Контент-анализ больших качественных данных

А.Н. Олейник

***Аннотация*— Потребность в извлечении, классификации и сохранении информации при работе с большими данными возникает как в науке (банки исследовательских данных, обзоры литературы), так и в повседневной жизни (осмысление и обсуждение новостей). Методология контент анализа в его различных формах, качественной (ручное кодировние), количественной (автоматический подсчет частотности и совместной встречаемости слов и словосочетаний) и комбинированной (использование специально созданных словарей), способна стать существенным подспорьем. Контент анализ позволяет преобразовать качественные данные (текст, изображение) в цифровой формат и впоследствии манипулировать уже с информацией в матричной форме. Преобразование качественных данных в цифровую информацию создает предпосылки и для машинного обучения, как с «учителем», так и без него. Существующие компьютерные программы для контент-анализа (QDA Miner, Atlas TI, NVivo и другие) имеют ряд ограничений, препятствующих их использованию для работы с большими данными. Так, число пользователей (кодировщиков), работающих с конкретными документами, заведомо ограничено. Создание он-лайн платформ для контент-анализа позволяет снять это и ряд других ограничений. В статье обсуждаются возможные параметры он-лайн платформы для контент-анализа.**

***Ключевые слова*— качественные данные, разработка данных, контент-анализ.**

I. Введение: от данных к информации

Термин «большие данные» вошел в широкий обиход в 2008 г., с выходом специального номера журнала Nature, посвященного вопросам увеличения объемов доступных данных [21; 22]. Данное словосочетание использовалось и раньше, в том числе и в русскоязычных публикациях. Например, в 2006 г. было проанализировано использование слов «война» и «агрессия» в русском языке на массиве всех публикаций в сми, включенных в базу данных «Интегрум» 14]. Однако частотность упоминания больших данных и в научном дискурсе (Рис. 1), и в повседневной жизни (Рис. 2) стала резко возрастать именно после дискуссии в Nature.

Большие данные имеют четыре отличительные черты, которые, собственно, и легли в основу их выделения в особый концепт. К этим чертам относится большой объем, значительное разнообразие, высокая скорость изменений и особая ценность [10, С. 144; 25, P. xx; 13; 22, С. 7]. Разнообразие больших данных можно проиллюстривать тем, что они принимают и цифровую форму (например, данные о геолокации пользователя сотовым телефоном и номера входящих/исходящих звонков), и визуальную (медиафайлы, которыми обменивается пользователь), и текстовую (смс сообщения, которые получает и отсылает пользователь). Иными словами, большие данные включают в себя как количественные (цифры), так и качественные (текст, изображение) данные.

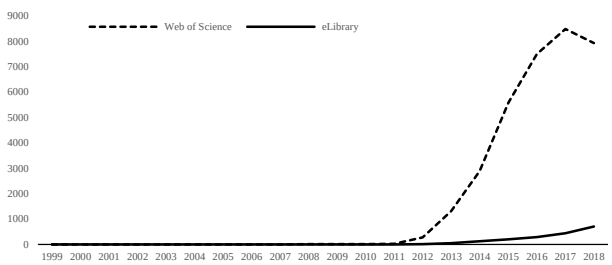


Рис. 1 «Частотность упоминания больших данных с научных публикациях, 1999-2018» Источник: Web of Science, eLibrary по состоянию на 11.05.2019

Статья получена 14 мая 2019.

Олейник А.Н., д.э.н., PhD, профессор Университета «Мемориал» (Канада) и в.н.с. ЦЭМИ РАН (Москва) (email: aoleynik@mun.ca)

Разнородный характер больших данных делает малопригодными многие статистические методы – в той мере, в которой их применение требует, например, нормального распределения данных. Взять данные в текстовой форме. В них наблюдается распределение Ципфа (Zipf): частотность встречаемости слова в тексте обратно пропорциональна его месту в проранжированном в порядке убывания по частоте списке всех слов. Наиболее часто употребляемые в тексте слова и составляют основную часть его объема [25, P. 115-116; 19, С. 75]. Качественными данными по этой причине труднее манипулировать, чем количественными, даже в небольших объемах с помощью стандартных в естественных науках методов. Если диагноз о неготовности обществоведов к работе с большими данными верен [5], то он еще в большей степени применим в отношении разработки больших качественных данных. Дальнейшее обсуждение будет посвещено именно работе с качественными данными.

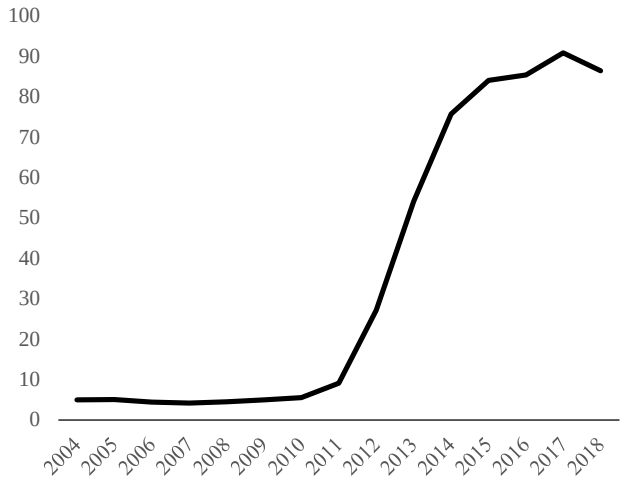


Рис. 2 «Относительная популярность термина big data в поисковых запросах Google, 2004-2018»
Источник: Google Trends (https://trends.google.com/trends/explore?date=all&q=big%20data, по состоянию на 12.5.2019). 100 соответствует пику популярности запросов

Манипулирование данными требуется ввиду их непригодности для использования в сырой, необработанной форме. Ценность представляют не столько сами данные, сколько информация, которую из них можно извлечь. Например, после разработки (mining) данных о перемещениях пользователей сотовых телефонов полученная информация может быть задействована для оптимизации управления дорожными и транспортными потоками. Разработка данных позволяет их упорядочить и структурировать. Использование навеянных горным делом терминов не случайно: данные можно сравнить с рудой, которую прежде чем использовать требуется добыть, обогатить и переработать. «Информация появляется в результате анализа обработанных данных человеком, этот анализ придает данным смысл и обеспечивает им потребительские качеств» [21; см. так же 25, P. 2; 11, С. 58]. Именно преобразование больших данных в информацию представляется наиболее проблематичным из всех возможных операций с ними – хранение, передача и так далее [29, P. 18].

