Практический маркетинг. – 2007. – №4. С. 2–8.

ПРАКТИЧЕСКИЕ ВОПРОСЫ ОБРАБОТКИ АНКЕТНЫХ ДАННЫХ

С.Н. Мартышенко, Н.С. Мартышенко

Условиям экономически развитого общества отвечает стремление принимать решения в социальной и экономической сфере с учетом мнений населения, чьи интересы они затрагивают. Основным источником информации, отражающей мнения населения, был и остается анкетный опрос.

Обработать данные опросов невозможно без использования компьютерной техники. Сегодня такая техника доступна практически каждому исследователю. Однако в специализированных средствах обработки анкетных данных на компьютере ощущается острый дефицит [4]. Необходимость использования специализированных средств обусловлена спецификой данных, получаемых в ходе анкетных опросов,которая заключается в том, что они содержат большое количество нечисловой информации, порождаемой использованием в анкетах разнообразных измерительных шкал [3]. Распространенные компьютерные программы, напротив, нацелены на обработку числовой информации.

При обработке достаточно большого количества данных, полученных в ходе различных анкетных опросов, невозможно обойтись без специальных программных средств, учитывающих специфику данных опросов. В течение последних лет мы уделяли большое внимание разработке комплекса программных средств по обработке анкетных данных [2] в связи с тем, что многие задачи не могли решить с помощью стандартных программных средств, таких как EXCEL, SPSS, Statistica и других. При разработке программных продуктов мы исходили из реальных проблем, с которыми сталкивались при обработке больших пакетов анкетных данных. Поэтому в нашем программном комплексе уделено большое внимание проблеме повышения качества данных и обработке многомерных данных различной природы.

Многие признаки, полученные по анкетным данным, носят нечисловой характер. В рамках пакета, в частности, были реализованы методики обработки нечисловых признаков, которые получаются при компьютерном представлении ответов на открытые и открытые составные вопросы [2]. Открытый или неструктурированный вопрос [1] наиболее сложен с точки зрения компьютерной обработки. В отличие от закрытых, такой вопрос не содержит подсказок, не “навязывает” тот или иной вариант ответа и рассчитан на получение неформализованного мнения. Еще чаще чем открытый вопрос, встречается полузакрытый вопрос, который кроме определенного числа вариантов ответа, содержит позицию “другое – укажите какое (что, где, как)”. Известны и иные формы открытого вопроса: “завершение предложения”, ”подбор ассоциации” и другие.

Большинство исследователей не применяют компьютерную обработку открытых вопросов, а используют их в поисковых целях для получения информации для будущих исследований. Между тем, ответы на эти вопросы могут оказаться очень информативными.

При открытой форме вопроса можно было бы ожидать, что респонденты не дадут одинаковых ответов. На практике, перечень действительно различных по сути, а не по форме ответов на такие вопросы анкет ограничен. Уже при выборке порядка 700 анкет можно выделить всего 30-40 различных возможных вариантов ответов. При увеличении объема выборки картина практически не изменяется. Выделенные варианты ответов можно интерпретировать, как значения признака, измеренного в номинальной шкале.

Наличие 30-40 вариантов значений тоже слишком большое количество для анализа измерений в номинальной шкале. Поэтому исследователь после формирования приемлемого списка действительно различных вариантов ответов, должен сгруппировать эти ответы, рассматривая их как некоторые характеристики непересекающихся классов (типов) респондентов. То есть, в соответствии с его ответом, каждому респонденту можно сопоставить некоторый идентификатор группы ответов.

Конечно, объединение ответов в группы будет носить субъективный характер, но, тем не менее, оно совершенно оправдано с точки зрения социологической теории личности, которая выделяет определенное количество типов личности. Это подтверждается большим количеством независимых исследований ученых из различных стран, которые приходили не более чем к 7-8 типам. В реальных исследованиях каждому из выделенных типов респондентов присваивается определенное название (идентификатор), ассоциированное с темой исследования. С математической точки зрения название не имеет никакого значения, а имеет смысл только операция объединения ряда значений признака в одну группу. Поэтому типы (классы) респондентов могли бы быть просто пронумерованы в произвольном порядке.

