Александр Докукин, Юрий Журавлев, Олег Сенько, Дмитрий Стефановский

Докукин Александр Александрович — старший научный сотрудник Федерального исследовательского центра «Информатика и управление» РАН, кандидат физико-математических наук.

Журавлев Юрий Иванович — главный научный сотрудник Федерального исследовательского центра «Информатика и управление» РАН, академик РАН.

Сенько Олег Валентинович — ведущий научный сотрудник Федерального исследовательского центра «Информатика и управление» РАН, доктор физико-математических наук.

Стефановский Дмитрий Владимирович — руководитель центра систем больших данных Международного научно-исследовательского института проблем управления, доцент РАНХиГС, кандидат технических наук.

Aleksandr A. Dokukin — Federal Research Center “Informatics and Management” of the Russian Academy of Sciences.

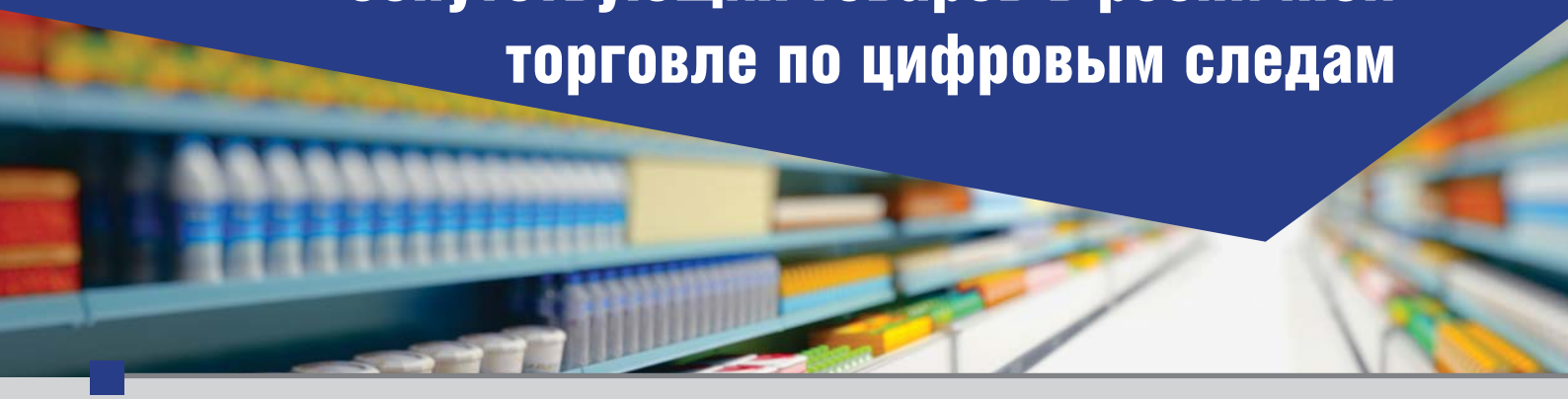
Yurii I. Zhuravlev — Federal Research Center “Informatics and Management” of the Russian Academy of Sciences.

Oleg V. Sen’ko — Federal Research Center “Informatics and Management” of the Russian Academy of Sciences.

Dmitrii V. Stefanovskii — International Research Institute for Advanced Systems.

Работа выполнена при частичной финансовой поддержке РФФИ, проект 18-29-03151.

Математическая модель выделения групп сопутствующих товаров в розничной торговле по цифровым следам



УДК 004.94

Важной управленческой технологией развития торгового бизнеса является категорийный менеджмент, ориентированный на выделение групп сопутствующих или замещающих товаров с учетом анализа покупательских предпочтений. Традиционный способ, включающий опрос покупателей у выхода из супермаркетов, является дорогостоящим и не очень достоверным из-за охвата только относительно небольших групп. Альтернативный подход — анализ кассовых чеков, по сути являющихся цифровыми следами покупок, сделанных при посещении гипермаркета. Для выделения групп сопутствующих товаров в работе предлагается применять методы иерархического кластерного анализа с использованием бинарных мер сходства, приводятся результаты экспериментов на реальных данных, сравнивается эффективность мер Жаккара и χ^2, анализируются методы оптимизации числа кластеров.

Ключевые слова
Категорийный менеджмент, сопутствующие товары, цифровой след, кластерный анализ, бинарные меры сходства.

Доля цифровых активов, например детализированной информации о покупках клиентов в розничных сетях, достигла той точки развития, когда эти ресурсы могут обеспечить получение преимуществ, не связанных только с поиском особых мест продажи или получением лучших условий поставок [1]. Пристальное внимание в этой связи уделяется категорийному менеджменту, то есть управленческой технологии развития торгового бизнеса, ориентированной на классификацию товаров, получение наборов показателей для оценки эффективности торговли за счет анализа новых агрегатов товаров [2].

