Юлия Павлова
Олеся Кольцова
НИУ-ВШЭ, СПб, 2011

К методологии сбора Интернет-данных для социологического анализа

Не смотря на то, что социальные науки всегда обращались к анализу текстовых данных, никогда еще социология не стояла перед таким беспрецедентным вызовом, который представляет собой появление Интернета. Уже сейчас Интернет стал и индикатором общественных процессов, и – шире – электронным вместилищем социальной жизни в ее самых разнообразных проявлениях, новым «домом» социального. Важность Интернета даже для обществ с невысокой долей Интернет-пользователей наглядно продемонстрирована недавними арабскими революциями. При этом, как набор электронных сигналов, Интернет одновременно является архивом социальной жизни, собранным массивом данных, доступных для анализа разнообразными автоматизированными методами. Это, с одной стороны – уникальная возможность для социологии, испытывающей все больше проблем со сбором данных опросными методами. Но с другой стороны, работа с этими данными требует совершенно новых подходов и компетенций и, что существенно, междисциплинарного сотрудничества. К этому социологи пока не совсем готовы. Как правило, для анализа ими используются традиционные методы – контент-анализ и дискурс-анализ, а проблема сбора данных вообще обходится. Исследуются конкретные сайты, полностью или выборочно, и вопрос о представительности данных не ставится. Таким образом, огромные возможности Интернет-данных служить индикаторами макро-процессов почти не используются. Масштабные исследования Интернет-данных проводятся теми, кто обладает компетенциями написания программного обеспечения и математических моделей для сбора, первичной обработки и анализа больших текстовых данных – физиками, computer scientists, биоинформатиками, иногда с привлечением компьютерных лингвистов. Все эти исследования, как правило, не ставят социологических задач и не содержат социологических выводов. Можно констатировать, что выполнение масштабных социологических исследований Интернета возможно только в междисциплинарных командах, для формирования которого требуется время, ресурсы и психологическая готовность исследователей разных дисциплин к сотрудничеству.

В данной статье мы коротко охарактеризуем основные проблемы сбора данных в социологических исследованиях Интернета и сфокусируемся на проблемах выделения и получения генеральных совокупностей и выборок и способах решения этих проблем, примененных в нашем исследовании. Задача нашего исследования – разработать методику, позволяющую определять, каким образом освещается та или иная социальная проблема или тематика в блогах, какие мнения и настроения по поводу нее доминируют в блогосфере, имеются ли по поводу нее социальные напряжения. Когда исследование начиналось, основной методологической сложностью нам представлялись вопросы анализа большого объема текстовых данных. Однако сейчас, к середине проекта стало ясно, что не меньшую сложность представляют собой вопросы сбора данных. Хотя, казалось бы, данные «там лежат» и их надо «только взять», каждый социолог, который ставит своей задачей получить срез мнений блогосферы по какой-либо тематике, становится перед рядом вопросов:

1. Как получить представление о структуре всей блогосферы для того, чтобы понять, какая ее часть является релевантной генеральной совокупностью для вашего исследования, и сделать из нее обоснованную выборку? За десятилетия исследований социологи (при том, что, конечно, они работали не в одиночестве) получили богатые данные о структурах населений разных обществ, на основании чего определение границ генеральной совокупности и построение выборок стало гораздо менее проблематичным. Также, исследователям традиционных медиа известны общие закономерности распределений различных параметров в газетных или телевизионных текстах, общая статистика по «массиву» этих текстов. Но по Интернету такой статистики попросту нет; а те данные, которые есть, как правило, находятся в распоряжении крупных поисковых систем и не доступны академическому сообществу. Скачать в домашний компьютер конкретного социолога «всю блогосферу» невозможно а) технически и б) т.к. не известны границы понятия «вся».

2. Как определить границы генеральной совокупности для своего исследования? В обычных опросных исследованиях такие вопросы считаются непроблематичными: например, генеральная совокупность «население России 16+» не требует особых методик, позволяющих отличить лиц «16+» от лиц «16–». В традиционном контент-анализе за генеральную совокупность обычно принимается какой-либо готовый корпус текстов (например, все публикации газеты “The New York Times” или все законодательство России). Но вопрос гораздо сложнее, если нужно провести исследование – сплошное или выборочное – всех текстов по определенной тематике. Как определить, удовлетворяют ли тексты критерию соответствия теме? В особенности, если определение в силу больших

объемов должно быть автоматизировано с тем, чтобы затем автоматически извлекать из генеральной совокупности выборки.