Таким образом, с содержательной точки зрения операция преобразования открытого вопроса к номинальной шкале или иначе операция типизации не так уж и сложна. Решение задачи типизации значений признака, порожденного открытым вопросом, можно производить с помощью стандартных средств EXCEL, используя функции сортировки и корректировки данных. Однако при больших объемах выборки такой способ будет весьма трудоемким. Один и тот же ответ можно выразить десятками способов. Даже различие в одном символе компьютер воспринимает, как различный ответ.

Достаточно поменять порядок слов ответа и один и тот же ответ окажется в различных частях отсортированного списка.

Для решения этой задачи нами было разработано специальное инструментальное средство, которое позволяет автоматизировать деятельность исследователя при поиске типологий по большим спискам первичных ответов на открытый вопрос. Разработанная программа позволяет решать не только задачу типизации в простейшем случае, которая была рассмотрена выше, но и допускает решение более сложных задач, встречающихся на практике.

В начале рассмотрим работу программы при решении простой задачи типизации. Поскольку программа предназначена для работы в программной среде EXCEL, то и принцип работы и возможности программы должны демонстрироваться в этой среде.

Учет в программе всех особенностей задачи позволяет на порядок сократить время получения конечного результата по сравнению с решением задачи стандартными средствами EXCEL. Кроме того, неискушенный пользователь в процессе работы со стандартными средствами может допускать ошибки на каждом этапе многоходовой операции.

Работа с программой начинается с отбора признака, подлежащего типизации. Затем программа формирует на отдельном листе EXCEL рабочую таблицу типизации, включающую четыре столбца. В первом содержится список неповторяющихся значений признака (уникальных значений), второй отведен для ввода названий классов, третий для ввода названий подклассов и в четвертом выводятся частоты повторяемости уникальных ответов. В исходном состоянии второй и третий столбцы не заполнены (рис. 1).

При запуске программы выводится панель управления типизацией (рис. 2). На все время активности программы типизации к таблице уникальных значений признака могут быть применимы все средства EXCEL.

B	C	D	E
Занимаюсь на море	Класс	Подкласс	Частота
вкусно покушать			61
вязать и вышивать			4
дайвинг			58
дискотека			12
загарать и купаться			801
заниматься с детьми			8
заниматься сексом			67
заниматься спортом			72
знакомиться			8
играть в бадминтон			26
играть в баскетбол			5
играть в волейбол			367
играть в карты			23
играть в мяч			29
играть в тенис			6
играть в футбол			21
играть на гитаре			5

Рис. 1. Фрагмент таблицы типизации уникальных значений признака “занимаюсь на море” анкетного опроса по пляжно-оздоровительному отдыху

Первоначально этот список может содержать от 500 до 700 строк. После серии корректировок списка записей с целью его унификации пользователь может выполнить команду “Сжать”. По этой команде все повторяющиеся записи “сжимаются” в одну, а соответствующие частоты уникальных значений признака пересчитываются.

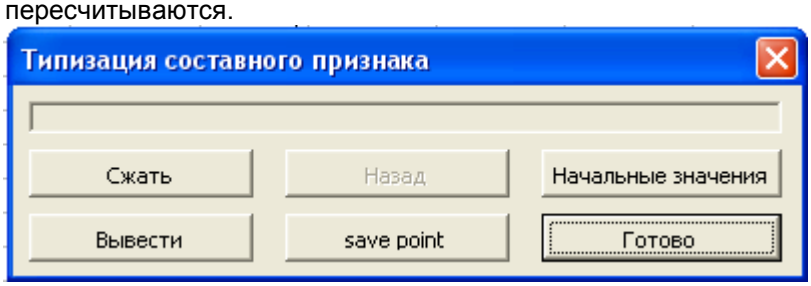

Рис. 2. Панель управления программой типизации

Корректировка одной записи таблицы уникальных значений эквивалентна корректировке множества связанных с ней записей исходной таблицы данных. При повторении нескольких циклов выполнения действий “корректировка – сжатие“, список уникальных значений быстро сокращается. По мере сокращения списка, время на обдумывание исследователем очередных корректировок возрастает, поскольку ему приходится анализировать все более и более сложные ситуации. Вместе с тем, при сокращении списка существенно сокращается время, затрачиваемое исследователем на поиск однотипных ответов.