Речь идет об управлении ассортиментом через товарные категории, а не через единичные бренды производителей, в основе которого лежит глубокое понимание трендов развития категории. Сегодня полной информацией о предпочтениях потребителя в целом по стране (или странам) владеют производители продукции. Они имеют полную картину своих продаж с учетом различных аспектов и поэтому могут делать содержательные выводы только о производимых ими товарах. При этом розничные продавцы могут на основе своих данных проанализировать категориальные и межкатегориальные связи в покупках. Производители этого сделать не могут без дополнительного сбора информации.

Развитие категориального менеджмента видится в изучении отношений «товар — товар» и двух отношений: «категория — категория» и «производитель — производитель».

Постановка задачи и вычислительный эксперимент

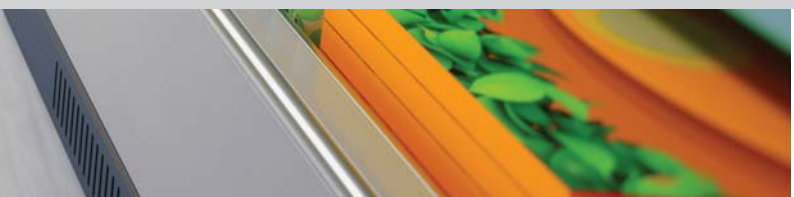
Сегодня данные о разнообразных предпочтениях покупателей собираются в основном за счет опроса людей, входящих в магазин и выходящих из него. Такие опросы проводятся хорошо обученными аспирантами или студентами, изучающими маркетинговые исследования [3]. С появлением чеков в электронном виде покупатели стали оставлять цифровой след о продажах, что делает целесообразным применение методов машинного обучения для получения информация о структуре предпочтений клиентов. Однако при анализе больших объемов цифровых следов есть ряд трудностей, которые необходимо преодолеть за счет разработки новых подходов к структуризации чеков. К таким трудностям относятся:

- отсутствие в ряде чеков связи между идентификатором покупателя и содержанием чека;
- большие размерности матриц и сильная разреженность бинарных матриц, описывающих отношение чек — товар;
- большие объемы информации, которые требуют больших вычислительных мощностей.

Mathematical Model for Selecting Groups of Related Products in the Retail by Digital Traces

An important management technology for trading business development is category management, focused on identifying groups of related or replacement products, based on consumer preferences analysis. Traditional method, including a survey of customers at the supermarket exit, is expensive and not very reliable due to the coverage of only relatively small groups. An alternative approach is the analysis of cash checks, which are essentially digital traces of purchases made while visiting a hypermarket. For identifying groups of related products, it is proposed to apply methods of hierarchical cluster analysis with the use of binary similarity measures, results of experiments on real data are presented, effectiveness of the Jacquard and χ^2 measures are compared, methods for optimizing the number of clusters are analyzed.

Keywords
Category management, related products, digital trace, cluster analysis, binary measures of similarity.



Следует отметить, что агрегация информации по относительно небольшим группам клиентов ничем не отличается от проведения исследований по социологическим методикам. Все вышесказанное требует разработки новой методики анализа данных о цифровых следах покупки, которая позволит сформировать формальное описание структур следующих типов:
1) «товар — товар»;
2) «категория — категория»;
3) «поставщик — поставщик».

Анализ каждой из структур — это отдельное направление, которое будет реализовано в дальнейших исследованиях. В этой статье описан способ работы с отношением товар — товар для получения дополнительной информации о подгруппах покупательских предпочтений на основе кластеризации данных о продажах. Полученные результаты станут основой для формирования профиля предпочтений покупателя и типизации его цифровых двойников. Иначе говоря, анализируя различные цифровые следы, можно создавать разные классификаторы, которые относят данного покупателя к тому или иному типу. Традиционное деление на несколько типов должно быть расширено, поскольку его недостаточно для предсказаний взаимодействий с покупателями, хотя и достаточно для работы с товарными категориями.

Возможные метрики и описание данных

Для анализа чеков будем использовать следующие соглашения. Для каждого чека входящий в него товар считается один раз, то есть формируется матрица размерностью: количество чеков на количество продаваемых товаров (ассортимент). Каждая строка в такой матрице — это один чек, а на пересечении строки или столбца — 1, если товар имеется в чеке, и 0, если товара нет в чеке. В современных торговых точках ассортимент может превышать несколько тысяч артикулов, и матрица товар — чек в этом случае является разряженной. Для кластеризации необходимо выбрать одну из мер близости, учитывающих эту особенность.

Пусть n — количество признаков (товаров) или размерность вектора функции. Определения бинарного сходства и дистанционные меры выражаются через элементы таблицы размерностью 2 × 2 [4], в которой a — количество функций, где значения i и j равны 1 (или присутствию), что означает «положительные соответствия», b — количество атрибутов, где значения i и j равны 0,1, что означает «отсутствие i», c — количество атрибутов, в которых значения i и j равны 1,0, что означает «несоответствие отсутствия j», и d — количество атрибутов, где оба i и j имеют 0 (или отсутствие), что означает «отрицательные соответствия». Диагональ sum $a + d$ представляет общее количество совпадений (*табл. 1*).

