УДК 004.838.2:316.334.3

КАЧЕСТВЕННЫЙ АНАЛИЗ СОЦИОЛОГИЧЕСКОГО ИССЛЕДОВАНИЯ «ОТНОШЕНИЕ ГРАЖДАН УКРАИНЫ К РЕФОРМАМ»

В.А. САЛАМАТОВ, Т.А. ТАРАН, А.Е. ТКАЧЕВ, С.Н. КОПЫЧКО

Рассматривается новый подход к анализу социологических данных, основанный на автоматической обработке логических закономерностей. Предложена стратегия поиска наиболее информативных логических правил, которая проиллюстрирована на примере социологического исследования «Отношение граждан Украины к реформам».

ВВЕДЕНИЕ

Социологические исследования направлены на выявление глубинных причин явлений, которые происходят в обществе, — так называемых «детерминант социального поведения» [1]. Они помогают понять их, опровергнуть или подтвердить гипотезы, выдвинутые об этих явлениях. На первом этапе исследования опрашиваются респонденты (граждане), количество которых должно быть репрезентативным, для того чтобы получить достаточно достоверные и надежные выводы. На втором этапе собранная информация анализируется (ищутся существенные связи и закономерности с целью выработки рекомендаций для принятия управленческих решений).

В настоящее время в социологических исследованиях широко используются статистические методы анализа данных [2]. Однако при этом происходит усреднение ответов по всем респондентам, в результате чего теряется информация об их индивидуальных особенностях. Поэтому с помощью статистической обработки сложно вскрыть закономерности и причины явлений, происходящих в социуме. В отличие от статистических методов, качественные методы анализа данных не используют усреднение по выборке, но, тем не менее, позволяют обнаруживать существенные закономерности в исходных данных. В социологии ситуация осложняется тем, что сами исходные данные носят преимущественно качественный характер и плохо поддаются числовой обработке. В таких условиях применение качественных методов анализа является альтернативой традиционным количественным методам, а иногда и единственно возможным средством анализа социологических данных.

В этой работе предлагается методика качественного анализа социологических данных на примере исследования отношения граждан Украины к реформам.

© *В. А. Саламатов, Т. А. Таран, А. Е. Ткачев, С. Н. Копычко,* 2003

ОСОБЕННОСТИ СОЦИОЛОГИЧЕСКИХ ДАННЫХ

К качественным методам анализа относят обнаружение формальных логических закономерностей, связей и ассоциаций в исходных данных [3]. Прежде всего, они применяются для исследования данных, для которых наиболее естественной является нечисловая (качественная) форма представления. В социологических анкетах в основном используются именно такие данные. Это, во-первых, все номинальные данные, например, ответы на вопросы о семейном положении, образовании, месте жительства и т.д.; во-вторых, порядковый тип данных (оценочные суждения типа «слабый», «средний», «сильный»). На множестве значений этого типа данных определено отношение «больше — меньше», но не определено расстояние между соседними положениями. Поэтому, хотя таким суждениям и могут быть поставлены в соответствие числа, например «слабый» = 1, «средний» = 2, «сильный» = 3, однако для этих чисел будет оправдано использование только операции сравнения. Другие математические операции над ними могут привести к плохо интерпретируемым результатам.

Количественные данные — это различные количественные характеристики объектов. Например, для социологических исследований это могут быть: *возраст*, *стаж*, *уровень дохода* и т. д. Как показано в работе [3], качественные методы применимы и к количественным данным, когда множество числовых значений разбивается на классы. В социологии такие разбиения общеприняты. Например, по уровню дохода граждан можно разбить на категории: *бедные*, *со средним достатком*, *состоятельные* и *богатые*. Таким образом, социология является хорошим объектом для применения качественных методов анализа.

Для применения анализа как количественного, так и качественного, исходные данные должны быть представлены в формализованном виде, например, в виде реляционной базы данных. Реляционную таблицу можно интерпретировать как многозначный контекст вида «объект — свойство», в котором каждому объекту соответствует набор свойств, заданных определенными значениями [1, 4–5]. В социологических исследованиях, в частности при опросах общественного мнения, объектами выступают респонденты, а в качестве свойств — их ответы на поставленные в ходе исследования вопросы. Вопросы можно разбить на три группы.

