Всероссийский научно-практический симпозиум

«СОЦИАЛЬНЫЕ КОММУНИКАЦИИ:

УНИВЕРСУМ ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ»

Санкт-Петербург, 9-10 ноября 2011 г.

Кинчарова Анастасия Владимировна,

Равлик Мария Васильевна

Некоторые методические проблемы анализа больших социальных сетей в Интернете и способы их решения

Аннотация. Доклад посвящён некоторым методическим проблемам анализа больших социальных сетей в Интернете, решаемых в рамках проекта «Разработка методологии сетевого и семантического анализа блогов для социологических задач» (осуществляется при поддержке Программы «Научный фонд НИУ ВШЭ», 2011-2012 гг.). Содержательным фокусом исследования является концептуализация ислама в блогах (семантический анализ) и анализ социальных сетей блогеров, пишущих и комментирующих посты об исламе.

В рамках исследования такого рода возникает ряд социологических и технических проблем.

Во-первых, необходимо отобрать посты (а также их авторов и комментаторов), в которых присутствует «исламский дискурс». Во-вторых, выбрать методы анализа социальных сетей, которые позволяют отразить структуру сетей блогеров, пишущих и/или комментирующих посты об исламе (желательно выделить их сообщества, «сгустки»). При этом выбор методов должен быть таким, чтобы результаты сетевого анализа (сообщества или группировка, выполненная другим способом) были сопоставимы с результатами группировки постов при семантическом анализе.

Технические проблемы связаны с необходимостью обращения к большим массивам данных, находящихся в Интернете, и подбора программного обеспечения, которое может не только реализовать выбранные методы анализа социальных сетей, но и «переварить» значительный объём данных.

В докладе освещаются используемые нами способы решения указанных проблем.

Annotation. The report addresses some methodological problems of analysis of large social networks on the Internet which are solved within the project "Development of methodology of semantic and social network analysis of blogs for sociological purposes" (supported by "Science Foundation NRU HSE", 2011-2012). Ontological focus of the project is conceptualization of Islam in blogs (semantic analysis) and social network analysis of bloggers, writing posts and commenting on Islam.

A study of this kind raises a number of sociological and technical problems.

First, it is necessary to sample posts (and their authors and commentators), containing "Islamic discourse". Second, we need to select methods of social network analysis that can capture the structure of the networks of bloggers who write and/or comment on posts about Islam (it is desirable to distinguish their communities or places of concentration). The choice of the methods should be such that the results of the social network analysis are comparable with the results of grouping of posts in the semantic analysis.

Technical problems are connected with the need to access to a large amount of data stored on the Internet, and require software which can not only implement the selected methods of social network analysis, but also to "digest" a significant amount of data.

The paper discusses methods that we are using to solve these problems.

Идея исследования

Исследование посвящено концептуализации ислама в блогах и анализ социальных сетей блогеров, пишущих и комментирующих посты об исламе. Описание концептуализации ислама в блогах решается с помощью семантического анализа, социальные сети блогеров изучаются посредством методов анализа социальных сетей.

Формирование массива данных

Первая задача, которую необходимо решить на методологическом уровне – это отбор постов, а вместе с ними их авторов и комментаторов, которые полностью или частично содержат какие-либо высказывания об исламе.

В методологическом плане данная задача практически идентична задаче формирования выборочной совокупности для контент-анализа, например, какого-либо СМИ на определённую тему.

Сначала нужно уточнить, во-первых, какой именно носитель информации будет анализироваться, и, во-вторых, за какой период. В нашем случае «география» исследования ограничена Живым Журналом (http://livejournal.ru/), временной диапазон для окончательного анализа пока не определён, поскольку исследование находится в процессе реализации.

Единый, относительно небольшой временной интервал, в рамках которого будут браться тексты, позволит получить «моментальный снимок» ЖЖ и картины концептуализации ислама в нём. Выбор небольшого отрезка времени для анализа (одна или несколько недель) наиболее оправдан в силу динамичности блогосферы и короткого периода актуальности постов.

Сложную задачу на этапе формирования выборочной совокупности представляет собой определение того, что такое «текст об исламе» и отбор таких текстов. В исследованиях, выполняемых на относительно небольших массивах данных, такой отбор можно произвести вручную, путём просматривания всех постов, попадающим в целевую выборку по датам размещения и другим «внешним» критериям, заданным исследователями. Вместе с тем, сформированный массив данных, полученный в результате такого отбора, несёт на себе отпечаток субъективного взгляда о том, что такое «ислам» и «пост об исламе» того, кто отбирал посты. Например, среди участников нашей исследовательской группы такой спор возник относительно того, считать ли «исламскими» постами посты об арабских революциях. Однако и эту проблему можно решить путем выполнения отбора текстов не одним, а несколькими кодировщиками и последующей триангуляцией результатов их работы.