3. Как построить выборку? Как правило, в интернет-исследованиях выборочная совокупность строиться случайным образом из посетителей какого-либо одного сайта или постов (блогов) какого-либо блог-хостинга. При этом не ставится вопрос о репрезентативности такой выборки, так как отсутствует правильной понимание того, что должно являться генеральной совокупностью в подобных исследованиях. Поэтому возникают вопросы: какой метод будет являться адекватным для построения выборочной совокупности? Как определить качественные и количественные характеристики выборки? Каковы критерии репрезентативности выборки?

4. Как определить единицу анализа? Особую сложность представляет собой в Интернете понятие текста: традиционно, он определяется как законченное «произведение», которое обладает смысловой самостоятельностью, т.е. может быть понято без отсылки к другим текстам. Интернет-текст, в отличие от большинства традиционных текстов – это мета-текст, со множеством ссылок, а некоторые тексты (такие, как комментарии в блогах) принципиально не вписываются в классификацию текстов на самостоятельные и несамостоятельные.

5. Как технически осуществить выгрузку из Интернета всех требуемых данных? Поисковые системы не отдают всех результатов поиска, не позволяют задавать надежные, с точки зрения требований социологического исследования, критерии поиска или тем более делать из них разные виды выборок. Поисковики предназначены для быстрого поиска ограниченного количества максимально полезных текстов, а не для сбора репрезентативных данных. Единственной альтернативой является создание собственного ПО для сбора социологических данных, которое, как показывает опыт, есть новый вид социологической работы, т.е. работы именно социолога в тесном сотрудничестве с программистом и его постановщиком задач.

6. Как подготовить огромное количество данных к анализу? Если социолог хочет использовать автоматизированные методы, он сталкивается с необходимостью такой предварительной подготовки данных, с которой он никогда не имел дела не только в опросах, но и в традиционном контент-анализе. К ним могут относиться: очистка от технического шума (напр., обрывки html-кодов), парсинг (определение некоторых элементов грамматической структуры текста), лемматизация (сведение слов к начальным формам для корректного частотного анализа), корректное выделение N-грамм (словосочетаний), переведение текста в векторную форму (для кластеризации) и т.д. По форме, это работа компьютерного лингвиста, однако, как правило, как сами лингвисты,

так и ПО, ими разрабатываемое, не заточено под социологические задачи, поэтому и здесь также требуется совместная работа социолога, лингвиста и программиста.

При исследованиях Интернета меняются многие представления социологов о возможностях методов исследования и привычной для них методологии. Одна из проблем, с которой сталкивается социолог, - это определение того, что является генеральной совокупностью. В основном социологи привыкли работать с людьми, поэтому и объем, и структура, и основные характеристики, и методы определения генеральной совокупности знакомы исследователям. Например, в исследовании, посвященном формированию информационной культуры у студентов, какими бы ни были выбраны методы исследования – опрос, анкетирование, наблюдение, параметры генеральной совокупности задать несложно: возраст, ВУЗ, факультет и прочее. При анализе блогосферы исследователь имеет дело с электронными текстами, презентующими определенные стороны социальной жизни людей, и определить, какие параметры необходимо выделить для включения того или иного текста в генеральную совокупность – сложная и важная задача. Сколько и какие блоги брать в качестве генеральной совокупности, какими характеристиками они должны обладать, чтобы быть в нее включенными? В Интернет-исследованиях с использованием привычных для социолога методов опроса, интервью генеральная совокупность обычно представляется произвольно самим исследователем в качестве посетителей определенного сайта или вообще пользователей сети Интернет. Однако исследования проводятся не по всей сети Интернет, а с определенной генеральной совокупностью и выборкой. Если каким-то образом определены границы генеральной совокупности и составлена выборка, то понимание того, на какую генеральную совокупность могут быть распространены результаты исследования должно обязательно присутствовать. В настоящее время трудно говорить о какой-то единой методологии, с помощью которой исследователь мог сформировать генеральную совокупность из блогов для дальнейшего анализа. Стоит обратить внимание на те трудности и проблемы, которые возникают при определении границ генеральной совокупности блогосферы.

В настоящее время социологической и статистической литературы в области проблематизации определения границ генеральной совокупности нет. В основном методологическая литература описывает, как моделировать генеральную совокупность по выборке или формировать выборку по генеральной совокупности.

