ТЕЛЕВИДЕНИЕ И РАДИО

Вартанов Сергей Александрович, кандидат физико-математических наук, доцент кафедры эконометрики и математических методов экономики Московской школы экономики МГУ имени М. В. Ломоносова; e-mail: sergvart@gmail.com

ТЕЛЕВИЗИОННЫЕ ИЗМЕРЕНИЯ В ЭПОХУ BIG DATA: КОНЦЕПЦИИ И ПРИМЕРЫ

В последнее время одним из актуальных трендов в маркетинговых и медиаисследованиях является все более широкое использование и обработка так называемых «больших данных» (Big Data). В медиаиндустрии Big Data и характерные методы работы с ними находят применение не только в маркетинговых и рекламных проектах, но и в медиаизмерениях. В этом случае одним из важнейших каналов получения информации являются приставки цифрового телевидения (digital set-top boxes), обладающие возможностью передачи данных по каналам обратной связи (return path data, RPD). Использование этих данных позволяет значительно повысить точность измерений телевизионной аудитории, получаемых традиционным методом – на основе пиплметровых панелей. Обзору методов Big Data-ориентированных методов измерения телевизионной аудитории и реальных примеров их использования посвящена настоящая работа.

Ключевые слова: *большие данные, измерения телевизионной аудитории, медиаизмерения, цифровое телевидение.*

Sergei A. Vartanov, PhD in Physics and Mathematics, Associate Professor at the Chair of Econometrics and Mathematical Methods in Economics, Moscow School of Economics, Lomonosov Moscow State University; e-mail: sergvart@gmail.com

TELEVISION MEASUREMENTS IN THE AGE OF BIG DATA: CONCEPTS AND EXAMPLES

One of the current trends in marketing and media research is the growing intensity of Big Data use. In the media industry, Big Data can be used not only for advertising and market research but also for audience measurements. The most important type of Big Data in this case is the return path data (RPD) gathered from set-top boxes (STBs) used for digital television broadcasting. The proper analysis of these data makes it possible to improve the accuracy of the traditional media measurement methods based on peoplemeter panels. This article provides a review of Big Data-oriented methods of television audience measurements and some real-life examples of their implementation.

Key words: *Big Data, television audience measurements, media measurements, digital television.*

Введение

Возрастающее проникновение в нашу повседневную жизнь различных гаджетов, растущая цифровизация всего нашего быта приводят к тому, что практически вся наша деятельность может быть тем или иным образом зафиксирована и сохранена в виде логов тех устройств, с которыми мы взаимодействуем. Речь здесь идет не только о традиционных журналах посещений сайтов в Интернете, которые ведутся любым браузером, или о cookies, по которым можно восстановить почти всю информацию о деятельности пользователя в Сети. Появление компьютеризированной бытовой техники, способной подключаться к Интернету и записывать диагностическую информацию о любых манипуляциях с ней, появление connected cars, внедрение GPS и ГЛОНАСС едва ли не в каждый смартфон – все это позволяет ежедневно создавать гигантский массив информации о поведении практически любого человека. И более того, даже общение с близкими и коллегами превращается в постоянный поток данных, которые при правильно организованном процессе сбора и обработки позволяют делать необходимые выводы о любом человеке.

По оценкам IBM, в 2012 г. в мире ежедневно создавалось более 2,5 эксабайтов (примерно 2,5 миллиарда гигабайт) новой информации. Сегодня объем ежегодно фиксируемой информации еще выше, ведь количество данных, фиксируемых на Земле, ежегодно удваивается. Все описанные ранее виды данных и представляют собой Big Data в широком смысле (автором термина является Клиффорд Линч, редактор журнала Nature, предложивший его в 2008 г. (Lynch, 2008)).

Все более широкое использование и обработка «больших данных» (калька с big data, прижившаяся в литературе) стали в последнее время одними из наиболее актуальных трендов в маркетинговых и медийных исследованиях. Само по себе понятие «больших данных» в разных источниках определяется по-разному, однако все определения согласованы с концепцией трех V: volume, velocity, variety (объем, скорость, разнообразие). В отдельных источниках к ним добавляется и четвертая «V» – veracity (достоверность). В широком смысле, Big Data – это разнородные неструктурированные данные крайне большого объема, увеличение которого происходит ежедневно с большой скоростью.

Очевидно, что с точки зрения социологических и медиаисследований задачи анализа и извлечения таких данных представляют

огромный интерес. Кроме того, они могут быть использованы в рамках любых коммуникационных кампаний на любом этапе: как – стратегически – при анализе рыночной ситуации, так и на этапе оценки эффективности (Назаров, Виговская, 2013). Отдельно следует упомянуть таргетирование рекламных обращений и профилирование наиболее важной для рекламодателя аудитории, являющиеся сегодня, наверное, одними из самых востребованных сфер использования «больших данных».

Профилирование аудитории на основе больших данных в том или ином виде применяется практически во всех проектах, связанных не только с маркетинговыми и рекламными исследованиями, но и с медиаизмерениями. И здесь одним из важнейших каналов получения информации являются приставки цифрового телевидения (digital set-top boxes, DSTB или просто STB), обладающие возможностью передачи данных по каналам обратной связи (return path data, RPD). Использование этих данных позволяет значительно повысить точность измерений телевизионной аудитории, получаемых традиционным методом – на основе пиплметровых панелей.

Big Data и традиционные методы ТВ-измерений: синергия вместо взаимоисключения

Литературу, посвященную использованию Big Data, родственных концепций и методов работы, можно разделить на два направления. Первое – описание понятия «больших данных» в контексте ТВ-индустрии и применение характерных для нее методов и алгоритмов извлечения информации к данным, полученным от ресиверов цифрового телевидения (set-top boxes, STB) по каналам обратной связи (return path data, RPD). Второе направление затрагивает вопрос обработки уже подготовленных данных о телесмотрении на каждом отдельно взятом ресивере для определения профилей телесмотрения домохозяйств и отдельных зрителей (Chang, Kauffman, Son, 2012)[1].

