FORUM STATISTICUM SLOVACUM Vol XVII (2021) No 2 pp 12-27

Aplikácia analýzy nákupného koša v oblasti CRM pomocou softvéru R
Application of market basket analysis in the field of CRM using the R software

Tadeáš Chujac, Martina Jantová, Iveta Stankovičová
Univerzita Komenského v Bratislave, Fakulta managementu, Odbojárov 10, 820 05 Bratislava, Slovenská republika
Comenius University in Bratislava, Faculty of Management, Odbojárov 10, 820 05 Bratislava, Slovak Republic
chujac1@uniba.sk, jantova9@uniba.sk, iveta.stankovicova@fm.uniba.sk

Abstrakt: *Pri rozhodovaní marketingových manažérov je veľmi dôležité vedieť odhadnúť nákupné správanie svojich zákazníkov. Aby informácie o zákazníkoch boli relevantné, je potrebné sa oprieť o dáta a vyťažiť z nich správne informácie, ktoré vedia firme pomôcť v konkurenčnom boji. Jednou z metód data miningu, ktorá dokáže správanie zákazníkov odhadnúť pomocou analýzy transakčných dát je analýza nákupného koša. Cieľom článku je popísanie využitia metódy analýzy nákupného koša v oblasti predaja elektro tovaru. Výsledky tejto analýzy vedia byť užitočné v oblasti CRM a pri rozhodovaní o marketingovej stratégii.*

Abstract: *When making marketing managers' decisions, it is very important to be able to estimate the purchasing behaviour of their customers. To be customer information relevant, it is necessary to rely on data and extract the right information from it, which can help the company in its competition. One of the data mining methods that can estimate customer behaviour using transaction data is market basket analysis. The aim of the article is to describe the use of the market basket analysis method in the field of electronics sales. The results of this analysis can be useful in the field of CRM and in deciding on marketing strategy.*

Kľúčové slová: *analýza nákupného koša, data mining, asociačné pravidlá, apriori algoritmus, transakčné dáta*

Key words: *market basket analysis, data mining, associations rules, apriori algorithm, transaction data*

1 Úvod

V dnešnom modernom svete firmy zhromažďujú veľké množstvá dát a informácií o svojich zákazníkoch, ktoré sú využiteľné v mnohých oblastiach vrátane marketingu či riadenia vzťahu so zákazníkmi (CRM). Pre získanie trhového podielu či zvýšenie tržieb firmy je nevyhnutné tieto dáta využiť čo najefektívnejšie. Na extrakciu skrytých a užitočných informácií z veľkých databáz sa používajú rôzne techniky data miningu. Následná analýza takto získaných informácií je nevyhnutná pre udržanie postavenia spoločnosti v konkurenčnom prostredí.

Analýza nákupného koša je jednou z najpopulárnejších metód hĺbkovej analýzy údajov (data miningu), ktorá sa zameriava na odhalenie „nákupných

vzorcov“ správania zákazníkov. Cieľom je identifikovať z transakčných dát zaujímavé asociácie medzi viacerými produktami, ktoré zákazníci často spolu nakupujú, čiže kladú spolu do nákupného košíka.

Cieľom príspevku je analyzovať transakčné dáta predajcu elektroniky a pomocou analýzy nákupného koša odhaliť, aké produkty si zákazníci kupujú spolu najčastejšie. Tieto informácie vedia byť nápomocné pre marketingových manažérov pri zvyšovaní efektivity marketingových stratégií. Článok je zároveň návodom, ako očistiť, spracovať a analyzovať transakčné dáta v bezplatnom štatistickom programe R.

V prvej časti je zhrnutý prehľad literatúry vo vzťahu k technikám analýzy nákupného koša. Následne je spísaná metodika spracovania príspevku, na ktorú nadväzujú výsledky analýzy nákupného koša na základe reálnych údajov predaja elektro tovaru. Na spracovanie údajov a výpočty bol použitý voľne dostupný štatistický softvér R.

2 Analýza nákupného koša

Cieľom analýzy nákupného koša je nájdenie asociačných pravidiel, teda častých vzorov pri nákupe. Ide o produkty, ktoré zákazníci často kupujú spoločne (Zendulka, 2006). Napríklad pri nákupe nového notebooku si zákazník často kúpi aj novú tašku, či myš, prípadne k mobilu si pravdepodobne dokúpi obal či ochranné sklo. Pri rozhodovaní v marketingu je analýza nákupného koša silným nástrojom na podporu implementácie stratégií krížového predaja. Napríklad, ak nákupný profil konkrétneho zákazníka zapadá do identifikovaného nákupného koša, navrhne sa ďalšia položka (Cavique, 2007). Na základe asociačných pravidiel vieme potom odhadnúť, že ak si zákazník kúpi varnú dosku a chladničku, pravdepodobne bude zariaďovať byt a vieme mu odporučiť aj umývačku riadu a tým zvýšiť tržby obchodu.

Pri využití tejto techniky sa vychádza z takzvaných transakčných dát, ktorých najvýznamnejšími atribútmi sú identifikácia transakcie a identifikácia položiek v transakcii. Položky sú zvyčajne napojené na tabuľku charakteristík produktov. Nákup sa viaže na určitého zákazníka (ID zákazníka), ktorý môže mať aj niekoľko zákaziek a preto sa ID zákazníka v transakčných dátach opakuje na viacerých riadkoch dátovej tabuľky (Stankovičová, 2009).

Táto analýza si našla uplatnenie v rôznych oblastiach manažmentu, najmä marketingu a riadenia vzťahov so zákazníkmi (CRM). Keďže CRM sa zameriava na zvyšovanie spokojnosti zákazníkov ale aj zvyšovanie rozsahu podnikania (Osmanbegovic, 2015), práve analýza nákupného koša vie odporučiť zákazníkovi produkty, o ktoré by mohol mať reálny záujem a tým zvýšiť aj tržby obchodu.