> Основная нагрузка получения знаний о предпочтениях потребителя ложится на производителей продукции.

Первые бинарные меры сходства были предложены Пирсом, Юле и Пирсоном в 1900-х годах. Мера Жаккара, предложенная в 1901 г., все еще широко используется в различных областях, таких как экология и биология [5]. В работе [5] показано, что многие меры близости во многом сходны друг с другом.

Наиболее широко известные метрики Хэмминга или метрики Эвклида, фактически подсчитывающие число совпадений между компонентами, очевидно, нельзя использовать для описания сходства покупательской активности, поскольку в этом случае высокое сходство будет вычисляться для товаров, присутствующих в минимальном числе чеков. Учитывая вышесказанное, для наших исследований мы выбрали две меры (метрики) — меру Жаккара и χ^2, которые являются наиболее популярными внутри соответствующих групп сходных мер.

Предположим, что нами рассматривается множество из n товарных чеков, выданных за какой-то промежуток времени.

При этом в магазине выставляется на продажу m товаров, пронумерованных некоторым образом $t[1],..., t[m]$. Каждому товару $t[i]$, очевидно, может соответствовать вектор $T[i]$ длины n, который вычисляется по следующему правилу: ком-

Таблица 1

Таблица соответствий

Показатель	Присутствует	Отсутствует	Сумма
Присутствует	$a = i \cdot j$	$b = i \cdot j$	$a + b$
Отсутствует	$c = i \cdot j$	$d = i \cdot j$	$c + d$
Сумма	$a + c$	$b + d$	$a + b + c + d$

понente $T[i][j]$ присваивается значение 1, если в чеке с номером j присутствует товар $t[i]$, компоненте $T[i][j]$ присваивается значение 0, если в чеке с номером j товар $t[i]$ отсутствует.

Меру сходства внимания покупателей к товарам $t[i']$ и $t[i'']$ естественно считать мерой сходства между бинарными векторами $T[i']$ и $T[i'']$. Для индексов a, b, c, d касательно задачи сравнения $t[i']$ и $t[i'']$, очевидно, будут справедливы следующие соотношения:

$$a=\sum_{j=1}^{n}T[i'][j]T[i''][j];$$

$$a+b=\sum_{j=1}^{n}T[i'][j];$$

$$a+c=\sum_{j=1}^{n}T[i''][j].$$

Мера сходства Жаккара S_{jac} определяется как отношение числа чеков, содержащих товары $t[i']$ и $t[i»]$ одновременно, к числу чеков, содержащих по крайней мере один из этой пары товаров:

$$S_{jac}(t[i'],t[i''])=\frac{a}{a+b+a+c-a}=\frac{a}{a+b+c}.$$

Альтернативным подходом для определения меры может быть вычисление χ^2, часто используемое при оценке гипотезы о независимости двух бинарных переменных с помощью одноименного теста. Обозначим через $v_1[i]$ и $v_Q[i]$ доли компонент вектора, соответствующих товару $t[i]$, равных 1 и 0 соответственно. Можно выделить четыре ячейки, соответствующие различным комбинациям значений компонент векторов $t[i']$ и $t[i»]$: (1,1), (1,0), (0,1), (0,0).

Обозначим через $m_{k,k'}$ число парных комбинаций компонент с одинаковым номером, попавших в ячейку (k, k'), $k, k' \in \{0, 1\}$. Мера близости вычисляется по формуле:

$$S_{\chi^2}(t[i'],t[i''])=\sum_{k\in\{0,1\}}\sum_{k'\in\{0,1\}}\frac{(mp_{kk'}-m_{kk'})^2}{mp_{kk'}}.$$

В том случае, если $t[i']$ и $t[i'']$ независимы, то $p_{kk'}=v_k[i']\,v_{k'}[i'']$,

где

$$v_1[i'']=\frac{a+b}{a+b+c+d};$$

$$v_0[i'']=\frac{c+d}{a+b+c+d};$$

$$v_1[i'']=\frac{a+c}{a+b+c+d};$$

$$v_0[i'']=\frac{b+d}{a+b+c+d}.$$

Учитывая также, что $m_{11}=a$, $m_{10}=b$, $m_{01}=c$, $m_{00}=d$, $m_{10}=b$, после некоторых преобразований получаем формулу:

$$S_{chi^2}=\frac{n(ad-bc)^2}{((a+b)(a+c)(c+d)(b+d))}.$$

Преимуществом метрики Жаккара является ее простота и естественность. Вместе с тем число 1 внутри сравниваемых векторов при фиксированной пропорции не учитывается. С этой точки зрения предпочтительнее может выглядеть мера χ^2.