Ряд публикаций, принадлежащих к первому направлению (Вартанов, 2014; Каменская, 2011; Shababb, Taylor, 2005)[2], посвящен сравнению возможностей, которые предоставляют данные с ресиверов и данные, полученные более привычным путем – с помощью пиплметров. В этом плане как российские авторы, так и зарубежные придерживаются единой точки зрения. С одной сто-

роны, данные с ресивера обладают большей точностью (длительности сессий можно отслеживать с точностью до секунды) и могут быть доступны мгновенно, в отличие от данных с пиплметров, которые становятся доступны в лучшем случае на следующий день и характеризуют смотрение поминутно. Кроме того, данные с ресивера дают информацию об использовании дополнительных опций при просмотре (запись, два экрана и т.д.), взаимодействии с интерактивными сервисами оператора, дополнительных сервисах (например, VOD). С другой стороны, пиплметровые данные точно характеризуют зрителей и домохозяйства, в то время как ресиверы неспособны фиксировать какие-либо данные, кроме тех, которые связаны с их использованием.

При всем кажущемся различии исходных данных и способов их обработки, противопоставление двух методик исследования аудитории – панельной с помощью пиплметров и обработки Big Data, получаемой с приставки, – было бы неэффективным. Эти два подхода должны применяться вместе и взаимно дополнять друг друга.

Помимо основной цели – измерения и анализа телеаудитории – важной прикладной задачей, которая решается с помощью данных, получаемых с STB, является таргетирование рекламы. Для этого большая часть источников предлагает так или иначе интегрировать базы данных, содержащие информацию о покупках и покупателях, с базами данных с STB. На примитивном уровне подобная интеграция выглядит следующим образом. Заинтересованные субъекты (в первую очередь, рекламодатели в лице производителей и продавцов) на основе данных о совершенных покупках их товара создают профили покупателей. Аналогичные профили создаются на основе информации о телесмотрении для каждого STB и каждой сессии смотрения на нем. Таким образом, становится возможным показывать в каждую сессию телесмотрения рекламу тех товаров, профиль покупателей которых наиболее близок профилю текущей сессии. Такое таргетирование предполагается проводить, минуя стадию определения типа домохозяйства и его основных характеристик. Кроме того, профили сессий не обязательно должны соответствовать только одному зрителю, это могут быть агрегированные профили совместного смотрения двух или более членов домохозяйства.

Определенные работы (Chang, Kauffman, Son, 2012) проливают свет и на промежуточную задачу профилирования зрителей: по характеристикам телесмотрения по данным с ресивера определить характеристики домохозяйства, использующего его, а также по-

строить прогнозные профили зрителей на каждую сессию. Таким образом, в дальнейшей части настоящего обзора предполагается описать предлагаемые в литературе алгоритмы решения двух смежных задач. С одной стороны, это задача выделения отдельных сессий телесмотрения на каждом ресивере и построение для этих сессий демографических профилей. В качестве подзадачи сюда входит определение среди всех сессий тех из них, которые соответствуют одному типу смотрения (например, смотрению ТВ одним и тем же человеком или одной и той же группой). С другой стороны, для профиля каждой сессии необходимо подобрать наиболее подходящий ей товар с точки зрения историй продаж, и наоборот.

Return path data: основной тип больших данных в телеизмерениях

Большая часть современных приставок цифрового ТВ обладает возможностью фиксации с точностью до миллисекунд всех событий, происходящих с ней, – связанных как с активностью зрителя (смена канала, изменение громкости, и т.д.), так и с внутренними событиями приставки (смена телепередачи на одном и том же канале, системные события). Таким образом, данные обратной связи, получаемые с приставок, довольно разнородны. Если учесть, что данные о событиях STB снимаются с миллисекундной точностью, а диагностическая информация с каждого STB поступает с определенной регулярностью (как правило, не реже, чем раз в пять минут), то при обработке RPD приходится иметь дело с огромным объемом данных. Это позволяет отнести задачи, связанные с обработкой данных с STB, к Big Data-задачам.

Опишем подробно, что может регистрироваться приставками, кроме служебной информации, например качество связи и техническое состояние устройства. В первую очередь, записывается и сохраняется вся информация о нажатии кнопок пульта управления приставкой и также кнопок на самой приставке, если они есть (clickstream). Помимо событий низкого уровня вроде нажатий кнопок пульта, отслеживаются и события более высокого уровня. Во-первых, фиксируются моменты включения и выключения приставки. Как правило, моменты включения и выключения телевизора не могут фиксироваться приставкой за редкими исключениями[3]. Во-вторых, фиксируется просмотр телеканалов в прямом

эфире: номера этих каналов, моменты переключения, передачи, которые шли на этих каналах во время просмотра.

Важной особенностью цифровых приставок является то, что они фиксируют не только линейный просмотр телеканалов, но и отложенный просмотр, и просмотр по запросу. Так, среди данных RPD, как правило, фиксируется просмотр time-shift телепередач с жесткого диска приставки или сервера, а также постановка вещания на паузу и последующее возобновление просмотра. Кроме того, многие STB предоставляют данные о просмотре видео по запросу (video-on-demand, VOD), как со встроенного жесткого диска, так и из Сети с помощью приложений интерактивного ТВ. Также в логах приставок фиксируются другие способы их использования, например прослушивание радио, просмотр интернет-сайтов через приложения STB, караоке и т.п. Кроме того, некоторые приставки сохраняют данные о взаимодействии пользователя с меню, навигации по электронной программе передач (EPG), о записи телепередач с эфира, о просмотре в режиме «картинка в картинке» и т.д.

Поскольку данные, получаемые от приставок цифрового ТВ, могут быть собраны со всех абонентов, преимущества их использования в качестве основы для системы медиаизмерений очевидны (естественно, при условии их правильного форматирования, очистки и обработки). Во-первых, размеры потенциальной выборки во много раз превышают размеры любой из существующих панелей, используемых при измерении ТВ-аудитории. Более того, в определенных случаях речь может идти о сплошных данных (census level data). Во-вторых, данные STB обладают несравнимо более высоким уровнем детализации, так как позволяют отслеживать всю зрительскую активность с точностью до секунды.

