Мартышенко С.Н., Мартышенко Н.С.

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ ПОВЫШЕНИЯ КАЧЕСТВА ДАННЫХ, ПОЛУЧЕННЫХ ПРИ ИССЛЕДОВАНИИ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ СИСТЕМ

Новые методы анализа информации, характеризующей состояние социально экономических систем, выдвигают более строгие требования к качеству данных. Среди таких данных особое место занимают данные анкетных опросов. Анкетный опрос один из основных источников информации, отражающей:

- реакцию населения на решения и действия, предпринимаемые органами государственного управления всех уровней;
- мнения потребителей, определяющие выбор различных товаров и услуг.

Устремления России к созданию общества демократического согласия и условий для здоровой конкуренции производителей в борьбе за потребителей неуклонно повышают интерес к исследованиям, основанным на анкетных опросах. Сегодня редкий университет не выпускает специалистов по маркетингу и социологии, у которых одним из основных инструментариев исследования является анкетный опрос. И все эти специалисты ощущают острую потребность в методах и компьютерных технологиях обработки статистических данных анкет.

Обработка анкетных данных переходит с уровня научного исследования на уровень практики повседневной работы многих предприятий. Преимущество смогут получить те, кто будет использовать более совершенные методы обработки статистических данных, основанные на последних достижениях в области прикладной статистики и компьютерных технологий.

Данные анкетных опросов имеют ряд существенных отличий от классических статистических данных учетного характера. Поэтому для их обработки необходима разработка специальных методов и программного обеспечения.

Можно выделить ряд особенностей анкетных данных.

Первая особенность заключается в том, что эти данные включают признаки различной природы. Многие признаки являются нечисловыми и качественными. В работе известного статистика И.И. Елисеевой [1] возрастание доли нечисловой информации в собираемых статистических данных объясняется следующими причинами:

- стремлением учесть человеческий фактор, выявить ориентации и предпочтения людей;
- сбором информации в форме нечисловых данных с тем, чтобы не затронуть количественные показатели, составляющие коммерческую тайну;
- использованием рейтингов (банков, предприятий, учебных заведений, политических деятелей и т.д.).

Большое количество нечисловой информации, порождается использованием в анкетах разнообразных измерительных шкал [7 и 8]. Наличие разнообразных шкал вызвано не прихотью исследователей, а их стремлением получить от респондентов более достоверную информацию. Поскольку не респондент, а исследователь заинтересован в получении информации, ему и приходится подстраиваться под респондента, предоставляя респонденту вопросы в такой форме при которой он сможет или пожелает ответить. Исследователь всегда вынужден искать компромисс между желаемой информацией и информацией, которую он может получить. Качественная информация часто является гораздо более содержательной. Однако для ее обработки нужно использовать свои методы. Большинство распространенных компьютерных программ, напротив, нацелено на обработку числовой информации.

Анкетные данные содержат от 70% до 90% нечисловой информации. Даже информация, представленная в числовом виде, таковой является весьма условно. Это, как правило, экспертные оценки респондентов средних значений каких-либо характеристик или показателей изучаемого явления или процесса. При анализе любых статистических данных не обойтись без содержательного анализа данных и результатов их обработки. Преобладание в анкетах качественных данных приводит к тому, что роль содержательного анализа намного выше, чем при обработке числовых данных (**вторая особенность**).

Присутствие в процессе формирования данных человеческого фактора в виде респондентов, которые являются далеко не квалифицированными экспертами и, как правило, привлекаются к этому виду деятельности в разовом случае, накладывает свой отпечаток на всю систему сбора данных.

Таким образом, анкетный опрос представляет собой некоторый специфический способ измерения. Специфика этого способа измерения состоит в высокой степени неопределенности оценок достоверности данных, которую можно выделить в качестве **третьей особенности данных анкетных опросов**. Неопределенность обусловлена тем, что данные имеют множество источников ошибки (рис. 1).

Среди ошибок в данных можно выделить особый вид ошибок – это ошибки не наблюдения или пропуска в данных. Эти ошибки могут быть настолько значительными, что их присутствие можно обозначить, как **четвертую особенность анкетных данных.**

При разработке информационных технологий обработки анкетных данных необходимо учитывать еще ряд особенностей присущих реальным исследованиям, основанным на таких данных. Реальные таблицы данных содержат очень большое количество признаков. Количество признаков может достигать ста и более единиц. Для получения надежных оценок по различным подмножествам выборки требуется значительное количество наблюдений (количество наблюдений может достигать нескольких тысяч). Большую размерность данных можно выделить как **пятую особенность.**

Шестая особенность тоже является в большой степени следствием размерности данных. Но при разработке технологии обработки данных должна быть выделена отдельным пунктом. Анализ данных включает очень большое количество задач и может занимать значительные отрезки времени. Процесс обработки может растянуться на месяцы и более. Таким образом, длительность периода обработки данных – **шестая особенность.**

Седьмая особенность состоит в том, что процесс обработки данных часто строится как поисковая задача. До получения данных мы можем только предполагать схему обработки, но результаты обработки могут порождать все новые и новые задачи. Обработка данных носит творческий характер.