Возникает вопрос, как проводить отбор текстов об исламе для больших массивов данных, большого количества постов, которые невозможно обработать вручную. В качестве решения было выбрано использование поисковых запросов по определённым словам и сочетаниям слов, которые являются индикаторами наличия в тексте «исламского» содержания. В свою очередь это требует построения списка слов и словосочетаний-индикаторов. Для реализации этой задачи использовалась комбинация экспертного опроса (экспертов просили назвать события, связанные с исламом) и качественного анализа (затем происходил поиск текстов, связанных с этими событиями, и из них отбиралась лексика, ассоциирующаяся с исламом).

Необходимость использовать формализованную процедуру (автоматизированный поиск определенных слов и словосочетаний) отбора текстов вызывает несколько методических вопросов, которые могут остаться незамеченными при использовании для отбора живых кодировщиков.

Во-первых, оказывается, что исламская лексика может использоваться в контекстах, никак не связанных с исламом, или связанных слабо (например, «Аллах акбар» как экспрессивное выражение в тексте, не связанном в остальном с исламской тематикой). Такие варианта использования можно считать случайными, и игнорировать при формировании коллекции текстов для исследования, но возможно, их следует рассматривать как проникновение исламских смысловых конструкций в неисламские контексты, а это может являться значимым социальным фактом. В исследовании более оправданным следует считать второй подход, поскольку он исключает ситуацию, когда мы можем проигнорировать, возможно, значимый вариант концептуализации ислама в широком социокультурном контексте.

Во-вторых, необходимо ответить на вопрос, достаточно ли использовать одно слово или словосочетание для того, чтобы определить текст как «исламский», или в тексте должно присутствовать не менее двух или трёх и т.д. «исламских» слов или словосочетаний. Заметим, что данная проблема отчасти связана с предыдущей, так как исламские слова или словосочетания, использованные в неисламских контекстах, часто, хотя и не всегда, бывают единичными.

Наиболее методологически оправданным представляется вариант включения в выборочную совокупность текстов, содержащих исламскую лексику в неисламских контекстах и поиск по одному словосочетанию, поскольку в этом случае мы в меньшей степени рискуем упустить из виду некий специфический класс (или классы) высказываний об исламе. Вместе с тем, такой подход увеличивает

массив данных, с которыми приходится работать, что в нашем случае является значимым фактором. Однако полнота исходного массива данных более предпочтительна, чем некоторое сокращение его объёма.

Формирование данных для анализа связано также с решением такой технической проблемы, как необходимость обращения к большим массивам данных, находящихся в Интернете. Для извлечения информации из Интернета могут использоваться программы-кроулеры, которые, однако, в большинстве случаев не позволяют адекватно оценить качество их работы.

Ещё одной методической проблемой является фиксация всей информации, которая необходима для проведения семантического и сетевого анализа и последующего сопоставления их результатов. Для проведения семантического анализа нужны только тексты постов, тогда как для проведения сетевого анализа нужны маркеры постов (наиболее очевидным маркером являются URL-адреса), а также маркеры комментаторов. Кроме того, значимой информацией является маркер автора поста. Все эти маркеры должны быть соотнесены. Данная задача является технической, однако на ней стоит заострить внимание, поскольку без её решения невозможно будет соотнести результаты сетевого и семантического анализа.

В нашем исследовании для решения этой проблемы было создано специальное программное обеспечение, которое позволяет закачивать посты Живого Журнала и их атрибуты (автор, комментаторы, дата), которые необходимы для нашего исследования.

Выбор методов и инструментов анализа социальных сетей

Следующий блок задач связан с выбором методов анализа социальных сетей, которые позволяют отразить структуру социальной сети блогеров Живого Журнала, пишущих об исламе. Эта задача сформулирована исходя из предположения о том, что в сети, образованной блогерами и их связями комментирования постов друг друга, существуют «сгустки», причём эти сгустки совпадают, хотя бы отчасти, с семантически однородными общностями. Иначе говоря, гипотеза состоит в том, что группировка блогеров в рамках семантического анализа по крайней мере частично совпадает с группировкой, полученной в результате сетевого анализа.

Задачей сетевого анализа, таким образом, является отразить структуру сетей блогеров, пишущих и/или комментирующих посты об исламе таким образом, чтобы можно было выделить относительно крупные «сгустки» блогеров, связанных друг с другом.

Для анализа структуры социальных сетей существует ряд методов и понятий.

В перспективе «снизу вверх» сеть рассматривается как «вырастающая» из соединений минимальных возможных групп – диад и триад, при такой перспективе определяются такие типы подструктур сети, как клики, n-клики, n-кланы, n-клубы, k-плексы, K-ядра, f-группы (cliques, n-cliques, n-clans, n-clubs, k-plexes, K-cores, f-groups).