В данной статье обсуждаются перспективы использования контент-анализа для преобразования больших качественных данных в информацию. Использование контент-анализа не требует нормально распределенных данных. Контент-анализ достаточно широко используется в общественных науках, прежде всего в социологии, политологии, лингвистике и коммуникативистике (communication studies), однако малознаком в науке о данных. Именно науке о данных приходится решать связанные с большими данными проблемы. В этом смысле статья имеет две целевые аудитории: обществоведы и специалисты по работе с данными. Первым будет небезынтересно узнать, что нужно предпринять для превращения контент-анализа в методологию работы с большими данными. А вторые, возможно, найдут некоторые подсказки для решения стоящих перед ними задач в области разработки больших данных.

II. Большие качественные данные в науке и повседневной жизни

С большими качественными данными сталкиваются как в науке, так и в повседневной жизни. Примером больших качественных данных в контексте исследовательской деятельности будут

этнографические материалы и транскрипты углубленных интервью, а так же корпус научных публикаций. В качестве примера больших качественных данных в повседневной деятельности будут рассмотрены новостные публикации.

A. Исследовательские данные

С большими количественными данными ученые работают как в естественных, так и в общественных науках. Мировой Банк создал и поддерживает банк количественных данных, широко используемый макро-экономистами [42]. Аналогичные банки экономических и социологических количественных данных, собранных на национальном уровне, существуют во многих странах, в том числе в России [8] и Украине [9]. Продолжающееся накопление количественных данных делает возможным мета-анализ – обзор, сопоставление и обработку полученных из различных источников данных [38, P. 306; 29, 2015, P. 25]. Мета-анализ является одним из способов превращения больших данных в пригодную для исследований информацию.

С качественными данными ситуация иная. Банки качественных данных можно пересчитать по пальцам. Один из наиболее известных – собрание этнографических материалов, в том числе результатов полевых исследований и подготовленных на их основе публикаций, созданное на базе Йельского Университета [27]. Собирание качественных исследовательских данных из разных источников и о разных культурах было начато еще в 1930-1940-ые годы. Для разработки больших качественных данных применялся целый ряд элементов контент-анализа, таких как параллельное кодирование как минимум двумя исследователями, о чем ниже [26]. Вторичная разработка качественных данных возможна и перспективна с точки зрения решения более широкого круга задач, чем те, что решались при первичном сборе данных [15]. Однако накопленный значительный объем качественных данных по-прежнему хранится, в основном, в разрозненном виде и за редким исключением недоступен для вторичной разработки. Как представляется, вторичная разработка тормозится и недостаточным развитием соответствующих методов, в частности, слабой приспособленностью контент-анализа для работы в большими качественными данными.

B. Обзоры литературы

Увеличение числа научных периодических изданий и, соответственно, публикаций в них позволяет охарактеризовать термином большие данные и обзоры литературы, требуемые для подготовки исследовательских работ. Включающий несколько десятков названий список цитируемых источников стал сегодня нормой. К примеру, в опубликованных в находящемся в средней части списка экономических периодических изданий, проранжированных по их влиятельности,[1] Journal of Economic Issues статьях цитируется в среднем 37.5 источников (стандартное отклонение – 28). Причем наблюдается тенденция к неуклонному росту числа процитированных источников. Если в конце 1960-ых годов (журнал издается с 1967 года) их среднее число составляло 18.8, то в 2016-2018 гг. – уже 45, что соответствует росту в 2.4 раза (Рис. 3). Статьи с более чем сотней цитируемых источников встречаются все чаще. Рекордсменом для данного журнала стала статья с 186 процитированными источниками.

Рис. 3 «Среднее количество цитируемых в опубликованных в Journal of Economic Issues статьях источников, 1967-2019»
Источник: Web of Science и расчеты автора

[1] Импакт-фактор данного издания в 2017 г. составил 0.58 по данным Journal Citation Reports (журнал включен в Web of Science), а пятилетний импакт-фактор – 0.768.

Возникновение в конце 1990-ых годов баз научных публикаций – Web of Science, Scopus, eLibrary и прочих – облегчило поиск в постоянно растущем объеме больших качественных данных по ключевым словам и аннотациям. Например, в Web of Science Core Collection по состоянию на май 2019 г. проиндексировано более 20 тысяч периодических изданий, 1.4 миллиарда статей [40]. Однако использование баз научных публикаций не решает вопрос с разработкой больших качественных данных и сохранением ее результатов. Цитирование источника не гарантирует собственно его прочтения. Цитаты могут использоваться для подтверждения компетентности, обозначения аффиляции в определенной сети, позиционирования по отношению к другим ученым и много другого, не предполагающего тесного знакомства с цитируемым источником [28]. Разработка источника (внимательное прочтение и конспектирование) требуется в случае развития идей и методов цитируемых авторов, а это весьма трудоемкий и связанный с существенными затратами времени процесс [37]. Сохранение информации из уже разработанных источников представляет собой отдельную задачу, равно как и обеспечение доступа к этой информации в будущем.