«Обратной стороной» всех описанных выше плюсов данных STB является тот факт, что они характеризуют использование только самого STB и представляют собой в лучшем случае телесмотрение, измеренное «в приставках». Иными словами, если оператор цифрового ТВ зафиксировал, что из 1 000 установленных приставок 500 были включены на канале N, то из этого не следует, что этот канал смотрел каждый второй зритель или даже каждая вторая семья. Приставка не в состоянии зафиксировать не только кто именно смотрел телевизор в каждый момент времени, но и то, сколько всего человек было перед экраном. Более того, ситуация, когда зритель выключил ТВ и перестал смотреть что-либо, но за-

был выключить приставку, будет воспринята как продолжающаяся сессия смотрения последнего включенного канала. Также данные с STB не могут учитывать другую аппаратуру, подключенную к телевизору (DVD-плеер, видеомагнитофон), как и наличие других телевизоров в домохозяйстве и характеристики их смотрения. Таким образом, RPD в исходном виде не дает возможности судить об аудиторных показателях в традиционном смысле.

Кэппинг как основной метод очистки и подготовки данных обратной связи

Чтобы привести данные STB в вид, пригодный для определения различных аудиторных показателей, надо решить последовательно две задачи. Во-первых, необходимо превратить clickstream-данные в статистику включений телевизора. Иными словами, необходимо избавиться от ситуации, когда ресивер был включен, но телевизор был выключен. Во-вторых, по уже очищенным от подобных ситуаций данным необходимо получить оценку характеристик аудитории: кто именно был перед телевизором, сколько было зрителей и каковы их социально-демографические показатели.

Решение первой задачи связано с выделением сессий – промежутков времени, в течение которых телесмотрение имело постоянный с точки зрения какого-либо параметра характер. В частности, интервал, в течение которого приставка была непрерывно включена, является сессией включения. В свою очередь период, в течение которого на приставке был выбран определенный канал, – сессия смотрения этого канала. Зная время и длительность всех сессий смотрения, можно рассчитать все требуемые статистические показатели аудитории ТВ.

Согласно данным различных пиплметровых панелей (например, британской BARB или российской ТВ-панели TNS), бóльшая часть сессий имеет небольшую длину. Так, Шабабб и Тейлор (Shababb, Taylor, 2005) приводят данные, полученные на основе наблюдения за 500 домохозяйствами панели BARB, оснащенными цифровыми приставками. Согласно полученным ими результатам, среди всех сессий, на протяжении которых на телевизоре был включен один и тот же канал (сессия типа «ТВ-канал»), доля длинных сессий (дольше 60 минут) не превышает 5%.

Из приведенной статистики следует самый простой и в то же время самый распространенный способ борьбы с «фиктивным»

телесмотрением – это обрезка длинных сессий (capping, кэппинг). Идея метода крайне проста: определяется пороговое значение максимальной длительности сессии смотрения, после чего все сессии, длина которых превышает данное значение, «обрезаются» до этой длительности. На рисунке 1 приведен результат применения к данным STB из панели BARB простейшего кэппинга с порогом в 45 минут. Все сессии смотрения длиной более 45 минут сокращались, и вместо момента окончания, зафиксированного STB, фиксировался момент через 45 минут после начала сессии. Подобная процедура позволила в значительной степени приблизить значение HUT, полученное на основе RPD, к реальному значению HUT в тот же момент времени (см. рис. 1).

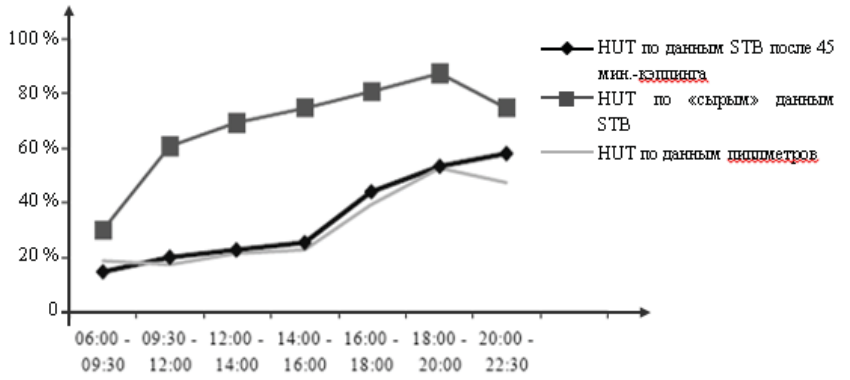


Рисунок 1. Доля домохозяйств с включенным ТВ (HUT) согласно «сырым» STB-данным, данным пиплметров и STB-данным после кэппинга с порогом 45 мин.

Следующим шагом после проведения кэппинга является уже профилирование данных, получаемых с STB, с учетом уже имеющихся результатов исследований. Под профилированием данных с STB понимаются две операции. Во-первых, это социально-демографическое профилирование домохозяйства, которому принадлежит приставка цифрового ТВ. Очевидно, что большая часть владельцев приставок не являются участниками проектов по измерению медиапотребления (дневниковых, пиплметровых), и нам apriori ничего не известно об их социально-демографических характеристиках.

Во-вторых, даже после профилирования домохозяйства остается открытым вопрос о том, кто находится перед телевизором в мо-

менты включения ресивера. Члены домохозяйства могут смотреть телевизор в любых комбинациях, и техническая возможность определения зрительского состава отсутствует. Конечно, некоторую информацию можно извлечь из RPD от ресиверов и без особой процедуры профилирования. Например, оценить в первом приближении общую аудиторию интересующих нас эфирных событий. Конечно, не зная точного количества зрителей, нельзя судить о точном количестве контактов, однако возможно провести оценку на уровне домохозяйств (HUT).