Существуют публикации, статьи отдельных отраслей, разрабатывающих свой методологический аспект при исследованиях блогосферы. С методологической точки

зрения следует сосредоточить внимание на том, что было взято исследователями за генеральную совокупность и как составлялась выборка. Так, например, существуют исследования блогосферы с использованием традиционных социологических методов [e.g. Papacharissi, 2004]. В данном исследовании границы генеральной совокупности З. Папачарисси были определены как все блоги, расположенные на сайте blogger.com. Она объясняет свой выбор тем, что это наиболее популярный и большой по числу блоггеров англоязычный сайт, который предоставляет возможности для персональных публикаций в стиле любительской журналистики. Любой блог, по мнению исследователя, размещенный на сайте blogger.com представлял собой единицу анализа, отвечающую по своим характеристикам признакам принадлежности к генеральной совокупности. К сожалению, данный факт невозможно подтвердить или опровергнуть, потому что признаки принадлежности единицы анализа к генеральной совокупности не оговорены З.Папачарисси в докладе по исследованию. В любом случае стоит отметить, что выборочная совокупность может являться репрезентативной только в отношении данной платформы, но не всей блогосферы в целом.

Так как blogger.com являлся основой для проведения исследования, выборка составлялась случайным образом из пользовательских директорий сайта, используя случайный выборочный интервал и случайную отправную точку. Таким образом, исследователем был применен обычный способ составления случайной выборочной совокупности, но не уточнены такие моменты: какие именно данные вводились в поисковую систему для поиска релевантных блогов и какие именно блоги считались релевантными, сколько блогов входило в генеральную совокупность, и почему было отобрано именно 260 блогов. В этом случае недостаток информации именно методологической стороны исследования относительно определения границ генеральной совокупности и формирования выборки не позволяет сделать вывод о репрезентативности данного исследования.

В российском исследовании, проведенном исследовательским агентством Social Media Research в 2009 году и посвященном анализу имиджа Патриарха РФ Кирилла, также следует обратить внимание на методологическую сторону относительно генеральной и выборочной совокупностей [Social Media Research, 2009]. Исследователями рассматривались только блоги, расположенные на платформе Live Journal. При помощи поисковой программы, включающей в себя также синтаксический анализатор текстов, были выбраны все записи в блогах, размещенные за определенный период, которые содержали слова «патриарх» и «Гундяев». Далее происходил отсев блогов, которые, по версии Яндекс, имели авторитетность менее 1000 баллов, а потом в ручную удалялись те

блоги, комментарии в которых не содержалось релевантной задачам исследования информации. Из выбранных блогов также ручным образом были выделены те комментарии, которые касались выборов нового патриарха. Таким образом, анализу подверглись 5 с лишним тысяч записей, которые исследовались как качественным контент-анализом с использованием техник сплошного и осевого кодирования, так и количественным - анализ соответствий и частот распределения. Стоит обратить внимание, что единицей анализа исследователями был выбран не пост, и даже не комментарий, а утверждение - высказывание по определенной теме, содержащееся в комментарии. В одном комментарии могло содержаться одно или несколько утверждений.

На примере данного исследования можно наиболее ярко продемонстрировать проблему определения границ генеральной совокупности и различения генеральной и выборочной совокупностей. Если бы исследователей интересовали все посты про Гундяева, безотносительно к времени, популярности и эксплицитности, то все процедуры, которые использовались для составления коллекции утверждений, в конечном итоге привели бы к формированию выборочной совокупности. В данном случае исследователи стремились получить все комментарии с эксплицитным упоминанием Гундяева за определенный период, поэтому полученная в результате коллекция утверждений и является генеральной совокупностью, так как именно она представляет собой набор утверждений на тематику исследования. Принадлежность к генеральной совокупности определялась принадлежностью к теме. В данном случае исследователями было выбрано два термина: «патриарх» и «Гундяев», по которым с помощью поисковой системы были найдены необходимые утверждения. Однако можно ли ограничиться только двумя терминами для формирования генеральной совокупности и можно ли утверждать, что во всех комментариях, которые отражают в себе какую-то оценку или мнение об имидже патриарха будет встречаться одно из этих двух слов. И наоборот, не попали ли в генеральную совокупность утверждения, которые не касаются имиджа патриарха, но содержат в себе данные термины?