К сбору анкет привлекаются временные сотрудники – интервьюеры, которые по-разному относятся к порученной работе. Необходимость учета личности интервьюера, предоставляющего данные является **восьмая особенность.**

Рис. 1. Источники ошибок при проведении анкетного опроса

Кроме того, анкетные опросы, производимые на профессиональной основе, не проводятся как единичная акция. По мере анализа данных анкета постоянно совершенствуется, как по содержанию, так и по форме. Удачную анкету целесообразно использовать в нескольких опросах – распространение процесса во времени. Сбор и обработка данных по одной анкете происходят на фоне опросов по другим анкетам – параллельные процессы. Отработанные блоки вопросов могут быть включены в виде модулей, связывающие различные опросы. Многие базы данных анкетных опросов, кроме как информация для обоснования управленческих решений могут быть использованы для научной работы многих других исследователей (препарирование).

Постоянное совершенствование системы сбора и накопление знаний в процессе обработки данных является **девятой особенностью**. Информационная технология должна быть рассчитана на системное накопление знаний в виде базы знаний.

Дсятая особенность состоит в очень высокой степени зависимости системы сбора информации от того, какими методами анализа данных владеет исследователь и того, какие средства компьютерной обработки данных ему доступны. Какой смысл собрать информацию, если исследователь не в состоянии ее обработать. Это приводит к тому, что с одной стороны значительная часть информации, содержащейся в данных, не используется или недостает какой-то малости, что исключает применение современных методов анализа данных, или очень затрудняет их применение.

Статистический анализ конкретных данных включает в себя целый ряд процедур и алгоритмов, выполняемых последовательно, параллельно или по более сложной схеме. В работе известного отечественного ученого А.И. Орлова – автора большого количества работ по прикладным вопросам статистики отмечается, что в научной литературе вопросам рассмотрения технологий обработки статистических данных уделяется явно недостаточное внимание [6]. Обычно все внимание сосредотачивается на том или ином элементе технологической цепочки, а переход от одного элемента к другому остается в тени.

Автор настоящей работы часто сталкивается с мнением, что разработка технологии не является наукой и вся наука заключается в методе. Между тем, многие методы обработки данных, сталкиваясь с проблемами реальных данных, не дают должного эффекта, и еще чаще не “стыкуются” с другими методами и алгоритмами. Тоже самое можно сказать и о программном обеспечении, реализующем эти методы.

Тот же А.И. Орлов утверждает, что говорить о полной автоматизации всего процесса анализа статистических данных говорить преждевременно, потому что слишком много нерешенных проблем, вызывающих дискуссии среди статистиков.

Однако снизить проблему дефицита технологий обработки реальных данных необходимо и возможно. Решение вопросов разработки и исследования возможностей компьютерных технологий обработки специфических данных анкетных опросов является предметом научных изысканий автора последних лет.

Основное внимание при разработке технологии обработки данных было сосредоточено на блоке проблем, связанных с повышением качества данных. То есть разрабатывалась не вся технология, а отдельный блок, который должен создать предпосылки использования методов многомерного статистического анализа, пока не получившего достаточного распространения при обработке анкетных данных, хотя эти данные по своей сути являются многомерными.

Основу технологии составляют методы обнаружения и подавления грубых ошибок. Отличие разработанных методов состоит в рассмотрении не отдельных признаков, а их совокупности, то есть многомерный подход.

Методы повышения качества данных неотрывно связаны с понятием грубой ошибки. Этому понятию невозможно дать однозначное формализованное определение. Поэтому попытаемся уточнить его через некоторые его свойства. Грубой ошибкой можно считать многомерное наблюдение, которое резко отличается на фоне всех остальных. Совокупность значений признаков можно считать грубой ошибкой, если они совместно воссоздают абсурдный, с содержательной точки зрения, объект или его поведение. При этом значения одномерных признаков могут быть вполне правдоподобными. Определить грань, за которой наступает абсурдность объекта, может только сам исследователь в процессе содержательного анализа многомерного объекта. Размытое определение грубой ошибки приводит к обобщенной схеме выявления грубых ошибок (рис.2).

