zbiory częste (k) algorytm wyszuka wszystkie podzbiory (l) znalezionych zbiorów częstych. Następnie zbada występowanie reguły, jeżeli l to

$(k-l)$. Dla podanych reguł algorytm wylicza poziom wsparcia i ufności. Od badacza zależy, jaki poziom wsparcia i ufności uzna on za minimalny dla danego badania.

Tabela 2. Wyniki analizy asocjacji z wykorzystaniem metody *a priori*

Następnik	Poprzednik	Identyfikator	Ilość	Wsparcie %	Poziom ufności %	Wsparcie	Dźwignia	Wdrażalność
sery, serki	• masła • jogurty	105	36 174	15,148	90,253	13,672	2,434	1,477
sery, serki	• desery	6	37 228	15,59	85,433	13,319	2,304	2,271
jogurty	• desery	5	37 228	15,59	82,981	12,937	3,054	2,653
sery, serki	• jogurty	58	64 882	27,17	81,133	22,044	2,188	5,126
sery, serki	• masła	53	57 812	24,21	79,143	19,16	2,134	5,049
sery, serki	• margaryny	33	52 184	21,853	78,25	17,1	2,11	4,753
sery, serki	• mleka, maślanki, kefiry	41	55 689	23,321	77,89	18,164	2,1	5,156
jogurty	• śmietany • sery, serki	99	40 674	17,033	74,254	12,648	2,733	4,385
jogurty	• mleka, maślanki, kefiry • sery, serki	91	43 376	18,164	73,73	13,393	2,714	4,772
ciastka	• wafelki • cukierki, pastylki, draże	133	46 163	19,332	72,89	14,091	2,025	5,241
sery, serki	• śmietany	49	55 929	23,421	72,724	17,033	1,961	6,388
jogurty	• margaryny • sery, serki	81	40 834	17,1	71,45	12,218	2,63	4,882
jogurty	• masła • sery, serki	106	45 754	19,16	71,356	13,672	2,626	5,488
przetwory warzywne	• majonez, ocet, musztarda	21	48 639	20,368	71,132	14,489	2,185	5,88
ciastka	• chrupki, paluszki, • chałwy, sękacze, murzynki • cukierki, pastylki, draże	115	38 559	16,147	69,691	11,253	1,937	4,894

Objaśnienia: patrz tabela 1.

Źródło: obliczenia własne za pomocą programu SPSS Clementine 11.1.

Przy założonym: minimalnym pokryciu procentowym na poziomie 15%, najniższym poziomie ufności wynoszącym 50% oraz przy liczbie poprzedników

nieprzekraczającej pięciu otrzymaliśmy 146 reguł asocjacyjnych, które zostały obliczone na podstawie 238 796 istotnych transakcji. Maksymalne pokrycie procentowe reguły wyniosło 37,08%, najwyższy poziom ufności reguły osiągnął 90,25%, dźwignia reguł zawierała się między 1,35 a 3,46%.

Zdecydowanie dominującym następnikiem wśród wyników z najwyższym poziomem ufności (powyżej 78%) są sery, serki, następnie jogurty. Można zaobserwować głównie wpływ takich grup produktowych jak: masło, jogurty, desery, margaryny. Wpływ masła i jogurtu widać na wyniku o najwyższym w tej tabeli poziomie ufności 90,235%, gdzie mamy przykład wpływu na zakup serów, serków zakupu łącznego obu produktów. Wpływ ten jest widoczny przy zakupie tychże produktów pojedynczo, jak w wynikach z poziomem ufności 81,133% oraz 79,143%. Wpływ deserów, wynik z poziomem ufności 85,433% na zakup serów, serków i z poziomem ufności 82,981% na zakup jogurtów. Wpływ margaryny – na przykładzie wyniku z poziomem ufności 78,250%. Występuje charakterystyczny wzajemny wpływ serów, serków i jogurtów na siebie, co widać w regułach 105, 5 i 58, dla których każdorazowo poziom ufności przekracza 80%. Metoda ta dała bardzo jednoznaczne wyniki, jednakże daje mało informacji w stosunku do potrzeb przy planowaniu kompleksowej kampanii reklamowej.