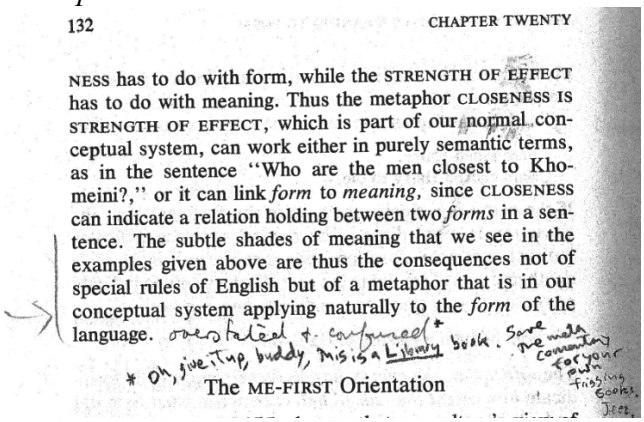
C. Новости

Чтение новостей, хотя и не требует конспектирования, тоже связано с существенными затратами времени. С распространением интернета количество источников новостей резко возросло. На смену практике просматривания утренней и/или вечерней газеты пришло «сканирование» в режиме реального времени множества сайтов. При чтении газеты читатель выделял наиболее значимые для себя элементы ее содержания [18, С. 61]. Просмотр новостей в интернете тоже предполагает выделение требующего внимания материала. Отличия заключаются в объеме доступных для разработки новостей, их разнообразии и обновлении в режиме реального времени.

Для облегчения разработки больших новостей появились новостные агрегаторы. Новостной агрегатор собирает информацию параллельно из нескольких источников. Новостные сообщения разбиваются на кластеры по критерию сходства между ними. Кластеры образуют новости, освещающие предположительно одно и то же событие. После построения кластера создается его название и краткая аннотация. В результате пользователю агрегатора достаточно ознакомится именно с названием и краткой аннотацией, вместо того, чтобы тратить время на просмотр всех исходных новостей [24, С. 36-38]. Наряду с новостными агрегаторами, ориентированными на массового потребителя – Google News, Яндекс.Новости, Rambler.Новости и др. – созданы и индивидуализированные, учитвающие особенности предпочтений конкретного пользователя. Для этого принимается во внимание история просмотров этим пользователем новостей из разных источников [33]. Учитывая тот факт, что социальные сети тоже превратились в источник новостей, появились новостные агрегаторы, извлекающие данные из социальных медиа [12; 34; 20]. Как и в случае с базами научных публикаций, вопросы глубины чтения больших новостей и сохранения результатов такого чтения остаются открытыми.

III. Контент анализ

A. Определение


132 CHAPTER TWENTY

NESS has to do with form, while the STRENGTH OF EFFECT has to do with meaning. Thus the metaphor CLOSENESS IS STRENGTH OF EFFECT, which is part of our normal conceptual system, can work either in purely semantic terms, as in the sentence "Who are the men closest to Khomeini?," or it can link *form* to *meaning*, since CLOSENESS can indicate a relation holding between two *forms* in a sentence. The subtle shades of meaning that we see in the examples given above are thus the consequences not of special rules of English but of a metaphor that is in our conceptual system applying naturally to the *form* of the language.

The ME-FIRST Orientation

Рис. 4 «Пример несистематическим образом проделанного контент-анализа фрагмента книги Жоржа Лакоффа и Марка Джонсона *Метафоры, которыми мы живем*»

Источник: экземпляр книги Lakoff G., Johnson M., Metaphors We Live By, принадлежащей библиотеке Университета «Мемориал» (Канада). Расшифовка заметок: «преувеличенно и неясно. – дружок, хватит писать заметки на библиотечной книге. Делай

это на своих собственных экземплярах. Посмейся»

Герой одной из пьес Мольера и не подозревал, что вот уже более сорока лет говорил прозой. Аналогичным образом, большинство исследователей и потребителей новостей используют в своих целях те или иные элементы контент-анализа, даже не подозревая об этом. В простейшем случае читатель как научной, так и иной литературы выделяет наиболее значимые моменты в прочитанном в зависимости от своих интересов и контекста прочтения в целом. Иногда это принимает форму подчеркивания соответствующих фрагментов и делания пометок на полях. Подчеркивание и заметки отражают результаты разработки качественных данных, сохраняемые надолго (по этой причине использование для «несистематического» контент-анализа библиотечных экземпляров подвержено критике и запретам; Рис. 4).

Даже эта простейшая иллюстрация позволяет выделить несколько принципиальных моментов. При контент-анализе особое внимание уделяется контексту разработки качественных данных. Согласно стандартному определению, «контент-анализ представляет собой технику исследования, позволяющую делать надежные и верные выводы из текстов (или любого другого значимого материала) относительно контекстов их использования» [31, P. 18]. В этом смысле контент-анализ создает предпосылки для восхождения от текста к внетекстовой реальности [2, С. 12; 3, С. 23]. Фрагменты одного и того же текста, скажем, уже упомянутой книги «Метафоры, которыми мы живем», отмеченные в качестве значимых читателем в Канаде и в России, могут отличаться.

Другой важный момент заключается в вовлеченности двух и более участников контент-анализа. Именно при этом условии возможна оценка надежности получаемых результатов, без чего данную исследовательскую технику было бы очень легко критиковать и отбросить как слишком субъективную и потому «ненаучную». Результаты участников контент-анализа сопоставляются с помощью коэффицентов согласия – Альфы Криппендорфа, Пи Скотта и других [36]. Только при достижении этими показателями критических величин можно говорить о надежности извлеченных из качественных данных информации. Шансы на достижение приемлимого уровня согласия выше, если контекст анализа четко задан, например, вопросом исследования.

