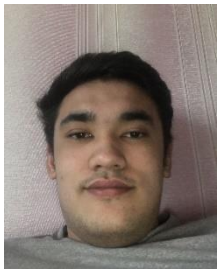
УДК 004.67

СПЕЦИФИКА РАБОТЫ С "БОЛЬШИМИ ДАННЫМИ" В СОВРЕМЕННЫХ СМИ



С.С. Курбанов
Студент БГУИР

С.Н. Нестеренков
Кандидат технических наук, доцент, декан факультета компьютерных систем и сетей

А.Н. Марков
Старший преподаватель, магистр технических наук, заместитель начальника Центра информатизации и инновационных разработок БГУИР

Центр информатизации и инновационных разработок Белорусского государственного университета информатики и радиоэлектроники, Республика Беларусь
Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: seyitjan0306@gmail.com, s.nesterenkov@bsuir.by, a.n.markov@bsuir.by

С.С. Курбанов
Студент 4 курса специальности "Программное обеспечение информационных технологий" БГУИР.

С.Н. Нестеренков
Кандидат технических наук, декан факультета компьютерных систем и сетей Белорусского государственного университета информатики и радиоэлектроники, доцент кафедры Программного обеспечения информационных технологий. Автор публикаций на тему машинного обучения, алгоритмов принятия решений, искусственных нейронных сетей и автоматизации

А.Н. Марков
Магистр технических наук, старший преподаватель кафедры ПИКС, заместитель начальника Центра информатизации и инновационных разработок Белорусского государственного университета информатики и радиоэлектроники.

Аннотация. «Большие данные» или Big Data на сегодняшний день являются источником информации, который журналисты уже не могут игнорировать. Наборы данных, базы данных, несистематизированные сведения на официальных сайтах государственных или коммерческих структур являются ресурсом для работы журналиста. Big Data является не только источником информации, но и доказательной базой. Используя в качестве аргументов для своих тезисов количественные, статистические показатели, журналист повышает уровень лояльности аудитории, степень доверия к публикации. Анализ данных, выявление взаимосвязей, корреляций, составление прогнозов, рейтингов позволяет создать эксклюзивный привлекательный и достоверный контент, привлекающий аудиторию и способствующий улучшению репутации издания. Статья посвящена определению места «больших данных» в деятельности информационных и аналитических отделов редакции, а также анализу роли Big Data в освещении общественно важных тем и выявлению тенденций дальнейшей работы журналисты с подобными сведениями. За период исследования были использованы следующие методы исследования: метод описания, сравнительного анализа и обобщения. В рамках выводов, прежде всего, необходимо обозначить, что работа по сбору и обработке данных требует от журналиста серьезного подхода и крайней внимательности. Используя в качестве аргументов для своих тезисов количественные, статистические показатели, журналист повышает уровень лояльности аудитории, степень доверия к публикации.

Ключевые слова: Большие данные, дата-журналистика, базы данных, СМИ, социальные сети, визуализация, исследование, цифровизация, верификация

Введение.

Современное медиапространство любой страны тесно связано с процессами цифровизации, дигитализации, конвергенции, которые в совокупности представляют собой единую тенденцию – переводить все существующие данные в электронный формат, доступный в том числе в режиме онлайн. Подобное явление приводит к экспоненциальному росту доступных любому пользователю данных. И теперь все чаще в рамках журналистской работы происходит обращение к «большим данным» как к источнику информации. Очевидным является причинно-следственная связь, в рамках которой постоянное увеличение оцифрованной информации обо всех сферах жизнедеятельности и активное обращение все большего количества людей к социальным сетям приводит к увеличению уже и без того значительного количества информации практически о чем угодно. В период, когда пользователи по личной инициативе охотно делятся персональными данными с аудиторией (делая отметки своей геолокации, ежедневно публикуя фотографии дома, на улице, на месте работы/учебы, открывая доступ к комментариям всем желающим, регулярно делясь своими мыслями, чувствами, событиями жизни посредством постов), журналисты имеют неограниченные возможности относительно поиска тем или важной информации для подтверждения своих тезисов. Журналист и медиа исследователь Эм Кунтце отмечает, что «мир больших данных избавляет журналистов от необходимости выдвигать гипотезы – не надо строить теории и искать данные для доказательств. Все, наоборот» [1]. Другими словами, теперь журналист получает огромное количество данных (уже подтвержденных фактов) и ищет в них тему для своего материала.