Также стоит обратить внимание на исследования блогосферы с помощью неотрефлексированных компьютерных технологий [Bruns, 2007]. Именно они отражают проблематику зависимости социологов от программистов и других междисциплинарных отраслей. Исследование австралийских социологов «Methodologies for Mapping the Political Blogosphere: An Exploration Using the Issue Crawler Research Tool» [Bruns,2007] посвящено методологическим возможностям использования Issue Crawler в анализе блогосферы. Задачей исследователей была использовать Issue Crawler для выявления взаимосвязи между собой блогов, темой обсуждения которых было содержание в

Гуантанамо австралийского гражданина Дэвида Хикса, основной тип таких блогов (политические, социальные) и их отношение к СМИ в Австрии. В данном исследовании использование Issue Crawler еще более остро ставит проблему определения границ генеральной совокупности. Все доступные описания программного продукта Issue Crawler не содержат информации о том, каким именно образом он формирует генеральную совокупность. На основе вводных данных, семян (seeds) программа выстраивает сеть взаимосвязанных между собой единиц (сайтов, блогов), при этом совершенно не ясно, каким образом Issue Crawler определяет, когда нужно заканчивать поиск, то есть ограничить сеть, по каким параметрам он определяет, какие единицы включать в сеть, а какие нет.

В исследовании А.Брюнса исследователями с помощью Issue Crawler была получена карта взаимосвязанных между собой образцов, по которой и делались выводы об обсуждаемости темы заключения Д.Хикса. В качестве вспомогательного инструмента исследователи использовали агрегатор-накопитель блогов Technorati, который отслеживает более 70 млн блогов во всем мире. Функция поиска Technorati использовалась в начале февраля 2007, чтобы идентифицировать сто новых постов, содержащих фразу "Дэвид Хикс" или в названии или в самом тексте поста. Именно URL этих постов и были загружены по очереди в Issue Crawler. Исследователь изучал данную карту, делал выводы, не задаваясь вопросом, каким именно образом Issue Crawler сформировал генеральную совокупность для визуализации этих данных. Изменение картины сети во времени А.Брюнс интерпретирует как изменение в блогосфере, тогда как вполне может быть, что Issue Crawler просто не способен выдавать стабильный результат. Публичных и доступных тестов Issue Crawler не проводилось или они не опубликованы.

Одним из возможных путей формирования выборочной совокупности может послужить исследование по анализу электронных текстов с целью извлечения необходимой информации о терроризме и террористических событиях [Z. Sun et al, 2005]. Это единственная проблема, которая разрабатывается в социологическом ключе, но с применением компьютерной лингвистики. Авторами исследования предложено несколько стратегий, в каждой из которых один за другим выбирается документ с целью максимизировать необходимую информацию. При этом для составления выборки предполагается, что ряд образцов в виде имен, названий, слов и словосочетаний, отражающих суть этого события, уже заданы экспертом. То есть изначально представлен некоторый набор слов и словосочетаний, которые вводятся в программу для поиска необходимых документов. В дальнейшем отобранные документы подвергаются

экспертной оценке, то есть проверка наличия всех релевантных элементов осуществляется через человека вручную.

Все стратегии направлены на нахождения минимального множества документов с более полной информацией. Оптимальным множеством текстов считается такое множество, если:

1. Набор всех классов релевантных элементов, в нем содержащихся, и есть набор всех искомых классов релевантных элементов

2. Нет другого множества текстов с таким же свойством, только меньшего.

Стоит обратить внимание на то, что исследователями формировалась минимально возможная коллекция текстов, отражающая тему исследования. Для социологических исследований поиск минимальной коллекции не подходит. Обычно социологи хотят видеть полную Генеральную Совокупность, чтобы затем изучать вариации внутри нее. Таким образом, такой метод построения генеральной и выборочной совокупностей может быть применен после модификации. Поэтому несмотря на все преимущества данной отрасли, стоит отметить, что использование data mining для анализа блогосферы, а тем более для формирования выборки, возможно только в совокупности с другими, вспомогательными методами. Также обязательно участие эксперта, способного дать набор исходных признаков и в процессе оценить полученные результаты на релевантность.

Наиболее близкими с методологической точки зрения являются анализ русскоязычной блогосфер [Kelly et al., 2008], осуществленные в Berkman Center for Internet and Society в Гарварде, которые в своем исследовании использовали как социологические, так и методы из смежных областей. Анализируя методологию данного исследования по определению генеральной совокупности и формированию выборки, стоит обратить внимание на несколько моментов. Анализ русскоязычных блогов проводился исследователями с целью выделения дискуссионных сетей, в которых обсуждаются политические и общественные темы. Исследование начиналось с 5 миллионов блогов, которые были найдены с помощью индексов Яндекс блогов в мае 2009 года. Далее был использован метод «снежного кома», с помощью которого количество блогов значительно возросло. Данный метод осуществлялся посредством включения в выборку упоминания тех или иных блоггеров, блогов друзей данного блоггера, ссылок на другие блоги. Далее был применен метод отсеивания неактивных блогов. К сожалению, исследователями не было прописано, по каким параметрам блог считался активным или нет. В итоге процедуры отсеивания остался 1 млн блогов. С помощью анализа социальных сетей, исследователями были определены 17000 наиболее связанных между собой

блоггеров, из которых с помощью установленного порога, а именно содержания в блоге ссылок на 10 из 4000 наиболее цитируемых сайтов, было исключено определенное количество блогов. Оставшиеся 11 тысяч были кластеризованы и нанесены на карту для визуального представления данных.