Первая — специальные вопросы, соответствующие целям исследования.

Вторая — личностные вопросы, которые касаются самого респондента (*пол*, *возраст*, *образование*, *национальность*, *профессия*, *должность*, *семейное положение*, *уровень дохода* и т. д.).

Третья — проверочные вопросы. Они касаются психологических и социальных характеристик респондента (*уровень компетентности*, *заинтересованность*, *мотивация*, *особенности характера* и т. д.).

Качественный анализ этих вопросов помогает установить отношение граждан к тем или иным социальным явлениям, выделить похожие группы респондентов и установить глубинные причины, которые оказывают влияние на их мнения.

ОБНАРУЖЕНИЕ ЛОГИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ В СОЦИОЛОГИЧЕСКИХ ДАННЫХ

Основой качественного анализа является индуктивный поиск логических закономерностей, которые выражены в форме логических правил (импликаций) [6].

Импликацией $A \to B$ называется условное суждение вида: «если A, то B», где A — посылка импликации, B — заключение. В контексте «объект — свойство» импликация $A \to B$ имеет смысл: набор свойств A влечет набор свойств B. Импликации называются строгими, если они выполняются для всех объектов рассматриваемого контекста, в противном случае — нестрогими.

На практике строгое выполнение импликации на всех объектах рассматриваемого контекста еще не означает, что оно будет таким же строгим для объектов всей генеральной совокупности (в нашем случае для всех членов исследуемого социума). С другой стороны, социальные закономерности не являются такими же жесткими, как физические законы, а имеют характер общих тенденций, на фоне которых бывают и исключения, поэтому такие закономерности адекватно описываются именно нестрогими импликациями.

Для характеристики нестрогих импликаций вводятся следующие оценки:

- точность,
- надежность,
- прирост точности,
- уровень поддержки правила,
- полнота.

Эмпирической оценкой строгости импликаций является точность. Пусть G_x — множество объектов, обладающих набором свойств X, тогда точность импликации

$$P(A \to B) = |G_{A \cup B}| / |G_A|,$$

где $|G_{A \cup B}|$ — количество объектов, на которых выполняется импликация (объем импликации); $|G_A|$ — количество объектов, приходящееся на посылку импликации.

Точность импликации находится в интервале [0, 1], причем $P(A \to B) = 1$, если $G_A \subseteq G_B$. В остальных случаях $P(A \to B) < 1$ и $P(A \to B) = 0$, если $G_A \cap G_B = \varnothing$. Таким образом, чем больше объектов со свойствами A обладают также свойствами B, тем выше точность импликации.

Надежность импликации оценивается с помощью показателя $R(A \to B)$, который отражает нашу уверенность в том, что найденное правило неслучайно. Если α — вероятность того, что импликация $A \to B$ могла появиться случайно, т.е она ошибочна, то ее надежность $R = 1 - \alpha$.

Пусть n — количество всех объектов в контексте; $k = |G_A|$ — количество объектов в посылке импликации; $l = |G_B|$ — количество объектов в заключении импликации; $m = |G_{A \cup B}|$ — количество объектов, общих и для посылки, и для заключения. Тогда надежность можно оценить по формуле

$$R = \sum_{i=m}^{k} C_l^i C_{n-l}^{k-i} \Big/ \sum_{i=0}^{k} C_l^i C_{n-l}^{k-i} .$$

Величина R также находится в интервале [0, 1]. $R = 0,5$ означает полную неопределенность. Значения R, близкие к 1, говорят о высокой надежности правила, а близкие к 0 — о надежности противоположного правила: $A \to \neg B$, где $\neg$ — символ отрицания.

