DISCOVERING ASSOCIATION RULES IN MARKET BASKET ANALYSIS

OTKRIVANJE PRAVILA ASOCIJACIJE U ANALIZI POTROŠAČKE KOŠARICE

RADOVANOVIC, Lazar; PETROVIC M., Teodor & MATANOVIC, Slavko

***Abstract**: The paper presents an algorithm for discovering association rules based on database transactions at a point of sale of a trading company. The analysis reveals a list of ten products which are most usually purchased, the couples of products from the list bought together and values of probability and support to the discovered rules. The rules can be used for marketing purposes by organizing promotional activities and arranging sales facilities and combining items that have a high rate of probability of coupled occurrence.*

***Key words**: association rule, data mining, market basket data, market basket analysis*

***Sažetak**: U radu je prezentiran algoritam za otkrivanje pravila asocijacije na temelju baze podataka o transakcijama na prodajnom mjestu unutar trgovinskog poduzeća. Analizom je utvrđena lista od deset proizvoda koji se najčešće kupuju, parovi proizvoda iz liste koji se kupuju skupa i vrijednosti vjerojatnosti i potpore pravilima. Otkrivena pravila mogu poslužiti u marketinške svrhe organiziranjem promidžbenih aktivnosti, te za uređivanje prodajnih objekata i kombiniranje artikla koji imaju visoku stopu vjerojatnosti skupnog pojavljivanja.*

***Ključne riječi**: pravilo asocijacije, data mining, podaci i analiza podataka iz potrošačke košarice*

Autors´ data: Lazar **Radovanovic**, PhD, University of East Sarajevo, Faculty of Economics in Brčko, l.radovanovic@efbrcko.ba, Teodor M. **Petrovic**, PhD, University of East Sarajevo, Faculty of Economics in Brčko, t.petrovic@ efbrcko.ba, Slavko **Matanovic**, MSc, My Software, d.o.o., Brčko, Bosnia and Herzegovina, s.matanovic@live.com.

1. Uvod

Analiza potrošačke košarice omogućava odgovor na jedno od najzanimljivijih pitanja: koje se skupine proizvoda često pojavljuju zajedno. Takve skupine proizvoda su vrlo korisne za davanje preporuka kupcima; na primjer, kupci koji su kupili neki proizvod mogu biti zainteresirani za kupnju drugih proizvoda. Pravila asocijacije mogu pružiti odgovore na ova pitanja. Primjenom tehnika *data mininga* moguće je odrediti koji proizvodi u kupnji sugeriraju kupnju i drugih proizvoda.

Cjelokupni spisak kupnja (sve transakcije iz baze podataka) koje obave kupci sadrži vrijedne informacije za trgovinu na malo – koju robu potrošači najviše kupuju, pojedinačno ili u kombinaciji s drugim artiklima, i u koje vrijeme.

Svaki kupac kupuje različit skup proizvoda, različitih količina i u različito vrijeme. Analiza potrošačke košarice koristi se informacijama o tomu što potrošač kupuje i daje uvid u to tko su kupci i zašto obavljaju određene kupnje. Analiza potrošačke košarice kazuje koji proizvodi imaju tendenciju da se kupuju skupa i koji bi se mogli skupa kombinirati u promidžbenoj kampanji. Ove informacije su značajne jer mogu sugerirati novi izgled prodavaonice, koje proizvode izdvojiti posebno, indicirati kada izdavati potrošačke kupone, davati popuste i sl. Podaci postaju vrijedniji ako se dovedu u svezu s konkretnim kupcima, preko kartice lojalnosti ili registriranjem pojedinačnih kupaca.

U analizi potrošačke košarice može poslužiti tehnika *data mininga* za automatsko generiranje pravila asocijacije i otkrivanje obrazaca ponašanja kupaca prigodom kupnje. Međutim, imaju li obrasci smisla, prepušteno je da ljudi sami zaključe i interpretiraju.

2. Data mining baze podataka o transakcijama na prodajnom mjestu

Jedno od područja primjene tehnika *data mininga* je otkrivanje pravila asocijacije u analizi podataka iz potrošačke košarice [1].