Таким образом, начав с изучения более пяти миллионов блогов, они использовали анализ социальных сетей для того, чтобы выделить весьма активное "дискуссионное ядро", состоящее из более чем 11 тысяч блогов. Логично предположить, что эти 11 тысяч блогов и являлись генеральной совокупностью, так как тематически принадлежали к исследуемой проблематике. Поэтому выводы исследования можно распространять только на это количество блогов, но не на всю русскоязычную блогосферу в целом. Существенным вопросом является, почему исследователи остановились на 11 тысячи блогов и что явилось критерием остановки ограничения числа единиц. Скорее всего, это объясняется тем, что больше 10-15 тысяч блогов визуализировать практически невозможно, и программное обеспечение, выбранное исследователями, не может проводить кластерный анализ на большем числе элементов. Поэтому следует еще раз обратить внимание на тот факт, что исследователи-социологи находятся в зависимости от технических возможностей программного обеспечения, компьютерных программ при исследованиях блогосферы. Методология, применяемая в исследовании русскоязычной блогосферы Центром Беркмана сочетает в себе разнообразные методы, каждый из которых имеет свои слабые и сильные стороны, но большой проблемой все-таки остается то, что социолог не до конца вникает в процесс проведения исследования, доверяя часть работы программистам, для которых не имеет значения социологическая значимость поступивших данных, а только техническая сторона дела.

С какими же еще проблемами сталкивается исследователь при анализе и изучении блогосферы. Первая и основная проблема состоит в том, что именно считать блогом. В настоящее время сложилась такая ситуация, что социолог не решает, что именно считать блогом, когда начинает свое исследование по анализу блогосферы. За него это делает либо поисковая система, у каждой из которых есть свои параметры отнесения текста к разряду блогов, поэтому по определенному запросу поисковая система выдает результаты из той совокупности, которая представляется ей как совокупность блогов, либо это решает блог-хостинг, если обращаться к нему напрямую. Самое большее, что может сделать социолог – это эксплицировать ограничения той или иной системы. Здесь уместно упомянуть и о проблеме зависимости исследователя в ходе формирования Генеральной Совокупности при анализе блогосферы от программистов, компьютерщиков и математиков. Такой зависимости при опросах или исследованиях другими общеизвестными методами у

социологов нет, они с ней не сталкивались и не умеют работать. Социологи не представляют, что алгоритмы, заложенные в программном обеспечении, влияют на социологические результаты, и тем более не представляют, как именно влияют.

Вторая проблема, с которой сталкивается исследователь – это выбор единицы анализа, то есть из чего именно формировать генеральную совокупность – из блогов, постов, комментариев. Если исследователя интересует именно тематическая подборка блогов, то за единицу анализа стоит брать пост, поскольку он чаще всего посвящен только одной теме, в отличие от блога, который содержит в себе высказывания на различные темы различных жанров. Тематически разделяются именно посты, а не блоги.

Отсюда возникает еще одна проблема - проблема тематического разнообразия. Когда происходит характеристика генеральной совокупности в обычных социологических исследованиях – это не вызывает трудности, так как исследователь четко определяет для себя параметры генеральной совокупности – пол, возраст, семейное положение, поэтому и самой задачи формирования генеральной совокупности не стоит перед исследователем. В исследовании блогосферы основные параметры генеральной совокупности, как правило, определяются принадлежностью блогов к тематике, выбранной исследователем для анализа. Определение наличия у блога этой характеристики далеко на так самоочевидно, как определение половозрастных характеристик человека, и требует разработки специальной процедуры.

В рамках проекта «Разработка методологии сетевого и семантического анализа блогов для социологических задач» был разработан многоступенчатый механизм для определения границ генеральной совокупности. Основная цель проекта[1] - разработать комплексную методику социологического анализа русскоязычной блогосферы, а именно извлечь все тексты, соответствующие тематике исследования (теме ислама) за определенный период, разделить корпус текстов на семантически близкие кластеры и сопоставить их с группами, полученными методом сетевого анализа, и, таким образом, обнаружить сети/сообщества, освещающие тему ислама в русскоязычных блогах. Конкретно наша задача – изучить мнение блоггеров на определенную тему, тему ислама, поэтому брать за генеральную совокупность все блоги не является релевантным. Необходимо принимать за генеральную совокупность только те блоги, которые отвечают тематике исследования.