4. METODA Z WYKORZYSTANIEM MODELU CARMA

Model CARMA pozyskuje zestaw reguł z danych bez potrzeby specyfikacji pól wejścia (predyktor) oraz wyjścia (cel). W porównaniu do metody *a priori* i algorytmu GRI - CARMA dysponuje wsparciem zarówno dla poprzednika, jak i dla następnika, a nie tylko dla poprzednika. Oznacza to, że wygenerowane reguły mogą być wykorzystane do szerszego zastosowania, na przykład, aby znaleźć listę produktów lub usług (poprzednik), których następnikiem jest obiekt.

Przy założonym: minimalnym pokryciu procentowym na poziomie 10%, najniższym poziomie ufności wynoszącym 50% otrzymaliśmy 120 reguł asocjacyjnych, które zostały obliczone na podstawie 238.796 istotnych transakcji. Maksymalne pokrycie procentowe reguły wyniosło 37,08%, najwyższy poziom ufności reguły osiągnął 90,66%, dźwignia reguł zawierała się między 1,41% a 3,35%.

Bezsprzecznie widoczny jest wpływ zakupu jogurtów na zakup serów, serków, co ilustrują przykładowo wyniki z identyfikatorami reguł: 1 (poziom ufności 90,662%), 2 (poziom ufności 90,253%), 3 (poziom ufności 89,839%), 5 (poziom ufności 89,049%) oraz 7 (poziom ufności 88,722%). Zgodnie ze specyfiką tej metody otrzymujemy również wachlarz produktów towarzyszących poprzednikom (grupa produktowa zawierająca mleka, maślanki, kefiry – wyniki o identyfikatorach reguły 1, 4, 6, 14; masła – wyniki o identyfikatorach reguły 2, 6, 8, 10; margaryny – wyniki o identyfikatorach reguły 3, 4, 10; desery – wyniki o identyfikatorach reguły 5, 11, 13, 15; śmietany – wyniki o identyfikato-

rach reguły 7, 8, 14). Dzięki tym wynikom oraz informacjom otrzymanym z poprzednich metod można zaplanować zintegrowane kampanie reklamowe par produktów zwiększając również sprzedaż serów, serków.

Tabela 3. Miary asocjacji z wykorzystaniem metody CARMA

Następnik	Poprzednik	Identyfikator	Ilość	Wsparcie %	Poziom ufności %	Wsparcie	Dźwignia	Wdrażalność
sery, serki	• Jogurty • mleka, maślanki, kefiry	1	35 275	14,772	90,662	13,393	2,445	1,379
sery, serki	• jogurty • masła	2	36 174	15,148	90,253	13,672	2,434	1,477
sery, serki	• jogurty • margaryny	3	32 476	13,6	89,839	12,218	2,423	1,382
sery, serki	• mleka, maślanki, • kefiry i margaryny	4	27 227	11,402	89,235	10,174	2,406	1,227
sery, serki	• jogurty • desery	5	30 892	12,937	89,049	11,52	2,401	1,417
sery, serki	• mleka, maślanki, • kefiry i masła	6	31 443	13,167	88,729	11,683	2,393	1,484
sery, serki	• jogurty • śmietany	7	34 041	14,255	88,722	12,648	2,392	1,608
sery, serki	• śmietany • masła	8	30 430	12,743	87,272	11,121	2,353	1,622
sery, serki	• przetwory warzywne • jogurty	9	31 411	13,154	86,82	11,42	2,341	1,734
sery, serki	• masła • margaryny	10	32 394	13,566	86,791	11,774	2,34	1,792
jogurty	• sery, serki • desery	11	31 805	13,319	86,493	11,52	3,183	1,799
sery, serki	• jogurty • ciastka	12	28 616	11,983	86,455	10,36	2,331	1,623
sery, serki	• desery	13	37 228	15,59	85,433	13,319	2,304	2,271
sery, serki	• mleka, maślanki, kefiry • śmietany	14	32 123	13,452	84,892	11,42	2,289	2,032
jogurty	• desery	15	37 228	15,59	82,981	12,937	3,054	2,653

Objaśnienia: patrz tabela 1.

Źródło: obliczenia własne za pomocą programu SPSS Clementine 11.1.