Если $A = A_1 \cup A_2$, то имеет смысл рассматривать надежность правила $A \to B$ относительно правил $A_1 \to B$ и $A_2 \to B$. Например, для правила $A_1 \to B$ рассматривается подконтекст, который состоит из объектов G_{A_1} и вычисляется надежность правила $A \to B$ относительно этого контекста. Относительная надежность связана с приростом точности [1].

$$\Delta P_{A_1} = P(A \to B) - P(A_1 \to B).$$

При добавлении свойств в посылку импликации точность может увеличиться, уменьшиться или остаться на том же уровне. Если точность не меняется, то относительная надежность (для объектов G_{A_1}) $R_{A_1} = 0,5$. Если точность увеличилась, то надежность $R_{A_1} > 0,5$. Если же точность уменьшилась, то надежность $R_{A_1} < 0,5$. То, насколько существенно надежность будет отклоняться от положения неопределенности, зависит от количества объектов, на которых выполняется импликация. Для обеспечения требуемого уровня надежности (например, $0,95$) оно должно быть не слишком мало ($m > 10$). Поэтому на практике вместо показателя надежности используется анализ прироста точности, при этом задается ограничение на минимальное количество объектов m, для которых выполняется указанное правило. Для больших контекстов в зарубежных исследованиях используется относительный показатель $supp = m / n$, который называется уровнем поддержки правила [7].

Для анализа логических правил используется также характеристика, называемая полнотой [1]. Полнота импликации

$$C(A \to B) = |G_{A \cup B}| \,/\, |G_B|,$$

где $|G_{A \cup B}|$ — количество объектов, на которых выполняется импликация; $|G_B|$ — количество объектов, приходящееся на заключение импликации. Очевидно, что $C(A \to B) = P(B \to A)$ и $0 \le C(A \to B) \le 1$. Эта характеристи-

ка говорит о том, какую часть всех случаев, при которых имеет место заключение импликации, «объясняет» данная импликация. Чем меньше значение полноты, тем большая часть объектов заключения импликации остается «необъясненной». С другой стороны, если значение полноты будет большое и сопоставимое со значением точности импликации, то точность импликации $A \to B$ будет примерно равна точности импликации $B \to A$, и в таком случае связь между свойствами A и B можно интерпретировать как ассоциацию.

СТРАТЕГИИ ПОИСКА ЗАВИСИМОСТЕЙ

Анализ логических импликаций строится по принципу, предложенному М. Бонгардом в алгоритме КОРА [8]. Согласно этому принципу, рассматриваются все импликации, удовлетворяющие заданным критериям (в нашем случае — объем, точность, полнота), что позволяет выявить все значимые импликации. Количество выявленных импликаций существенно зависит от уровня задаваемых критериев. С ослаблением требований к этим уровням число логических правил стремительно увеличивается, и появляется необходимость в разработке методики индуктивного поиска зависимостей, учитывающих информативность правил и позволяющих отбирать наиболее информативные. Для этого предлагается следующая стратегия (более подробно см. в работе [9]).

Для строгих логических правил (точность которых равна 1) правило считается неинформативным, если оно может быть получено из других правил с помощью дедуктивного вывода. Для нестрогих правил информативность зависит от расхождения между реальной точностью правила и ожидаемой оценкой точности $P_{\text{ож}}$. Реальную точность правила $P(A \to B)$ можно понимать как условную вероятность $P(B/A)$. Расчет ожидаемой точности строится на гипотезе о статистической независимости свойств контекста. Как известно, если свойства A и B независимы, то условная вероятность $P(B/A)$ равна безусловной вероятности $P(B)$. Поэтому ожидаемая точность правила $A \to B$ равна $P_{\text{ож}} = P(B) = P(\varnothing \to B)$. Следовательно, информативность зависит от прироста точности $\Delta P = P(A \to B) - P(\varnothing \to B)$. Для сложной импликации $A_1 \cup \ldots \cup A_s \to B$ ожидаемая точность вычисляется на основе импликаций $A_1 \to B, \ldots, A_s \to B$ по методу Байеса [10, 11].