Определение того, что есть принадлежность к теме и отнесение элементов к генеральной совокупности и представляет собой главную методологическую проблему.

---

[1] проект «Разработка методологии сетевого и семантического анализа блогов для социологических задач», рук. Е.Ю.Кольцова, грант Научного Фонда ГУ-ВШЭ в рамках конкурса «Учитель-Ученики 2011-2012 гг.»

Решение данной проблемы возможно с помощью различных методов и методик. В качестве основы такой процедуры трудно предложить что-либо, кроме многоступенчатой экспертизы. В начале исследования в рамках проекта была проведена операционализация понятий, которые использовались при работе с экспертами. Были даны определения, что считать исламом и исламским событием. Таким образом, эксперту понятно, какие именно тексты считать релевантными теме исследования, и, тем самым, он может дать исходную информацию в виде списка слов, терминов или событий, относящихся к данной теме.

На этом этапе важной задачей является работа с экспертами. В рамках работы над проектом перед началом поиска предварительно были составлены письмо к эксперту и лист вопросов. Письмо содержало в себе обращение к эксперту с пояснением целей и задач исследования, а также операционализацию основных используемых понятий. В листе вопросов экспертов просили назвать несколько событий за последние два года, которые, с их точки зрения, имеют отношение к Исламу и которые могли бы обсуждаться в русскоязычном Интернете, указать важные слова и понятия. Также экспертам предложено было указать ссылки на сайты или блоги, где обсуждаются эти или другие события и контакты тех людей, которые могли бы выступить экспертами в подобном исследовании. Ответы экспертов на предложенные вопросы явились основой для составления исходного списка терминов и событий, необходимых для выкачивания текстов.

. Проблема подбора экспертов является одной из наиболее сложных в теории и практике экспертных исследований. Очевидно, что в качестве экспертов необходимо использовать тех людей, чьи суждения наиболее помогут принятию адекватного решения. Экспертами должны являться люди, обладающие достаточным опытом и компетенцией, работами и достижениями в исследуемой области. Поиск экспертов осуществляется несколькими методами, не исключающими друг друга:

- с помощью Интернета и поисковых систем, профессиональных связей участников проекта. Для начала был составлен примерный список из нескольких видных представителей науки, политики, занимающихся исламом в профессиональной или общественной жизни. Данным экспертам были высланы бланки анкеты, но поиск также продолжался по упоминаниям других людей на их личных страницах в социальных сетях, на сайтах мест профессиональной деятельности, в публикациях, совместных выступлениях.

- В бланке анкеты экспертам предлагалось посоветовать людей, к которым можно обратиться по данному вопросу. Таким образом, методом «снежного кома» также велся поиск экспертов для участия в проекте.

- Одновременно поиск осуществлялся по спискам конференций и круглых столов, посвященных исламу: «Россия и исламский мир», «Мир ислама*:* история, общество, культура», «Ислам и женщина», "Ислам*:* толерантность, мир и неприятие насилия".

- Также время подбора экспертов совпало с проходившими в НИУ-ВШЭ Публичными лекциями ведущих востоковедов страны про страны Средней Азии и Кавказа, с некоторыми из докладчиков также удалось установить контакт и привлечь к работе проекта.

- Поиск велся и в социальных сетях vkontakte.ru, facebook.com. Электронные адреса некоторых из потенциальных экспертов отсутствуют в открытом доступе, поэтому связаться с ними удается только посредством данных сайтов. Если страница деятеля находится в закрытом доступе или ему нельзя отправить сообщение, то чаще всего открыт список друзей или групп, в которых он состоит. Это также помогло в поиске экспертов, потому что, как правила, друзья и участники группы связаны с изначальным объектом профессиональными связями, поэтому сами являются компетентными в теме ислама.

В результате было разослано 35 писем к экспертам с предложением о сотрудничестве. Двое из них отказались от заполнения бланка, объясняя это спецификой своей работы по исламу, которая не может быть вполне адекватной задачам проекта. Пятеро экспертов согласились на участие в проекте и прислали заполненные бланки с ответами. Экспертами стали люди, компетентные в области ислама, занимающиеся этой темой в профессиональной и общественной деятельности:

Ответы экспертов на предложенные вопросы явились основой для составления исходного списка событий, необходимых для выкачивания текстов. Таким образом, на основе предложенного экспертами списка событий была выкачена коллекция текстов. Эти тексты были переданы кодировщикам для анализа с целью:

1. проверки качества списка слов с помощью ручной классификации постов данной коллекции на релевантные или нерелевантные.