Выделение групп сходных по какой-либо метрике объектов может производиться с помощью метода иерархической кластеризации. На первом шаге использования данного метода выбирается мера близости между кластерами. Например, в качестве меры близости между кластерами Cl и Cl' может быть использовано минимальное расстояние между двумя векторами, принадлежащими разным кластерам:

$$\rho_{NN}(Cl_l,Cl_{l'})=\min_{s'\in Cl_l,s''\in Cl_{l'}},\rho_{\chi^2}(s',s'').$$

В качестве другой естественной меры близости между двумя кластерами использовалась средняя мера сходства между объектами двух кластеров:

$$\rho_{AV}(Cl_l,Cl_{l'})=\frac{1}{|Cl_l|}\frac{1}{|Cl_{l'}|}\sum_{s'\in Cl_l}\sum_{s''\in Cl_{l'}}\rho_{\chi^2}(s',s'').$$

В настоящем исследовании использовалась мера сходства между кластерами ρ_W, являющаяся средним по двум упомянутым выше мерам:

$$\rho_W(Cl_l,Cl_{l'})=\frac{1}{2}\left[\rho_{AV}(Cl_l,Cl_{l'})+\rho_{NN}(Cl_l,Cl_{l'})\right].$$

Таблица 2

Сравнение перечисленных методов по усредненным условным вероятностям *P(CK(x)/x) и P(x/CK(x))*

Метод	*P(CK (x)/x)*	*P(x/CK (x))*
Метод Хи-квадрат A	0.197	0.021
Метод Хи-квадрат B	0.18	0.019
Метод Жаккара	0.163	0.029
Базовый метод A	0.171	0.017
Базовый метод B	0.156	0.014

Иерархический алгоритм кластеризации устроен следующим образом.

• Выбирается порог 5 для меры сходства P_W, определяющий образование новых кластеров.

• На первом шаге все отдельные объекты (бинарные векторы) объявляются кластерами, содержащими только один объект.

• На шаге r просматриваются всевозможные пары кластеров, полученные на шаге r — 1. В том случае, если минимальное сходство ρ_W между кластерами превышает порог 5, то из пары, соответствующей достигнутому минимуму, образуется новый кластер и осуществляется переход к шагу $r+1$. Если минимальное сходство оказывается ниже порога, то процесс прекращается.

Подход к анализу данных

Одним из способов выделения профиля покупателя является поиск групп товаров, которые достаточно систематически покупаются одновременно, то есть в ходе одного посещения конкретного магазина. Такие товары должны попадать в одни и те же товарные чеки, которые отражают сам факт осуществления покупки, и поэтому к ним не так просто применить методы машинного обучения с учителем. Однако методы обучения без учителя, например кластеризацию, можно применить, и мы будем использовать только методы иерархической кластеризации.

В качестве меры качества кластеризации будем использовать внешнюю меру. Для этого в каждом кластере выделялось ядро, включающее пять товаров, которые покупаются чаще других товаров кластера. Ядра кластеров будем обозначать *CK*. Через *P(CK/xj)* обозначим вероятность покупки товара *Xj* при условии одновременной покупки хотя бы одного товара из *CK*. Через $CK_m(xj)$ обозначим ядро кластера, для которого вероятность *P(CK/xj)* максимальна. Условная вероятность $P(CK_m(xj)/xj)$, очевидно, описывает вероятность покупки товара *Xj* при условии покупки хотя бы одного товара из $CK_m(xj)$, то есть из группы востребованных товаров, более всего связанных с *xj*.

Сравнение перечисленных методов по усредненным условным вероятностям $P(CK(x)/x)$ и $P(x/CK(x))$ представлено в *табл. 2*.

Через *P(xj/CK)* обозначим вероятность покупки хотя бы одного товара из *CK* при условии покупки товара *xj*. Условная вероятность $P(xj/CK_m(xj))$, очевидно, описывает вероятность покупки хотя бы одного товара из $CK_m(xj)$ при условии покупки товара *xj*. Иначе говоря, $P(xj/CK_m(xj))$ обозначает вероятность покупки хотя бы одного из востребованных товаров, более всего связанных с *xj*, при условии покупки *xj*. Через $P(CK_m(x)/x)$ и $P(x/CK_m(x))$ обозначим соответствующие условные вероятности, усредненные по всем товарам.

Описание данных и результаты экспериментов

Для эксперимента был взят условный массив данных, основанных на кассовых чеках. В расчетах использовались номер чека, название и код товара и количество товаров в чеке. Кластеризация проводилась при помощи таблицы товар — чек, в строках которой были указаны товары, а в столбцах — чеки. Размерности такой матрицы — тысячи на десятки тысяч. На основе этой таблицы реализовано несколько методов формирования ядер кластеров.