Типология BigData в журналистской практике.

Сотрудник Института Рейтер М. Стоун в своем исследовании «Данные для медиа» отмечает важные свойства «больших данных», так называемые 4Vs: объем данных (volume), скорость передачи данных (velocity), разнообразие структурированных и неструктурированных данных (variety) и потенциальную ценность с точки зрения бизнеса и получения дохода (value) [1]. Эти же свойства выделяет С. А. Вартанов, отмечая при этом, что «Big data – это разнородные неструктурированные данные крайне большого объема, увеличение которого происходит ежедневно с большой скоростью» [2].

В рамках журналистской работы «большие данные» используются в качестве источника информации для разнообразного контента: начиная от новостных заметок и заканчивая масштабными расследованиями [3]. В связи с этим, можно типологизировать «большие данные» на пять основных категорий (виды данных, к которым обращаются журналисты в рамках своей профессиональной деятельности чаще всего): базы данных, социальные сети, видеозаписи, фотоизображения и научные исследования.

Самым распространенным и наиболее востребованным видом «больших данных» являются базы данных, которые можно разделить на открытые/частично открытые и закрытые, а также на платные и бесплатные. Кроме ресурсов с данными от некоммерческих организаций, фондов, благотворительных центров и отдельных энтузиастов, есть некоммерческие расследовательские организации, которые создают свои базы данных по результатам проведенных расследований. Так у «Международного консорциума журналистов-расследователей» (ICIJ) есть база данных, посвященная архивным данным по документам, составляющим «Панамские архивы» и «Архивы Райских островов».

Согласно оценкам Т. Беттса (2018), привычные несколько лет назад паттерны (устоявшиеся практики и способы) потребления контента аудиторией Financial Times претерпели значительные изменения. Основным изменением стал переход львиной доли аудитории в цифровую среду: по состоянию на 2018 г. число подписчиков онлайн-версии издания FT.com значительно превысило число подписчиков бумажной газеты Financial Times [4]. Это превратило сбор данных о пользователях сайта FT.com из сугубо технической

задачи IT-отдела в стратегическую: анализ этих данных помогает увеличивать аудиторию и количество подписчиков за счет лучшего понимания их потребностей.

На основе собираемых данных о социально-демографических характеристиках и активности пользователей в Financial Times создаются так называемые «сигнатуры цифрового потребления». В эти сигнатуры включаются данные о потреблении читателем контента из разных разделов сайта (Companies, Markets, World News, Management, Weekend и т.д.). Далее эти сигнатуры используются в различных целях: для изучения контент-предпочтений аудитории, для улучшения обратной связи между редакцией и читателями, для персонализации контента, а также для таргетирования рекламных материалов.

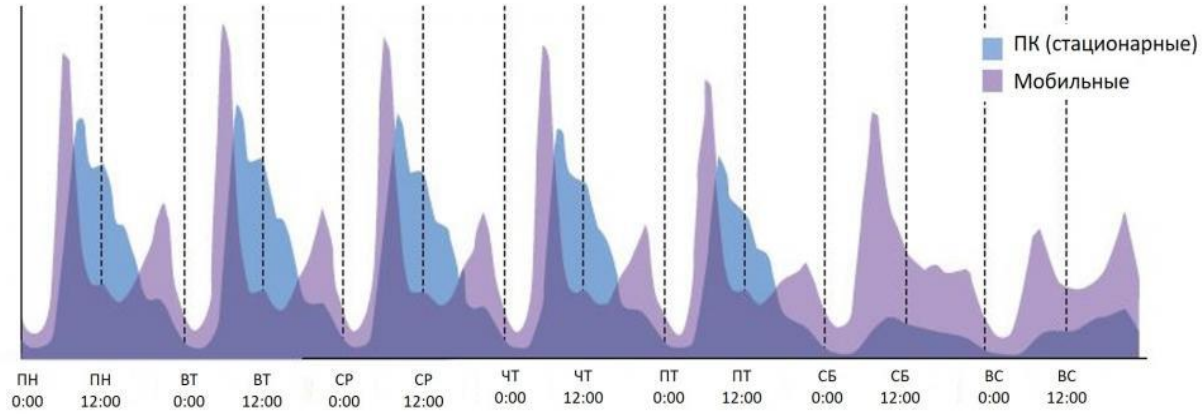


Рисунок 1 – Недельная динамика аудитории онлайн-версии Financial Times в зависимости от типа платформы

Появление мобильных платформ добавило к данным, собираемым Financial Times, еще один уровень, который также подвергается анализу. В частности, одним из результатов этого анализа стало понимание того, что контент из разных разделов потребляется с помощью разных устройств и в разное время. Например, раздел о досуге (Weekend) пользователи предпочитают читать с помощью мобильных устройств и делают это, как правило, по выходным. В то же время, бизнес-разделы (Management, Finance) собирают наибольшую аудиторию с помощью компьютеров и по будням.