Частоты повторения уникальных ответов (четвертый столбец таблицы) служат весьма полезной информацией для логических рассуждений исследователя. Исследователь, в первую очередь, сосредотачивает свое внимание на ответах, имеющих высокие частоты, и пытается свести к

ним все остальные ответы, если это не приводит к искажению смысла ответов. В конечном итоге список удается сократить в десять и более раз, причем без искажения информации.

После завершения операции типизации признака пользователь может, либо заменить значения исходной выборки, либо, в случае сомнений в корректности действий, разместить столбец признака с замещенными значениями на новом месте. В частном случае, эта программа может быть использована для корректировки любого признака или построения частотных рядов признака. Кроме того, пользователь и сам может находить другие ситуации использования программы.

При выполнении операции типизации в полном объеме, исследователь объединяет ответы в группы, вводя названия (или номера) классов во второй столбец. В простом случае третий столбец повторяет первый. Однако, при выполнении операции на реальных данных, возникает необходимость внесения в третий столбец значений, более общих, чем в первом столбце. В реальной ситуации могут встретиться очень близкие по смыслу, но все-таки различные ответы. Например, ответы “пробки на дорогах” и “отсутствие автостоянок”, можно было бы заменить одним обобщенным ответом – “транспортные проблемы”. Создавать два подкласса по очень близким по смыслу ответам бывает нецелесообразно, поскольку это может привести к чрезмерному количеству вариантов с крайне низкой частотой встречаемости. С другой стороны, иногда не желательно терять информацию при замене двух вариантов ответов одним обобщенным. При пополнении базы данных за счет новых анкет может оказаться, что один из этих ответов достигнет такого уровня встречаемости, когда его будет целесообразно выделить, как вполне самостоятельный вариант.

Поэтому для сохранения информации “на будущее”, используется следующий подход. В строки таблицы уникальных значений, соответствующие приведенным выше ответам, вносят следующие значения: “транспортные проблемы (пробки на дорогах)” и “транспортные проблемы (отсутствие автостоянок)”, а в столбец подкласс для обоих ответов вносят обобщенное значение “транспортные проблемы”. Определив названия классов и подклассов, исследователь может вывести результаты типизации в форме таблиц частот и создать новые признаки в таблице данных, составленные из значений, ассоциированных с названиями классов или подклассов.

Операция типизации допускает обобщение на случай, когда респондент на один вопрос может дать не один ответ, а несколько. При этом ответы записываются в одном столбце таблицы данных, соответствующем вопросу. Несколько простых ответов разделяются каким либо знаком (“; “ или “,”). Такой признак мы определяем как составной. Например, на вопрос о любимых занятиях в пляжной зоне респондент может ответить: “осматривать достопримечательности; играть в бадминтон; читать”. В этом случае ответ содержит три простых ответа. Составной признак – это некоторая форма записи или компьютерного представления ответов на вопрос, допускающий несколько вариантов ответа.

Для обработки таких множественных ответов применяется многошаговая типизация. В этом случае в таблицу уникальных значений включаются все возможные варианты простых ответов. В результате типизации составного открытого ответа будут получены и составные признаки в номинальной шкале измерения.

Составной признак может быть получен не только при записи ответов на открытый вопрос, но и при записи ответов на любой другой вопрос, в котором респондент может выбирать из списка вариантов ответа на вопрос анкеты не один, а несколько вариантов. Причем, различные респонденты могут выбирать различное количество вариантов.

Формально, составной признак можно определить как последовательности, составленные из нескольких возможных вариантов ответа или идентификаторов классов. Список возможных вариантов обозначим, как $\mu = (\mu_1, \mu_2, \ldots, \mu_j \ldots, \mu_k) \quad j = \overline{1,k}$, k – количество возможных вариантов ответа (или групп ответов). Операция типизации как раз и позволяет сформировать такие списки.