• Метод Хи-квадрат *A* — кластеры формируются методом иерархической кластеризации с использованием меры близости Хи-квадрат.

• Метод Хи-квадрат *B* — кластеры формируются методом иерархической кластеризации с использованием меры близости Хи-квадрат. При этом используются только кластеры с наиболее тесной связью.

• Метод Жаккара — кластеры формируются методом иерархической кластеризации с использованием меры близости Жаккара.

• Базовый метод сравнения *A* — товары из кластеров, формируемых методом Хи-квадрат *A*, случайным образом заменяются произвольными товарами.

Таблица 3

Примеры кластеров

Номер кластера	Список товаров ядра кластера
Кластер № 27	Тушь для ресниц VIVIENNE SABO ARTISTIC VOLUME Cabaret тон 01 (черный), Помада для губ ESSENCE MATT MATT MATT тон 02 коричневый нюд (матовая), Карандаш для губ ESSENCE тон 05, 3... Помада-карандаш для губ ESSENCE VELVET STICK тон 05 (сливовый) матовая... Помада для губ ESSENCE MATT MATT MATT тон 03 розово-коричневый (матовая)... Помада для губ ESSENCE LONGLASTING LIPSTICK стойкая тон 04, Тушь для ресниц BOURJOIS VOLUME GLAMOUR PUSH UP black serum объемная (black), Консилер для лица ESSENCE CORRECTING LIQUID CONCEALER тон 20 (светло-желтый)
Кластер № 115	Пилинг для лица SHINETREE 12 мл, Маска для лица SHINETREE Жемчуг (очищающая) 15 мл... Маска для лица SHINETREE грязевая (сужающая поры) 15 мл, Сыворотка для лица SHINETREE с экстрактом красных водорослей (сужающая поры) 8 мл... Гель для умывания SHINETREE оливковый 12 мл, Лосьон для тела и рук SHINETREE 12 мл, Скраб для лица SHINETREE энзимный мягкий 20 г, Маска-сорбет для лица SHINETREE 20 г, Шампунь для волос SHINETREE 2 в 1 20 г, Мусс тональный для лица ESSENCE SOFT TOUCH тон 02
Кластер № 241	Щетка зубная COLGATE ZIG ZAG Древесный уголь... Паста COLGATE зубная Защита от кариеса 100 мл FBR14910 K48... Паста зубная COLGATE АЛТАЙСКИЕ ТРАВЫ с фторидом и кальцием 100 мл, Паста зубная BLEND-A-MED Анти-кариес 100 мл, Паста зубная детская COLGATE ДОКТОР ЗАЯЦ со вкусом клубники 50 мл, Пена для ванн ORGANIC SHOP STRAWBERRY 500 мл, Паста зубная ЛЕСНОЙ БАЛЬЗАМ Тройной эффект 75 мл, Паста зубная НОВЫЙ ЖЕМЧУГ Сила моря 100 мл, Щетка зубная CJ LION Crystal (мягкая), Щетка зубная CJ LION Dr. Sedoc (супертонкая)

• Базовый метод сравнения *B* — товары из кластеров, формируемых методом Хи-квадрат B, случайным образом заменяются произвольными товарами.

В результате получено более 300 кластеров совместно покупаемых товаров. Примеры кластеров представлены в *табл. 3*.

Далее на *рис. 1* и *2* представлены характеристики, оценивающие результаты кластеризации. Для каждого кластера были рассчитаны следующие характеристики:
• *n* — число объектов в кластере;
• *nst1* — число различных пар в кластере, то есть попросту $n(n-1)/2$;
• *nst* — число пар, связанных по статистике Хи-квадрат, вероятность ошибки нулевой гипотезы: $(1 - p - value)$ порог — > 0.98;
• *fra = nst/nst1;*
• min — минимальное сходство по мере Хи-квадрат между объектами в кластере (чаще 0, но может быть отрицательное –1);
• max — максимальное сходство по Хи-квадрат между объектами в кластере;
• *Quality* — среднее сходство по мере Хи-квадрат.

Как видно из гистограмм, кластеры с мерой χ^2 распределены сложнее, чем с мерой Жаккара. Это говорит о том, что мера Жаккара менее чувствительна к особенностям структуры совместной покупки товаров. Также у меры Жаккара дисперсия и среднее количество товаров в кластерах меньше, чем для меры χ^2, что косвенно показывает преимущество последней меры. Нормальное распределение указывает на предрасположенность меры к тому, что количество кластеров имеет одинаковое значение, что не совсем соответствует действительности с экспертной точки зрения.

Сходная картина наблюдается для параметра *Qualiti* с мерой χ^2, который распределен сложнее, чем тот же параметр с мерой Жаккара. Это также говорит о том, что мера Жаккара менее чувствительна к особенностям структуры совместной покупки товаров.