2.1 Asociačné pravidlá

Zďaleka najbežnejšou praxou v analýze nákupného koša je identifikácia asociačných pravidiel (Brijs, 2004). Asociačné pravidlá reprezentujú schémy, resp. vzory v dátach bez toho, aby bola definovaná závislá premenná (Stankovičová, 2009). Tieto pravidlá môžu maloobchodníkom pomôcť pochopiť nákupnú povahu zákazníkov (Manpreet, 2016). Každá dvojica produktov A a B (pričom A a B môže označovať aj viacero položiek) sa hodnotí na základe troch parametrov (Kamakura, 2012):

- ***Support (podpora)*** – spoločná pravdepodobnosť nájdenia páru tovarov A a B vo všetkých košoch, čiže $\mathrm{P}(\mathrm{A} \cap \mathrm{B})$. Ak hodnota vyjde nízka znamená to, že pár A a B nie je relevantný, pretože sa táto kombinácia produktov nenakupuje dostatočne často.
- ***Confidence (spoľahlivosť)*** – vyjadruje pravdepodobnosť, že nákup produktu A povedie k nákupu aj produktu B. Ide o podmienenú pravdepodobnosť, čiže miera sa vypočíta podľa vzťahu: $\mathrm{P}(\mathrm{B}|\mathrm{A}) = \mathrm{P}(\mathrm{A} \cap \mathrm{B})/\mathrm{P}(\mathrm{A})$
- ***Lift (miera závislosti)*** - pomer medzi spoločnou pravdepodobnosťou produktov A a B v košíku a pravdepodobnosťou spoločného výskytu pri ich nezávislosti a vzorec je nasledovný: $\mathrm{P}(\mathrm{A} \cap \mathrm{B})/(\mathrm{P}(\mathrm{A}) \times \mathrm{P}(\mathrm{B}))$. Tento parameter udáva mieru závislosti medzi dvoma produktami A a B. Hodnota tohto parametru hovorí, či je závislosť pozitívna (hodnota lift je väčšia ako 1), negatívna (hodnota lift je menšia ako 1) alebo sa závislosť nevyskytuje (hodnota lift je rovná 0).

Firmám sa oplatí pri nákupe odporučiť produkt B tým zákazníkom, ktorí si práve pridali produkt A do svojho nákupného košíka, ak majú produkty A a B vysokú tendenciu vyskytovať sa v košíku spoločne (t.j. hodnota support je vysoká) a taktiež je vysoká pravdepodobnosť, že nákup produktu A povedie k nákupu produktu B (t.j. vysoká hodnota confidence) (Kamakura, 2012).

V neposlednom rade dôležitým parametrom, ktorý napovie, či sa firme oplatí odporučiť zákazníkovi pri nákupe produktu A aj produkt B, je hodnota lift. Ak vyjde táto hodnota vyššia ako 1, napovie to, že miera závislosti medzi produktami A a B je u daného zákazníka vysoká (napr. hodnota lift = 2 znamená, že u zákazníka, ktorý si kúpil produkt A, je až dvakrát vyššia pravdepodobnosť kúpy aj produktu B ako u iných zákazníkov). Závislosť medzi produktmi však môže byť i negatívna (vtedy hodnota lift je nižšia ako 1), prípadne závislosť nemusí existovať, ak hodnota lift je rovná nule (Stankovičová, 2009).

2.2 Apriori algoritmus

Algoritmus apriori slúži k vyhľadávaniu základných asociačných pravidiel, pričom berie do úvahy všetky transakcie v databáze pri definovaní nákupného

koša. Ako vstup používa tento algoritmus tabuľku s transakciami, pričom každá obsahuje svoju identifikáciu a nakupované položky nasledovne (transakcia, položka) (Cavique, 2007).

Pri tejto metóde vstupujú tri parametre:

- maximálny počet položiek v nákupnom košíku (max_n),
- minimálna podpora (min_sup),
- minimálna spoľahlivosť (min_conf).

Apriori algoritmus postupne generuje množiny, v ktorých sa nachádzajú frekventované položky, ktoré sa často kupujú spolu, pričom algoritmus postupuje od množín s najmenším počtom prvkov k tým najväčším (veľkosti množín sú ohraničené hodnotou max_n). Algoritmus používa apriori vlastnosť frekventovaných množín: Všetky neprázdne podmnožiny frekventovanej množiny položiek sú (musia byť) tiež frekventované. Súbor frekventovaných množín, ktoré majú n prvkov sa označuje L_n. Následne algoritmus odfiltruje množiny, ktoré nesĺňajú minimálnu hodnotu podpory (min_sup). Pokiaľ je to možné, z L_n množín sa generujú L_{n+1} množiny s n+1 prvkami (až kým n nedosiahne max_n). Výstupom algoritmu sú všetky silné n-množiny, z ktorých je možné generovať asociačné pravidlá, ktoré spĺňajú minimálnu hodnotu podpory (min_sup) a spoľahlivosti (min_conf), pričom pre každé z pravidiel vypočíta aj hodnoty support a confidence.

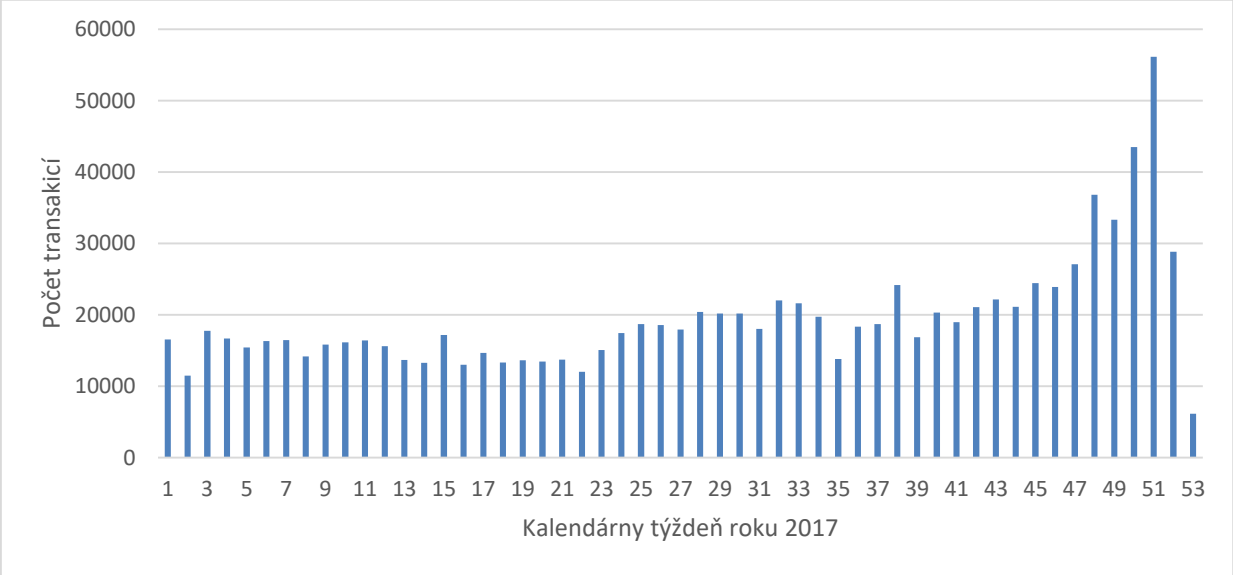
Za výhody Apriori algoritmu možno považovať množstvo vzorov, ktoré sú výstupom a zároveň sú tieto vzory ľahko pochopiteľné a interpretovateľné. Za nevýhodu možno považovať dlhší výpočtový čas.

