6. Туманова О.И. Страх как инструмент политики // Вестник ТвГТУ. Серия «Науки об обществе и гуманитарные науки». 2020. № 4 (23). С. 27–31.

7. Чанышева З.З. Этологические аспекты управления массовым поведением // Политическая лингвистика. 2020. №1 (79). С. 16-26.

8. Эко У. Отсутствующая структура. Введение в семиологию / У. Эко. М.: Петрополис, 1998. 432 с.

9. Эпштейн О.В. Лексические показатели политики страха: на материале заголовков англоязычной прессы //Филологические науки. Вопросы теории и практики. Тамбов: Грамота, 2012. № 6. С. 215-218.

10. Callanan, Valerie J. (March 1, 2012). «Media Consumption, Perceptions of Crime Risk and Fear of Crime: Examining Race/Ethnic Differences». Sociological Perspectives. 55 (1): 93–115.

11.CNN,UKlabelsRussiatopsecuritythreat,issueswarningonChina,andpromisestobuildmore nuclear warheads[Электронный ресурс] / [Режим доступа]:https://edition.cnn.com/2021/03/16/europe/uk-security-defense-review-intl-hnk-gbr/index.html#:~:text=UK%20labels%20Russia%20top%20security,to%20build%20more%20nuclear%20warheads&text=To%20accomplish%20its%20goals%2C%20the,it%20spent%20in%202019%2D2020 (дата обращения 22.03.2021).

12. Sorokin P.A., Horowitz I.L. Man and Society in Calamity [Электронный ресурс]/ P.A.Sorokin. New Brunswick (USA); London (UK):Transaction Publishers, 2017. p.362.[Режим доступа]:https://books.google.ru/books?id=4R0uDwAAQBAJ&pg=PT25&hl=ru&source=gbs_toc_r&cad=4#v=onepage&q&f=false (дата обращения 20.02.2021).

Т.Н. Новосильцева (Россия, Санкт-Петербург)

ОБРАБОТКА БОЛЬШИХ ДАННЫХ В ДАТА-ЖУРНАЛИСТИКЕ

В статье рассматриваются понятия больших и открытых данных, определение дата-журналистики и основные принципы работы с большими данными в дата-журналистике. Большие данные описываются путем сопоставления понятий из словарей и определения их основополагающих признаков. Процесс работы с большими данными разделен на этапы обработки и визуализации. На этапе обработки описываются основные журналистские приемы и компьютерные технологии, а на этапе визуализации приводятся примеры инструментов для работы журналистов.

Ключевые слова: *большие данные, дата-журналистика, журналистика данных, медиатекст.*

В цифровую эпоху количество открытой информации для хранения и распространения в интернете растет с каждым годом, образуя массивы данных. В связи с этим появились особые термины – «big data» и «большие объемы данных». Большие данные лежат на стыке между технической стороной, которая собирает и анализирует информацию, и журналисткой, которая исследует, трактует и представляет результаты.

Расширение информационных возможностей привело к появлению новых задач современной журналистики. С усложнением процесса сбора и обработки

огромных объемов информации, появляется все больше вопросов о работе с большими данными в журналистике. Столкновение с необходимостью анализировать сырые данные вызывает у многих журналистов ступор, хотя современному журналисту необходимо понимать основы работы с большими данными для создания актуального медиатекста. Это определило потребность в новых журналистских инструментах для работы с большими данными. Таким образом, возможность исследовать и создавать сюжет из массива информации становится новым этапом в развитии журналистики сегодня.

Большие данные и открытые данные

Дата-журналистика использует в своих исследованиях крупные массивы данных и определенные методы и инструменты их обработки, в связи с чем необходимо дать определение «больших данных». Само появление больших объемов данных было связано с увеличением объемов и типов данных, а также скорости их поступления. Определения разнятся в зависимости от источника. Согласно Кембриджскому словарю, большие данные (англ. Big data) – это очень большие массивы данных, производимые человеком через интернет, которые могут храниться, исследоваться и использоваться с помощью специальных инструментов и методов [5]. В данном случае выделяется центральная роль человека в группировке информации с помощью технологий. Оксфордский словарь фокусируется на доле влияния технологии на обработку информации. Он предлагает определение больших данных как очень больших массивов данных, которые могут быть проанализированы компьютером для выявления трендов, тенденций и ассоциаций, особенно связанных с человеческим поведением и взаимодействиями [7].