Алгоритмам кэппинга и методам борьбы с false positives-данными посвящена также работа Peng Wang "Modeling TV Rating by Channel Switching in the Set-Top Box"[4]. Основная задача работы ставится следующим образом. Имеется набор данных, описывающий владельцев STB, являющихся также членами панели NPM. Возможно ли, обладая лишь данными пиплметров, построить математическую модель, определяющую рейтинги эфирных событий на основе только данных от STB (т.е. корректно определять интервалы реального телесмотрения)? И можно ли применить такую модель не только к данным панельных STB, но и ко всей генеральной совокупности – всем домохозяйствам США с STB?

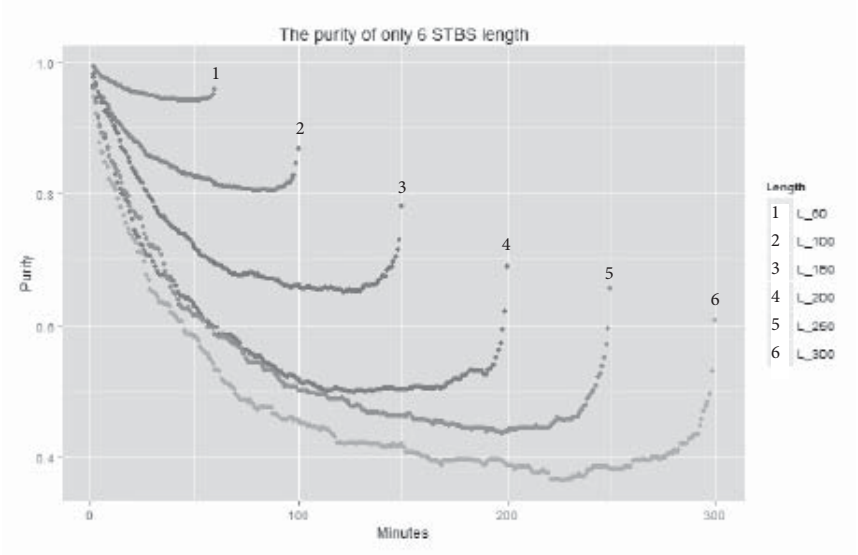


Рисунок 2. Длительность непрерывной сессии просмотра и доля реального телесмотрения

Рассматриваемый в этой работе подход к кэппингу основан на анализе сессий включения ресивера – интервалов между его включениями и выключениями. В каждый момент такой сессии вероятность «обнаружить» зрителя перед телевизором зависит от двух параметров: времени, которое прошло с момента включения ресивера, и общей длительности сессии. На рисунке 2 приведены 6 совокупностей точек на плоскости «Время работы STB» vs «Вероятность реального смотрения» (purity). Совокупность точек одного и того же цвета соответствует одной сессии телесмотрения (цвет соответствует длительности сессии работы STB). Координата **x** показывает время, прошедшее от включения STB, а координата **y** – долю раз, когда на соответствующей минуте фиксировалось реальное телесмотрение (респондент реально присутствовал у ТВ). Например, розовая точка (130; 0.42) означает, что среди всех ситуаций в панели, когда STB был непрерывно включен в течение 200 минут, только в 42% случаев на 130 минуте включенному STB соответствовало реальное телесмотрение.

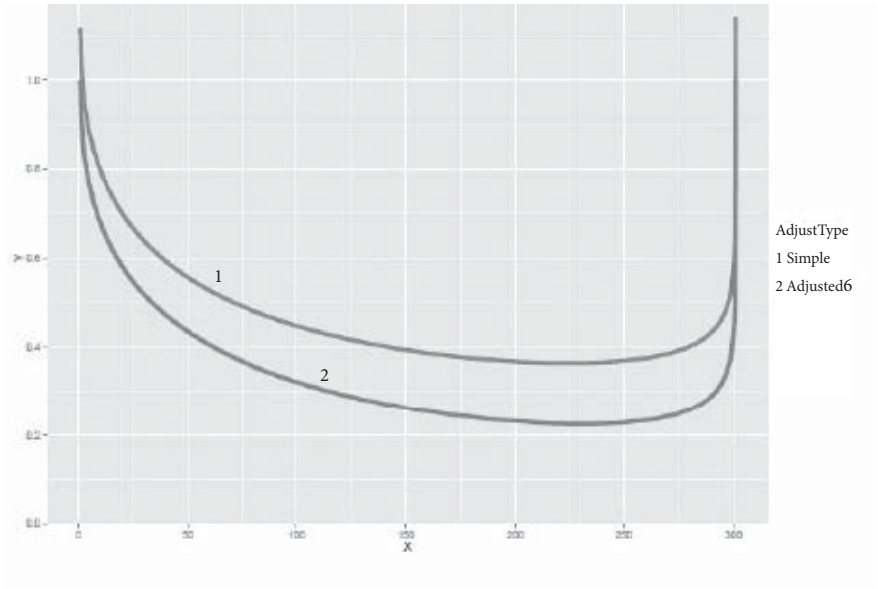


Рисунок 3. Графики простого и скорректированного двойного распределений Вейбулла

Как доказано в обсуждаемой работе, форма эмпирических распределений зрительского внимания с высокой точностью соответствует форме экспоненциального вейбулловского распределения. Примерный график такого распределения приведен на рисунке 3. Для того, чтобы это распределение можно было использовать для описания интересующей нас характеристики, проводилась его нормализация: значения purity в моменты включения и выключения STB должны равняться 1.