3 Metodológia a popis dátového súboru

Analyzovaný dátový súbor, na ktorom budeme demonštrovať možnosti využitia analýzy nákupného koša pozostával z reálnych transakcií, resp. nákupov uskutočnených u jedného z najväčších predajcov elektroniky v Slovenskej republike za rok 2017. Dáta aj napriek staršiemu dátumu neznižujú relevantnosť výstupov, keďže nákupné správanie spotrebiteľov, týkajúce sa asociácií produktov, ktoré sú kupované spoločne, sa za ostatné 4 roky výrazne nezmenilo.

Pre analýzu sme spracovali dátový súbor s viac ako 1 milión transakciami, v ktorých viac ako 400 tisíc rôznych zákazníkov s vernostnými kartami zakúpilo takmer 1,7 milióna položiek. Na základe týchto transakčných údajov bolo prostredníctvom tohto výskumu možné simulovať reálne prostredie z praxe, keďže dostupné dáta netvorili vzorku transakcií, ale celú populáciu transakčných údajov za rok 2017 (tab. 1).

Spomínaný obchod s elektro tovarom predáva takmer 16 tisíc unikátnych tovarov. Predajca má vo svojom portfóliu produktov nielen elektroniku, ale aj drogériu, potreby do domácnosti, či športové potreby. Napriek tomu, priemerne sa v jednej transakcii nachádza iba 1,64 položky a dokonca viac ako 62 % transakcií tvoria iba jednopoložkové nákupy (tab. 1). Takúto štruktúru počtu tovarov v jednotlivých transakciách sme však čakali vzhľadom na to, že oblasť elektroniky sa nevyznačuje vysokým počtom položiek v jednej transakcii, keďže väčšina tovarov patrí medzi drahšie tovary, elektroniku si zákazníci kupujú s nižšou frekvenciou ako iné lacnejšie produkty.



Graf 1 Počet transakcií uskutočnených v jednotlivých týždňoch roka 2017 (*Zdroj*: *vlastné spracovanie*)

Tab. 1 Počet tovarov v jednotlivých transakciách *(Zdroj: vlastné spracovanie)*

Počet produktov v transakcii	**Počet transakcií**	**f_i**
1	644605	62,40%
2	236912	22,93%
3	87075	8,43%
4	36452	3,53%
5	14840	1,44%
6	6525	0,63%
7	3092	0,30%
8	1616	0,16%
9	773	0,07%
10 a viac	1201	0,12%
SPOLU	**1033091**	**100,00%**

Keďže predmetom výskumu je analýza nákupného koša, ktorá má za cieľ skúmať vzťahy medzi položkami, ktoré boli do „košíka“ vkladané v rámci jednej

transakcie, pre vyššiu výpovednú hodnotu uskutočnenej analýzy bolo potrebné transakčné dáta upraviť. Po bližšom preskúmaní dátového súboru sme zistili, že v analyzovaných transakčných dátach sa pomerne často vyskytuje skutočnosť, že daný zákazník uskutoční v ten istý deň viacero nákupov. Takmer 100 tisíc zákazníkov urobilo v deň nákupu aj nejaký iný nákup pod odlišným ID nákupu, preto tieto nákupy neregistrujeme v databáze ako jeden (tab. 2).

Tab. 2 Počet transakcií na jedného zákazníka počas jedného dňa *(Zdroj: vlastné spracovanie)*

Počet transakcií zákazníka v jeden deň	n_i
1	819600
2	79421
3	10853
4 a viac	4311

Nákup elektroniky býva často ale spojený aj s dodatočným nákupom rôzneho príslušenstva. Pri mobile si často zákazník uvedomí až dodatočne, že mu napríklad do balenia nepribalili nabíjačku, slúchadlá alebo mobil používa iné zakončenie, ktoré nie je kompatibilné s vybavením, ktoré zákazník vlastní z predchádzajúceho modelu. S tým sú spojené dodatočné nákupy, ktoré obsahujú položky nevyhnutné pre to, aby zakúpený tovar mohol zákazník využívať naplno. Preto sme sa pozreli aj na to, či daný zákazník spravil nejaký nákup aj v rozsahu ďalších 7[1] dní od prvotného nákupu. Nakoniec sa ukázalo, že dokonca až pri viac ako 30 % z celkového počtu transakcií existuje takáto príbuzná transakcia, ktorú urobil daný zákazník do 7 dní. Tieto informácie sú práve pre analýzu nákupného koša veľmi cenné, keďže jedným z využití tejto metódy v praxi je ponúknuť zákazníkovi ďalšie vhodné produkty, ktoré by si mohol k už existujúcemu produktu v nákupnom koši prikúpiť a nedovoliť tak, aby si príslušenstvo zakúpil u konkurenčných predajcov elektroniky. Tieto transakcie, ktoré boli uskutočnené zákazníkov v rozmedzí maximálne 7 dní boli spojené do jednej, čím sme redukovali počet unikátnych transakcií a zvýšili sme počet produktov v jednom nákupe.

Výsledný dátový súbor pozostáva zo 6 premenných: ID_zákazníka, ID_nákupu, Dátum_nákupu, ID_tovaru, Kategória_tovaru (tab. 3). Zahŕňa v sebe viac ako 730 tisíc transakcií, v ktorých bolo zakúpených takmer 1,5 milióna tovarov. Pokles v počte transakcií a produktov je spôsobený spájaním viacerých transakcií do

[1] Voľba 7 dní je subjektívneho charakteru, keďže bolo v úmysle zachytiť produkty, ktoré bezprostredne nasledovali po kúpe iných produktov. Nebol preto zvolený dlhší interval, napr. 30 dní, lebo pri takejto dĺžke intervalu už môžu byť vo väčšej miere zachytené aj také nákupy, ktoré spolu nijakým spôsobom nesúvisia. Na druhej strane nebol zvolený ani kratší interval, keďže pri elektronike, ktorá je zložitejšia na používanie môže zákazníkom trvať aj niekoľko dní, kým si uvedomia, že dané príslušenstvo k nej potrebujú. Preto voľba intervalu 7 dní bola kompromisná voľba, ktorá by mala zachytiť práve tie nákupy, ktoré medzi sebou súvisia.

jednej a v rámci jedného nákupu postačuje, aby sa daný tovar v nej nachádzal iba jedenkrát. Štatistická jednotka výsledného datasetu reprezentuje jeden zakúpený tovar.