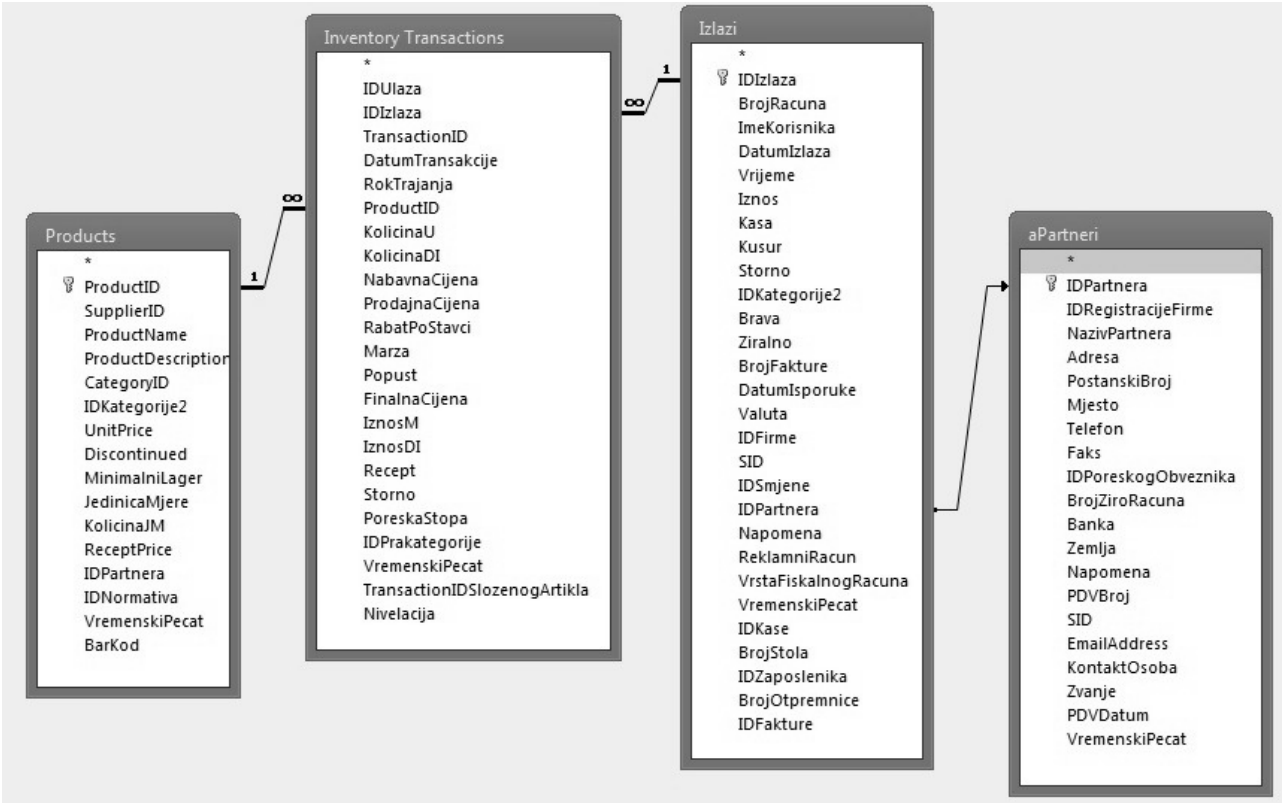
U ovom radu za analizu poslužit će baza podataka trgovinskog poduzeća, koja sadrži podatke o transakcijama na prodajnom mjestu (POS – *point of sale*). Svaka transakcija sastoji se od skupa artikala iz pojedinačne kupnje. Pravilima asocijacije otkrivaju se relacije između kupljenih artikala. Jednim od pravila asocijacije moguće je otkriti koja se dva artikla (A i B) pojavljuju skupa u jednoj kupnji. Pravilo asocijacije ima oblik $A \Rightarrow B$.

