УДК 004.62

МЕТОДИКА ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА СЛАБОСТРУКТУРИРОВАННЫХ МНОГОМЕРНЫХ ДАННЫХ СОЦИОЛОГИЧЕСКИХ ОПРОСОВ

Руденко А. И.
д-р техн. наук, проф. Арсирий Е. А.
Одесский Национальный Политехнический Университет, УКРАИНА

АННОТАЦИЯ. Рассмотрены существующие проблемы сбора и обработки данных социологических исследований.Показано что данные соцопросов отображаются в многомерное признаковое пространство, являются слабоструктурированными и нуждаются в дополнительной кодировке, кроме того содержат пропуски, противоречия и не всегда являются статистически достаточными, поэтому применение статистических методов для их обработки и анализа не является эффективным. Предложено использовать методы интеллектуального анализа.

Введение. Проведение социологических опросов (исследований)является одним из лучших, а иногда и единственно возможнымспособом,получить ценные знания о целевой (опрашиваемой) аудиториив виде количественных метрик или качественных оценокповеденческих, социально-демографических, географических, психофизических, медицинских или любых других характеристик. Такие знания позволяют организациям принимать обоснованные решения при планировании стратегии своего развития илиулучшения рабочих процессов.Современный уровень развития информационных технологий для извлечения знаний о целевой аудитории из эмпирических, количественных и качественных данных социологических исследований позволяет успешно использовать механизмы интеллектуального анализа данных (ИАД) в виде моделирования в том числе нейросетевого и методов визуализации данных.

Цель работы: разработка методики интеллектуального анализа слабоструктурированных многомерных данных социологических опросов, которая позволит снизить трудоемкость их обработки и анализа,автоматизировав процесс извлечения знаний о целевой (опрашиваемой) аудитории.

Основная часть работы. Как правило, для проведения социологического исследования разрабатывается анкета или опросный лист, состоящий из ряда вопросов и пространства для ответов. Они могут сочетать в себе вопросы с "вариантным" выбором ответа, вопросы с численным или балльным ответом и вопросы, предполагающие текстуальный ответ[1]. Если первые два типа вопросов направлены на получение метрик, то текстуальные ответы предоставляют дополнительную информацию для истолкования получаемых численных статистик.Опросы могут проводиться во многих режимах, в том числе онлайн, а анонимность соцопросов позволяет респондентам отвечать более откровенно и правдиво.При этом отборреспондентов в целевую аудиторию осуществляется по уровню дохода возрасту, полу, образованию, может учитываться стиль жизни или другие особенности. Полученные таким образом социологические эмпирические данные, т.е. данные, характеризующие конкретные социальные факты, могут представать перед исследователем в виде: чисел, характеризующих те или иные объекты; индикаторов определенных отношений между рассматриваемыми объектами; совокупности определенных высказываний; текстов документов; каким-либо способом зафиксированных результатов наблюдения за невербальным поведением каких-либо людей и т.п.

Для получения пригодного для проведения дальнейшего компьютерного анализа датасета полученные эмпирические данные подвергают предварительной обработке, которая, как правило, включает следующие процедуры [2]: удаление не валидных записей; присвоение записям уникальных идентификаторов; приведение номинальных полей к категориальным; приведение категориальных полей к ординальным; введение данных в базу данных.

При проведении предварительной обработки перед исследователем,как правило, возникают следующие трудности [2]: при проведениипредобработки вручную, затраты по времени будут довольно существенными, кроме того повышается вероятность допустить

ошибки и опечатки; при приведении категориальных полей к ординальным, мы следуем определённому алгоритму, который мы многократно повторяем для каждого нового поля, и данную операцию необходимо автоматизировать, инструменты такие как GNU PSPP позволяют несколько ускорить данный процесс, однако всё еще остаётся большая доля ручного труда; при выполнении процедуры введения данных в базу данных, необходимо выбрать способ их хранения и соответствующую СУБД, что сложно сделать,не имея достаточных знаний в области организации баз данных, а также импорта/экспорта данных из них.

Далее необходимо заметить, что для получения глубинных знаний из предварительно обработанных данных соцопросов методы статистического анализа оказываются малопригодными по следующим причинам:

– если исследователи собирают данные с использованием ошибочных или предвзятых процедур, полученныерезультат статистического анализа будет вводить в заблуждение;

– исследователи часто находят доказательства того, что две переменные сильно коррелированы, но это не доказывает, что одна переменная вызывает другую;

– статистический анализ является средством использования агрегированных измерений для составления выводов, но если исследователи не измеряют правильную вещь, анализ потерпит неудачу;

– последней проблемой статистического анализа является его склонность давать чрезмерно простые ответы на сложные вопросы.

Статистический анализ хорош тогда, когда необходимо получить некоторые поверхностные знания,например, частотные или корреляционные показатели, либо для обработки результатов более сложных методов/алгоритмов, так называемых метаданных [2].

В данной работе предлагается применить методы ИАД после предварительной очистки и обработки [3]. Например, для уменьшения признакового пространства провести кластерный анализ данных по выделенным подмножествам полей датасета, присвоив каждой записи определённые метки (*male*, *sportsman*, *does_not_drink_alcohol*, и т.д.)Процедуру кластеризации можно проводить не однократно, всё более сужая признаковое пространство и образуя более высокоуровневые информационные объекты (метаданные). Далее к полученным метаданным представляется возможным применить дискриминантный анализ с целью обучения классификатора относить метаданные к определенному классу метаданных. Таким образом, можно построить методику отнесения любого респондента на основе полученных его метаданных к определенному классу. Как показывают исследования, применение разработанной методики позволяет автоматизировать процесс анализа данных социологических исследований, сократив долю ручного труда в 14 раз.

Выводы. Показано, что данные соцопросов отображаются в многомерное признаковое пространство, являются слабоструктурированными и нуждаются в дополнительной кодировке, кроме того содержат пропуски, противоречия и не всегда являются статистически достаточными для проведения дальнейшего анализа с целью извлечения глубинных знаний и скрытых закономерностей. Предварительная обработка таких данных может занимать у исследователя достаточно много времени. Статистические методы малопригодны для анализа эмпирических данных соцопросов [2]. Поэтому предложено разработать методику интеллектуального анализа слабоструктурированных многомерных данных социологических опросов, которая позволит снизить трудоемкость их обработки и анализа, автоматизировав процесс извлечения знаний о целевой (опрашиваемой) аудитории [3].

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Good practice in the conduct and reporting of survey research [Електронний ресурс]. – Режим доступу: URL: https://academic.oup.com/intqhc/article/15/3/261/1856193

2. Methods for Analysis of Complex Surveys [Електронний ресурс]. – Режим доступу: URL: https://www.nap.edu/read/11990/chapter/9

3. An Overview of Data Mining Techniques [Електронний ресурс]. – Режим доступу: URL: http://www.thearling.com/text/dmtechniques/dmtechniques.htm