Tab. 3 Príklad štruktúry výsledného dátového súboru *(Zdroj: vlastné spracovanie)*

ID_zákazníka	ID_nákupu	Dátum_nákupu	ID_tovaru	Kategória_tovaru
2358304	94417219	2017-09-12	2002144	Hry na herné konzoly
2358304	94417219	2017-09-12	2002142	Herné konzoly
2355514	94417064	2017-09-12	1106806	Pamäťové karty

Prostredníctvom zlučovania transakcií s jedným produktom do početnejších skupín sa podarilo znížiť zastúpenie jednopoložkových transakcií o viac ako 9 %. Priemerný počet produktov na jednu transakciu dosiahol hodnotu 2, hoci mediánový počet produktov v jednej transakcií zostal stále na hodnote 1.

Tab. 4 Počet tovarov v jednotlivých transakciách po úprave datasetu *(Zdroj: vlastné spracovanie)*

Počet produktov v transakcii	Počet transakcií	f_i
1	388690	53,07%
2	180251	24,61%
3	82143	11,22%
4	39303	5,37%
5	18815	2,57%
6	9484	1,29%
7	4992	0,68%
8	2807	0,38%
9	1550	0,21%
10 a viac	4348	0,59%

Celú analýzu sme vykonávali v prostredí bezplatného štatistického nástroja R s využitím knižníc pre manipuláciu s dátami *tidyverse* a *data.table* a knižnicou *arules* určenou v prostredí softvéru R pre analýzu nákupného košíka, manipuláciu s transakčnými dátami a pre tvorbu asociačných pravidiel. Výsledné grafy a tabuľky boli spracované pomocou MS Office.

4 Analýza nákupného koša v R

Analýza nákupného koša v prostredí štatistického softvéru R vyžaduje ešte jeden krok v rámci prípravy. Analyzované dáta je nevyhnutné pretransformovať do formátu transakčných dát, v ktorých dôležitými premennými už budú iba *ID_nákupu* a *ID_tovaru*, resp. kategória tovaru, ak sa analýza nákupného koša robí na viacerých úrovniach databázy produktov. Vyššie spomínaný finálny dátový súbor *„transakcie“* po všetkých úvodných úpravách, ktorý mal pre

potreby úprav štruktúru „data.frame“ sme nasledujúcim príkazom modifikovali na transakčné dáta:

>trans<-as(split(transakcie[,ID_tovaru], transakcie[,ID_nákupu]), 'transactions').

Na rozdiel od finálneho dátového súboru, kde jeden riadok označoval jednu zakúpenú položku, tak v prípade transakčných dát jeden riadok označuje jednu faktúru, resp. jeden nákup a následne zakúpené tovary sú zapísané v množinovej zátvorke s oddeľovačom v podobe čiarky (obr. 1). Takže vieme povedať, že v rámci transakcie číslo 12101449, ktorá má v transakčných dátach poradové číslo 1 boli zakúpené 3 rôzne položky s číslami 1006687, 1020027 a 1020028, ktoré označujú postupne vrecká do vysávača, tablety na vodný kameň a čistiace tablety.

```
> inspect(head(trans, 10))
      items                          transactionID
[1]   {1006687,1020027,1020028}      12101449
[2]   {1076055,1106806}              12101455
[3]   {1065481}                      12101467
[4]   {1104819}                      12101468
[5]   {1075944,2000400}              12101473
[6]   {1108586}                      12101483
[7]   {1049584,1052336}              12101484
[8]   {1020029,1106702}              12101492
[9]   {1114860}                      12101497
[10]  {1033932,1107250}              12101501
```

Obr. 1 Štruktúra transakčných údajov v prostredí softvéru R (*Zdroj: vlastné spracovanie*)

```
> summary(trans)
transactions as itemMatrix in sparse format with
 732383 rows (elements/itemsets/transactions) and
 15919 columns (items) and a density of 0.0001256365

most frequent items:
1025665 1102192 1085679 2009031 1011730 (Other)
  19528   13519   11549    9287    8527 1402361

element (itemset/transaction) length distribution:
sizes
     1      2      3      4      5      6      7      8      9     10     11     12     13     14     15     16     17     18     19     20     21     22     23     24
388695 180246  82144  39304  18814   9484   4991   2807   1551    976    623    437    278    236    185    159    117     96     85     75     74     60     59     40
    25     26     27     28     29     30     31     32     33     34     35     36     37     38     39     40     41     42     43     44     45     46     47     48
    37     44     36     35     37     30     28     24     24     23     20     16     19     19     14     18     21     12     18     16     13      7     17     12
    49     50     51     52     53     54     55     56     57     58     59     60     61     62     63     64     65     66     67     68     69     70     71     72
     9      8     12     13     13      5     13     10      8      5      7     11     10      7      3      6      1      6      6      5      4      5      5      3
    73     74     75     76     77     78     79     80     81     82     83     84     85     86     87     88     89     90     91     92     93     94     96     97
     3      6      3      5      7      2      3      3      6      5      2      2      5      1      2      1      3      1      2      1      4      2      1      2
    98    102    103    104    105    106    108    110    111    113    114    115    118    119    124    126    127    130    131    133    135    139    140    142
     3      2      1      2      2      2      1      1      2      1      2      3      1      1      1      2      2      1      1      1      1      1      2      1
   143    144    149    150    151    155    158    160    178    183    186    197    203    207    209    217    229    255    396    598
     1      1      2      1      1      1      1      2      1      2      1      1      1      1      1      1      1      1      1      1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      1       1       1       2       2     598

includes extended item information - examples:
   labels
1 1000582
2 1000584
3 1000594

includes extended transaction information - examples:
  transactionID
1      12101449
2      12101455
3      12101467
```

Obr. 2 Informácie o transakčných dátach (*Zdroj: vlastné spracovanie*)

Transakčné dáta sú v pamäti uložené prostredníctvom riedkej matice, ktorej počet riadkov označuje počet unikátnych transakcií a počet stĺpcov označuje

počet unikátnych tovarov nachádzajúcich sa v analyzovaných transakciách. Pomocou funkcie *summary()* môžeme v R-ku získať informácie o tom, že v našich transakčných dátach sa nachádza presne 732 383 transakcií obsahujúcich 15 919 jedinečných tovarov. Päť najčastejšie zakúpených položiek tvoria tovary ako Sodastream náplne, vrecká do vysávača, či služba lepenia fólie na smartfón. V dátach sa nachádzajú aj „extrémne" faktúry, ktoré obsahovali viac ako 200 tovarov, či dokonca až 598. Tieto extrémy sú spojené so zlučovaním transakcií a nami nastavené pravidlá zachytili aj niekoľkých veľkoodberateľov, ktorí spôsobujú isté skreslenie dát. Nakoniec súčasťou výstupu sú aj popisné štatistiky, ktoré už boli vyššie popísané ako priemer, či medián (obr. 2).