База данных поделена на рубрики: Offshore Leaks, Panama Papers, Bahamas Leaks и Paradise Papers. На сайте также представлена масштабная база данных «Международная база данных медицинского оборудования», в которой содержится более 120 000 напоминаний, предупреждений по технике безопасности и эксплуатации медицинских изделий и их связи с производителями.

Еще один некоммерческий центр, на протяжении нескольких лет регулярно публикующий собственные базы данных – ProPublica [5]. Их данные поделены на несколько рубрик (здравоохранение, политика, бизнес, правосудие, финансы, религия, транспорт, военная промышленность, образование, экология) и классифицируются как премиум-базы (платные) и обычные (бесплатные). Для доступа к базе необходимо заполнить небольшую регистрационную форму. В 2020 году на сайте предлагалось 104 набора данных по самым различным сферам жизнедеятельности, при этом 67 наборов данных, то есть большинство, можно получить бесплатно. Премиальные данные являются уже обработанными и эксклюзивными сведениями, которые журналисты получили с помощью запросов или вычислительных манипуляций. Бесплатные базы часто содержат необработанные, неструктурированные данные, которые журналисты брали за основу своих исследований и проводили дальнейшие операции с ними уже в рамках своей публикации.

Наряду с базами данных социальные сети также представляют собой неограниченный источник информации практически о любом человеке [6]. Постоянные отметки геолокации пользователей, непрерывная публикация фотоизображений, на которых отчетливо видны места посещения человека, круг друзей и знакомых, родственников, посты с публикациями мыслей, знаний, рассуждений человека и т.д. В рамках «больших данных» можно говорить об анализе твитов, постов, фотографий, результат которого демонстрируется в совокупности, без акцентирования внимания на конкретных персоналиях или личных аккаунтах.

Помимо многочисленных баз данных и персональной информации из социальных сетей, еще одним вариантом «больших данных» являются многочисленные записи с камер видеонаблюдения [7]. Они установлены практически во всех общественных местах: офисах, общественном транспорте, магазинах, салонах красоты, торговых центрах, на городских улицах и на приборных панелях автомобилей. «Так, в 2011 году общее количество камер видеонаблюдения в Великобритании оценивалось в два миллиона единиц – по одной на каждые тридцать жителей страны. Если эта пропорция верна для всего остального мира, получается цифра примерно в 100 миллионов камер круглосуточного наблюдения, установленных в общественных местах. Впрочем, это всего лишь десятая часть миллиарда камер в смартфонах». На сегодняшний день в Москве (данные на январь 2021 г.) насчитывается более 204 000 камер видеонаблюдения по всему городу, более 102 900 камер на подъездах, более 21 000 на придворовых территориях, более 6 000 камер установлено в местах массового скопления граждан [8].

Фотоизображения могут представлять собой «большие данные» при использовании фотогалерей, фотобанков, фотоархивов, то есть сбор и систематизация разрозненных файлов с целью выявления каких-то взаимосвязей или тенденций. Сегодня более 2,5 триллиона изображений ежегодно публикуются или хранятся в Интернете [9].

Научные исследования в категории «больших данных» выступают для журналистов отправной точкой для собственных исследований или расследований [10].

Заключение.

В целом, роль «больших данных» в журналистике сегодня трудно переоценить. Прежде всего, это важный источник для поиска тем, которые в таком массиве информации могли остаться незамеченными, это возможность выявления и доказательства злоупотреблений, нарушения закона, совершения преступления. Кроме того, Big Data является не только источником информации, но и доказательной базой. Используя в качестве аргументов для своих тезисов количественные, статистические показатели, журналист повышает уровень лояльности аудитории, степень доверия к публикации. Анализ данных, выявление взаимосвязей, корреляций, составление прогнозов, рейтингов позволяет создать эксклюзивный привлекательный и достоверный контент, привлекающий аудиторию и способствующий улучшению репутации издания.