При построении частотного ряда простых значений, входящих в составной признак, возникает неоднозначность, которая не может быть разрешена с помощью стандартных средств. Для того чтобы дать формализованное описание возможных способов построения частотных рядов, представим составной признак в обобщенном числовом формате (табл. 1)

Таблица 1

Числовая форма представления составного ответа

Номер анкеты	Номер группы ответов						
	1	2	3	…	j	…	к
1	r_{11}	r_{12}	r_{13}		r_{1j}		r_{1k}
2	r_{21}	r_{22}	r_{23}		r_{2j}		r_{2k}
3	r_{31}	r_{32}	r_{33}		r_{3j}		r_{3k}

…							
i	r_{i1}	r_{i2}	r_{i3}		r_{ij}		r_{ik}
…							
n	r_{n1}	r_{n2}	r_{n3}		r_{nj}		r_{nk}

В таблице 1 приняты следующие обозначения:

r_{ij} - количество простых ответов μ_j в составном признаке *i*- ой анкеты;

i – номер анкеты *i*=1,2,3,…*n;*

j - номер группы ответов *j*=1,2,3,…*k*.

По данным табл. 1 можно построить частотные ряды двумя способами или получить две модификации частотных рядов. Частоту встречаемости *j* – ого простого значения признака можно рассчитать по формуле:

$$P_j^{(1)} = \frac{\sum_{i=1}^{n} r_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}} \qquad (1)$$

и по формуле:

$$P_j^{(2)} = \frac{\sum_{i=1}^{n}\left(\frac{r_{ij}}{\sum_{j=1}^{k} r_{ij}}\right)}{n}. \qquad (2)$$

Обе эти формулы дают значения, отвечающие основному свойству частотного ряда:

$$\sum_{j=1}^{k} P_j^{(1)} = \sum_{j=1}^{k} P_j^{(2)} = 1 \qquad (3)$$

В каждом конкретном случае частотные ряды, рассчитанные по формулам (1) и (2), могут существенно отличаться. То есть, для составного признака имеет место неоднозначность расчета частотного ряда.

Предпочтение тому или иному способу отдается в зависимости от того, какой содержательный смысл имеют значения составного признака. Если значения имеют смысл типа личности, то встречаемость в одной строке исходной таблицы (табл. 1) нескольких различных значений мы можем интерпретировать как то, что конкретный респондент обладает чертами сразу нескольких типов личности. В этом случае для расчета частотного ряда предпочтительней использовать формулу (2).

Рассмотрим другой случай, приводящий к составному ответу. Например, если мы спрашиваем респондента о том, какие виды развлекательно-оздоровительных учреждений он посещает, то простые ответы из составного ответа “ресторан; фитнес-клуб” целесообразно учитывать по первой схеме. То есть такой потребитель создает нагрузку двум различным типам предприятий.

С формальной точки зрения составные ответы в двух рассмотренных случаях тоже имеют различия. В первом случае r_{ij} может принимать значения 0,1,2,3, …, а во втором только значения 0,1.

Программные модули построения модифицированных частотных рядов по составным признакам также включены в разработанный нами специализированный пакет обработки анкетных данных. Кроме того, пакет включает модули, позволяющие преобразовывать составные признаки к простым и обратно.

Расчеты на реальных данных показали очень высокую устойчивость числовых характеристик частотных рядов, построенных по данным, полученным в результате типизации ответов на открытые вопросы. Поэтому эти данные могут выступать в роли характеристик исследуемых совокупностей. Результаты типизации могут быть с успехом использованы для анализа структуры потребителей товаров и услуг. Апробацию, рассматриваемых в работе, методик анализа анкетных данных мы производили на данных опросов потребителей продуктов туристского комплекса региона.

При выборе стратегии развития туристской отрасли необходимо ориентироваться на сложившуюся структуру потребления. Выбор стратегии - это выбор действий, которые должны создать условия для изменения структуры потребления в желаемом направлении. Для исследования

структурных сдвигов потребления также использовались составные вопросы. Решение этой задачи мы производили на основе собственных маркетинговых исследований востребованности услуг туристского комплекса Оценка конъюнктуры, сложившейся на рынке туристских услуг, производилась на основе нескольких анкетных опросов.

Поскольку потребление товаров и услуг предприятий туристского комплекса население края, в основном, производит в отпускное время, мы предприняли попытку исследовать структурные характеристики времяпрепровождения отпускного периода жителей края. Один из анкетных опросов был предпринят для изучения времяпрепровождения отпусков. Опросы производились в течение последних четырех лет. За это время были опрошены более пяти тысяч человек. Анкета позволяет оценить тенденции структурных изменений в сфере потребления услуг комплекса. Такие оценки можно построить, поскольку в анкете имеются ряд вопросов требующих от респондентов предоставления информации за последние два года.