4.1 Dolovanie asociačných pravidiel z jednoduchých transakčných dát v R

Ďalším krokom po príprave transakčných dát pri analýze nákupného koša v R je ich vloženie do funkcie *apriori().* Apriori algoritmus sa používa na generáciu asociačných pravidiel, ktoré by mali odhaliť skryté vzorce v analyzovaných dátach. Táto funkcia má v R-ku aj svoje základné nastavenie, kedy ako parameter do funkcie stačí vložiť iba transakčné dáta. Ako je možné vidieť na obrázku nižšie (obr. 3), základné nastavenia obsahujú prednastavené hodnoty pre nasledovné parametre (Hahsler, 2006):

- *confidence* – minimálna hodnota confidence (preddefinovaná hodnota = 0.8)
- *smax* – maximálna hodnota support (preddefinovaná hodnota = 1)
- *support* – minimálna hodnota support (preddefinovaná hodnota = 0.1)
- *minlen* – minimálny počet položiek v transakcii (preddefinovaná hodnota = 1)
- *maxlen* – maximálny počet položiek v transakcii (preddefinovaná hodnota = 10)

Najproblematickejším zo základných nastavení, ktoré nedovolilo vygenerovanie ani jedného asociačného pravidla je nastavenie úrovne podpory (support) až na úrovni 0,1. Znamená to, že aby vôbec pravidlo mohlo byť vygenerované, položky sa musia spolu nachádzať aspoň na 10 % transakcií, čo by bolo v našom prípade aspoň 73 238 faktúr. Táto hranica podpory je v prípade tak veľkých databáz produktov, aké máme my k dispozícií na úrovni jednotlivých produktov (takmer 16 000 produktov) veľmi ťažko dosiahnuteľná a tým pádom si to vyžaduje buď zníženie tejto hranice podpory alebo zoskupenie viacerých produktov do kategórií (pozri podkapitola 4.2). Obr. 2 poskytuje základný prehľad o tom, na akú približnú úroveň by sme mali úroveň podpory minimálne posunúť. Keďže vieme, že piaty najpredávanejší produkt sa vyskytoval v transakciách 8 527-krát, support je na úrovni 8527/732383 = 0,01164, čiže vrecká do vysávača boli súčasťou košíka v 1,164 % transakcií.

```
> rules <- apriori(trans)
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
        0.8    0.1    1 none FALSE            TRUE       5     0.1      1     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 73238

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[15919 item(s), 732383 transaction(s)] done [0.85s].
sorting and recoding items ... [0 item(s)] done [0.02s].
creating transaction tree ... done [0.02s].
checking subsets of size 1 done [0.00s].
writing ... [0 rule(s)] done [0.00s].
creating S4 object  ... done [0.03s].
```

Obr. 3 Výsledky apriori algoritmu so základným nastavením (*Zdroj: vlastné spracovanie*)

Za účelom získania vyššieho počtu asociačných pravidiel sú v ďalšom kroku pozmenené 2 parametre, a to hladiny podpory a spoľahlivosti na hodnoty 0,0001, resp. 0,5. S takýmito vstupnými parametrami apriori algoritmus vygeneroval 190 pravidiel a celkové spracovanie dát aj napriek ich veľkému rozsahu trvalo iba niekoľko sekúnd (obr. 4).

```
> rules2 <- apriori(trans, parameter = list(supp = 0.0001, conf = 0.5))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
        0.5    0.1    1 none FALSE            TRUE       5   1e-04      1     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 73

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[15919 item(s), 732383 transaction(s)] done [1.49s].
sorting and recoding items ... [4155 item(s)] done [0.05s].
creating transaction tree ... done [1.20s].
checking subsets of size 1 2 3 4 done [0.29s].
writing ... [190 rule(s)] done [0.05s].
creating S4 object  ... done [0.25s].
```

Obr. 4 Výsledky apriori algoritmu s pozmenenými parametrami (*Zdroj: vlastné spracovanie*)

Prostredníctvom funkcie *summary()* je možné zobraziť základné informácie a štatistiky o vygenerovaných asociačných pravidlách. Na základe výstupu tejto funkcie je možné zistiť, že 88 pravidiel je zložených z 2 položiek, čiže aj na ľavej a aj pravej strane je iba 1 položka, 94 pravidiel má na ľavej strane 2 položky a na pravej 1 a 8 pravidiel je zložených až zo 4 položiek, čiže 3 položky na ľavej strane

a 1 na pravej[2]. Dostupné sú na obr. 5 aj popisné štatistiky piatich mier kvality asociačných pravidiel (hladina podpory, spoľahlivosti, hladina podpory ľavej strany (coverage), miera závislosti (lift), počet).

```
> summary(rules2)
set of 190 rules

rule length distribution (lhs + rhs):sizes
 2  3  4
88 94  8

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   3.000   2.579   3.000   4.000

summary of quality measures:
    support           confidence        coverage             lift             count
 Min.   :0.0001010   Min.   :0.5000   Min.   :0.0001024   Min.   :  32.04   Min.   :  74.00
 1st Qu.:0.0001150   1st Qu.:0.5758   1st Qu.:0.0001833   1st Qu.: 157.67   1st Qu.:  84.25
 Median :0.0001700   Median :0.7455   Median :0.0002342   Median : 529.98   Median : 124.50
 Mean   :0.0002706   Mean   :0.7500   Mean   :0.0003792   Mean   :1377.64   Mean   : 198.21
 3rd Qu.:0.0002850   3rd Qu.:0.9226   3rd Qu.:0.0004287   3rd Qu.:1741.92   3rd Qu.: 208.75
 Max.   :0.0032551   Max.   :1.0000   Max.   :0.0063423   Max.   :9508.13   Max.   :2384.00

mining info:
  data ntransactions support confidence
 trans        732383   1e-04        0.5
```