B. Техника или методология?

В отличие от спонтанно применяемого контент-анализа, его использование в исследовательских целях предполагает систематичность и следование процедуре. Некоторые исследователи даже призывают видеть в контент-анализе не только технику, а своеобразную методологию [3, С. 20]. В отличие от инструмента исследования, исследовательский метод предполагает наличие своего рода философии и комплексного подхода. Каковы же эпистемологические основания контент-анализа?

Базовой операцией в контент-анализе можно считать кодирование, или присвоение кода (тэга) фрагменту текста или изображения. Таким фрагментом, как правило, выступает предложение или абзац текста или область изображения. Во время дискуссии о контент-анализе в начале 1970-ых годов был сформулирован тезис о сходстве контент-анализа с распознаванием образов в прикладной математике [17, С. 93]. В тексте как массиве качественных данных требуется распознать коды-образы, которые позволяют его описать более сжато в соответствии со стоящей перед исследователем задачей. Автор текста при написании текста закладывает образы. Однако при прочтении текста многими читателями распознанные ими образы не обязательно сводятся к тем, что были первоначально заложены автором [37].

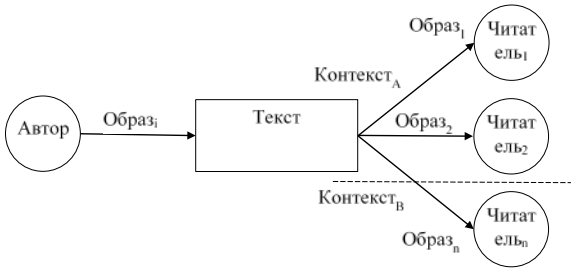


Рис. 4 «Контент-анализ как процесс распознавания образов»

Причем множество распознанных образов будет отличаться в зависимости от контекстов прочтения текста (Рис. 4). Так, Образ$_1$ и Образ$_2$ имеют больше шансов оказаться сходными, чем Образ$_1$ и Образ$_n$ так как Читатель$_1$ и Читатель$_2$ распознали их в одном и том же Контексте$_A$. Некоторые авторы даже утверждают что значимы только распознанные образы, а вот заложенные автором образы можно вывести за скобки контент-анализа [31, P. 22].

Предложенная схема подсказывает, что эпистемологические основания контент-анализа потенциально позволяют найти общий язык прикладным математикам и программистам, с одной сторонй, и социологам, с другой стороны. Сторонники понимающей социологии, берущей свое начало в работах Макса Вебера, видят свою основную задачу в интерпретации. «Отличительной чертой социологического подхода... является интерпретация действия в субъективных категориях» [41, P. 8]. Данный принцип применим и к тексту или изображению – в той мере, в которой они отражают социальное действие. Определение контент-анализа тогда можно уточнить. Контент-анализ нацелен на интерпретацию качественных данных в зависимости от специфического контекста, в котором эта интерпретация происходит [31, P. 24; 1, C. 21; 39, P. 3]. Результаты экспериментов, проведенных еще в начале 1970-ых годов, показали, что при расположении смысловых элементов текста в случайном порядке испытуемые, как правило, выстраивают из полученного таким образом новые по смыслу тексты [6, C. 92]. Иначе говоря, они находят новые интерпретации в зависимсти от особенностей ситуации, в которой оказались.

C. Типы контент-анализа

Присвоение кодов фрагментам текста делает возможным преобразование качественных данных в цифровой формат. Текст преобразуется в матрицу, строки которой соответствуют предложениям, а столбцы – кодам (образам) [25, P. 4; 39, P. 42]. Если в рассматриваемом предложении распознан тот или иной образ, то соответствующая клетка матрицы содержит «1». В противном случае эта клетка содержит «0». В результате контент-анализа происходит формализация – «совокупность познавательных операций, обеспечивающая отвлечение от значения понятий и смысла выражений [в тексте] с целью исследования [их] логических особенностей, дедуктивных и выразительных возможностей» [16, C. 40]. Последующие шаги в разработке данных, которые первоначально были качественными, принимают форму операций над векторами (строками и столбцами матриц), осуществлямых с использованием линейной алгебры, многомерного шкалирования и других широко используемых в естественных науках инструментов. Именно преобразование качественных данных в цифоровой формат и позволяет рассматривать контент-анализ в качестве перспективного метода для работы с большими данными.

Существует три основные разновидности контент-анализа: качественный («ручное» кодирование), количественный (подсчет частотности слов и выражений, а так же их совместной встречаемости в тексте) и со смешанными методами (mixed methods). Качественный контент-анализ невозможно масштабировать для работы с большими данными, так как он требует непосредстенного участия человека, выступающего в роли кодировщика. Количественный контент-анализ, с другой стороны, легко адаптировать к специфике больших данных, ибо подсчет частотности слов и их совместной встречаемости не обязательно требует человеческого участия. Однако при этом теряется принятая в прикладном программировании ориентация на вклад

человека, рассматриваемого в качестве «золотого стандарта» [см., например, 30]. Смешанные методы контент-анализа представляются с этой точки зрения оптимальной комбинацией. С их помощью возможно масштабирование контент-анализа с участием человека на большие данные.

Человек, будь то исследователь или потребитель новостей, на первом этапе вручную кодирует небольшую по объему выборку качественных данных. Результаты ручного кодирования затем используются для машинного обучения. В случае его успеха контент-анализ остального массива качественных данных может быть делегирован компьютеру и осуществлен без непосредственного участия человека [32]. Выражение «обучение с учителем», в роли которого как раз и выступает человек, лучше всего отражает специфику масштабирования контент-анализа на большие качественные данные [21].

