УДК 336.74

АФФИНИТИВНЫЙ АНАЛИЗ ДАННЫХ ПОТРЕБИТЕЛЬСКОЙ КОРЗИНЫ С ПОМОЩЬЮ АЛГОРИТМА APRIORI

К.А. Кот
магистрант кафедры математического и информационного обеспечения экономических систем УО ГрГУ им. Я.Купалы

Н.В. Марковская
доцент кафедры математического и информационного обеспечения экономических систем УО ГрГУ им. Я.Купалы

УО Гродненский государственный университет имени Янки Купалы
E-mail: kristinakot.gsl@gmail.com, n.markovskaya@grsu.by

К.А. Кот
Окончила Гродненский государственный университет имени Янки Купалы. Магистрантка кафедры математического и информационного обеспечения экономических систем УО ГрГУ им. Я. Купалы.

Н.В. Марковская
Доцент кафедры математического и информационного обеспечения экономических систем УО ГрГУ им.Я.Купалы, кандидат физико-математических наук, доцент.

Аннотация. Ассоциативные правила являются удобным механизмом для обнаружения логических взаимосвязей среди набора объектов. В связи с развитием маркетинга и торговли в целом очень важно понимать, какие тенденции наблюдаются в покупательском поведении, чтобы увеличить продажи, помочь покупателям в поиске необходимых товаров и избежать нежелательных ситуаций, а также для общего анализа ситуации с целью проведения грамотной маркетинговой политики. В работе проанализированы покупательские транзакции на примере одного магазина. Предложный способ анализа с помощью алгоритма Apriori и возможностей языка Python позволит облегчить и ускорить процесс поиска и получения результатов.

Ключевые слова: потребительская корзина, ассоциативные правила, алгоритм Apriori, поддержка, достоверность, транзакции.

Критерии оценки ассоциативных правил.

Ассоциативные правила представляют собой механизм нахождения логических закономерностей между связанными элементами (событиями или объектами).

Выделяют три вида правил:

- полезные правила, содержащие действительную информацию, которая ранее была неизвестна, но имеет логическое объяснение;
- тривиальные правила, содержащие действительную и легко объяснимую информацию, отражающую известные законы в исследуемой области, и поэтому не приносящие какой-либо пользы;
- непонятные правила, содержащие информацию, которая не может быть объяснена.

Для оценки полезности и продуктивности перебираемых правил используются различные частотные критерии, анализирующие встречаемость кандидата в массиве экспериментальных данных. Важнейшими из них являются поддержка (support) и

достоверность (confidence).

Правило A→T имеет поддержку s, если оно справедливо для s% взятых в анализ случаев:

$$support(A \rightarrow T) = P(A \cup T). \quad (1.1)$$

Например, пусть поддержка правила "если покупатель приобретет муку, то он купит и яйца" ($X \rightarrow Y$, X − мука, Y − яйца) равна 25%. Это говорит о том, что в 25% сделанных покупок мука и яйца ($X \cup Y$) присутствовали одновременно.

Достоверность правила показывает, какова вероятность того, что из наличия в рассматриваемом случае условной части правила следует наличие заключительной его части (т.е. из A следует T):

$$confidence(A \rightarrow T) = P(A \cup T)/P(A) = supp(A \rightarrow T)/supp(A). \quad (1.2)$$

Например, пусть достоверность правила "если покупатель приобретет яйца, то он купит и муку" равна 70%. Это говорит о том, что в 70% случаев встречи яиц в списке сделанных покупок присутствует и мука.

Кроме этого используются и другие показатели - подъемная сила, или лифт (lift), которая показывает, насколько повышается вероятность нахождения T в анализируемом случае, если в нем уже имеется A" [3]:

$$lift(A \rightarrow T) = confidence(A \rightarrow T)/support(T). \quad (1.3)$$

"Усиление (leverage) отражает, насколько интересной может быть более высокая частота A и T в сочетании с более низким подъемом" [3]:

$$leverage(A \rightarrow T) = support(A \rightarrow t) - supp(A) \times supp(T). \quad (1.4)$$

Начало работы с алгоритмом.

Учитывая вышеизложенные факты, проведем анализ потребительской корзины магазина и выделим ассоциативные правила с помощью алгоритма Apriori. На первом этапе работы осуществляется просмотр базы данных и ее преобразование в вид, удобный для последующей работы. В используемой в данной работе базе данных были удалены товары "пакеты" и "мешки" из всех транзакций, поскольку в ходе предварительного анализа было установлено, что данный товар наиболее часто покупается и, соответственно, большинство правил его содержат, в то время как нас интересуют другие взаимосвязи. Также все названия объектов были заменены на более короткие эквиваленты.