Obr. 5 Popisné štatistiky vygenerovaných asociačných pravidiel (*Zdroj: vlastné spracovanie*)

Jednotlivé asociačné pravidlá je možné zobraziť v prostredí R zoradené podľa jednotlivých mier kvality asociačných pravidiel, napríklad miery závislosti (obr. 6). Prvé a druhé pravidlo hovorí o tom, že ak si niekto kúpi HP azúrovú, (resp. žltú) atramentovú kazetu do tlačiarne, tak má 9508-krát vyššiu pravdepodobnosť, že si taktiež zakúpi HP žltú, (resp. azúrovú) atramentovú kazetu do tlačiarne. Tretie pravidlo sa tiež týkalo náplní do tlačiarní, konkrétne, že ak zákazník nakúpi EPSON čiernu, žltú a purpurovú atramentovú náplň do tlačiarne, tak má až 5126-krát vyššiu pravdepodobnosť, že si zakúpi aj azúrovú náplň rovnakej značky. Tieto 3 pravidlá môžeme všetky zaradiť medzi asociačné pravidlá, ktoré sú samozrejmé, keďže často tlačiarne vyžadujú výmeny všetkých náplní naraz a preto ich zákazníci často nakupujú spoločne v jednom nákupe.

```
> inspect(head(rules2, by = "lift", 3))
    lhs                           rhs       support      confidence coverage     lift     count
[1] {2005638}                  => {2005636} 0.0001010400 0.9866667  0.0001024054 9508.130 74
[2] {2005636}                  => {2005638} 0.0001010400 0.9736842  0.0001037708 9508.130 74
[3] {1056942,1056944,1056945} => {1056943} 0.0001024054 0.9868421  0.0001037708 5125.861 75
```

Obr. 6 Tri najlepšie asociačné pravidlá na základe najvyššej miery závislosti (*Zdroj: vlastné spracovanie*)

[2] Softvér R a funkcia apriori() z knižnice arules generuje iba také asociačné pravidlá, ktoré majú na pravej strane len 1 položku.

Asociačné pravidlá založené na jednoduchých transakčných dátach majú niekoľko nevýhod najmä pri databázach zložených z veľkého množstva produktov. Hodnotu hladiny podpory je potrebné znížiť na veľmi nízke hodnoty, aby aspoň niektoré položky z databázy spĺňali dané kritérium. Vzhľadom na to, že v našich transakčných dátach sa nachádza skoro 10 tisíc produktov, ktoré sa za celý rok predali maximálne 10-krát, tak je potrebné pre lepšiu analýzu nákupného koša vytvoriť hierarchické transakčné dáta.

4.2 Dolovanie asociačných pravidiel z hierarchických transakčných dát v R

Hierarchické transakčné dáta obsahujú nielen ID tovaru a nákupu, ale k ID tovaru je priradená aj kategória, do ktorej tento produkt patrí. V analyzovanej databáze predajcu elektroniky sa nachádzajú 4 preddefinované úrovne kategórií, do ktorých patria jednotlivé produkty. Napr. LED TV patrí do kategórie Čierna elektronika -> TV -> LED TV -> TV LED 52“ + 01. Takže postupne je produkt zaraďovaný do menších a konkrétnejších kategórií. Nevýhodou jednoduchých transakčných údajov v predchádzajúcej podkapitole boli málopočetné, resp. málo zastúpené produkty v transakciách. Tento istý problém nastáva ale aj pri niektorých kategóriách produktov, ktoré nie sú častým artiklom v nákupných košíkoch zákazníkov. Vtedy je potrebné zlučovanie kategórií do väčších, aby sme dosiahli rovnomernejšie prerozdelenie tovarov v jednotlivých kategóriách. Vďaka tejto úprave sa podarilo obmedziť počet kategórií, v ktorých bol zakúpený iba jeden tovar. Navyše touto úpravou sa aj znížil počet jednotlivých kategórií z 488 na 226 (tab. 5).

Tab. 5 Popisné štatistiky početností nákupu pôvodných a novovytvorených kategórií *(Zdroj: vlastné spracovanie)*

Popisná štatistika	Pôvodné	Nové
Priemer	3001,60	6481,34
Medián	589	2732
Modus	1	833
Štandardná odchýlka	8130,44	11646,78
Roztyp	66104073,68	135647376,30
Rozpätie	95760	100634
Minimum	1	73
Maximum	95761	100707
Sum	1464782	1464782
Počet	488	226
5 - ty najväčší	34969	34969
5 - ty najmenší	1	194

Po pretransformovaní do štruktúry transakčných dát, s ktorými pracuje knižnica *arules* je prvým krokom skontrolovanie popisných štatistík analyzovaných dát. Najčastejšie zastúpenými kategóriami v transakciách sú

Batérie, Káva, Vrecká do vysávačov, Mobilné telefóny a Služba poistenie. V transakciách sa najčastejšie nachádzajú produkty iba jednej kategórie, čo je spôsobené najmä tým, že celkovú prevahu v transakciách majú jednopoložkové faktúry.

```
> summary(trans_multi)
transactions as itemMatrix in sparse format with
 732383 rows (elements/itemsets/transactions) and
 226 columns (items) and a density of 0.007804913

most frequent items:
            Batérie               Káva Vrecká do vysávačov    Mobilné telefóny    Služba poistenie             (Other)
              77686              56466                48089               46669               34143            1028805

element (itemset/transaction) length distribution:
sizes
     1      2      3      4      5      6      7      8      9     10     11     12     13     14     15     16     17     18     19     20     21
425640 180904  71393  29475  12358   5456   2677   1481    735    455    319    219    165    126     96     75     67     61     54     56     43
    22     23     24     25     26     27     28     29     30     31     32     33     34     35     36     37     38     39     40     41     42
    35     37     36     34     41     34     19     28     17     20     17     15     13     20     14     10     10      9      4      7     11
    43     44     45     46     47     48     49     50     51     53     54     55     56     57     58     59     60     61     62     64     65
     7      2      6      3      7      2      6      4     11      2      4      2      1      5      3      4      1      1      2      1      2
    66     67     68     70     72     73     74     75     79     80     82     84     89     94     99    114    132
     2      1      3      1      1      1      2      1      1      1      1      1      1      1      1      1      1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.764   2.000 132.000

includes extended item information - examples:
                labels
1          Alkoholtestery
2   Atramentové tlačiarne
3 Auto zvukové komponenty

includes extended transaction information - examples:
  transactionID
1      12101449
2      12101455
3      12101467
```