Для анализа структурных изменений спроса потребителей на услуги туристской индустрии были использованы две программы разработанного комплекса программных средств. Рассмотрим принцип работы этих программ.

Первая программа выполняет вспомогательные функции. Ее назначение состоит в преобразовании компьютерного представления некоторых видов данных, получаемых в результате анкетных опросов.

Например, для ввода данных по вопросу анкеты “Как вы проводите отпуск?”, оператор быстрее всего вводит данные в форму, приведенную на рис. 3. Для компьютерного представления данных ответов на такой вопрос необходимо зарезервировать на каждый вариант ответа и каждый год отдельный бинарный признак, принимающий два значения: 1 - “истина” или 0 - “ложь”. Таблица значений признаков, описывающая ответы на вопросы, при такой форме представления, будет в основном состоять из нулей. Однако, если преобразовать данные из бинарного представления к форме составного вопроса, то можно добиться компактности и наглядности представления данных. В составном признаке несколько ответов на один вопрос считаются одним значением. Отдельные варианты ответа отделяются друг от друга знаком разделителя (как правило, используется “;”)

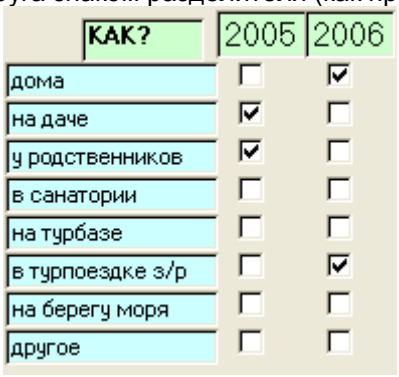

Рис. 3. Форма ввода таблицы данных вопроса анкеты для оператора

Для хранения данных, представленных на рис. 3 потребуется два составных признака: “Как? 2005” “Как? 2006”. Например, значение составного признака “Как? 2005” для данных формы рис. 3 будут иметь вид “на даче; у родственников”. Функция первой из рассматриваемых программы обработки данных – это преобразование нескольких бинарных признаков в один составной.

Вторая программа рассчитывает по двум составным признакам, относящимся к двум различным временным этапам, матрицу структурных переходов M размерности $k \times k$. Рассмотрим методику расчета элементов матрицы M.

В простейшем случае, значения двух сравниваемых составных признака включают только по одному ответу. Например, если признак, ассоциированный с первым временным этапом, принял значение μ_i, а признак, связанный со вторым временным этапом, принял значение μ_j, то зафиксировать переход из состояния μ_i в состояние μ_j можно прибавлением единицы к элементу m_{ij} матрицы переходов M.

Если значение первого составного признака включает q_1 простых значений, а второго составного признака включает q_2 простых значений, то можно составить $v = q_1 \times q_2$ различных вариантов переходов. Для каждой пары простых значений (μ_i, μ_j) индексы элементов определяют свой элемент m_{ij}, но теперь к нему будем прибавлять не единицу, а некоторый весовой коэффициент:

$$\varphi = \frac{1}{\nu}. \quad (4)$$

В результате просмотра и сравнения всех значений двух составных признаков, относящихся к двум временным этапам, можно рассчитать элементы матрицы M. Сумма всех элементов матрицы M будет равна количеству анкет N. Определим сумму элементов матрицы M по строкам:

$$n_r = \sum_{s=1}^{k} m_{rs}, \quad r = \overline{1,k}. \quad (5)$$

Очевидно, будет выполняться условие:

$$\sum_{r=1}^{k} n_r = N. \quad (6)$$

Разделив построчно элементы матрицы M на величину $n_r (r = \overline{1,k})$, получим матрицу переходов F, измеряемую в относительных единицах. Элементы матрицы F рассчитываются по формуле:

$$f_{rs} = \frac{m_{rs}}{n_r}, \quad r = \overline{1,k}; s = \overline{1,k}. \quad (7)$$