Для анализа параметра *Qualiti* воспользуемся методом выделения естественных границ Дженкса [6]. Метод основан на поиске естественного деления на классы на основе значений данных. Классы выделяются на основе максимизации различий, то есть наборы делятся на классы, чьи границы устанавливаются там, где имеются относительно большие различия в значениях данных. Это метод Дженкса, который также называют методом естественного разлома Дженкса. В целом можно сказать, что этот метод стремится уменьшить дисперсию внутри классов и увеличить дисперсию между классами. С использованием метода естественных границ было рассчитано количество кластеров, которые относятся к каждому интервалу. Результаты представлены на *рис. 3* и в *табл. 4*.

Хотя мера Жаккара показывает иной результат, ее преимуществом является простота и естественность. Мера Хи-квадрат является несколько более сложной. Об эффективности данной меры свидетельствует практика ее использования в качестве статистики соответствующего теста. Таким образом, высокие значения сходства по мере Хи-квадрат свидетельствуют о не-

Рисунок 1

Гистограмма размеров кластеров: а — с мерой Хи-квадрат; б — с мерой Жаккара

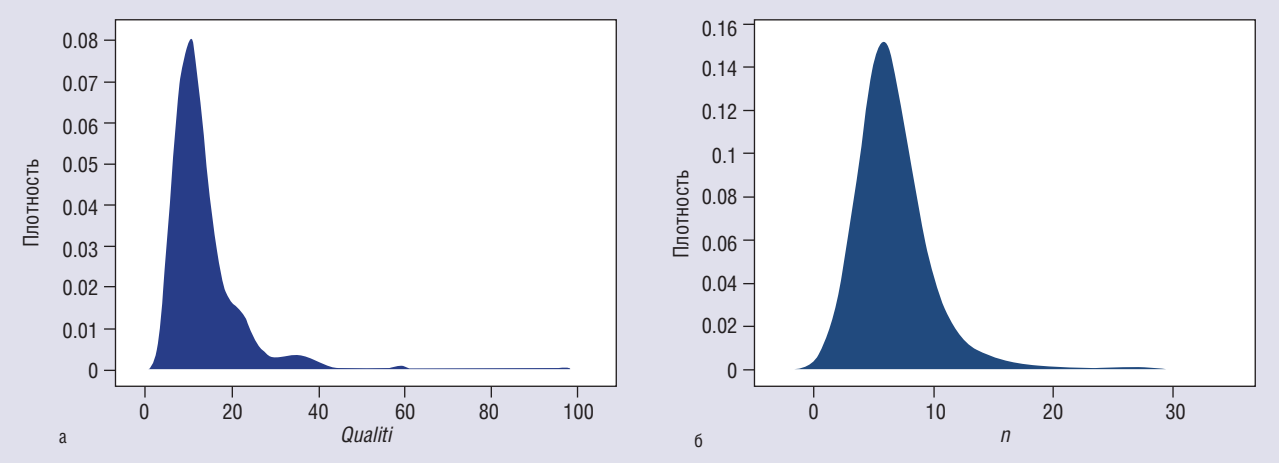


Рисунок 2

Гистограмма *Qualiti* кластеров: а — с мерой Хи-квадрат; б — с мерой Жаккара

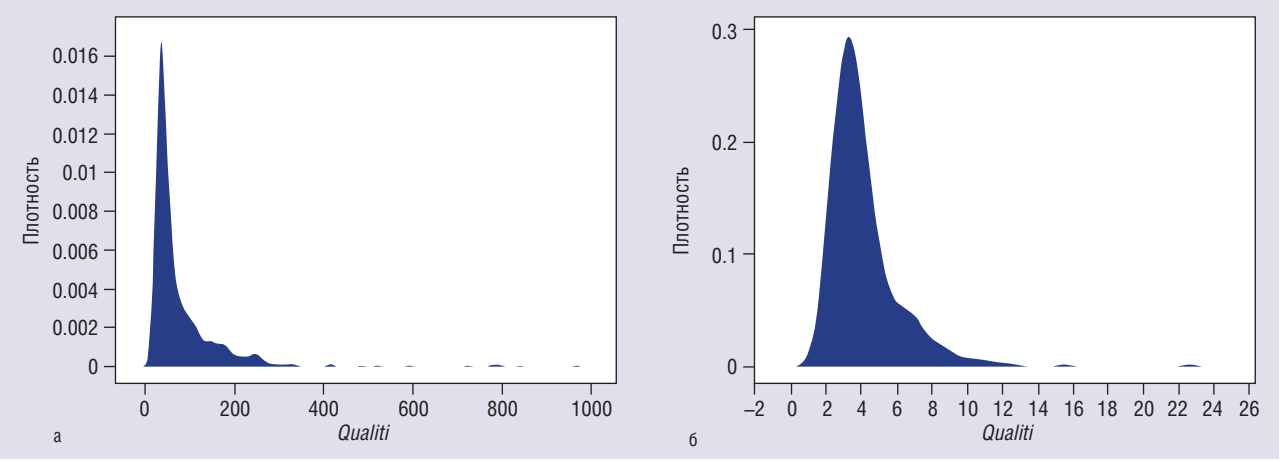


Рисунок 3

Количество кластеров в интервалах *Qualiti*: а — мера χ^2; б — мера Жаккара

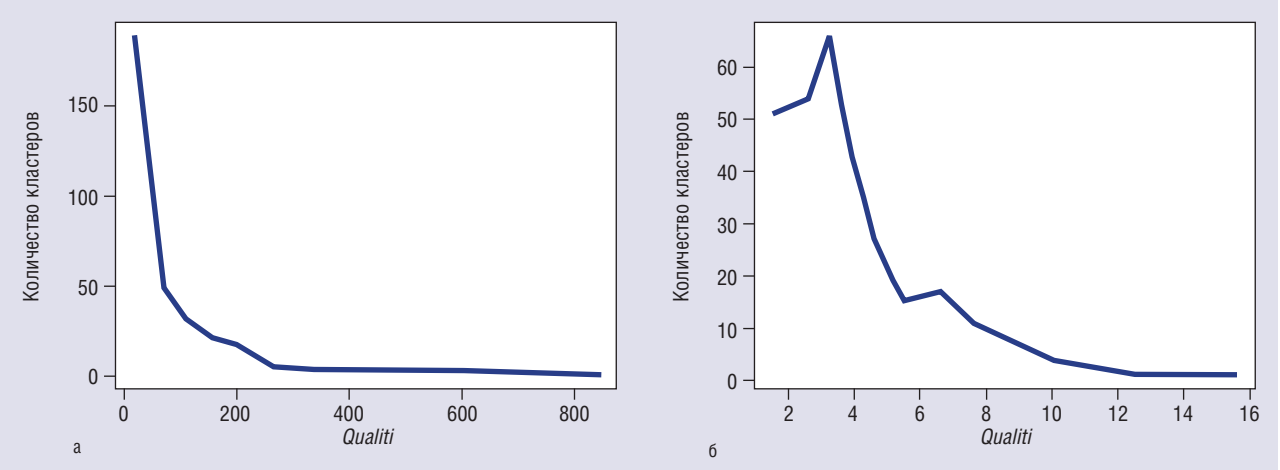


Таблица 4

Количество кластеров в интервалах *Qualiti* мера χ^2 и Жаккара

Интервал (мера χ^2)	Количество кластеров (мера χ^2)	Интервал (мера Жаккара)	Количество кластеров (мера Жаккара)
01. [18.27:43.66]	189	[1.53:2.59]	51
02. [43.66:71.85]	112	[2.58:3.25]	54