Применение Python.

Исходные данные были получены в формате MS Excel, и были экспортированы в Python. Все преобразования над ними выполнены с помощью инструментов языка Python для работы с объектами DataFrame и библиотеки pandas.

Алгоритм действий выглядит следующим образом:

1) удаляются все данные за исключением номера чека, даты и полного наименования товара;
2) полные названия заменяются на краткие;
3) данные упорядочиваются по номеру чека и дате;
4) товары с одинаковой датой и номером чека группируются в один чек.

```
import pandas as pd
df = pd.read_excel('C:\\Users\\User\\Downloads\\m4.xlsx')
```

```
del df['Заголовок чека. Код кассы']
del df['Заголовок чека. Код Z-отчета']
del df['Заголовок чека. Время чека']
df['Карточка товара. Название короткое'] = df['Карточка товара. Название короткое'].astype(str)
df['Карточка товара. Название короткое'] = df['Карточка товара. Название короткое'].str.partition(' ')[0]
df.groupby('Заголовок чека. Дата чека')
df.head(50)
del df['Заголовок чека. Дата чека']
df["Чек"] = ""
```

	Чек
0	СПИЧКИ, СПИЧКИ
1	ВИНО, БАТОН, МОЛОКО, МЯСО
2	КОРМ, МОЛОКО, МОЛОКО, РЫБА
3	КОЛБАСА, ХЛЕБ, САЛАТ, ВОДА
4	НАПИТОК
5	СОЛЬ, ПЕРЕЦ, КРУПА
6	КЕФИР
7	БАТОН, КЕФИР, МОЛОКО, ВОДА, ЗЕЛЕНЬ, ЯБЛОКО
8	МЯСО, КОЛБАСА
9	БУЛКА, БУЛКА

Рисунок 1. Результат работы алгоритма

Результаты работы в RStudio.

Дальнейшая работа в среде разработки RStudio проводился с файлом, конвертированным в csv-формат.

Применяя алгоритм Apriori, мы можем определить товары, купленные вместе – то есть установить ассоциативные правила. Рассмотрим первый магазин и описательные статистики его транзакций:

```
>library(arules)
>library(arulesViz)
>groceries=read.transactions(file.choose(), sep=",")
>G<-groceries
>inspect (head (G, 10))
>summary(G)
transactions as itemMatrix in sparse format with
18620 rows (elements/itemsets/transactions) and
531 columns (items) and a density of 0.02107276
most frequent items:
ХЛЕБ  МОЛОКО   БАТОН СМЕТАНА КОЛБАСА (Other)
11447  9764          6553       6006          5821          168760
Min. 1st Qu.  Median    Mean   3rd Qu.    Max.
1.00    7.00    11.00       11.19   15.00       56.00
```

Таким образом, мы видим, что в наборе данных присутствуют 531 продукт и 18620 транзакций (чеков). Наиболее часто встречающиеся товары:

- хлеб;
- молоко;
- батон;
- сметана;
- колбаса.

Максимальное количество товаров в чеке – 56, а минимальное – 1. Наиболее часто транзакции с большим количеством товаров выпадают на праздничные периоды.

Визуально отобразим данные в виде диаграммы частоты. В качестве типа шкалы используем относительный, то есть, какова частота встречаемости того и иного товара по отношению к другим товарам в выборке:

>itemFrequencyPlot(G,topN=20,type="relative",col=brewer.pal(8,'Pastel2'), main="Relative Item Frequency Plot")