Для каждой строки матрицы F будет выполняться условие:

$$\sum_{s=1}^{k} f_{rs} = 1, \quad r = \overline{1,k}. \quad (8)$$

Для того чтобы выделить только наиболее существенные переходы введем некоторое пороговое значение $0 < d < 1$ и рассчитаем элементы матрицы P по формуле:

$$P_{rs} = \begin{cases} 1, & если\ f_{rs} \geq d \\ 0, & если\ f_{rs} < d \end{cases} \quad r = \overline{1,k}; s = \overline{1,k}. \quad (9)$$

Структурные переходы, описываемые с помощью матрицы P, удобно представить в виде ориентированного графа (рис. 4), вершины которого соответствуют номерам вариантов возможных ответов на исследуемый вопрос, стрелками соединены вершины графа, для которых элементы $P_{rs} = 1, \quad r = \overline{1,k}; s = \overline{1,k}$.

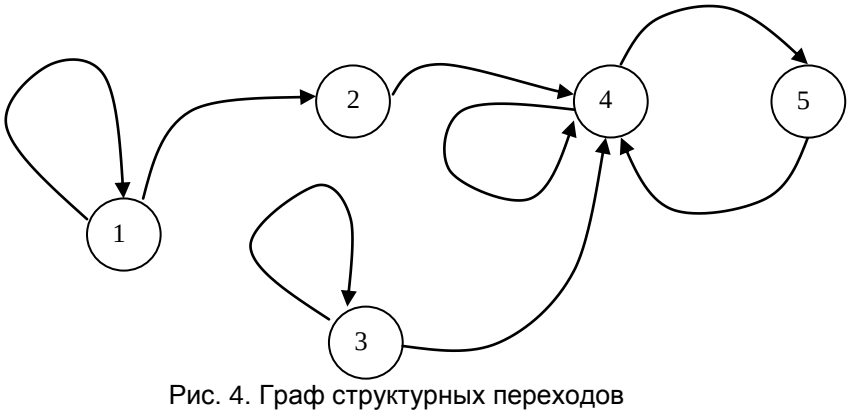

Рис. 4. Граф структурных переходов

Граф на рис.4. был построен по значениям конкретной матрицы (10):

$$P_{rs} = \begin{vmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{vmatrix}. \quad (10)$$

В практической работе целесообразно сравнить графы для различных социально-демографических групп потребителей. Величина порогового значения подбирается экспериментально так, чтобы обеспечить наибольшую наглядность графов. Построение графов

переходов особенно важно на предварительных этапах исследования поведения потребителей. Анализ графов позволяет сформулировать гипотезы, объясняющие происходящие изменения в структуре потребления туристских услуг.

Наличие средств по обработке открытых вопросов обеспечивает широкому кругу исследователей новые возможности сбора первичного материала методом анкетного опроса.

К числу достоинств, разработанных программных модулей, мы относим то, что даже при очень больших выборках они позволяют получать результаты в реальном времени, что открывает большие возможности для экспериментальной работы исследователя.

Рассмотренные программные средства входит в состав разработанного нами специализированного комплекса программных средств обработки анкетных данных, предназначенного для работы в среде EXCEL [2]. Подход, состоящий не в разработке собственного автономного пакета программных средства, а в расширении функций распространенного среди широкого круга практиков пакета, на наш взгляд наиболее отвечает сегодняшнему уровню использования программных средств по обработке данных. Разрабатывая собственную технологию решения специфических задач по обработке анкетных данных, мы можем использовать всю мощь пакета EXCEL, как при выполнении отдельных промежуточных операций, так и при оформлении результатов.

Литература

1. Малхотра Нэреш К. Маркетинговые исследования. – М.: Вильямс,2002. – 960с.
2. Мартышенко С.Н. Совершенствование математического и программного обеспечения обработки первичных данных в экономических и социологических исследованиях / С.Н. Мартышенко, Н.С. Мартышенко, Д.А. Кустов // Вестник ТГЭУ. – 2006. – № 2 – С. 91–103.
3. Орлов А.И. Нечисловая статистика / А.И.Орлов. – М.: МЗ-Пресс, 2004. – 513 с.
4. Толстова Ю.Н. Анализ социологических данных. Методология, дескриптивная статистика, изучение связей между номинальными признаками. – М.: Научный мир, 2000. - 352с.