Тем не менее, определение больших данных как очень больших массивов не всегда характеризует их достаточно четко. Во избежание противоречий, для определения больших данных используется формула «семь V». Наиболее ключевыми являются первые три V: Volume (объем), Velocity (скорость прироста) и Variety (разнообразие информации) [6]. С развитием и увеличением уже достаточно больших объемов данных начали возникать проблемы ограничения концепции больших данных по трем названным аспектам. Поэтому для их определния были выделены еще четыре V: Veracity (достоверность), Variability (изменчивость), Visualization (визуализация), Value (ценность) [6]. Таким образом, помимо очевидного значительного объема, большим данным необходимо иметь некоторые постоянные признаки для их обработки и точного определения. Эти характеристики также подтверждают представления о сложности обработки и анализа таких данных.

Традиционным источником больших данных являются интернет и социальные медиа, в которых миллиарды пользователей взаимодействуют друг с другом ежесекундно. Помимо информации пользователей, компании и организации также обладают большим объемом информации, который они могут выкладывать в открытый доступ. Данные могут непрерывно поступать с различных измерительных устройств для дальнейшей обработки. Примерами источников больших данных могут быть информация о транзакциях и покупках, логи и сигналы с датчиков и другие. Такие данные активно используются не

только в научно-исследовательской деятельности, но и в журналистике для аргументации и предоставления фактологической информации.

Подобная информация в открытом доступе называется открытыми данными. Открытые данные – это данные, которые могут свободно использоваться и распространяться кем угодно [1]. Помимо этого открытые данные не должны иметь ограничений на использование и распространение. Таким образом, не любая информация, выложенная в интернет, является открытой, так как может иметь ограничения, следовательно, не может быть использована в журналистской истории.

Отдельный интерес представляют открытые государственные данные, которые дают возможность журналистам и исследователям изучить материал, преобразовать его и использовать в дальнейшей работе. Подобный доступ предоставляется в рамках программы «Открытое государство». Различные субъекты России, например, администрации Санкт-Петербурга, Москвы и регионов, предоставляют открытые данные по многим направлениям общественной деятельности, на их основе составляются журналистские истории.

Что такое дата-журналистика

Распространение компьютеров и баз данных с 1970 годов ознаменовало появление дата-журналистики [4, с. 82]. Появление массивов с открытыми данными и инструментов их обработки и визуализации выделило дата-журналистику в отдельный жанр журналистики. Само понятие больших данных появилось в 1995 году, а их использование в журналистике датируется серединой 2000-х годов [2, с. 185]. С помощью больших данных, технических инструментов и навыков программирования журналисты получили возможность разрабатывать новые сюжеты и варианты их представления аудитории.

Дата-журналистика – это раздел журналистики, который использует большие данные для создания информационных поводов или уточнения сведений, приведенных в качестве исследования. Дата-журналистика подразделяется на несколько направлений - журналистика данных (работа с текстовой информацией) и инфографика (визуализация данных). Базы данных могут стать базой для журналистской истории или послужить инструментом для ее раскрытия. Дата-журналистика включает не только поиск информации по результатам журналистских запросов, но и ее обработку и анализ. Именно грамотная обработка данных и проработка журналистских историй отражает главную компетенцию современного журналиста. В этом журналистам помогают основные приемы работы с большими данными: очистка массива данных, анализ, трактовка и визуальная репрезентация для аудитории.

Основные приемы обработки

Анализ больших данных наиболее часто осуществляется посредством таких техник, как:

• Data mining (Добыча данных) – поиск и классификация разнородных данных, основанные на анализе массивов информации;

• машинное обучение и нейронные сети – самостоятельное нахождение компьютером ассоциаций для статистического анализа и составления прогно-

зов с помощью искусственного интеллекта на основе специальных моделей и алгоритмов;

• краудсорсинг – привлечение большой группы людей для обработки больших данных;

• смешение и интеграция данных – процесс привлечения информации из разных источников в один комплекс;

• статистический анализ – сбор статистики по определенным критериями для получения конкретного результата или подтверждения выводов;

• визуализация аналитических данных – использование графиков, диаграмм и рисунков для репрезентации данных визуально.

Для осуществления подобных техник также используются определенные технологии и программы, такие как системы управления базами данных NoSQL, алгоритмы MapReduce и программные каркасы Hadoop [1]. Подобные компьютерные системы значительно облегчают и ускоряют процесс работы с огромными объемами информации.