Baza podataka sadrži oko 1.800.000 zapisa, gdje svaka transakcija obuhvaća skup artikala. Pravilo „*Ako A, onda B*“ ($A \Rightarrow B$) označava da kad god kupnja sadrži artikal A, postoji stanovita vjerojatnost da će sadržavati i artikal B. Pravilo ima potporu p u transakcijama ako $p\%$ svih transakcija sadrži i A i B. Za otkrivanje pravila asocijacije

potrebno je postaviti minimalne vrijednosti, ili prag, potpore i vjerojatnosti. Pravila koja imaju vrijednosti potpore i vjerojatnosti veće ili jednake minimumu nazivaju se pravila zanimljivih asocijacija. Pored uporabe pravila zanimljivih asocijacija, pravila asocijacija koja ne zadovoljavaju minimalne zahtjeve, tj. imaju vrijednosti potpore i vjerojatnosti manje od minimalnih vrijednosti koje definira korisnik, također se uzimaju u razmatranje u toku procesa donošenja poslovnih odluka [2].

Prije nego što se pristupi analizi postavlja se pitanje: Što su podaci iz potrošačke košarice? To su transakcijski podaci koji opisuju tri bitno različita entiteta (1) kupce, (2) kupnje ili košarice i (3) artikle [3].

Transakcija, ili kupnja, je osnovna struktura za podatke iz potrošačke košarice. Kupnja predstavlja jedan događaj – košaricu namirnica koja obuhvaća ukupni asortiman i kupljenu količinu, ukupan iznos, način plaćanja kao i druge relevantne podatke o transakciji. U relacijskoj bazi podataka tablica transakcija povezana je s drugim tablicama putem odgovarajućih spoljnih ključeva (Slika 1.)



Slika 1. Relacije u bazi podataka o transakcijama

Pojedinačni artikli u kupnji predstavljaju se posebno kao stavke kupnje. Relevantni podaci kupnje su: cijena i količina artikla, PDV i mogući troškovi koji utječu na visinu marže. Tablica o artiklima (ovdje *Products*) po pravilu sadrži više opisnih informacija o svakom proizvodu. Ove informacije mogle bi uključiti i kategorije proizvoda i druge podatke relevantne za analizu [3].

Tablica *Kupci* (ovdje *aPartneri*) neophodna je za identifikaciju kupca, na primjer, kada kupac koristi kreditne ili debitne kartice sa posebnim pogodnostima – kartice povoljnosti. Iako tablica *aPartneri* ima i druga zanimljiva polja, najvrijedniji element je *IDPartnera*, jer služi za povezivanje transakcija tijekom vremena. Analizom je moguće dokučiti i navike kupaca i otkriti koji kupci ponavljaju kupnju. To ukazuje na povoljnu poslovnu priliku za povećanje obima prodaje po kupcima, kao i *cross-selling*. Također, korisno je međusobno usporediti ove pokazatelje da se vidi širina odnosa s kupcima (broj jedinstvenih artikala koji su bilo kada kupljeni) po dubini odnosa (broj kupnji) za kupce koji su kupili više od jednog artikla [4].

Ono što je blagodet pravila asocijacije jeste jasnoća i primjenljivost rezultata. U tome se sastoji intuitivna privlačnost pravila asocijacije, jer ona odražavaju načine na koje je moguće grupirati proizvode. Pravilo poput: "Ako kupac kupuje proizvod *A*, onda će kupiti i proizvod *B*", je jasno i sugerira konkretno djelovanje – pakiranje proizvoda *A* i proizvoda *B* u jedan paket.

Iako je pravila asocijacije lako razumjeti, nisu sva pravila uvijek korisna. Dakle, postoje *primjenljiva*, *trivijalna* i *neobjašnjiva* pravila. Korisno pravilo sadrži kvalitetne, djelotvorne informacije. Otkrivanje obrazaca vodi ka spoznaji i sljedstvenoj akciji. Mnogi rezultati analize potrošačke košarice su često trivijalni ili neobjašnjivi. Trivijalna pravila su općepoznata, pa je uzaludno trošenje napora i vremena za njihovo otkrivanje, a neobjašnjiva pravila su nejasna i ne ukazuju na smjer djelovanja [3].