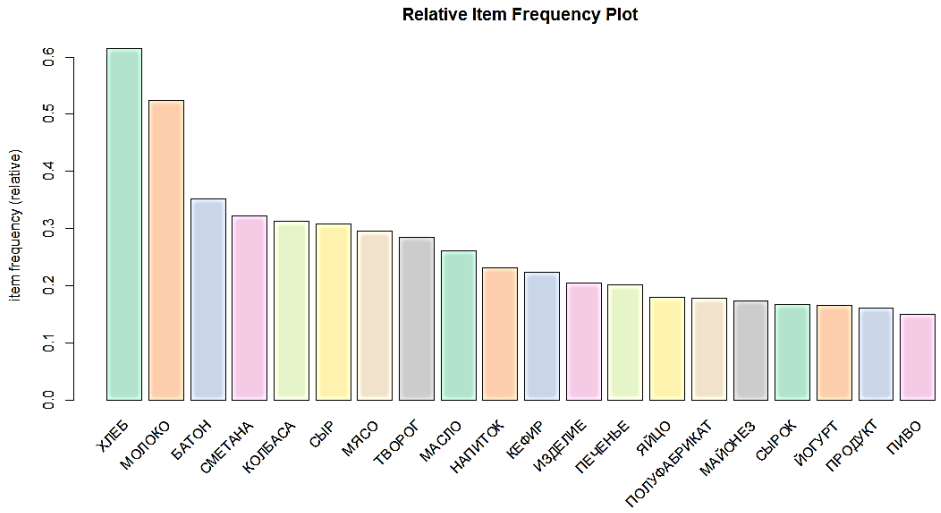

Рисунок 2. График 20 наиболее встречаемых товаров в первом магазине

Применим алгоритм Apriori, где supp – минимальный уровень поддержки, а conf – достоверность, и отсортируем выборку по уменьшению уровня достоверности. В конечную выборку попало 308816 ассоциативных правил.

Рассмотрим первые 10 из них:

```
> inspect(head(G_rules, n = 10, by ="confidence"))
lhs                                  rhs          support     confidence
[1]{СОДА,СЫРОК}                      => {ХЛЕБ}    0.001181525 1
[2] {МАКАРОНЫ,ПЕЛЬМЕНИ}              => {ХЛЕБ}    0.001181525 1
[3] {ГРАНАТ,МОЛОКО,СЫР}              => {ХЛЕБ}    0.001020408 1
[4] {БАТОН,ВЫРЕЗКА,МЯСО}             => {ХЛЕБ}    0.001127820 1
[5] {БАТОН,ВЫРЕЗКА,КОЛБАСА}          => {ХЛЕБ}    0.001235231 1
[6] {МОЛОКО,СОДА,СЫРОК}              => {ХЛЕБ}    0.001020408 1
[7] {КЕФИР,Колбаса,СЫР}              => {ХЛЕБ}    0.001074114 1
[8]{КОЛБАСА,ПЛЕТЕНКА,П/Ф}            => {ХЛЕБ}    0.001020408 1
[9]{КОЛБАСА,ПЛЕТЕНКА,ЯЙЦО}           => {ХЛЕБ}    0.001074114 1
[10]{БАТОН,ПЛЕТЕНКА,СЫР}             => {МОЛОКО}  0.001503759 1
```

Полученные результаты можно трактовать следующим образом (на примере первого ассоциативного правила): если покупатель приобрел питьевую соду и сырок, то с достоверностью 1 он приобретет и хлеб. Поддержка 0,011 означает, что на 11 случаев из 10000 правило сработает.

Теперь, когда правила сформированы, для одного из самых часто покупаемых товаров в первом магазине можно узнать:

- что покупатели скорее всего приобрели перед этим;
- что приобрели после покупки данного товара.

Для этого зададим функции apriori новые параметры – lhs (left hand side – товары, которые приобрели) и rhs (right hand side – товары, которые приобретут впоследствии):

```
>G_rules<-apriori(data=G, parameter=list(supp=0.001,conf = 0.7),
          appearance = list(default="lhs",rhs="МОЛОКО"),
          control = list(verbose=F))
>G_rules<-sort(G_rules, decreasing=TRUE,by="confidence")
>inspect(G_rules[1:10])
lhs                                                    rhs          support    conf
[1] {БАТОН,ПЛЕТЕНКА,СЫР}                            => {МОЛОКО} 0.001503759  1
[2]  {ЙОГУРТ,КЕФИР,Форель}                           => {МОЛОКО} 0.001074114  1
[3]  {ЙОГУРТ,ТВОРОГ,Форель}                         => {МОЛОКО} 0.001288937  1
[4]  {ВАФЛИ,КЕФИР,МАНДАРИНЫ}                        => {МОЛОКО} 0.001235231  1
[5]  {БУМАГА,МОРОЖЕНОЕ,ЯБЛОКИ}                      => {МОЛОКО} 0.001074114  1
[6]  {БАТОН,КОЛБАСА,ПЛЕТЕНКА,СЫР}                   => {МОЛОКО} 0.001074114  1
[7]  {БАТОН,ПЛЕТЕНКА,СЫР,ХЛЕБ}                      => {МОЛОКО} 0.001288937  1
[8]  {БАТОН,КОРЖИ,СЫР,ТВОРОГ}                       => {МОЛОКО} 0.001020408  1
[9]  {БАТОН,МАРГАРИН,ПОЛУФАБРИКАТ,ХЛЕБ}=>{МОЛОКО} 0.001074114  1
[10] {КЕФИР,МАРГАРИН,МАСЛО}                         => {МОЛОКО} 0.001074114  1
```