В данной работе построен алгоритм нахождения наиболее значимых импликаций с учетом приведенных выше оценок. Суть алгоритма состоит в том, что вначале в качестве потенциальных посылок импликаций исследуются все наборы свойств длины 1, затем длины 2 и т. д. до достижения определенной длины l_o. Из всех наборов свойств генерируются импликации, посылками которых являются эти наборы свойств, а заключениями — свойства из некоторого заданного набора (возможно, все свойства контекста). Из этих импликаций отбираются значимые, т. е. удовлетворяющие заданным критериям объема, точности и полноты. Поскольку генерация импликаций

происходит от более общих к менее общим, то это позволяет также вычислять информативность правил и отбирать из них наиболее информативные. При добавлении новых свойств к их некоторому набору количество объектов, имеющих все эти свойства, может только уменьшаться, так как $\forall A_1, A_2 \;\; G_{A_1 \cup A_2} \subseteq G_{A_1}$. Отсюда следует, что с ростом количества свойств в посылке правила его объем и полнота уменьшаются, в то время как точность может либо расти, либо уменьшаться, либо оставаться неизменной. Поэтому, если на очередном этапе некоторые наборы свойств дают импликации с недопустимыми значениями объема и полноты, то они исключаются из дальнейшей генерации импликаций.

Можно выделить два варианта логического анализа: разведывательный и подтверждающий. Первый вариант обычно используется, когда структура данных неясна и с помощью логического анализа исследователь хочет ее прояснить. При этом посылки и заключения импликаций формируются на одном и том же наборе свойств, в результате чего могут получаться самые разнообразные импликации. Второй вариант используется, когда имеется модель предметной области, дающая представление о том, какие свойства могут быть причинами, а какие следствиями. В этом случае объясняющие и объясняемые свойства разделяют на два непересекающихся множества, из которых формируются соответственно посылки и заключения импликаций.

Предложенная методика и алгоритмы поиска импликаций реализованы в программной системе качественного анализа «Сизид» [12]. Система предназначена для поиска значимых импликаций в больших массивах данных. Она обеспечивает весь цикл анализа многозначных и однозначных контекстов, в том числе автоматическую сегментацию числовых признаков, и сочетает эффективность работы с удобным пользовательским интерфейсом.

РЕЗУЛЬТАТЫ АНАЛИЗА СОЦИОЛОГИЧЕСКИХ ДАННЫХ

Методика поиска наиболее значимых импликаций с помощью системы «Сизид» была применена для анализа данных социологического исследования «Отношение граждан Украины к реформам», которое проводилось в Украинской академии общественного управления. В проводимом исследовании опрошено 154 респондента, которым задавались вопросы о ситуации в стране в целом, об их работе и личных характеристиках. Ответы на вопросы представляли собой оценки от 1 до 7, где 1 соответствовала отрицательному полюсу шкалы, 7 — положительному. Вопросы были разбиты на блоки, в каждом требовалось оценить несколько параметров по одному и тому же критерию.

Для исследования данных проводился двухуровневый логический анализ. На первом уровне анализировались отдельно блоки вопросов, чтобы выявить типовые взгляды, представления и зависимости. На втором — устанавливался социальный портрет тех респондентов, которые представляли различные по своим взглядам группы. В табл. 1 приведен один из вопросов, в табл. 2 — варианты ответов по каждому пункту вопроса.

Таблица 1. Вопрос Q1: состояние сфер национальных интересов

Номер	Оцениваемый параметр
Q1_1	Формирование гражданского общества
Q1_2	Внутриполитическая стабильность
Q1_3	Социальная защита населения
Q1_4	Структурная перестройка экономики
Q1_5	Ход приватизации
Q1_6	Обеспеченность инвестициями
Q1_7	Реформа налогообложения
Q1_8	Обеспечение национальных интересов в сфере международных отношений
Q1_9	Состояние науки
Q1_10	Состояние культуры
Q1_11	Состояние национального самосознания
Q1_12	Ответственность должностных лиц

Таблица 2. Варианты ответов на вопрос Q1

Оценка	Значение оценки
1	Крах
2	Кризис
3	Некоторое ухудшение
4	Отсутствие изменений
5	Некоторое улучшение
6	Улучшение
7	Устойчивое хорошее состояние

В табл. 3 приведены наиболее интересные значимые импликации (с объемом $m \geq 10$ и точностью P≥75%). Посылка и заключение импликации записываются в таблице отдельно, для каждой импликации приводится ее объем, точность P и полнота C. Импликации, в которых участвует вариант ответа «отсутствие изменений», не приведены, так как обычно средние значения шкалы выбираются, когда человек не уверен в своей оценке.