3. Otkrivanje pravila asocijacije

Otkrivanje pravila asocijacije počinje analizom transakcija koje sadrže jedan ili više proizvoda. Svaka od ovih transakcija pruža informacije o tome koji su proizvodi kupljeni i sa kojim drugim proizvodima. U otkrivanju pravila uzeta su u obzir tri važna aspekta: (1) izbor pravog skupa artikala, (2) generiranje pravila izračunom brojeva u matrici međusobnog pojavljivanja i (3) prevladavanje praktičnih ograničenja nametnutih velikim brojem (nekoliko tisuća ili desetaka tisuća) artikala.

Code	Product Name	Count	Probability	Rank
P1	MLIJEKO 2,8 KOZ.DUB.	11092	0,00658	1
P2	JELEN PIVO 0,5L	10905	0,00647	2
P3	VECERNJE NOVOSTI	9415	0,00559	3
P4	RUM PLOCICE 50	9047	0,00537	4
P5	MIN. VODA VITINKA 1,5L.	8679	0,00515	5
P6	PROLOM VODA 1,5	8497	0,00504	6
P7	KISELO VRHNJE ZA SIR 300 GR.	8313	0,00493	7
P8	MLIJEKO DOMACE TRAJNO 2.8%MM 1L.	8195	0,00486	8
P9	SMOKI SOKO STARK 50 GR.	7964	0,00473	9
P10	JOGURT DUKAT 0,2	6549	0,00389	10

Slika 2. Deset najprodavanijih artikala

U konkretnom primjeru izdvojeno je deset najprodavanijih artikala uopće, i to primjenom nekoliko faznih SQL upita i metodom kresanja (*pruning*). U prvom koraku je izdvojeno 100 artikala, u drugom u odgovarajuće polje dodan ukupan broj transakcija u bazi podataka, zatim je izračunata vjerojatnost pojavljivanja tih artikala u svim transakcijama. Potom je sljedećem koraku iz privremene tablice *temp_Top100_Products* upitom izdvojeno deset najprodavanijih artikala uopće:

U sljedećem kodu prikazan je akcijski SQL upit u kojem se kreira tabela *temp_Top100_Products*, te *select* upit gdje je odabrano prvih 10 artikala, isključujući, primjenom metode kresanja, neke artikle koji bi vodili otkrivanju trivijalnih pravila, ili pravila koja nemaju praktičnu primjenu.

```
SELECT TOP 100 [Inventory Transactions].ProductID, Products.ProductName,
Count([Inventory Transactions].TransactionID) AS CountOfTransactionID, CDbl(0)
AS Probability, CLng(0) AS CountOfInventoryTransactions
INTO temp_Top100_Products
FROM Products INNER JOIN (Izlazi INNER JOIN [Inventory Transactions] ON
Izlazi.IDIzlaza = [Inventory Transactions].IDIzlaza) ON Products.ProductID =
[Inventory Transactions].ProductID
GROUP BY [Inventory Transactions].ProductID, Products.ProductName, CDbl(0),
CLng(0)
ORDER BY Count([Inventory Transactions].TransactionID) DESC;

SELECT TOP 10 temp_Top100_Products.ProductID,
temp_Top100_Products.ProductName,
temp_Top100_Products.CountOfProductTransactions,
temp_Top100_Products.Probability
FROM temp_Top100_Products
WHERE (((Left([temp_Top100_Products].[ProductName],3))<>"CIG"))
ORDER BY temp_Top100_Products.CountOfProductTransactions DESC;
```

Otkrivanje pravila asocijacija u ovom primjeru seže do razine parova artikala, ali je analizu moguće proširiti na kombinacije više artikala, čime se eksponencijalno uvećava broj mogućih kombinacija. U odgovarajućoj programskoj proceduri izračunane su vrijednosti u tablici međusobnog pojavljivanja (*co-occurrance table*), koje kazuju koliko se puta svaki od tih parova proizvoda kupuje zajedno, na temelju koje je jednostavno izračunati vjerojatnost stavljanjem u odnos broja međusobnog pojavljivanja i ukupnog broja transakcija u bazi podataka (Slika 3).