03. [71.85:112.3]	49	[3.24:3.89]	66
04. [112.3:156.9]	31	[3.88:4.56]	44
05. [156.9:205.9]	21	[4.56:5.53]	27
06. [205.9:268.7]	17	[5.53:6.57]	15
07. [268.7:337.7]	5	[6.57:7.63]	17
08. [337.7:486.1]	3	[7.63:10.06]	11
09. [486.1:593.5]	2	[10.06:12.74]	4
10. [593.5:842.4]	4	[12.73:15.60]	1
11. [842.4:967.5]	1	[15.60:22.69]	1

возможности объяснения такого сходства простой случайностью. В отличие от меры Жаккара мера Хи-квадрат существенно зависит не только от доли взаимных совпадений при покупках, но также от частоты покупок каждого из двух товаров. При этом сходство между товарами по мере Хи-квадрат существенно возрастает, если каждый из них покупается относительно редко.

Как видно из графиков и таблиц, в нашем случае можно взять все кластеры, кроме тех, что находятся в первых двух интервалах, то есть кластеры со значением *Qualiti* с 43.6643 по 967.5333 по мере χ^2. Иначе говоря, всего 107 кластеров являются интересными, их стоит предъявить экспертам. Отбрасывание кластеров, принадлежащих первым двум интервалам, обусловлено тем, что существуют два больших разрыва в количестве кластеров, в которых резко теряется качество и повышается количество: в шестом и во втором интервалах. Кластеры с шестого по одиннадцатый будем считать кластерам с сильной связью, кластеры со второго по пятый имеют среднюю связь, а остальные слабую. Такой подход позволяет сфокусировать мнения экспертов на следующих вопросах:

• Почему между товарами сильная или слабая связь?

• Есть ли среди товаров с сильной связью товары со слабой связью?

• Есть ли среди товаров со слабой или средней связью товары с сильной связью?

Такого рода вопросы дадут возможность получить новую информацию об особенностях (патернах) продаж.