Список литературы

[1] Вайгенд, А. Big Data. Вся технология в одной книге / Андреас Вайгенд. – М.: Бомбора, 2018. – 384 с.

[2] Гусева, А. А. «Большие данные»: понятие, источники, возможности // Master’s Journal. – Пермь: Пермский национальный исследовательский политехнический университет. – 2016. – № 1. – С. 320-324

[3] Нестеренков, С.Н. Применение больших данных в электронном образовании / С.Н. Нестеренков, М.И. Макаров, Н.В. Ющенко, А.Д. Радкевич // BIG DATA and Advanced Analytics = BIG DATA и анализ высокого уровня: сб. материалов V Междунар. науч.-практ. конф. (Республика Беларусь, Минск, 13-14 марта 2019 года). В 2 ч. Ч. 2 / редкол.: В. А. Богуш [и др.]. - Минск: БГУИР, 2019. - С. 242-245.

[4] Бабурин В. А., Яненко М. Е. (2018) Технологии Big Data в сервисе: новые рынки, возможности и проблемы // ТТПС. № 1 (27). С. 100–105.

[5] Исаев Е. А., Корнилов В. В. (2019) Проблема обработки и хранения больших объемов научных данных и подходы к ее решению // Математическая биология и биоинформатика. Т. 8. № 1. С. 49–65

[6] Мультимедийная журналистика / под общ. ред. А. Г. Качкаевой, С. А. Шомовой. – М.: Изд. дом Высшей школы экономики, 2017. – 413 с.

[7] Харин Ю. С. Теория вероятностей, математическая и прикладная статистика: учебник / Ю. С. Харин, Н. М. Зуев, Е. Е. Жук. - Минск: БГУ, 2011. – 463 с.

[8] Митрович, С. Рынок «больших данных» и их инструментов: тенденции и перспективы в России // МИР (Модернизации. Инновации. Развитие). – 2018. – Т. 9. – № 1. – С. 74-85.

[9] Daniell, P.J., Discussion on the paper by M. S. Bartlett "On the theoretical specification and sampling properties of autocorrelated time-series". – Suppl. J. R. Stat. Soc. 8(1), 1946. P. 88–90.

[10] Михнев И.П. Технологии Big Data и их применение в сфере современного высшего образования / И.П. Михнев, А.Д. Челнокова, А.Д. Реут // Развитие современного образования: от теории к практике: Материалы IV Междунар. науч.-практ. конф. (Чебоксары, 19 март 2018 г.) / Редкол.: О.Н. Широков [и др.]. – Чебоксары: ЦНС «Интерактив плюс», 2018.

THE SPECIFICS OF WORKING WITH "BIG DATA" IN MODERN MEDIA

S.S. KURBANOV
Student of Belarusian State University of Informatics and Radioelectronics

S.N. NESTERENKOV
PhD, Associate Professor
Dean of Faculty of Computer Systems and Networks

A.N. MARKOV
Senior lecturer of the department, Deputy head of the Center for Informatization and Innovative Developments

Center for Informatization and Development of the Belarusian University of State Informatics and Radioelectronics, Republic of Belarus.
Belarusian State University of Informatics and Radioelectronics, Republic of Belarus.
E-mail: seyitjan0306@gmail.com, s.nesterenkov@bsuir.by, a.n.markov@bsuir.by

Abstract. As of today, Big Data is the source of information that journalists can no longer neglect. Data sets, databases, unstructured data on the official websites of the government or commercial institutions are a resource for the work of journalists. Big Data is not only the source of information, but also the evidence base. Using quantitative and statistical indicators as the arguments for their theses, the journalists increase the level of audience loyalty and trust to the publication. Data analysis, establishment of correlations, making forecasts and ratings allows creative exclusive, attractive and reliable content that attracts the audience and improves reputation of the publisher. This article is dedicated to determination of the role of "big data" in the work of information and analytical departments of the publisher, as well as in coverage of the socially relevant topics and outlining trends in further work of the journalists with such information. In conclusion, the author notes that the collection and processing of the data

Keywords: research, visualization, fact-checking, social networks, massmedia, databases, data journalism, Big Data, digitalization, verification