1qryCoOccurrenceCrossTab

ProductName	JELEN PIVO	JOGURT DU	KISELO VRH	MLIJEKO 2,8	MLIJEKO DO	PROLOM	RUM PLOCI	SMOKI SOK	VECERNJE N	VITINKA 1,5
JELEN PIVO 0,5L		10	6	82	31	29	54	40	58	202
JOGURT DUKAT 0,2	10		68	105	58	154	36	40	274	54
KISELO VRHNJE ZA SIR 300 GR.	6	68		212	162	12	23	31	62	44
MLIJEKO 2,8 KOZ.DUB.	82	105	212		43	195	77	92	96	282
MLIJEKO DOMACE TRAJNO 2.8%MI	31	58	162	43		29	27	30	42	121
PROLOM VODA 1,5	29	154	12	195	29		30	31	71	46
RUM PLOCICE 50	54	36	23	77	27	30		125	12	44
SMOKI SOKO STARK 50 GR.	40	40	31	92	30	31	125		43	55
VECERNJE NOVOSTI	58	274	62	96	42	71	12	43		114
VITINKA 1,5L.	202	54	44	282	121	46	44	55	114	

Slika 3. Tablica međusobnog pojavljivanja

Ove jednostavne tablice pojavljivanja podcrtavaju i neke jednostavne obrasce:

- P1 i P5 imaju najveću vjerojatnost da se kupuju skupa u odnosu na bilo koja druga dva artikla,
- P2 i P7 se najmanje kupuju skupa

Artikal	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P1	0,17530	0,04175	0,01555	0,01421	0,03139	0,02887	0,00637	0,01214	0,01362	0,01140
P2	0,04175	0,14243	0,00800	0,01688	0,00651	0,00681	0,01792	0,02991	0,00814	0,00651
P3	0,01555	0,00800	0,11830	0,04057	0,01007	0,02280	0,00859	0,00148	0,00592	0,00533
P4	0,01421	0,01688	0,04057	0,11430	0,00918	0,01051	0,00622	0,00859	0,00637	0,00178
P5	0,03139	0,00651	0,01007	0,00918	0,09180	0,00178	0,02399	0,00089	0,00459	0,00341
P6	0,02887	0,00681	0,02280	0,01051	0,00178	0,08839	0,00429	0,00429	0,00459	0,00444
P7	0,00637	0,01792	0,00859	0,00622	0,02399	0,00429	0,08040	0,00459	0,00444	0,00400
P8	0,01214	0,02991	0,00148	0,00859	0,00089	0,00429	0,00459	0,07581	0,00592	0,00800
P9	0,01362	0,00814	0,00592	0,00637	0,00459	0,00459	0,00444	0,00592	0,07211	0,01851
P10	0,01140	0,00651	0,00533	0,00178	0,00341	0,00444	0,00400	0,00800	0,01851	0,06337

Tablica 1. Vjerojatnosti skupnog pojavljivanja

1. artikal	2. artikal	Count	Probability
P1	P5	282	0,04175
P3	P10	274	0,04057
P1	P7	212	0,03139
P2	P5	202	0,02991
P1	P6	195	0,02887
P7	P8	162	0,02399
P6	P10	154	0,02280
P4	P9	125	0,01851
P5	P8	121	0,01792
P3	P5	114	0,01688
a			

1. artikal	2. artikal	Count	Probability
P2	P7	6	0,00089
P2	P10	10	0,00148
P3	P4	12	0,00178
P6	P7	12	0,00178
P4	P7	23	0,00341
P4	P8	27	0,00400
P6	P8	29	0,00429
P2	P6	29	0,00429
P8	P9	30	0,00444
P4	P6	30	0,00444
b			

Tablica 2. Parovi artikala s najvećom (a) i najmanjom (b) vjerojatnosti skupnog pojavljivanja