Таблица 3. Список значимых импликаций вопроса Q1

Объем	*P*	*C*	Посылка	Заключение
13	77	29	Формирование гражданского общества — 2 Состояние национального самосознания — 1	Социальная защита населения — 1
11	91	29	Структурная перестройка экономики — 1	Социальная защита населения — 1
10	90	26	Формирование гражданского общества — 2 Состояние науки — 1	Социальная защита населения — 1
10	80	31	Формирование гражданского общества — 2 Состояние науки — 1	Состояние национального самосознания — 1

Первая импликация читается так: «ЕСЛИ *Формирование гражданского общества = Кризис* и *Состояние национального самосознания = Крах*, ТО *Социальная защита населения = Крах*». Таким образом, значительная груп-

па респондентов (13 из 154) связывает *крах* в сфере *социальной защиты населения* с плохим состоянием *национального самосознания* и с *кризисом* в *формировании гражданского общества*. Следующая импликация говорит о том, что *кризис* в области *структурной перестройки экономики* влечет за собой также *кризис* в сфере *социальной защиты населения*. Следующие две импликации показывают, что *кризис* в *формировании гражданского общества* и *крах* в *науке* означают *крах*, как в сфере *национального самосознания*, так и в сфере *социальной защиты населения*.

Менее значимые импликации также представляют интерес для более глубокого анализа, так как показывают совместное появление признаков, что говорит об их связи. Например, в ответах на данный вопрос было выявлено совместное появление признаков «*Внутриполитическая стабильность*» и «*Обеспеченность инвестициями*», что говорит о наличии ассоциации между ними.

Обращает на себя внимание то, что чаще других в посылках импликаций фигурируют такие атрибуты, как «*Структурная перестройка экономики*», «*Состояние культуры*», «*Формирование гражданского общества*», «*Состояние национального самосознания*» и «*Социальная защита населения*». Частота появления тех или иных атрибутов зависит от того, насколько они значимы для респондентов и насколько ясное и стереотипное мнение имеется у респондентов относительно этих атрибутов. Стереотипность оценок приводит к тому, что множество людей отвечают одинаково на одни и те же вопросы, что повышает значимость импликаций.

Другой вопрос анкеты был составлен для исследования результативности административной реформы, которая также оценивалась по шкале от 1 до 7. Список наиболее значимых импликаций, выявленных в ответах на данный вопрос, приведен в табл. 4.

Таблица 4. Список значимых импликаций вопроса Q6

Объем	P	C	Посылка	Заключение
26	88	40	6.11 Децентрализация властных полномочий — 1	6.10 Коррумпированность управления — 1
25	96	41	6.9 Использование передового опыта в государственном управлении — 1	6.10 Коррумпированность управления — 1
24	75	31	6.6 Формы и методы управления в органах государственной власти — 1	6.10 Коррумпированность управления — 1
22	82	31	6.5 Кадровое обеспечение государственного управления — 1	6.10 Коррумпированность управления — 1
22	77	29	6.13 Профессионализм управленческой элиты — 1	6.10 Коррумпированность управления — 1
20	75	26	6.7 Дееспособность исполнительной власти — 1	6.10 Коррумпированность управления — 1
16	75	50	6.8 Устранение функционального и организационного дублирования в управлении — 1	6.6 Формы и методы управления в органах государственной власти — 1

Больше всего импликаций приходится на объяснение *коррумпированности управления*. Если есть проблемы хотя бы в одном из направлений административной реформы, находящихся в посылках импликаций 1 — 6, то это негативно влияет на решение проблемы *коррумпированности управления*. Таким образом, решение проблемы *коррупции* следует начинать с реформирования в области *децентрализации власти*, при этом использовать *передовой опыт государственного управления*, совершенствовать *формы* и *методы управления* в *органах исполнительной власти*, решать проблему *кадрового обеспечения государственного управления* и заботиться об уровне *профессионализма управленческой элиты*. Из полученных импликаций следует, что положительное влияние на преодоление *коррупции* может оказать повышение *дееспособности исполнительной власти*.

Как видим, анализ логических импликаций позволяет получить простые и конкретные рекомендации и, что наиболее важно, устанавливает направление связей от причины к следствию.