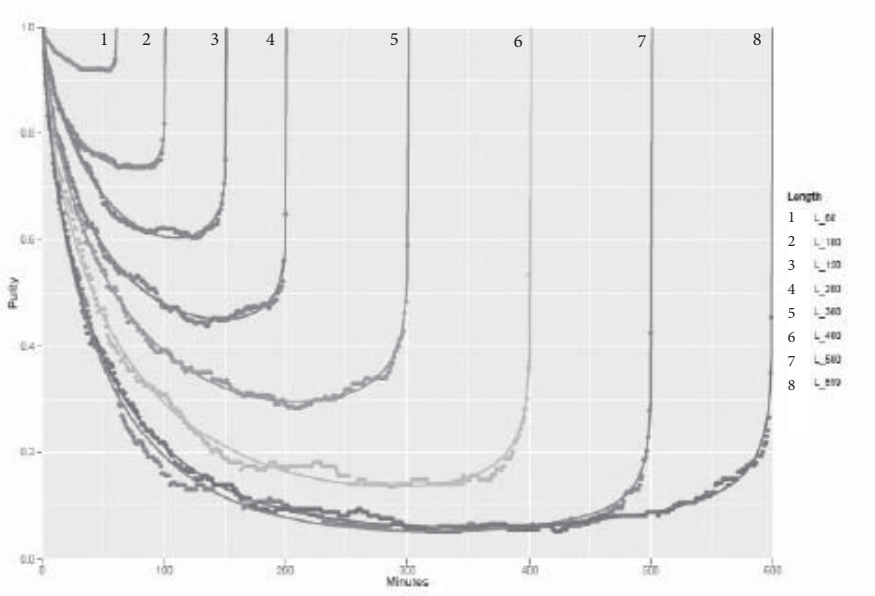


Рисунок 4. Аппроксимация зависимости реального смотрения от длительности сессий с помощью скорректированного экспоненциального распределения Вейбулла

Для любой длительности сессии работы STB на основе пиплметровых данных можно построить соответствующее нормализованное экспоненциальное распределение Вейбулла, показывающее вероятность реального телесмотрения (см. рис. 4). После этого рейтинг каждого эфирного события в рамках сессии включения определяется довольно просто. Считается, что зритель присутствовал перед ТВ, если вероятность реального смотрения на соответствующем STB в момент эфирного события больше 0.5. В противном случае считается, что перед ТВ никого не было, и соответствующего контакта события с представителем аудитории не

было. Дополнительный результат, который приводится в работе, заключается в том, что разработанная двойная Вейбулловская модель оказалась более точной с точки зрения определения рейтингов, чем данные панели NPM (Nielsen Peoplemeter Panel).

Подход Nielsen: Data fusion и методы моделирования телезрителей

О том, как принято работать с данными, получаемыми с ресиверов, можно судить по отчетам и пресс-релизам крупнейших медиаметрических компаний мира. В первую очередь это Nielsen, TNS, Rentrak и Kantar Media. Именно эти компании в последнее время наиболее активно развивают направление STB-исследований. Причем к данным есть два подхода: интеграция с другими базами данных, описывающими те же домохозяйства (например, данные переписей, данные покупок и т.д.), и моделирование отсутствующих данных на основе различных панельных исследований.

Первый подход – data fusion, интеграция баз данных, описывающих одни и те же домохозяйства, но отражающих характеристики, описание которых получено в результате разноплановых исследований. Прежде всего речь идет об интеграции данных государственных органов, ответственных за статистику, и данных о покупках и транзакциях, совершаемых членами домохозяйства. В этом случае данные часто объединяются на основе адресов и ZIP-кодов или демографических характеристик (если они присутствуют).

Как поясняется в брошюре “Introduction to Nielsen Data Fusion”[5], задача интеграции заключается в том, чтобы создать взаимно-однозначное соответствие между элементами первой и второй баз данных исследований на основе общих характеристик. Под ними в Nielsen понимают географические и демографические данные, данные о медиапотреблении и т.д. В брошюре приводится пример подобной интеграции между данными ТВ-панели Nielsen peoplemeter и данными о покупках.

В патенте Питера Доу[6] (Peter Campbell Doe, Nielsen SVP по интеграции данных) описываются методы и алгоритмы, приписывающие всем владельцам цифровой приставки, подлежащим мониторингу, демографические характеристики на основе данных по участникам панели. На вход подобные алгоритмы получают две базы данных, представляющих информацию о телесмотрении двух неравных групп домохозяйств. Меньшая из этих групп представ-

ляет собой панельные домохозяйства, о членах которых имеется вся необходимая демографическая информация. Вторая группа представляет собой непанельные домохозяйства, оснащенные только приставкой, и никакой информации о ее демографии (количество членов семьи, их пол, возраст, доход и т.д.) нет.

Ресиверы цифрового телевидения, установленные в домохозяйстве, имеют возможность отслеживать зрительское поведение и привычки его членов. Тем не менее медиаисследователи и участники рынка не имеют права получать персональные демографические данные членов домохозяйства, за исключением случаев, когда они дают явное и осознанное согласие на это. Таким образом, возможно отслеживать зрительское поведение, «очищенное» от всякой персональной информации: все подобные данные не собираются ресивером, хотя и привязаны к его серийному номеру. В то же время, такие параметры, как просматриваемые каналы, время их просмотра, изменение громкости, включение и выключение ресивера, как уже упоминалось ранее, отслеживаются и по обратному каналу передачи данных возвращаются провайдеру (return path data).

Использование домохозяйств, выбранных с помощью статистических и социологических методов построения панелей, позволяет медиаисследователям наблюдать за привычками телесмотрения интересующих групп. Участвующие в панели домохозяйства могут использовать специальную аппаратуру для фиксации просматриваемых каналов, времени просмотра и прочих характеристик зрительского поведения, в частности, в дополнение к ресиверу они могут быть оборудованы пиплметрами. Комбинация зрительского поведения и демографических характеристик в правильно спроектированной панели позволяет экстраполировать наблюдаемые в ней явления (тренды и шаблоны телесмотрения, эффективность рекламы) на всю генеральную совокупность.

В основе предлагаемой Питером Доу системы – множество всех домохозяйств с STB, разделенное на две части. Первая часть – это панельные домохозяйства, участники которых дали разрешение на манипуляцию их персональными (социально-демографическими, географическими) данными. Вторая часть состоит из домохозяйств, оснащенных цифровой ТВ-приставкой, но не имеющих пиплметра.

Данные, получаемые от непанельных STB, включают в себя дату и время просмотра каждого телеканала, статус STB (вкл/выкл), информацию о переключении каналов и изменении громкости. Как правило, эти данные включают в себя также серийный

номер устройства, однако подобная информация не сохраняется и не учитывается еще на этапе получения данных. В то же время члены панельных домохозяйств раскрывают свой возраст, количество детей, доход, образование, профессию, место жительства, индекс и т.д. Вся эта информация добавляется к данным от ресиверов, установленных в этих домохозяйствах.