Для оценки качества кластеризации в каждом кластере выделяется ядро из пяти наиболее востребованных товаров. Товар J относится в кластер i, для которого $P(J|C\,i\,ker)$ максимальна, где ker — событие, состоящее в покупке хотя бы одного из товаров ядра. Оценка качества кластеризации:

$$\frac{\sum_J P\left(C^{ker}_{i(J)}\middle|J\right)}{N},$$

где N — число товаров; $i(J)$ — номер кластера, в который был отнесен товар, вычисляется стандартно по формуле Байеса:

$$\sum_J P\left(C^{ker}_{i(J)}\middle|J\right).$$

Списки товаров по кластерам были предъявлены экспертам — опытным работникам торговли. Им был задан вопрос: можете ли вы описать покупателя, который купит эти товары, и отнести его к тому или иному типу; можно ли как-то одним словом назвать покупателя товаров этого кластера? По итогам опроса получилось следующее: есть кластеры, которые покупают все типы покупателей, а есть кластеры, которые покупают только определенные типы. Это может быть основой для формирования профиля клиентов, что позволит более точно управлять промоактивной деятельностью и ценообразованием в следующих аспектах.

• Товары, входящие в кластеры с небольшим размером, являются паттернами, которые в первую

Александр Докукин, Юрий Журавлев, Олег Сенько, Дмитрий Стефановский

➤➤ С появлением чеков в электронном виде покупатели стали оставлять цифровой след о продажах, что делает целесообразным применение методов машинного обучения для получения информация о структуре предпочтений клиентов.

очередь требуют пристального маркетингового изучения.

• Ядра кластеров позволят организовать кастомизацию продаж и различные автоматизированные рекомендации. Под ядрами кластеров мы понимаем первые пять популярных товаров.

Выводы

• Традиционные способы кластеризации, основанные на определении количества кластеров по внешним коэффициентам, не подходят.

• Для оценки качества кластеризации информации о продажах, характеризующейся сильной разряженностью и большими размерностями, целесообразно использовать меру χ^2 .

• Кластеризация на несколько сотен кластеров требует ранжирования результатов кластеризации, то есть существуют значимые и незначимые кластеры. Ранжирование позволяет повысить качество использования экспертных знаний за счет предъявления экспертами данных из разных диапазонов ранжировок.

• В будущих работах предстоит исследовать другие метрики, чтобы выбрать оптимальную. В этой работе было продемонстрировано, что качество кластеризации зависит от способа оценки расстояния между объектами, и показано, как выбирать из небольшого числа мер. Также предстоит понять, каким образом можно ускорить подбор способа оценки расстояния.

ПЭС 19004 / 05.01.2019

Источники

1. Festa J.P. New technologies and emerging threats: personnel security adjudicative guidelines in the age of social networking // Naval postgraduate school Monterey CA, 2012.

2. Ручьева А.С. Категорийный менеджмент в розничном канале продаж: сущность концепции и актуальные направления исследований // Вестник Санкт-Петербургского университета. Серия 8. Менеджмент. 2015. № 3.

3. Агеев А.И., Логинов Е.Л. Формирование организационных и информационных механизмов управления построением в России цифровой экономики // Экономические стратегии. 2018. № 3. С. 56–67.

4. Dunn G., Everitt B. An Introduction to Mathematical Taxonomy Cambridge University Press Cambridge // UK Google Scholar. 1982.

5. Choi S.-S., Cha S.-H., Tappert C.C. A survey of binary similarity and distance measures // Journal of Systemics, Cybernetics and Informatics. 2010. Т. 8. N 1. P. 43–48.

6. Jenks G.F. The Data Model Concept in Statistical Mapping // International Yearbook of Cartography. 1967. Vol. 7. P. 186–190.

References

1. Festa J.P. *New technologies and emerging threats: personnel security adjudicative guidelines in the age of social networking*. Naval postgraduate school Monterey CA, 2012.

2. Ruch'eva A.S. Kategoriinyi menedzhment v roznichnom kanale prodazh: sushchnost' kontseptsii i aktual'nye napravleniya issledovanii [Category Management in the Retail Sales Channel: the Concept Essence and Current Directions of Research]. *Vestnik Sankt-Peterburgskogo universiteta*, seriya 8, Menedzhment, 2015, no 3.

3. Ageev A.I., Loginov E.L. Formirovanie organizatsionnykh i informatsionnykh mekhanizmov upravleniya postroeniem v Rossii tsifrovoi ekonomiki [Forming Organizational and Information Management Mechanisms Building a Digital Economy in Russia]. *Ekonomicheskie strategii*, 2018, no 3, pp. 56–67.

4. Dunn G., Everitt B. An Introduction to Mathematical Taxonomy Cambridge University Press Cambridge. *UK Google Scholar*, 1982.

5. Choi S.-S., Cha S.-H., Tappert C.C. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 2010, t. 8, no 1, pp. 43–48.

6. Jenks G.F. The Data Model Concept in Statistical Mapping. *International Yearbook of Cartography*, 1967, vol. 7, pp. 186–190.