Ova zapažanju su primjeri asocijacija i mogu sugerirati formalno pravilo poput „Ako kupac kupuje *P1*, onda kupuje i *P5*“. Postavljaju se dva pitanja: prvo, kako pronaći pravilo automatski i, drugo, koliko je kvalitetno svako pravilo? U podacima, x (282) od y (6.754) transakcija uključuju i *P1* i *P5*. Ove transakcije podupiru pravilo. Potpora za pravilo je x od y ili 3,14% (x/y *100). Budući da transakcije koje sadrže *P1* također sadrže i *P5*, postoji visok stupanj pouzdanosti u pravilo kao dobro. Pravilo "*ako P1, onda P5*" ima pouzdanost 3,14% (282/6.754*100). Dakle, pouzdanost je omjer broja transakcija koje podupiru pravilo i broja transakcija u kojima se nalazi uvjetni dio pravila (dio pravila iza riječi „*ako*“). Kazano na drugi način, pouzdanost je omjer broja transakcija sa svim artiklima i broja transakcije sa samo "*ako*" artiklima.
Izračun potpore i pouzdanosti brzo izmiče kontroli kako broj artikala u kombinacijama raste. Ima skoro 50 milijuna mogućih kombinacija od dva artikla u tipičnoj prodavnici namirnica i više od 100 milijardi kombinacija od tri artikla. Iako kompjuteri postaju sve snažniji i jeftiniji, još uvijek je potrebno mnogo vremena da bi se izvršilo brojanje u ovako velikom broju kombinacija. Izračun broja za pet ili više

artikala je nedopustivo skup. Uporaba hijerarhije proizvoda smanjuje broj artikala do razumne veličine. Broj transakcija je također vrlo velik. U tijeku jedne godine, lanac supermarketa pristojne veličine generirat će na desetke ili stotine milijuna transakcija. Svaka od tih transakcija se sastoji od jednog ili više artikala, često nekoliko desetaka u isto vrijeme. Dakle, utvrđivanje da li se određena kombinacija artikala pojavljuje u pojedinoj transakciji može zahtijevati poprilično napora i sve to pomnoženo s milijunom za sve transakcije [5].

4. Zaključak

Proučavanje pravila asocijacije u *data miningu* je popularna i dobra istraživačka metoda za otkrivanje zanimljivih relacija između pojedinih artikala u velikoj transakcijskoj bazi podataka. Otkrivanje artikala koji se pojavljuju skupa u potrošačkoj košarici vršeno je primjenom SQL upita i programskih algoritama. Analizirana baza podataka sadrži transakcijske podatke u oko 1.800.000 zapisa na prodajnom mjestu jednog trgovinskog poduzeća. Primjenom tehnika *data mininga,* kao i metode kresanja podataka, dobijena je skupina od deset artikala koji se najviše kupuju, kao i parovi artikala koji se najčešće kupuju skupa. Također su izračunate i vrijednosti potpore za pojedina pravila asocijacija koja se odnose na kombinacije parova artikala. Otkrivena pravila mogu imati praktičnu primjenu kod uređenja unutrašnjeg prostora prodajnog mjesta, u smislu razmještaja artikala, kao i za promidžbene aktivnosti kombiniranjem onih artikala koji se najčešće kupuju skupa.

5. Literatura

[1] Matanović, S. & al. (2010). Association rule mining in market basket data. Proceedings of the 2nd International conference "Vallis Aurea", pp. 821-827. ISBN 978-953-7744-06-9, ISBN 978-3-901509-76-6, Pozega – Vienna, Croatia – Austria.
[2] Omari, A. & Conrad, S. (2006). On the usage of data mining to support website designers to have better designed websites, Proceedings of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services, p. 171, ISBN: 0-7695-2522-9, Guadeloupe, French Caribbean, February 2006, IEEE Computer Society, Washington, DC, USA.
[3] Berry, M. J. A. & Linoff, G. S. (2004). *Data mining techniques for marketing, sales, and customer relationship management*, Wiley Publishing, Inc., 287-319, ISBN 0-471-47064-3, Indianapolis, Indiana, USA
[4] Ohsawa, Y. & Yada, K. ed. (2009). *Data Mining for Design and Marketing*. Chapman & Hall/CRC Taylor & Francis Group, ISBN 978-1-4200-7019-4, NW, Boca Raton, FL, USA.
[5] Ohsawa, Y. and McBurney, (2003). *Chance discovery, Advanced Information Processing*, Springer, ISBN-13: 978-3540005490, New York.

Photo 103. Truck / Kamion