2. обнаружения упущенных ранее важных терминов или исключения нерелевантных. В рамках второй с этой же целью автоматически происходит подсчет частотности слов, позволяющий определить, насколько часто встречаются те или иные слова в тексте и обнаружить ранее не включенные в список, но достаточно часто встречающиеся в исламских текстах, слова и словосочетания. На основе полученных данных происходит формирование нового или пополнение старого списка терминов или событий до тех пор, пока не будет появляться новой информации о словах и событиях, то есть список не окажется исчерпывающим.

Этот процесс в дальнейшем автоматизируется и уже с помощью автоматического отбора текстов по составленному списку происходит формирование коллекции электронных текстов, то есть определение границ генеральной совокупности исследования. В рамках проекта генеральная совокупность текстов об исламе – это все тексты, содержащие высказывания об исламе.

Автоматизированный процесс отбора релевантных текстов как для тестовой коллекции, так и для окончательной генеральной совокупности также необходимо проверять вручную с помощью кодировщиков. Такая проверка тоже может носить итеративный характер. Для уменьшения субъективности кодировщиков обычно проводится их обучение, а затем проверка надежности интеркодирования. В ходе нее исследователь предлагает для анализа один и тот же текст нескольким кодировщикам и проверяет, насколько сходятся их результаты. Надежными считаются те результаты, в которых кодировщики сходятся. Если расхождение слишком большое, может происходить дополнительное обучение кодировщиков, удаление параметра или другая корректировка исследования. Если расхождение минимально, кодирование может быть продолжено без дублирования одним кодировщиком других на одном и том же массиве данных.

На основе разработанного алгоритма определения границ генеральной совокупностями нами был проведен пробный эксперимент. Перед началом работы был сформирован список терминов, на основе предположений о том, что эти слова и словосочетания должны выводить на различные исламские дискурсы в русскоязычной блогосфере. Список состоял из следующих слов: мусульмане, мусульманство, ислам, Аллах, мусульманский, исламский, исламисты, джихад, моджахед, Коран, хадисы, мечеть, имам, исламские террористы. Каждый термин искался через Яндекс-блоги (поиск проводился только в блогах, без комментариев; микроблоги на данном этапе тоже не рассматривались). Брались только первые 10 результатов поиска, то есть первая страница результатов поисковой системы. Эти результаты были занесены в таблицу. В таблице были представлены ячейки: номер (присваиваемый по порядку), url (ссылка на текст), сам текст (если текст поста был длинный, он сохранялся в текстовом файле с теми же номерами). Каждому тексту был присвоен номер, была составлена кодировочная таблица, в которой отображались по вертикали - номера текстов, по горизонтали – оценка кодировщиков. Составленная коллекция текстов вместе с кодировочной таблицей была передана 4 кодировщикам, в качестве которых на данном этапе выступали участники проекта. Релевантность оценивалась следующим образом: 3 – текст релевантен, 1 - нерелевантен, 2 – релевантность неопределена. Таким образом, было получено 4

кодировочных таблицы, в которой отображались оценки кодировщиков по постам блогов, выкаченных с помощью исходного списка терминов.

Результаты работы кодировщиков необходимо было проверить на степень надежности интеркодирования. По итогам анализа литературы и тестовых испытаний наилучшим показателем для такой проверки является Krippendorff's alpha, подсчет которого возможно провести в SPSS. Однако данный индекс рассчитан на большие коллекции от 1000 единиц анализа, все доступные макросы написаны именно для такого количества текстов [Hayes, Krippendorff, 2007]. Также стоит отметить, что данный индекс при подсчете степени надежности интеркодирования для трех и более кодировщиков не позволяет выявить того, чьи оценки расходятся с большинством. Однако общий показатель Krippendorff's alpha вычисляет с высокой точностью и будет использован в работе проекта уже на большой коллекции.

Для тестовой экспериментальной выборки текстов использовался другой показатель Cohen's kappa (k) (Cohen 1968), который ведет подсчеты как для сразу нескольких кодировщиков, так и попарно, с целью выявления того, чьи оценки расходятся с большинством. Степень надежности интеркодирования согласно показателю Cohen's kappa (k) = 0,6. Это среднее значение, которое говорит о том, что между кодировщиками не достигнута высокая степень согласованности, что влечет за собой дальнейшее обучение кодировщиков и выработки четких инструкций кодирования.