Одним из вариантов практического воплощения смешанных методов контент-анализа является использование основанных на замещении словарей. Закодированные вручную фрагменты текста анализируются на предмет характерных прежде всего для них слов и выражений. Эти слова и предложения должны встречаться в закодированных фрагментах значительно чаще, чем в остальных частях текста. Выявленные таким образом слова и выражения затем включаются в словарь, формирумый для конкретного кода. И так для каждого используемого в контент-анализе кода [15]. Затем большие качественные данные кодируются с использованием словаря в автоматическом режиме, без участия человека. Примеры подобного подхода можно найти в исследованиях как политической [1, С. 338-340], так и экономической [11, С. 61] направленности. Словари могут создаваться и на основе теоретических соображений, как это произошло в результате адаптации тематического апперцептивного теста (ТАТ) и некоторых других методик из прикладной психологии для использования в контент-анализе [23; 4; 19].

IV. Он-лайн платформа для контент-анализа

Для адаптации контент-анализа к специфике больших качественных данных требуется перенос соответствующего программного обеспечения с отдельных компьтеров в сеть интернет. С одной стороны, именно интернет стал одним из ключевых источников больших данных. С другой стороны, реализовать возможность параллельного кодирования текста сразу несколькими кодировщиками легче всего именно он-лайн. Как было замечено ранее, без параллельного кодирования текста несколькими кодировщиками достижение приемлимой надежности результатов проблематично.

Существует множество специализированных компьютерных программ для контент-анализа. Среди наиболее известных – NVivo, QDA Miner, Atlas TI, и другие. Однако ни одна из них не была создана именно как он-лайн платформа для контент-анализа, что отразилось и на функционале указанных программ. Набор наиболее базовых функций включает кодирование текстов и изображений, выдача закодированных фрагментов, анализ частотности кодов, анализ совместной встречаемости кодов, иногда – расчет коэффициентов согласия между кодировщиками. Однако последнее требует, либо чтобы файл проекта был последовательно закодирован несколькими людьми, либо чтобы результаты работы отдельных кодировщиков были соединены в один файл. Процедура трудоемка, сопряжена со сбоями и с ограничениями на количество участников.

В случае он-лайн платформы набор базовых функций остается прежним. Так, единожды прочитав и закодировав текст, пользователь может хранить его на сервере, получая доступ к интересующей его информации по мере надобности и вне зависимости от своего местонахождения. Наиболее популярные тексты (начиная с «горячих» новостей и заканчивая Библией и работами классиков – Шекспира или Достоевского), доступные на он-лайн платформе, прочитывает множество ее пользователей. Появляется возможность сопоставления образов, распознанных в одном и том же тексте, практически неограниченным кругом людей, действующих в самых разных

контекстах ввиду своего нахождения в разных точках земного шара.

Сопоставление распознаваемых пользователями образов может быть осуществлено двумя способами. В первом случае, пользователю предоставляется уже готовый список кодов. При использовании одних и тех же кодов сравнивается то, какие именно фрагменты кодируют пользователи. Поэтому важно, чтобы они не видели результатов кодирования (тэгов), присвоенных остальными участниками. Пользователи осведомлены лишь о результатах сопоставления своего кодирования с кодированием окружающих. Два пользователя интерпретируют текст одинаково, если присваивают коды тем же самым фрагментам. Количественным выражением сходства между пользователями являются, например, уже упомянутые коэффициенты согласия. Только в отличие от стандартной практики, низкий коэффициент согласия между двумя пользователями тоже значим – он указывает на их действия в разных контекстах. Ввиду большого количества кодировщиков всегда есть шанс найти себе «пару» или группу, распознающую образы сходным образом.

Общие книги кодов подобны предметным указателям книги или, говоря более широко – классификациям. Создание классификаций требует следования строгим логическим правилам. Классификации обычно имеют сложную структуру и несколько уровней [25, P. 37; 7]. Поэтому разработка общих книг кодов требует согласованных усилий пользователей и участия экспертов, например, в роли модераторов соответствующих обсуждений.

В качестве альтернативы, пользователь создает свою собственную книгу кодов. Никаких согласований и модерации экспертов при таком раскладе не требуется. Однако и следования строгим правилам в этом случае ожидать не следует. У одного пользователя книга кодов будет содержать 1-2 кода, а у другого – десятки. Речь тогда может идти не столько о претендующих на универсальность классификациях, сколько о частных онтологиях пользователей [43, P. 814]. Частные онтологии отражают индивидуальные и по определению менее строгие попытки формализации текста или совокупности текстов. При таком подходе сравнивается не то, где пользователь расставляет общие коды, а сами индивидуальные списки кодов. Задачу можно решить в том числе и используя адаптированный метод наименьших квадратов [37].

V. Заключение: необходимость масштабирования

При выполнении ряда условий контент-анализ может стать существенным подспорьем в разработке больших качественных данных. Одно из этих условий заключается в дальнейшем развитии смешанных методов контент-анализа. С их помощью большие качественные данные, будь то результаты исследования или новости, преобразуются в информацию с учетом контекста. При полностью машинной разработке качественных данных полноценный учет контекста затруднителен, если возможен вообще [35].

Другой предпосылкой адаптации контент-анализа к специфике разработки больших качественных данных является его перенос в режим он-лайн. В отличие от наличных програм для контент-анализа, устанавливаемых на пользовательском компьютере, он-лайн платформа позволит сравнивать образы, распознаваемые в тексте или в изображении, практически неограниченным кругом пользователей. На этой основе появляется возможность для создания социальных сетей нового типа. Члены таких сетей не обязательно работали, учились вместе или живут в одном населенном пункте. Они оказываются близкими друг другу по совершенно иным основаниям – ввиду сродства распознаваемых ими образов – вне зависимости от разделяющих их расстояний.