Существует множество вариантов проявления грубых ошибок. Поэтому для их обнаружения необходимо иметь набор инструментальных средств выделения ошибок. Такие алгоритмы работают по принципу многомерных фильтров. Программы позволяют выделить анкеты, которые являются “подозрительными” на содержание грубой ошибки или выброса. Исследователь должен подвергнуть выделенные критические анкеты углубленному содержательному анализу, после чего принять решение о том, как поступить с такими анкетами. Исследователь может оценить ситуацию, как допустимую или как недопустимую. В последнем случае он может либо отбросить данные анкеты как недостоверные и только искажающие конечный результат, либо попытаться восстановить отдельные значения признаков по многомерной выборке. В отдельных случаях он может интерпретировать необъяснимое значение, как ситуацию отсутствия данных или пропуск. Отбрасывание небольшой части данных низкого качества никак не сказывается на репрезентативности выборки. Тем более, что при недостатке данных мы можем произвести опрос дополнительной группы респондентов.

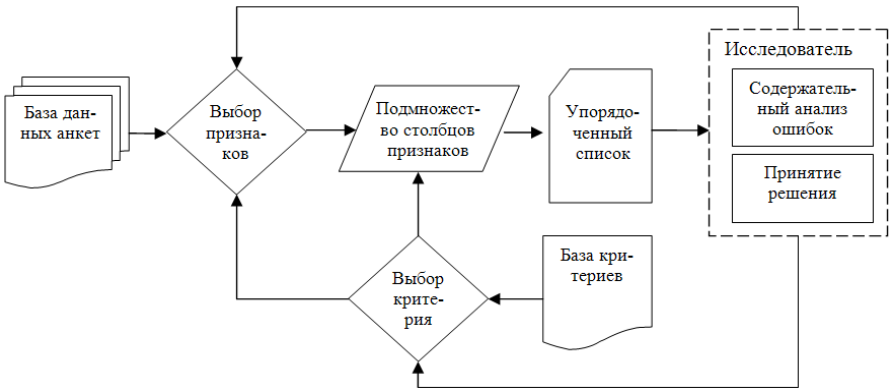

Рис. 2. Обобщенная схема выявления грубых ошибок

Разработанные методы обнаружения грубых выбросов условно можно разделить на статистические и логические. Вначале рассмотрим статистические методы. Анализ выбросов целесообразно начинать с анализа таблицы данных на отсутствие данных, рассматривая пропуски как ошибки.

Компьютерная технология анализа пропусков в данных основывается на правилах описания и компьютерного представления отсутствия данных. Необходимость единообразного описания ситуации отсутствия данных обусловлена требованиями системного подхода к разработке компьютерных технологий обработки данных, которые строятся с учетом некоторых общих свойств данных и специфики решаемых задач.

В состав технологии выявления грубых ошибок в настоящее время включены семь статистических алгоритмов:

Фильтр отсутствия данных (ФОД);
Фильтр экстремальных непрерывных значений (ФЭНЗ);
Фильтр ранжирования непрерывных значений (ФРНЗ);
Фильтр метрический непрерывных значений (ФМНЗ);
Фильтр частот кодированных значений (ФЧКЗ);
Фильтр замены кодированных значений(ФЗКЗ);
Фильтр эталонных значений (ФЭЗ).

В программной реализации каждый алгоритм представлен двумя модулями. Один модуль служит для обработки отдельных наблюдений, второй — для обработки пакетов анкет, представленных различными интервьюерами. При пакетной обработке в название фильтра добавляется буква «Г» (групповой).

Формализованное описание фильтров приводится в работе автора [3]. Список статистических алгоритмов постоянно пополняется новыми алгоритмами.

Кроме статистических фильтров в состав разработанного комплекса входят средства разработки и сопровождения логических фильтров. Логические фильтры органично дополняют статистические. Во многих случаях логические методы позволяют обнаружить противоречия в данных, которые не выявляют статистические методы. Идея поиска логических противоречий состоит в накоплении обнаруженных противоречий в виде логических выражений, которые реализуются в форме настраиваемых фильтров. Фильтры выделяют анкеты, в которых были обнаружены противоречия. В фильтрах может участвовать значительное количество признаков. Логические связи могут быть как между отдельными значениями признаков, так и диапазонами значений. В логических фильтрах могут быть использованы признаки разных типов. Многие противоречия очень затруднительно выявить без специальных программных средств. Окончательное решение о корректировке данных, как и в случае статистических фильтров, принимает исследователь на основе углубленного содержательного анализа конкретной ситуации.

Логические алгоритмы позволяют аккумулировать знания и опыт, полученные в ходе работы над проектом анализа анкетного опроса. Отличие этих алгоритмов состоит в активном участии исследователя в процессе работы программ [2]. Такие алгоритмы зависят от возможностей программной среды, в которой они реализованы. В нашем случае в процессе работы с программами пользователь может использовать весь арсенал средств обработки данных, предоставляемых EXCEL.