Таким образом, в качестве методологического решения основной проблемы определения границ генеральной совокупности в данной статье предлагается создание автоматизированного итеративного процесса отбора единиц анализа на основе первичной информации, выданной экспертами, который в каждом новом исследовании проверяется вручную с помощью кодировщиков, и которые сами также подвергаются проверке друг другом с помощью расчетов коэффициентов надежности интеркодирования.

Следует сказать, что итеративность методики позволяет расширить границы генеральной совокупности, однако в любом случае ограничением данной методики является зависимость результата от начальных представлений экспертов о тематике. Т.о. применимость такой методики к выявлению латентных дискурсов ограничена, хотя и не сводится к нулю; в большей степени она оправдана тогда, когда границы тематической области не проблематичны и ставится задача выявления каких-либо явлений внутри этих границ. Исходя из такого представления о возможных ограничениях, нами было проведено сопоставление генеральной совокупности постов об исламе, полученной предложенным способом из двухнедельного сплошного массива постов на все темы, и тематических кластеров, получаемых из того же массива с помощью одной из методик

выявления тем – латентной Дирихле-аллокации (LDA; Blei et al 2003). Последняя не выявила существования лингвостатистического единства текстов «про Ислам вообще»; социально-политические темы образовывались в основном вокруг событий или конкретных явлений. В то же время, LDA «не обратила внимания» на некоторые тексты с высоким содержанием ключевых слов, определенных экспертами, т.е. не сгруппировала их в отдельные темы. Т.о. LDA выделяет тематические группы, когда наблюдает в выборке много статистически похожих текстов; если тексты очень разнородны, то, даже если они при ручном кодировании определяются как «посвященные исламу», LDA их не замечает. Использование нескольких методик может компенсировать их ограничения и объединить их сильные стороны.

Список литературы:

1. Blei, David M.; Ng, Andrew Y.; Jordan, Michael I; Lafferty, John. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: pp. 993–1022. doi:10.1162, 2003.
2. Bruns A. Methodologies for Mapping the Political Blogosphere: An Exploration Using the IssueCrawler Research Tool. – 2006. [Электронный ресурс]. URL: http://eprints.qut.edu.au/7832/1/7832.pdf
3. Cohen, J. "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit". *Psychological Bulletin* **70** (4): 213–220, 1968.
4. Etling B., Alexanyan K., Kelly J., Faris R., Palfrey J., Urs Gasser « Public Discourse in the Russian Blogosphere: Mapping RuNet Politics and Mobilization». – 2010. [Электронный ресурс]. URL: http://cyber.law.harvard.edu/publications/2010/Discourse_Russian_Blogosphere
5. Hayes, Andrew F. & Krippendorff, Klaus . Answering the call for a standard reliability measure for coding data. Communication Methods and Measures. – 2007, p. 77-89.
6. Merriam-Webster Dictionary online. [Электронный ресурс]. URL: http://www.merriam-webster.com/
7. Papacharissi, Z. Audiences as Media Producers: Content Analysis of 260 Blogs. In M. Tremayne (Ed.), Blogging, Citizenship, and the Future of Media. New York: Routledge, 2007. [Электронный ресурс]. URL: tigger.uic.edu/~zizi/Site/Research.../TremayneChapterBlogs.pdf
8. Rogers R. Issue Crawler Web Network Mapping Software and Allied Tools. Issue Mapping Contextual Essay. [Электронный ресурс]. URL: http://www.issuecrawler.net, http://tools.issuecrawler.net

9. Sun Z., Lim E., Chang K., Ong T., Gunaratna R. Event-Driven Document Selection for Terrorism Information Extraction. In Proceedings of ISI'2005. [Электронный ресурс]. URL:svn.mosuma.com/r9999/doc/publications/sun2005isi.pdf
10. Woods A., Fletcher P. Huches A. Statistics in Language Studies. Cambridge Textbooks in Linguistic. - 2003.
11. Давыдов *А. А.* Социология изучает блогосферу // // Социологические исследования. - 2008. - № 11. - С. 92-101.
12. Ландэ Д.В., Снарский А.А., Безсуднов И.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы. - 2009.
13. Мангейм Дж. Б., Рич Р. К. Политология. Методы исследования: Пер. с англ. / Предисл. А.К. Соколова. – М.: Издательство “Весь Мир”, 1997. - 544 с
14. Хавенсон Т.Е.; Применение методов Data Mining в социологии. III Всероссийский социологический конгресс "Социология и общество: проблемы и пути взаимодействия". Москва, 21-24.10.2008