В чисто научном плане он-лайн платформа для контент анализа позволит наконец продвинуться к созданию банка качественных данных, доступных для вторичной разработки. Как и при кодировании текстов пользователями он-лайн платформы, исследователям стоит предоставить возможность выбирать между использованием

универсальных классификаций, как в случае eHarf и разработкой списков кодов (онтологий) под конкретный исследовательский проект. Разумеется, для создания он-лайн банка качественных данных потребуется предварительное изъятие из них всей идентифицирующей информации.

Библиография

[1] Аверьянов Л.Я. Контент-анализ. Учебное пособие. – М.: Кнорус, 2009.

[2] Алексеев А.Н. Контент-анализ в социологии и точки соприкосновения с другими отраслями знания // Проблемы контент-анализа в социологии: Материалы Сибирского социологического семинара / под ред. А.Н. Алексеева. – Новосибирск: Институт истории, филологии и философии СО АН СССР, 1970. – С. 11-18.

[3] Алексеев А.Н. Контент-анализ: техника или методология? (к постановке проблемы) // Методологические и методические проблемы контент-анализа (тезисы докладов рабочего совещания социологов). Вып. 1 / Здравомыслов А.Г., отв. ред. – Москва-Ленинград: Институт социологических исследований, 1973. – С. 19-28.

[4] Алмаев Н.А. Применение контент-анализа в исследованиях личности. – М.: Издательство «Институт психологии РАН», 2012.

[5] Берроуз Р., Севидж М. После кризиса? Big data и методологические вызовы эмпирической социологии // Социологические исследования. – 2016. – №3. – С. 28-35.

[6] Брудный А.А. О психологии понимания текста // Методологические и методические проблемы контент-анализа (тезисы докладов рабочего совещания социологов). Вып. 1 / Здравомыслов А.Г., отв. ред. – Москва-Ленинград: Институт социологических исследований, 1973. – С. 92-93.

[7] Джумайло Е.С., Баранюк В.В. Методика онтологического связывания объектов в автоматизированных системах с использованием классификаторов // International Journal of Open Information Technologies. – 2018. – vol. 6, no.6. – С. 97-102.

[8] Единый архив экономических и социологических данных. http://sophist.hse.ru/, проверено 13.05.2019 г.

[9] Київський архів: національний банк соціологічних даних. http://ukraine.survey-archive.com/, проверено 13.05.2019 г.

[10] Клеменков П.А., Кузнецов С.Д. Большие данные: современные подходы к хранению и обработке // Труды Института системного программирования РАН. – 2012. – том 23. – С. 143-158.

[11] Кононова О.В., Ляпин С.Х., Прокудин Д.Е. Исследование терминологической базы междисциплинарного научного направления «цифровая экономика» с использованием инструментов контекстного анализа // International Journal of Open Information Technologies. – 2018. – vol. 6, no.12. – С. 57-66.

[12] Махотина Н.В. Применение метода контент-анализа при исследовании большого массива информации (на примере «Федерального списка экстремистских материалов») // Труды ГПНТБ СО РАН (Государственной публичной научно-технической библиотеки СО РАН). Выпуск 10. Теория и практика научных исследований в библиотеках (Материалы межрегиональной научно-практической конференции, Абакан, 21-25 сентября 2015 г.) / Артемьева Е.Б., Лаврик О.Л., отв ред. – Новосибирск: ГПНТБ СО РАН, 2016. – С. 379-385.

[13] Намиот Д.Е., Куприяновский В.П., Николаев Д.Е., Зубарева Е.В. Стандарты в области больших данных // International Journal of Open Information Technologies. – 2016. – vol. 4, no. 11. – С. 12-18.

[14] Никипорец-Такигава Г.Ю. Агрессия в языке СМИ: опыт статистического анализа // Язык, сознание, коммуникация: Сб. статей / Отв. ред. В.

В. Красных, А. И. Изотов. – М.: МАКС Пресс, 2006. – Вып. 33. – С. 56-64.

[15] Олейник А.Н. Триангуляция в контент анализе: вопросы методологии и эмпирийная проверка // Социологические Исследования. – 2009. - №2. – С. 65-79.

[16] Понкин И.В., Редькина А.И. Цифровая формализация права // International Journal of Open Information Technologies. – 2019. – vol. 7, no.1. – С. 39-48.

[17] Самков Л.М. Проблема контекста в теоретической информатике // Проблемы контент-анализа в социологии: Материалы Сибирского социологического семинара / под ред. А.Н. Алексеева. – Новосибирск: Институт истории, филологии и философии СО АН СССР, 1970. – С. 89-93.

[18] Секерин В.П. Контент-анализ в комплекном изучении газеты (некоторые методические выводы) Методологические и методические проблемы контент-анализа (тезисы докладов рабочего совещания социологов). Вып. 2 / Здравомыслов А.Г., отв. ред. – Москва-Ленинград: Институт социологических исследований, 1973. – С. 58-62.

[19] Таршис Е.Я. Контент-анализ: принципы методологии. Построение теоретической базы. Отнология, аналитика и феноменология текста. Программы исследования. – М.: URSS, Либроком, 2013.

[20] Фролов А.А., Сильнов Д.С., Садретдинов А.М. Анализ механизмов обнаружения запрещенного содержимого в сети Интернет // International Journal of Open Information Technologies. – 2019. – vol. 7, no.1. – С. 90-96.

[21] Черняк Л.С. Большие Данные – новая теория и практика // Открытые системы. СУБД. – 2011. – №10. – С. 18.

[22] Чехарин Е.Е. Большие данные: большие проблемы // Перспективы Науки и Образования. – 2016. – №3(21). – С. 7-11.

[23] Шалак В.И. Контент-анализ. Приложения в области: политологии, психологии, социологии, культурологии, экономики, рекламы. – М.: Омега-Л, 2004.