Obr. 7 Popisné štatistiky hierarchických transakčných dát (*Zdroj: vlastné spracovanie*)

Asociačné pravidlá budeme hľadať s pozmenenými parametrami oproti základnému nastaveniu. Najskôr je potrebné znížiť hladinu podpory. Napriek tomu, že zlúčením sme získali početnejšie kategórie, hladina podpory na úrovni 0,1 by znamenala, že by sa daná kategória musela vyskytovať v transakčných dátach aspoň 73 238. Z obr. 7 však vieme, že túto podmienku spĺňa iba kategória batérií. Taktiež je potrebné znížiť aj hladinu spoľahlivosti, keďže z popisných štatistík a charakteru transakčných dát vieme, že väčšinou sa produkty nachádzajú samostatne a iba v menšine sa vyskytujú v analyzovaných košíkoch 2 produkty a viac. Posledným nastaveným parametrom, ktorý budeme využívať bude maximálna dĺžka asociačného pravidla. Apriori algoritmus za účelom nájdenia čo najvyšších hodnôt hladín podpory, spoľahlivosti, či miery závislosti bude hľadať zložité pravidlá zložené z niekoľkých produktov na ľavej strane, ktoré by mali podmieňovať kúpu produktu na pravej strane asociačného pravidla. Avšak, často v praxi marketingoví manažéri nehľadajú zložité vzory, ale jednoduché pravidlá, ktoré sú logické a budú mať za následok zvýšenie predaja. Preto sme maximálnu dĺžku pre účely výskumu znížili na maximálne 3 položky, čiže vo výsledku najdlhšie vydolované asociačné pravidlo bude mať maximálne 2 položky na ľavej strane a jednu na pravej. Apriori algoritmus s takto nastavenými pravidlami vygeneroval 126 pravidiel. Jednotlivé asociačné pravidlá je potom potrebné manuálne prezrieť a overiť, či sa v nich nenachádzajú aj

nejaké nelogické asociácie, ktoré mohli vzniknúť iba náhodným výskytom daných položiek spolu v jednom košíku (obr. 8).

```
> rules_multi <- apriori(trans_multi, parameter = list(supp = 0.001, conf = 0.2, maxlen = 3))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
        0.2    0.1    1 none FALSE            TRUE       5   0.001      1      3  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 732

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[226 item(s), 732383 transaction(s)] done [0.22s].
sorting and recoding items ... [183 item(s)] done [0.02s].
creating transaction tree ... done [0.60s].
checking subsets of size 1 2 3 done [0.02s].
writing ... [126 rule(s)] done [0.00s].
creating S4 object  ... done [0.11s].
```

Obr. 8 Výsledky apriori algoritmu s hierarchickými transakčnými dátami (*Zdroj: vlastné spracovanie*)

Ako môžeme vidieť na obr. 9, tak všetkých 10 najlepších asociačných pravidiel z pohľadu miery závislosti sa týka kuchynských spotrebičov a ich rôznych kombinácií. Všetky v podstate hovoria o tom, že ak si zákazník nakúpi nejaký kuchynský spotrebič, resp. ich kombináciu, tak je vysoko pravdepodobné, že si v tom momente zariaďuje novú kuchyňu a je potrebné mu odporučiť v nákupe ďalšie kuchynské spotrebiče. Napríklad prvé asociačné pravidlo hovorí o tom, že zákazník, ktorý si nakúpi tovar z kategórie odsávače a rúry, tak má až takmer 104-krát vyššiu pravdepodobnosť, že si kúpi aj varnú dosku oproti náhodne vybranému zákazníkovi. Nákupný košík s 3 položkami z týchto kategórií sa nachádzal v transakčných dátach až 1253-krát.

```
> inspect(head(rules_multi, by = "lift", 10))
     lhs                                  rhs                 support     confidence coverage    lift      count
[1]  {Odsávače,Rúry}                   => {Varné dosky}       0.001710853 0.8471941  0.002019435 103.80969 1253
[2]  {Rúry,Umývačky riadu}             => {Varné dosky}       0.001721777 0.8373174  0.002056301 102.59947 1261
[3]  {Chladničky,Rúry}                 => {Varné dosky}       0.001516966 0.8097668  0.001873337  99.22359 1111
[4]  {Mikrovlnné rúry,Rúry}            => {Varné dosky}       0.001581140 0.7882914  0.002005781  96.59213 1158
[5]  {Mikrovlnné rúry,Varné dosky}     => {Rúry}              0.001581140 0.8674157  0.001822817  90.49580 1158
[6]  {Umývačky riadu,Varné dosky}      => {Rúry}              0.001721777 0.8526031  0.002019435  88.95043 1261
[7]  {Chladničky,Varné dosky}          => {Rúry}              0.001516966 0.8157122  0.001859683  85.10167 1111
[8]  {Odsávače,Varné dosky}            => {Rúry}              0.001710853 0.8141650  0.002101360  84.94026 1253
[9]  {Mikrovlnné rúry,Umývačky riadu} => {Rúry}              0.001021324 0.7663934  0.001332636  79.95634  748
[10] {Mikrovlnné rúry,Rúry}            => {Umývačky riadu}   0.001021324 0.5091899  0.002005781  59.22218  748
```

Obr. 9 TOP 10 asociačných pravidiel na základe lift-u (*Zdroj: vlastné spracovanie*)

V ďalších 126 vygenerovaných asociačných pravidlách sa často vyskytovali ešte rôzne kombinácie kuchynských spotrebičov, ale napríklad aj zaujímavé poznatky o tom, že ak si zákazník zakúpi mobil, tak je výhodné ponúknuť mu k nemu puzdro a tvrdené sklo, k vysávaču vrecká do vysávačov, ku práčke odvoz,

k plnoautomatickému espresso kávovaru kávu, či k herným konzolám hry a príslušenstvo k nim v podobe joystickov a podobne. Tieto asociačné pravidlá sa javia byť na prvý pohľad očakávané, ale na druhej strane prinášajú do rozhodovania marketingových manažérov podklady na základe ktorých môžu robiť kvalitné rozhodnutia založené na dátach, v podobe umiestnenia týchto produktov v kamennej predajni v tesnej blízkosti vedľa seba, ponúknutia ďalších tovarov alebo služieb pri vložení daného tovaru do košíka, či správne nastavených akcií a zliav (obr. 10).