Логические фильтры оказываются наиболее полезными для исследователей, которые занимаются анкетным опросом на профессиональной основе. Такие исследования отличает то, что опросы с помощью одной и той же анкеты могут повторяться через какой-то промежуток времени. Одновременно могут проводиться опросы по нескольким анкетам. Различные анкеты могут включать блоки вопросов, которые уже использовались в других анкетах. Эти методы могут быть использованы для проверки корректности восстановления данных при использовании статистических методов.

Логические методы были использованы нами при обработке открытых вопросов [5]. Для обработки таких данных использовались словари замен, которые автоматически пополняются при накоплении данных.

Статистические и логические методы анализа реализованы в виде программного комплекса, который выполнен в форме надстройки к EXCEL. Такой путь был выбран в связи с тем, что большинство пользователей, занимающихся обработкой данных, используют в своей работе EXCEL и легко смогут освоить ряд новых функций. Пользователю совершенно необязательно сразу осваивать все возможности комплекса, он может осваивать их постепенно, переходя от более простых методов к более сложным.

Однако разработанный комплекс нельзя рассматривать как простой набор программ. Программы комплекса образуют единую технологию. Структура и принципы работы специализированного комплекса программных средств обработки анкетных данных представлены в работе [4]. Разработка программного комплекса основана на определении понятий «проекта анкетного опроса» и «модель данных опроса», которые приводят к определенным правилам компьютерного представления информации и доступа к программам комплекса. Структура проекта включает семь элементов: исходные данные по анкетному опросу, параметры проекта, даты изменений, логические фильтры, словари замены, отчеты, изъятые данные. В работе [4] обсуждается содержание и назначение этих элементов. Отдельные модули программного комплекса объединены в четыре раздела по функциональному признаку (рис.3).

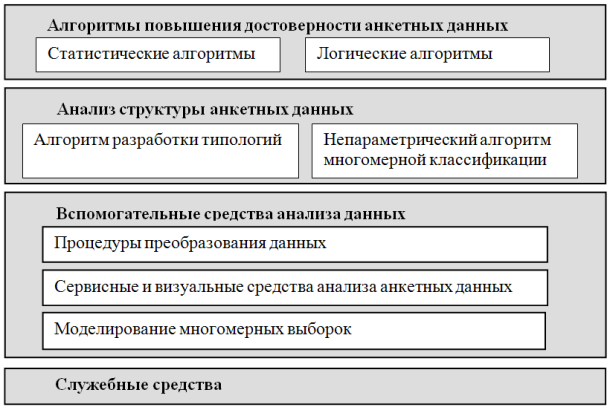

Рис. 3. Основные разделы программного комплекса

Разработанные программные средства прошли апробацию на нескольких крупных проектах анкетных опросов и показали высокую эффективность.

Актуальность работы подтверждается поддержкой гранта РФФИ – ДВОРАН № 06-05-96017 в рамках научно-исследовательской работы "Исследование взаимодействия в системе "биологический объект - внешняя среда" на основе моделирования и обработки данных статистики в условиях ограниченности и неопределенности исходной информации.

ЛИТЕРАТУРА

1. Елисеева И.И., Юзбашев М.М. Общая теория статистики: Учебник / Под. ред. И.И. Елисеевой. – 5 - е изд., перераб. и доп. – М.: Финансы и статистика, 2006. – 656 с.

2. Каневский Е.А., Саганенко Г.И., Гайдукова Л.М., Клименко Е.Н. Диалоговая система классификации и анализа текстов // Социология: 1997. № 9. - С. 198 - 216.

3. Мартышенко С.Н. Многомерные статистические методы повышения достоверности маркетинговых данных / С.Н. Мартышенко, Н.С. Мартышенко, Д.А. Кустов // Практический маркетинг – 2007. – № 119(1. 2007)– С. 20-30.

4. 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.

5. Мартышенко С.Н. Средства разработки типологий по данным анкетных опросов в среде EXCEL / С.Н. Мартышенко, Н.С. Мартышенко, Д.А. Кустов // Академический журнал Западной Сибири. – 2007. – № 1 – С. 75–77.

6. Орлов А.И. Нечисловая статистика / А.И.Орлов. – М.: МЗ-Пресс, 2004. – 513 с.

7. Татарова, Г.Г. Основы типологического анализа в социологических исследованиях: Учеб. пособие /Г.Г. Татарова; Федер. агенство по образованию, Нац. фонд подготовки кадров. - М.: Новый учебник, 2004. -206 с.;

8. Толстова Ю.Н. Анализ социологических данных. Методология, дескриптивная статистика, изучение связей между номинальными признаками. – М.: Научный мир, 2000. - 352с.