Основной частью описываемой системы является процедура моделирования данных. Эта процедура реализует многоэтапный алгоритм генерации данных о зрителях и вероятностей их нахождения перед телевизором. При этом используются как данные из пиплметровой панели (в первую очередь, социально-демографические), так и данные, полученные от STB (например, из соответствующей базы данных).

Процедура моделирования данных включает в себя: удаление ошибочных и незначимых сессий, приписывание отсутствующих характеристик и определение вероятностей нахождения телезрителей перед экраном. Процедуры удаления, приписывания характеристик и определения вероятностей используют данные о телесмотрении панельных и непанельных домохозяйств, в то время как менеджер отчетов оперирует данными, полученными только от процедуры определения вероятностей, что является средством организации уже обработанных данных.

Кратко опишем каждый из компонентов процедуры моделирования данных. Процедура кэппинга представляет собой применение одного или нескольких правил, позволяющих по их результатам применения удалить часть сессии телесмотрения (или всю сессию целиком). Правила удаления и сохранения сессий, основанные на показателе удаления, опираются на эмпирические наблюдения, получаемые с помощью пиплметров. Другие факторы, которые также учитываются при определении показателей удаления и порогов длительности сессий, включают в себя: время года, тип медиаконтента и так далее.

В случае, если длительность сессии превышает пороговое значение, то данные о ней редактируются. Помимо метода полного кэппинга, приведенного в качестве примера ранее, используются также методы частичного кэппинга. Метод полного обрезания сессий обладает одной неприятной особенностью, которая заведомо «огрубляет» результат обработки данных. Речь идет о том, что аномально долгие сессии обрезаются целиком. В то же время представляется очевидным, что в последние минуты любой сес-

сии, предшествующие ее окончанию, имеет место присутствие зрителя перед экраном (хотя бы для того, чтобы выключить STB или переключить канал). На этом предположении основаны методы частичного кэппинга, главная идея которых заключается в том, что вместо обрезки всего «хвоста» сессии вырезается только часть ее времени из середины. Таким образом, одна большая сессия после применения этого метода превращается в две более коротких, при этом начало первой из них и конец второй совпадают с началом и концом исходной сессии.

Следующим шагом является приписывание отсутствующих характеристик непанельным STB. Для интеграции данных STB и панелей домохозяйств, оборудованных пиплметрами, применяются методы data fusion (DF). Фактически DF – это процесс, объединяющий две базы данных на уровне их элементов на основе их близости по определенному набору общих переменных этих баз. В частности, применительно к рассматриваемому случаю соединения данных STB и пиплметров объединение может происходить на основе паттернов телесмотрения, включающих в себя любые проявления телезрительской активности. Происходит это следующим образом. Для каждого ресивера из числа непанельных проводится анализ особенностей его использования, в результате которого выделяются несколько основных паттернов связанного с ним телесмотрения. Среди панельных домохозяйств находится то, для которого на предыдущем этапе в рамках аналогичной процедуры анализа телесмотрения был выделен набор паттернов, наиболее близкий к полученному с проанализированного непанельного ресивера. Далее социально-демографические характеристики этого панельного домохозяйства переносятся в базу данных непанельных ресиверов и приписываются домохозяйству, использующему рассматриваемый ресивер. В ходе интеграции данных такие характеристики, как национальность, родной язык, расовая принадлежность, приписываются одновременно, что сохраняет корреляцию между ними и позволяет избежать несоответствия данных в объединенной базе.

DF дает возможность рассматривать одновременно произвольное количество переменных, что позволяет нам далее говорить обобщенно о «паттернах смотрения» и «демографических характеристиках». Наличие взаимосвязи между этими группами характеристик очевидно: паттерны телесмотрения могут хорошо предсказывать демографию, а демография – паттерны. Тем не менее традиционные реализации процессов DF в других сферах (напри-

мер, проект объединения Nielsen NPM + MRI – пиплметровые данные и данные о читательской активности) обычно используют демографические характеристики для определения зрительского поведения групп зрителей и отдельных индивидов. В случае телесмотрения процесс DF происходит в обратном порядке: методы, описываемые в работе Питера Доу, приписывают демографические характеристики данным о зрительском поведении.

Уже подготовленные и очищенные данные о зрительском поведении обрабатываются с целью распознать (или сформировать вновь) в них хотя бы один из заранее определенных паттернов зрительского поведения. Полученный паттерн затем сравнивается с паттернами поведения, хранящимися в базе данных. Они могут быть распознаны на основе библиотеки, которая содержит некоторое количество «шаблонных» паттернов, заранее определенных пользователем (программистом, социологом) на основе данных, получаемых с пиплметров. Извлеченные из данных STB и распознанные паттерны зрительского поведения поступают на вход в подпрограмму-классификатор типов смотрения участников пиплметровой панели. Эта подпрограмма просматривает базу данных пиплметров в поисках близких к полученным из данных STB паттернов, наблюдавшихся в панельных домохозяйствах. Когда такие домохозяйства находятся, их характеристики (в первую очередь, социально-демографические) передаются в подпрограмму интеграции данных. Эта подпрограмма объединяет обрабатываемые базы, находя общие переменные. Как правило, набор таких переменных включает в себя количество телевизоров в домохозяйстве, общее время просмотра ТВ, время просмотра конкретных каналов, программ конкретных жанров, а также время просмотра, приходящееся на отдельные интервалы дня.

Совпадения по переменным связи позволяют приписывать непанельным домохозяйствам соответствующие характеристики. На первой итерации процесса интеграции данных определяется тип домохозяйства, которому принадлежит соответствующий STB. Тем не менее более точной информации о том, кем были смотревшие каждую передачу в сессии зрители, на этом этапе нет. Для того, чтобы это определить, производится следующая итерация, позволяющая приписывать каждой сессии все более детальные характеристики. При этом те характеристики, которые на предыдущей итерации приписывались, на последующей воспринимаются уже как переменные связи.