[21]	{Herné konzoly}	=>	{Hry na herné konzoly}	0.001269827	0.3445721	0.003685230	39.461880	930
[22]	{Herné konzoly}	=>	{Príslušenstvo herné konzoly}	0.001227500	0.3330863	0.003685230	38.715563	899
[23]	{Púzdra na mobil,Služba inštalácie}	=>	{Tvrdené sklá}	0.005538086	0.7719833	0.007173842	38.561411	4056
[24]	{Mobilné telefóny,Služba inštalácie}	=>	{Tvrdené sklá}	0.008993928	0.7050198	0.012756986	35.216513	6587
[44]	{Púzdra na mobil,Služba poistenie}	=>	{Tvrdené sklá}	0.001054093	0.4225506	0.002494596	21.106868	772
[52]	{Práčky,Služba poistenie}	=>	{Služba odvoz}	0.002456365	0.3279256	0.007490616	17.658051	1799
[53]	{Mobilné telefóny,Služba poistenie}	=>	{Tvrdené sklá}	0.002093167	0.3517669	0.005950439	17.571141	1533
[54]	{Kamery}	=>	{Pamäťové karty}	0.001237058	0.4961665	0.002493231	16.847508	906
[55]	{Mobilné telefóny,Tvrdené sklá}	=>	{Púzdra na mobil}	0.005595433	0.4918387	0.011376561	16.768192	4098
[82]	{Videoprehrávače}	=>	{Káble - HDMI}	0.001112806	0.2196174	0.005067021	11.566519	815
[112]	{Vysávače}	=>	{Vrecká do vysávačov}	0.012016936	0.3916778	0.030680668	5.965151	8801
[114]	{Espresso - plnoautomatické}	=>	{Káva}	0.002648887	0.4259991	0.006218058	5.525352	1940

Obr. 10 Ďalšie asociačné pravidlá (*Zdroj: vlastné spracovanie*)

5 Záver

Analýza nákupného koša je užitočná metóda, ako odhaliť vzory v nákupnom správaní spotrebiteľov a tým predpovedať, o aký typ produktov môžu mať pri nákupe záujem. Takéto informácie sú využiteľné v mnohých oblastiach manažmentu a často sa využívajú v oblasti riadenia vzťahov so zákazníkmi. Po odhalení pravidiel, ktoré produkty si kupujú spotrebitelia spoločne, vedia marketingový manažéri v procese nákupu odporučiť v eshope produkty, u ktorých je vysoká pravdepodobnosť, že si ich kúpia tiež. Taktiež v kamenných predajniach sa oplatí usporiadať regály a tovar v nich tak, aby skupiny produktov, pri ktorých je vysoká pravdepodobnosť že budú kupované spolu, boli v regáloch vedľa seba.

V príspevku bolo na reálnych transakčných dátach z oblasti predaja elektro tovaru popísaný a vysvetlený proces čistenia a prípravy dát a následnej aplikácie analýzy nákupného koša vo voľne dostupnom a bezplatnom štatistickom programe R. Takýto postup môže slúžiť ako návod na použitie tejto metódy v praxi.

Pri využití apriori algoritmu v štatistickom softvéri R sme vygenerovali niekoľko skupín asociačných pravidiel s rôznymi nastaveniami parametrov tohto algoritmu. Významnejšie pre prax sa ukázalo byť použitie hierarchických transakčných dát, ktoré vedia zabezpečiť rovnomernejšie rozdelenie položiek v transakčných dátach. Nakoniec sme sformulovali aj niekoľko prakticky využiteľných asociačných pravidiel, ktoré môžu predajcom zvýšiť tržby, napr. ak by k mobilným telefónom ponúkali priamo aj kúpu puzdra na mobil, tvrdeného

skla a aj službu jeho inštalácie. Asociačné pravidlá odhalili vysokú pravdepodobnosť toho, že keď si zákazník zariaďuje novú kuchyňu, tak často sa vyberie cestou nákupu všetkých potrebných spotrebičov u jedného predajcu. Analýza nákupného košíka sa ukázala byť vhodnou a jednoduchou metódou, ktorej výsledky sú ľahko interpretovateľné a rýchlo implementovateľné do praxe.

6 Literatúra

Aguinis, H., Forcum, L. E., Joo, H. (2013). *Using market basket analysis in management research*. Journal of Management, 39(7), 1799-1824.

Brijs, T. et al. (2004). *Building an association rules framework to improve product assortment decisions*. Data Mining and Knowledge Discovery, 8(1), 7-23.

Cavique, L. (2007). *A scalable algorithm for the market basket analysis*. Journal of Retailing and Consumer Services, 14(6), 400-407.

Chandwani, M. H. (2018). *Market basket analysis using association rule*. International Journal of Advance Research, Ideas and Innovations in Technology.

Hahsler, M., Grün, B., Hornik, K. U. (2006). *Mining association rules and frequent itemsets.* R package version 0.4, 119 s.

Kamakura, W. A. (2012). *Sequential market basket analysis*. Marketing Letters, 23(3), 505-516.

Kaur, M., Kang, S. (2016). *Market basket analysis: Identify the changing trends of market data using association rule mining*. Procedia Computer Science, 85, 78-85.

Osmanbegovic, E., Suljic, M. (2015). *Selection of CRM applications*. International Scientific Conference Economy of Integration, 307-318.

Qisman, M., Rosadi, R., Abdullah, A. S. (2021). *Market basket analysis using apriori algorithm to find consumer patterns in buying goods through transaction data* (case study of Mizan computer retail stores). In Journal of Physics: Conference Series, 1722(1)

Stankovičová, I. (2009). *Možnosti extrakcie asociačných pravidiel z údajov pomocou SAS Enterprise Miner*. Informační a datová bezpečnost ve vazbě na strategické rozhodování ve znalostní společnosti, 1-7.

Trnka, A. (2010). *Market basket analysis with data mining methods*. In 2010 International Conference on Networking and Information Technology (pp. 446-450). IEEE.

Zendulka, J. et al. (2006). *Získávání znalostí z databází, studijní opora*. Brno: FIT VUT.

7 Poďakovanie

Príspevok je čiastkovým výstupom projektu VEGA 1/0737/20 Spotrebiteľská gramotnosť a medzigeneračné zmeny v preferenciách spotrebiteľov pri nákupe slovenských produktov a grantu UK/373/2021 Využitie veľkých objemov údajov v oblasti riadenia vzťahov so zákazníkmi.