[24] Храмова Н.Н. Спецификация генерации новостей через RSS на примере работы агрегатора Яндекс.Новости // Знак: проблемное поле медиаобразования. – 2015. – №3(17). – С. 36-41.

[25] Berman J.J. Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information. – Waltham, MA: Morgan Kaufmann, 2013.

[26] Beierle J.M., Witkowski S. HRAF Coding Reliability // Cross-Cultural Research. – 1974. – vol. 9, no. 1. – P. 57-65.

[27] eHarf World Cultures. https://ehrafworldcultures.yale.edu/ehrafe/, accessed 13.05.2019.

[28] Harwood N. An interview-based study of the functions of citations in academic writing across two disciplines // Journal of Pragmatics. – 2009. – vol. 41, no. 3. – P. 497-518.

[29] Hesse B.W., Hesse Moser R.P., Riley W.T. From Big Data to Knowledge in the Social Sciences // The ANNALS of the American Academy of Political and Social Science. – 2015. – vol. 659, no. 1. – P. 16-32.

[30] Jurafsky D., Martin J.H. Speech and Language Processing. – 2nd ed. – Upper Saddle River, NJ: Pearson-Prentice Hall, 2008.

[31] Krippendorff K. Content Analysis: An Introduction to Its Methodology. – Thousand Oaks, CA: SAGE, 2004. – 2nd edition.

[32] Lewis S.C., Zamith R., Hermida A. Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods // Journal of Broadcasting & Electronic Media. – 2013. – vol. 57, no. 1. – P. 34-52.

[33] Mannens E., Coppens S., de Pessemier T., Dacquin H., Van Deursen H., De Sutter R., Van de Walle R. Automatic news recommendations via aggregated profiling

// Multimedia Tools and Applications. – 2013. – vol. 63, no 2. – P. 407-425.
[34] Nikiporets-Takigawa G. 'Socio-Political Insider' system: promises and limitations for the political analysis and prognosis // PolitBook. – 2018. – №1. – С. 6-20.
[35] Oleinik A.N. What are neural networks not good at? On artificial creativity // Big Data & Society. – On-line first
[36] Oleinik A.N., Popova I., Kirdina S., Shatalova T. On the choice of measures of reliability and validity in the content-analysis of texts // Quality and Quantity. – 2014. – vol. 48, no. 5. – P. 2703-2718.
[37] Oleinik A.N., Kirdina-Chandler S., Popova I., Shatalova T. On academic reading: citation patterns and beyond // Scientometrics. – 2017. – vol. 113, no. 1. – P. 417-435.
[38] Vogt W.P. Quantitative Research Methods for Professionals. – Boston, MA: Pearson, 2007.
[39] Wiedemann G. Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences // Forum Qualitative Sozialforschung / Forum: Qualitative Social Research. – 2013. – vol. 14, no. 2. – Art. 13.
[40] Web of Science. Web of Science Core Collection. https://clarivate.com/products/web-of-science/web-science-form/web-science-core-collection/, accessed 14.05.2019.
[41] Weber M. Economy and Society: An Outline of Interpretative Sociology. – edited by Roth G., Wittich C. – New York: Bedminster Press, 1968.
[42] World Bank. World Bank Open Data. https://data.worldbank.org/, accessed 13.05.2019.
[43] Yang Q. A novel recommendation system based on semantics and context awareness // Computing. – 2018. – vol. 100, no. 8. – P. 809-823.

Content analysis of big qualitative data

Anton Oleinik

***Abstract*— When working with big data in science (research databanks, literature reviews) and everyday life (news aggregators), there is a need for mining, classifying and storing information. Information is defined as data in a processed form. The methodology of content analysis in its various forms, qualitative (manual coding), quantitative (words frequencies and co-occurrences) and mixed methods (creation of ad hoc dictionaries based on substitution), offers a tool to address this issue. Interest in content analysis emerged as early as in the 1970s, yet it remains relatively unknown outside of sociology, linguistics and communication studies. Content analysis allows converting qualitative data (texts, images) into digital format (vectors and matrices) and subsequent manipulating digital information using linear algebra, multidimensional scaling and other tools from natural sciences. The conversion into digital formal also paves the way to machine learning. Supervised machine learning looks particularly promising since it implies keeping focus on interpretation of data proper to interpretative sociology. Supervised machine learning is compatible with mixed methods content analysis. The existing program for computer-assisted content analysis (QDA Miner, Atlas TI, NVivo etc.) have several limitations. Restrictions on the number of their users (coders) refer to one of the limitations. The creation of on-line platforms for content analysis allows bypassing this and some other limitations. The idea of creating an on-line databank for qualitative data and a platform for content analyzing it is discussed. In contrast to quantitative data, qualitative research data is rarely available for secondary analysis.**

***Key words*— qualitative data, data mining, content analysis**

[1] Averyanov L.Ya. Content analysis: a text. Moscow: Knorus, 2009.

[2] Alexeev A.N. Content analysis in sociology and in relationship with other disciplines. In Problems in content analysis in sociology, Alexeev A.N. (ed.) Novosibirsk, 1970, P. 11-18.

[3] Alexeev A.N. Content analysis: a technique or a method? In Methodological and methodical issues in content analysis. Issue 1, Zdravomysov A.G. (ed.) Moscow-Leningrad., 1973, P. 19-28.

[4] Almayev N.A. Applications of content analysis to psychology of individuality. Moscow, 2012.

[5] Burrows R., Savage M. After the crisis? Big data и methodological challenges in empirical sociology. Socis, 3, 2016.

[6] Brydny A.A. On psychology of understanding a text. In Methodological and methodical issues in content analysis. Issue 1, Zdravomysov A.G. (ed.) Moscow-Leningrad, 1973. P. 92-92.