Приписанные каждому STB характеристики домохозяйства и приписанные каждой сессии характеристики зрителей сохраняются в соответствующей базе данных и передаются на вход процедуры определения вероятностей смотрения. На основе данных о сессиях смотрения определяются дни и части дня, когда эти сессии имели место. При этом каждый интервал времени связан с информацией о зрительском поведении, основанной на интегрированных данных. Для каждого домохозяйства рассчитываются средние значения количества зрителей в разные дни недели, а также средние значения количества зрителей в каждое время суток (утром, днем, в прайм-тайм и т.д.). Пример: на основе пиплметровых данных калькулятор определяет, что каждый будний день с 16.00 до 18.00 в домохозяйстве рассматриваемого типа включены два телевизора, размер этого домохозяйства – 3 человека и в среднем на 1 ТВ приходится по 1,8 зрителя.

На основе этих значений, а также на основе приписанных данных о социально-демографических характеристиках предполагаемых зрителей рассчитываются вероятности смотрения ТВ с помощью данного ресивера зрителями в зависимости от их пола, возраста, жанра передачи и времени суток. После завершения расчета вероятностей смотрения для всех предполагаемых членов домохозяйства полученные значения умножаются на поправочный коэффициент, равный отношению среднего числа зрителей данного ТВ в каждый промежуток времени к сумме всех вероятностей для этого промежутка времени. Уточненные значения вероятностей смотрения, наряду с приписанными характеристиками домохозяйств и социально-демографическими характеристиками предполагаемых зрителей, дальше могут использоваться для построения отчетов и являются финальным результатом работы всей подпрограммы восстановления данных STB.

Работа Питера Доу, в которой описан метод интеграции данных от приставок цифрового ТВ и пиплметровых панелей, является одной из немногих опубликованных в открытом доступе, которая бы содержала описание всей методики слияния данных зрительского поведения панельных и непанельных домохозяйств. Другие работы того же автора (написанные как индивидуально, так и в соавторстве) представляют собой различного рода презентации и пресс-релизы, описывающие результаты моделирования показателей STB с помощью данных панелей. Очевидно, что эти результаты основаны на моделях и методах, родственных приведенным в

указанном патенте. Так, в докладе Patrick Dineen & Pete Doe "Hybrid measurement goes local"[7] интересующий нас вопрос (моделирование профилей телезрителей STB на основе данных пиплметровых панелей) рассматривается как один из примеров гибридных методов измерений. В этой работе обосновываются причины, по которым интересующий нас способ измерения аудитории является необходимым направлением развития медиаизмерений в целом. Дело в том, что с ростом фрагментации аудитории относительная ошибка растет, и повышается вероятность зафиксировать нулевой рейтинг у тех эфирных событий, которые смотрело не слишком большое количество зрителей (так называемая проблема «длинных хвостов»). Есть вероятность, что среди панельных домохозяйств вообще никто не имел контактов с этим событием, в то время как в генеральной совокупности оно набрало пусть и небольшой, но ненулевой рейтинг.

В то же время, отмечают авторы, панельные измерения обладают высокой репрезентативностью, и демография всех участников измерений исследователям известна, что приводит к низкому уровню смещения (bias). Данные STB являются сплошными (census data), однако не связаны с демографией и дают существенное смещение за счет частого появления феномена false positives – ситуаций «STB on – TV off» (рис. 5).

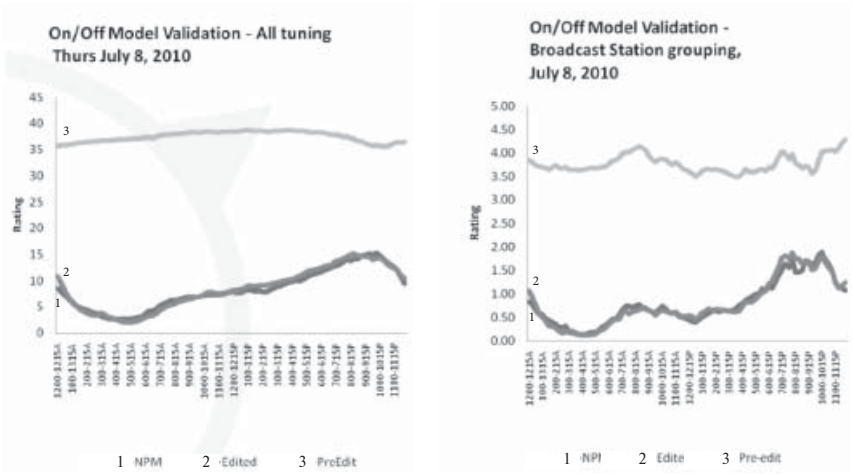


Рисунок 5. Рейтинги всего ТВ (слева) и отдельной передачи (справа), построенные по сырым данным STB (3), обработанным данным STB (2) и данным пиплметровой панели NPM (1)

Приведенные на рисунке 5 графики характеризуют точность работы алгоритма кэппинга, применяемого в Nielsen и частично отображенного в патенте Питера Доу в рамках описания процедуры удаления сессий. Из представленных графиков следует, что значения рейтингов, определяемых с помощью вышеизложенных методов, очень близки к значениям, полученным с NPM-панели. В то же время, за счет того, что аудитория STB составляет почти 2/3 всех домохозяйств США, величины абсолютной и относительной ошибок существенно ниже.