Таким же способом можно задать параметру lhs значение “МОЛОКО” и узнать, какие продукты часто берут после него. Значение достоверности понизим до 0,3, поскольку при conf=0.7 ассоциативные правила не обнаруживаются. Параметру minlen присвоено значение 2 во избежание пустых наборов объектов lhs.

```
>G_rules<-apriori(data=G, parameter=list(supp=0.001,conf = 0.3,minlen=2),
          appearance = list(default="rhs",lhs="МОЛОКО"),
          control = list(verbose=F))
>G_rules<-sort(G_rules, decreasing=TRUE,by="confidence")
inspect(G_rules[1:5])
lhs                 rhs          support    confidence coverage   lift     count
[1] {МОЛОКО} => {ХЛЕБ}       0.3727175 0.7107743  0.5243824 1.156165 6940
[2] {МОЛОКО} => {БАТОН}      0.2245435 0.4282057  0.5243824 1.216724 4181
[3] {МОЛОКО} => {СМЕТАНА} 0.2110634 0.4024990  0.5243824 1.247841 3930
[4] {МОЛОКО} => {СЫР}        0.1897422 0.3618394  0.5243824 1.172342 3533
[5] {МОЛОКО} => {ТВОРОГ}     0.1858754 0.3544654  0.5243824 1.247665 3461
```

Как видно из результатов работы алгоритма, довольно часто проявляется закономерность между покупкой колбасных изделий и алкогольной продукции.

Заключение.

По результатам, которые были достигнуты в ходе проведенного исследования, следует отметить, что наиболее часто встречающейся закономерностью является совместное приобретение молочных продуктов и выпечки. Следовательно, самым

рациональным решением будет расположить этим отделы рядом. Также следует отметить, что необходимо в первую очередь следить за доступностью продуктов, занимающих главные позиции на графике встречаемости товаров.

Список использованных источников

[1] Загребанцев А.Н. Использование алгоритма «Apriori» для поиска ассоциативных правил / Загребанцев А.Н., Марковская Н.В.// "Science and education in the modern world: challenges of the XXI century " атты III Халықар. ғыл.-тәж. конф. материалдары (II том)/ Қңраст.: Е. Ешім, Е. Абиев т.б.– Нур-Султан, 2019 – С.221-225.

[2] Залога А.Ю. Аффинитивный анализ данных. Поиск ассоциативных правил / Залога А.Ю., Марковская Н.В. // BIG DATA and Advanced Analytics = BIG DATA и анализ высокого уровня: сб. Материалов V Междунар. науч.-практ. конф. (Республика Беларусь, Минск, 13-14 марта 2019 года). В 2 ч. Ч. 2. / редкол.: В. А. Богуш [и др.]. ¬– Минск: БГУИР, 2019. – С. 20-26.

[3] 3). Acheampong F., Markovskaya N.V. Using apriori algorithm for investigation of Economic Data / Acheampong F., Markovskaya N.V. // BIG DATA Advanced Analytics: collection of materials of the fourth international scientific and practical conference. (Minsk, Belarus, May 3 – 4, 2018) / editorial board: M. Batura [etc.]. – Minsk, BSUIR, 2018. – P. 140-144.

AFFINITIVE ANALYSIS OF CONSUMER BASKET DATA USING THE APRIORI ALGORITHM

K.A.KOT
Master student of the Department of Mathematical and Information Support of Economic Systems, YKSUG

N.V. MARKOVSKAYA,
Associate professor of the Department of Mathematical and Information Support of Economic Systems

Yanka Kupala State University of Grodno, Republic of Belarus
E-mail: kristinakot.gsl@gmail.com, n.markovskaya@grsu.by

Abstract. Association rules are convenient mechanism for discovering logical relationships among a set of objects. In connection with the development of marketing and trade in general, it is very important to understand what trends are observed in consumer behavior in order to increase sales, help buyers find the right products and avoid unwanted situations, as well as for a general situation analysis in order to conduct a competent marketing policy. This work shows analyzes of transactions of one store. The proposed method of analysis using Apriori algorithm and the capabilities of the Python language will facilitate and speed up the process of searching and obtaining results.

Keywords: consumer basket, association rules, Apriori algorithm, support, reliability, transactions.