5. PODSUMOWANIE

Każda z prezentowanych metod badania asocjacji charakteryzuje się innymi założeniami, dlatego przedstawione wyniki analiz na tym samym zestawie da-

nych nie są identyczne. Obliczenia w uogólnionej metodzie indukcji reguł są oparte na *J-miarze*, natomiast metoda *a priori* posiłkuje się różnymi metodami selekcji reguł oraz indeksowaniem danych, z kolei metoda CARMA pozwala na zdefiniowanie wskaźnika wsparcia również dla następnika. Przedstawione wyniki zostały ustawione w kolejności od najwyższego procentowego wskaźnika poziomu ufności, który właściwie odzwierciedla siłę reguły i jest najlepszym źródłem informacji o zbiorze danych. Należy zwrócić uwagę, że na pierwszym miejscu w każdej metodzie znajduje się zależność: jeżeli zakupione zostają jogurty (z towarzyszeniem drugiego składnika), to następnikiem będą sery, serki. Część wspólna wyników potwierdza tylko właściwość zastosowania kilku metod oraz daje pełniejszy obraz sytuacji. Na podstawie powyższych badań można wyciągnąć wniosek, iż kampanie reklamowe powinny skupić się na takich produktach jak: jogurty, masła, margaryny, kefiry oraz desery, ponieważ mają one najczęściej wpływ na zakup serów, serków lub jogurtów. Natomiast zakup jogurtów lub serków wpłynie na siebie nawzajem, co spowoduje, iż dzięki odpowiedniej kampanii reklamowej możemy zwiększyć sprzedaż zarówno jogurtów jak i serów, serków.

Wyniki poszczególnych metod badania asocjacji są podobne, jednak nie identyczne. Wynika to z faktu, że ilość możliwych reguł asocjacyjnych rośnie wykładniczo wraz ze wzrostem liczby atrybutów. W przedstawionym przykładzie dysponujemy 118 produktami. W obliczeniach ograniczyliśmy się wyłącznie do atrybutów binarnych (dany produkt został zakupiony lub nie), dlatego liczba spodziewanych reguł asocjacyjnych jest rzędu: 118×2^{117}. W związku z tym, że poszczególne metody nie są w stanie sprawdzić wszystkich możliwych reguł, starają się zmniejszyć przestrzeń poszukiwań. Na przykład metoda a priori wybiera spośród zbioru zdarzeń - zbiory częste, przycina je, tworząc zbiory kandydujące do poszukiwań reguł asocjacyjnych. Zupełnie odmienny sposób selekcji potencjalnych atrybutów ma metoda GRI, która opiera swój wybór kandydatów na J-miarze. Ostateczny wynik obliczeń poszczególnych metod nie będzie taki sam z uwagi na różne metody selekcji potencjalnych kandydatów branych pod uwagę przy tworzeniu reguł asocjacyjnych.

Analizy opisane w artykule mają charakter statyczny, kolejnym etapem badania będzie wprowadzenie czynnika czasu związanego z cyklicznością tygodniową, roczną, wpływu zaistnienia świąt oraz promocji marketingowych na wyniki.

LITERATURA

Agrawal R., Imieliński T., Swami A. (1993), *Mining association rules between sets of items in large databases*, *Proceedings of ACM SIGMOD,* International Conference on Management of Data, Washington DC.

Han, Jiawei (2001), *Data mining: concepts and techniques*, Morgan Kaufman Publishers.

Hastie T., Tibshirani R., Friedman J. (2001), *The elements of statistical learning. Data mining, inference and prediction*, Springer Verlag.

Kita R. (2002), *Analiza sposobu poruszania się użytkowników po portalu internetowym,* [w:] *Data mining – metody i przykłady*, StatSoft Polska (artykuł dostępny na stronie www.statsoft.pl/czytelnia/dm/wstepdm.html).

Larose D. T. (2006), *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych*, Wydawnictwo Naukowe PWN, Warszawa.

Rauch J. (2005), *Logic of Association Rules,* [w:] *Applied Intelligence 22(2005),* Springer Science.

Taniar D. (2008), *Data Mining and Knowledge Discovery Technologies,* IGI Publishing, Hershey.

Westphal C., Blaxton T. (1998), *Data mining solutions. Methods and Tools for Solving Real-World Problems*, Wiley Computer Publishing.

THE ASSOCIATION MEASURES IN THE MARKET BASKET ANALYSIS – THE EMPIRICAL STUDY

A b s t r a c t. The following article is an attempt to present the use of the market basket analysis in the study of associations in the example of transaction data from the food wholesaler. Three analysis methods were presented: the GRI method, the a priori method and the CARMA method. The application of these rules and the results analysis will allow to obtain information necessary for marketing analysis and will give rise to the application of effective marketing strategies.

K e y w o r d s: market basket analysis, association rules, GRI, *a priori*, CARMA.