Выводы

В перспективе описанные выше методы интеграции больших данных и традиционных пиплметровых панелей позволят предоставлять результаты совместного RPD-исследования под интерфейсом привычного и понятного рынку программного обеспечения. Это, естественно, позволит рассчитывать все необходимые медиаиндустрии аудиторные показатели и строить медиапланы. Конечный продукт по своим качественным характеристикам ничем не будет уступать пиплметровым измерениям, а в чем-то может и превосходить их. Более того, если судить по опыту западных стран (в первую очередь, Великобритании и США), методы анализа данных обратной связи, описанные в настоящем обзоре и родственные им, в перспективе станут основой для новой «валюты» на рынке телерекламы. Подобная «валюта» будет лишена типичных для «старых» панельно-пиплметровых моделей недостатков, связанных с малым размером панели. В частности, это позволило бы избавиться от проблемы «длинных хвостов». Как бы ни были высоки качество построения панели и точность репрезентации ею генеральной совокупности, телесмотрение каналов с малыми рейтингами часто может не быть зафиксировано традиционными методами. При этом наличие сплошных данных обратной связи, в перспективе покрывающих всю генеральную совокупность, позволило бы получить информацию о зрителях даже тех каналов, размер аудитории которых может исчисляться лишь тысячами или даже сотнями домохозяйств. Очевидно, что с привычными размерами пиплметровых панелей (например, панель TNS в России имеет размер порядка 2500 домохозяйств) вероятность «поймать» такое телесмотрение и учесть его при составлении телевизионных рейтингов довольно невелика.

Тем не менее, несмотря на все очевидные достоинства, RPD-исследования пока не могут заменить общенациональные телевизионные измерения до тех пор, пока на рынке существует аналоговое телевещание. Поэтому на данном этапе развития медиаиндустрии России (как, впрочем, и многих других стран мира) полностью перейти на гибридные методы телевизионных измерений (пиплметры + RPD) пока невозможно. Однако использование больших данных о телесмотрении абонентов уже существующих операторов цифрового ТВ даже сейчас позволяет существенно повысить точность пиплметровых измерений.

Примечания

[1] Doe P.C. Methods and Apparatus to Model Set-Top Box Data. Patent US 20080300965 A1

[2] Roberts A.C., Wheatley M A., Wilcox P. Analyzing viewing data to estimate audience participation. Patent US 7954120 B2; Dineen P., Doe P.C. Hybrid Measurement Goes Local. Audience Measurement 6.0 Symposium. June 2011.

[3] Технически это возможно только в том случае, когда приставка подключена к телевизору по HDMI-кабелю.

[4] Wang P. Modeling TV Rating by Channel Switching in the Set-Top Box (STB). Accessed at: http://pengstats.macssa.com/download/notes/STB%20at%20Nielsen.pdf

[5] Introduction to Nielsen Data Fusion. Accessed at: http://www.nielsen.com/content/dam/corporate/us/en/docs/solutions/Nielsen-Introduction-to-Data-Fusion.pdf, The Nielsen Company, 2009.

[6] Doe P.C. Methods and Apparatus to Model Set-Top Box Data. Patent US 20080300965 A1.

[7] Dineen P., Doe P.C. Hybrid Measurement Goes Local. Audience Measurement 6.0 Symposium. June 2011.

Библиография

Вартанов С.А. Big Data в измерении телевизионной аудитории: методы очистки и обработки данных цифровых STB // Теория и практика медиарекламных исследований. Выпуск 4. Аналитический центр "Видео Интернешнл", ООО "НИПКЦ Восход-А Москва, 2014. С. 151–167.

Каменская М. Измерения аудитории цифрового телевидения // Теле-Спутник. 2011. С. 58–60.

Назаров М.М., Виговская Е.Г. «BIG DATA»: концепт и его приложения к практике маркетинговых коммуникаций // Интернет-маркетинг. 2013. № 5.

Chang R.M., Kauffman R.J., Son I. (2012) Consumer Micro-Behavior and TV Viewership Patterns: Data Analytics for the Two-Way Set-Top Box. *Proceedings of the 14th Annual International ACM Conference on Electronic Commerce*: 272–273.

Lynch C. (2008). Big Data: How do your data grow? *Nature* 455 (7209): 28–29.

Shababb G., Taylor A. (2005). From set top meters to set top boxes. Audience measurement for the digital world. In *WAM - Worldwide Audience Measurement 2005 - Cross Media*.

Notes

Dineen P., Doe P.C. Hybrid Measurement Goes Local. Audience Measurement 6.0 Symposium. June 2011.

Doe P.C. Methods and Apparatus to Model Set-Top Box Data. Patent US 20080300965 A1

Introduction to Nielsen Data Fusion. Accessed at: http://www.nielsen.com/content/dam/corporate/us/en/docs/solutions/Nielsen-Introduction-to-Data-Fusion.pdf, The Nielsen Company, 2009.

Roberts A.C., Wheatley M A., Wilcox P. Analyzing viewing data to estimate audience participation. Patent US 7954120 B2

Wang P. Modeling TV Rating by Channel Switching in the Set-Top Box (STB). Accessed at: http://pengstats.macssa.com/download/notes/STB%20at%20Nielsen.pdf

References

Chang R.M., Kauffman R.J., Son I. (2012) Consumer Micro-Behavior and TV Viewership Patterns: Data Analytics for the Two-Way Set-Top Box. *Proceedings of the 14th Annual International ACM Conference on Electronic Commerce*: 272–273.

Kamenskaya M. (2011) Izmereniya auditorii tsifrovogo televideniya [Digital Television Audience Measurements]. *Tele-Sputnik* 58-60.

Nazarov M. M., Vigovskaya E. G. (2013) «BIG DATA»: kontsept i ego prilozheniya k praktike marketingovykh kommunikatsiy ["BIG DATA": the Concept and Its Applications to the Practice of Marketing Communications]. *Internet-marketing* 5.

Lynch C. (2008). Big Data: How Do Your Data Grow? *Nature* 455 (7209): 28–29.

Vartanov S. A. (2014). Big data v izmerenii televizionnoy auditorii: metody ochistki i obrabotki dannykh tsifrovykh STB [Big Data in Television Audience Measurement: Methods of Cleaning and Processing Digital STB Data]. In *Teoriya i praktika mediareklamnykh issledovaniy. Vypusk 4*. Analiticheskiy tsentr "Video Interneshnl", OOO "NIPKTs Voskhod-A Publ. Moscow. P. 151–167.

Shababb G., Taylor A. (2005). From set top meters to set top boxes. Audience measurement for the digital world. In *WAM - Worldwide Audience Measurement 2005 - Cross Media*.

Поступила в редакцию
04.03.2017