Рассмотрим этапы обработки больших данных в модели MapReduce, которая является базовой. Всего работа осуществляется в 2 этапа:

1. Этап Map (карта) – предварительная обработка данных. На этом этапе данные загружаются на компьютер или несколько компьютеров, определяется задача, согласно которой данные разделяются между компьютерами [3, с. 23]. Кратко этот этап можно охарактеризовать как фильтрующий и выполняющий предварительную обработку.

2. Этап Reduce (сокращение) – данные обрабатываются согласно задаче и формируется результат [3, с. 24]. На выходе получается сгенерированная статистика без затрачивания времени на ручную обработку.

Для работы с большими данными программная модель Hadoop MapReduce выделяет следующие принципы:

• горизонтальная масштабируемость – с повышением количества поступающих данных система также должна расширяться [3, с. 25].

• Отказоустойчивость – сбои работы отдельных машин не должны выводить из строя весь процесс работы. Файлы дублируются для предотвращения возможных потерь при сбое.

• Локальность – данные должны храниться на той же машине, на которой они обрабатываются во избежание излишних трат на перенос данных с одного устройства на другое [3, с. 26].

Визуализация

Последним этапом обработки больших данных в дата-журналистике является визуализация как наиболее наглядный способ представления информации аудитории. Для выполнения визуализации необходимо достоверно трактовать полученные данные после машинной обработки. Именно трактовка данных является важнейшим элементом в цепочке обработки, так как журналист является посредником между информацией и аудиторией и от него зависит, как будет

воспринята информация. Сырые данные не всегда позволяют раскрыть картину и сделать последовательную журналистскую историю.

Для визуализации данных есть большое количество бесплатных инструментов, таких как Overview, Tableau, GIS [1]. Они предлагают стандартизированные варианты визуализации информации в виде инфографики, таблиц и графиков, однако не всегда подходят для конкретных визуальных журналистских проектов. Для решения таких задач важно создавать черновые варианты и вспомогательные графики для отбора лучшего варианта представления информации. На основе полученных результатов после обработки принимается решение о необходимости и лучших способах визуализации.

Журналистика данных является одним из наиболее перспективных направлений журналистики за последние годы благодаря стремительному распространению больших данных и журналистских материалов, основанных на их обработке и визуализации. В цифровую эпоху в распоряжении журналистов в интернете имеется огромное количество открытых источников и весь необходимый инструментарий для проведения полноценного журналистского исследования. В связи с этим журналистам важно уметь обрабатывать и анализировать большие объемы данных для получения новых журналистских сюжетов. Для этого необходимо уметь пользоваться источниками открытых данных и инструментами обработки и визуализации данных. Все это значительно предотвращает возможность представления информации в искаженном виде и повышает прозрачность публикуемой информации.

Литература

1. Абдыкаримова А.Т. Big Data: проблемы и технологии [Электронный ресурс] // Международный журнал гуманитарных и естественных наук. 2019. №5 (1). URL: https://cyberleninka.ru/article/n/big-data-problemy-i-tehnologii(дата обращения 24.03.2021).

2. Дарменова А.С., Мамыкова Ж.Д., Андерсен К.Н.Открытые данные: двадцатипятилетняя история развития[Электронный ресурс] // Вестник НГУЭУ. 2020. №2. С. 183-197. URL: https://cyberleninka.ru/article/n/otkrytye-dannye-dvadtsatipyatiletnyaya-istoriya-razvitiya (дата обращения 24.03.2021).

3. Ермолин Д.С., Варфоломеева И.Ф. Основы Bigdata. Mapreduce. Примеры задач. [Электронный ресурс] // Череповецкий государственный университет. 2016. С.22-26. URL: https://www.elibrary.ru/item.asp?id=27491575&pf=1(дата обращения 08.03.2021).

4. Корнев М.С. История понятия «большие данные» (Big Data): словари, научная и деловая периодика [Электронный ресурс] // Вестник РГГУ. 2018. №1 (34). С.81-85. URL: https://cyberleninka.ru/article/n/istoriya-ponyatiya-bolshie-dannye-big-data-slovari-nauchnaya-i-delovaya-periodika(дата обращения 08.03.2021).

5. Big Data [Электронный ресурс] // Cambridge Dictionary. URL: http://dictionary.cambridge.org/dictionary/english/big-data(дата обращения 08.03.2021).

6. Big Data [Электронный ресурс] // IT enterprise. URL: https://www.it.ua/ru/knowledge-base/technology-innovation/big-data-bolshie-dannye(дата обращения 08.03.2021).

7. Big Data [Электронный ресурс] // Oxford Dictionary. URL: https://en.oxforddictionaries.com/definition/big_data (дата обращения 08.03.2021).