[7] Djumailo E.S., Baranyuk V.V. A method for connecting ontologically objects in automated systems using classifiers. International Journal of Open Information Technologies. Vol. 6, no.6, 2018.

[8] United archive of economic and survey data. http://sophist.hse.ru/, accessed on 13.05.2019.

[9] Kyiv archive: a national bank of survey data. http://ukraine.survey-archive.com/, accessed on 13.05.2019.

[10] Klemenkov P.A., Kuznetsov S.D. Big data: contemporary approaches to mining and storage. Proceedings of the Institute of System programming. Vol. 23, 2012.

[11] Kononova O.V., Lyapin S.Kh., Prokudin D.E. A study of terms used in ‘digital economy’ with the help of content analysis. International Journal of Open Information Technologies. Vol. 6, no.12, 2018.

[12] Makhotina N.V. Using content analysis to study a large dataset. In Proceedings of the

GPNTB SO RAN. Issue 10. Novosibirsk, 2016. P. 379-385.
[13] Namiot D.E., Kupriyanovsky V.P., Nikolaev D.E., Zubareva E.V. Standards related to big data. International Journal of Open Information Technologies. Vol. 4, no. 11, 2016.
[14] Nikiporets-Takigawa G. Aggression in the language of mass media: a statistical analysis. In Language, consciousness, communication, edited by V.V. Krasnykh, A.I. Izotov. Issue 33. Moscow: Maks, 2006. P. 56-64.
[15] Oleinik A.N. Triangulation in content analysis: issues in methodology and empirical tests. Socis, №2, 2009.
[16] Ponkin I.V., Redkina A.I. Digital formalization of the law. International Journal of Open Information Technologies. Vol. 7, no.1, 2019.
[17] Samkov L.M. The issue of context in theoretical informatics. In Problems in content analysis in sociology, Alexeev A.N. (ed.) Novosibirsk, 1970. P. 89-93.
[18] Sekerin V.P. Content analysis as a tool to a complex study of a newspaper. In: Methodological and methodical issues in content analysis. Issue 2, Zdravomysov A.G. (ed.) Moscow-Leningrad, 1973. P. 58-62.
[19] Tarshis E.Ya. Content analysis: methodological issues. Moscow: URSS, Librocom, 2013.
[20] Frolov A.A., Silnov A.A., Sadretdinov A.M. Tools for detecting prohibited content in the internet. International Journal of Open Information Technologies. Vol. 7, no.1, 2019.
[21] Chernyak L.S. Big data: a new theory and practice. Open systems. SUBD. 10, 2011.
[22] Chekharin E.E. Big data: key issues. Prospects of science and education. №3(21), 2016.
[23] Shalak V.I. Content analysis and its applications. Moscow: Omega-L, 2004.
[24] Khramova N.N. Particularities of news generation via RSS using Yandex.News as an example. Symbol: the field of media-education. №3(17), 2015.
[25] Berman J.J. Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information. Waltham, MA: Morgan Kaufmann, 2013.
[26] Beierle J.M., Witkowski S. HRAF Coding Reliability. Cross-Cultural Research. Vol. 9, no. 1, 1974.
[27] eHarf World Cultures. https://ehrafworldcultures.yale.edu/ehrafe/, accessed 13.05.2019.
[28] Harwood N. An interview-based study of the functions of citations in academic writing across two disciplines. Journal of Pragmatics. Vol. 41, no. 3, 2009.
[29] Hesse B.W., Hesse Moser R.P., Riley W.T. From Big Data to Knowledge in the Social Sciences. The ANNALS of the American Academy of Political and Social Science. Vol. 659, no. 1, 2015.
[30] Jurafsky D., Martin J.H. Speech and Language Processing. 2nd ed. Upper Saddle River, NJ: Pearson-Prentice Hall, 2008.
[31] Krippendorff K. Content Analysis: An Introduction to Its Methodology. Thousand Oaks, CA: SAGE, 2004. 2nd edition.
[32] Lewis S.C., Zamith R., Hermida A. Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods. Journal of Broadcasting & Electronic Media. Vol. 57, no. 1, 2013.
[33] Mannens E., Coppens S., de Pessemier T., Dacquin H., Van Deursen H., De Sutter R., Van de Walle R. Automatic news recommendations via aggregated profiling. Multimedia Tools and Applications. Vol. 63, no 2, 2013.
[34] Nikiporets-Takigawa G. 'Socio-Political Insider' system: promises and limitations for the political analysis and prognosis. PolitBook, №1, 2018.
[35] Oleinik A.N. What are neural networks not good at? On artificial creativity. Big Data & Society. On-line first
[36] Oleinik A.N., Popova I., Kirdina S., Shatalova T. On the choice of measures of reliability and validity in the content-analysis of texts. Quality and Quantity. Vol. 48, no. 5, 2014.
[37] Oleinik A.N., Kirdina-Chandler S., Popova I., Shatalova T. On academic reading: citation patterns and beyond. Scientometrics. Vol. 113, no. 1, 2017.

[38] Vogt W.P. Quantitative Research Methods for Professionals. Boston, MA: Pearson, 2007.
[39] Wiedemann G. Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research. Vol. 14, no. 2, 2013.
[40] Web of Science. Web of Science Core Collection. https://clarivate.com/products/web-of-science/web-science-form/web-science-core-collection/, accessed 14.05.2019.
[41] Weber M. Economy and Society: An Outline of Interpretative Sociology. Edited by Roth G., Wittich C. New York: Bedminster Press, 1968.
[42] World Bank. World Bank Open Data. https://data.worldbank.org/, accessed 13.05.2019.
[43] Yang Q. A novel recommendation system based on semantics and context awareness. Computing. Vol. 100, no. 8, 2018.