Предлагаемая методика позволяет также составить портреты групп респондентов, которые придерживаются различных мнений относительно ситуации в стране и необходимости ее реформирования. Для этого анализируются такие черты респондентов, как *пол*, *возраст*, *уровень жизни* (шкала: *ухудшился*, *без изменений*, *улучшился*), *социальная принадлежность* (*государственная служба* или *частный бизнес*), *положение* (*руководитель* или *простой служащий*).

Например, всех респондентов можно разделить на две группы по их ответам на вопрос об уровне социальной защиты населения. В одну группу попадают респонденты, думающие, что ситуация с социальной защитой критическая, в другую — те, кто считает, что в этом отношении все обстоит нормально. Путем анализа импликаций была выявлена зависимость отрицательных ответов на этот вопрос от следующих характеристик:

1) молодые люди в возрасте 20 — 30 лет, находящиеся на государственной службе;

2) люди, которые не находятся на государственной службе и уровень жизни которых ухудшился.

Положительные ответы на вопрос о социальной защите населения давали в основном мужчины, уровень жизни которых улучшился за последнее время.

Таким образом, найдена вполне понятная зависимость положительных и отрицательных ответов респондентов на вопрос о социальной защите от их собственного материального положения. Те, у кого оно ухудшилось, склонны считать, что ситуация в этом отношении критическая, а те, кто улучшил свое материальное положение, считают, что в сфере социальной защиты населения происходят изменения к лучшему. Интересен также тот факт, что достаточно критично мыслят и проявляют наибольшую активность в реформах молодые люди (до 30 лет), находящиеся на государственной службе, — видимо, здесь сказывается присущее молодому поколению отсутствие консерватизма и желание изменить ситуацию.

ЗАКЛЮЧЕНИЕ

Качественный анализ позволил установить некоторые интересные логические зависимости по данным проведенного социологического исследования. С его помощью проанализировано текущее положение дел в социальной сфере и в сфере административного реформирования. Кроме того, найдены некоторые причинно-следственные связи между социальным статусом респондентов и их взглядами на ход реформ. Таким образом были выявлены мнения различных социальных групп. Учет этих мнений — залог построения гармоничного правового государства.

ЛИТЕРАТУРА

1. *Чесноков С. В.* Детерминационный анализ социально-экономических данных. — М.: Наука, 1982. — 168 с.
2. *Ядов В. А.* Стратегия социологического исследования. Описание, объяснение, понимание социальной реальности. — М.: Добросвет, 1999. — 596 с.
3. *Лбов Г. С.* Методы обработки разнотипных экспериментальных данных. — Новосибирск: Наука, 1981. — 157 с.
4. *Ganter B., Wille R.* Formal Concept Analysis. Mathematical Foundations. — Springer, 1999. — 284 p.
5. *Кузнецов С.О.* Автоматическое обучение на основе анализа формальных понятий // Автоматика и телемеханика. — 2001. — №10. — С. 3–27.
6. *Финн В. К.* Правдоподобные рассуждения в интеллектуальных системах типа ДСМ // Итоги науки и техники. Сер. Информатика. — 1991. — Т. 15. — С. 54–101.
7. *Agrawal R., Srikant R.* Fast algorithms for mining association rules in large databases // Proceedings of the 20-th VLDB Conference, Santiago, Chile. — 1994. — P. 487–499.
8. *Бонгард М. М.* Проблема узнавания. — М.: Наука, 1967. — 320 с.
9. *Ткачев А. Е.* Логический анализ разнородной информации // Искусственный интеллект. — 2002. — №2. — С. 268–276.
10. *Heckerman D.* Bayesian Networks for Data Mining // Data Mining and Knowledge Discovery. — 1997. — № 1. — P. 79–119.
11. *Дюк В., Самойленко А.* Data mining: учебный курс. — СПб: Питер, 2001. — 368 с.
12. *Таран Т.А., Евтушенко С.А., Ткачев А.Е.* Приобретение знаний методом концептуального анализа данных: алгоритмы и программные реализации // Искусственный интеллект. — 2000. — № 2. — С. 200–206.

Поступила 19.11.2002