РАЗРАБОТКА МЕТОДОЛОГИИ СОСТАВЛЕНИЯ ВЫБОРОК ЭЛЕКТРОННЫХ ТЕКСТОВ ДЛЯ СОЦИОЛОГИЧЕСКОГО АНАЛИЗА РУССКОЯЗЫЧНЫХ БЛОГОВ

Автор:
Павлова Юлия Валерьевна,
НИУ ВШЭ — Санкт-Петербург, факультет социологии,
1-й курс магистратуры,
E-mail: Julia.v.pavlova@gmail.ru

Научный руководитель:
Кольцова Олеся Юрьевна,
доцент кафедры социологии НИУ ВШЭ — Санкт-Петербург,
декан факультета социологии

В своих исследованиях социологи обычно имеют дело с людьми, поэтому объем, структура, основные характеристики и методы определения генеральной совокупности им хорошо знакомы. В работе с блогосферой определение границ генеральной совокупности затруднительно: сколько и какие блоги брать в качестве генеральной совокупности, какими характеристиками они должны обладать, чтобы быть в нее включенными — это вопросы, на которые необходимы ответы. Основная цель проекта[1], в рамках которого выполнена эта работа — 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. Задача автора доклада в рамках проекта — разработка методологии формирования генеральной совокупности для данного исследования. В этом отношении обнаружен ряд проблем.

Первая проблема состоит в том, что именно считать блогом. Эта проблема здесь не рассматривается, так как в настоящее время социолог, начиная исследование, не решает, что именно считать блогом. За него это делают поисковые системы, у каждой из которых есть свои параметры отнесения текста к разряду блогов, и самое большее, что может сделать социолог — это эксплицировать ограничения той или иной системы.

Вторая проблема — это выбор единицы анализа, то есть из чего именно формировать генеральную совокупность: из блогов, постов, комментариев. Так как нас 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.

Третья, и главная проблема, с которой сталкивается исследователь — это проблема операционализации критериев принадлежности выбранных единиц анализа к генеральной совокупности. В обычных социологических исследованиях, когда происходит формирование генеральной совокупности, эти критерии достаточно ясны или, во всяком случае, привычны: пол, возраст, семейное положение. В нашем случае основной характеристикой для включения в генеральную совокупность является принадлежность поста к исследуемой теме. Определение наличия у блога этой характеристики далеко не так самоочевидно, как определение половозрастных характеристик человека, и требует разработки специальной процедуры.

В качестве основы такой процедуры трудно предложить что-либо, кроме многоступенчатой экспертизы. В начале исследования дается операционализация понятий, которые будут использованы при работе с экспертами. В рамках осуществляемого проекта исследователями 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.

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

На основе предложенного ими списка событий или терминов выкачивается тестовая коллекция текстов, найденная с помощью поисковой системы среди постов блогов. Эти тексты в дальнейшем передаются тем же экспертам либо кодировщикам для двух задач: для проверки качества списка слов с помощью ручной классификации постов данной коллекции на релевантные или нерелевантные, и для обнаружения упущенных ранее важных терминов или исключения нерелевантных. В рамках второй задачи также производится автоматический анализ частот слов в тестовой коллекции постов, признанных релевантными, по сравнению со случайной коллекцией. Это позволяет выявить лексику, специфическую для данной тематики, и включить ее в список ключевых терминов. На основе полученных данных происходит пополнение списка терминов или событий до тех пор, пока не прекратит появляться новая информация.

Далее с помощью автоматического отбора текстов по окончательному списку происходит формирование генеральной совокупности исследования. 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.

Для проверки надежности интеркодирования разработаны различные коэффициенты: Percent agreement, Holsti's method, Scott's pi (*p*), Cohen's kappa (*k*), Krippendorff's alpha (*a*), Perrault & Leigh's Ir. По результатам пилотного анализа применяемых для проверки надежности интеркодирования коэффициентов можно сделать вывод, что самым приемлемым для анализа блогосферы является Krippendorff's alpha, так как он подходит для работы с большим объемом информации и возможен расчет для нескольких кодировщиков.

Таким образом, методологическим решением основной проблемы определения границ генеральной совокупности является создание автоматизированного интерактивного 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.