Клика – это максимальный набор вершин, каждая из которых непосредственно связана с каждой другой вершиной в наборе, N-клика – то же, но в ней вершины могут быть связаны не непосредственно, а на расстоянии N шагов. N-клан подобен N-клике с тем ограничением, что связь между вершинами в группе должна проходить через вершины, которые также принадлежат группе (в отличие от N-клики, где «связывающая» вершина могла находиться за пределами самой N-клики). N-клуб – это максимальный набор вершин, каждая из которых находится на расстоянии не более N шагов от другой, причем эти шаги должны проходить через вершины, находящиеся внутри группы. K-плекс – это набор вершин, каждая из которых связана со всеми, за исключением K вершин данного набора. K-ядро – это максимальное количество вершин, каждая из которых связана с K вершин в данном наборе. F-группы представляют собой группы, сформированные «сильно транзитивными» и «слабо транзитивными» триадами, которые в свою очередь определяются соотношением силы связи между участниками триад.

В рамках подхода «сверху вниз» сеть рассматривается целиком, и в ней выделяются области, внутри которых связи между вершинами более тесны, чем с вершинами, лежащими за её пределами, тем самым эта область выделяется в сети. Этот подход к определению структуры сети реализуется такими методами, как выделение компонентов, блоков и точек сочленения, LS-наборов, лямбда-наборов и мостов, фракций (components, blocks/cutpoints, LS-sets, Lambda sets and bridges, factions).

Для нас более продуктивным подходом является подход «сверху вниз», то есть рассмотрение её как целого, где есть места, где сплетение связей более редкое, чем в других. Эти места являются местами потенциального разбиения сети на части.

Компоненты – это части сети, которые связаны внутри себя, но отделены от других (в том числе могут состоять из одной вершины – изоляты). Точки сочленения – это вершины, при удалении которых сеть распадётся на несвязанные части, блоки (или би-компоненты) – части, на которые сеть распадается. Фракции – это алгоритм, который находит такую конфигурацию сети, при которой она максимально приближается к идеально-типической сети, в которой все вершины имеют связи только внутри своей подсети, то есть сети, полностью разделённой на компоненты. Предполагает знание количества компонент этой идеально-типической сети.

LS-набор (LS-set) – это набор составных частей, который имеет меньше связей с внешним окружением, чем каждая из его составных частей. То есть в отличие от выше перечисленных методов, он определяется по соотношению связей внутри подсети и связей, соединяющей подсеть с остальной сетью. Этот метод, однако, имеет слишком жесткие ограничения. Более общим случаем является лямбда-набор – набор связей в сети, удаление которых сильнее всего повлияет на связи между всеми остальными вершинами (с учётом значимости каждой связи для сети в целом).

Подход, связанный с выявлением составляющих сети, вершины внутри которых связаны сильнее, чем с другими вершинами сети, является наиболее продуктивным для исследования структуры больших сетей, каковой является сеть блогеров ЖЖ, пишущих об исламе. Однако указанные выше методы не могут решить наши задачи, поскольку не рассчитаны на использование на больших объемах данных.

Идея разделения сетей на некие имеющие содержательный смысл части развивалась также в рамках таких направлений, как разделение графа (graph partitioning) в рамках теории графов и компьютерных науках и метод иерархической кластеризации в социологии. Однако оба эти метода для полноценного выделения сообществ в сети были недостаточно хороши. Компьютерные науки ставят задачу разделения графа следующим образом: необходимо разделить некоторое количество вершин сети на известное количество групп определённого размера таким образом, чтобы количество связей между вершинами было минимальным. Такая постановка задачи актуальна при построении, например, компьютерных сетей. Однако так ставить задачу применительно к социальным сетям невозможно, потому что исследователь, как правило, не знает, сколько сообществ обнаружится в данной сети и какого они будут размера. Очень вероятно, что они будут не равны между собой по величине.

Группа методов иерархической кластеризации предоставляет более подходящие возможности анализа сообществ. В основе этих методов лежит измерение сходства между парами вершин в соответствии с некими значимыми критериями. Затем осуществляется последовательное объединение вершин, обнаруживших наибольшее сходство. Данная группа методов иерархической кластеризации называется агломеративной. Другой вариант иерархической кластеризации состоит, наоборот, в выделении в отдельный кластер вершин, имеющих наименьшие показатели сходства, при том, что первоначально вся сеть рассматривается как отдельный кластер. Данная группа методов называется дивизимными методами. Оба варианта иерархической кластеризации в большой степени зависят от того, как измеряется сходство между вершинами.

Не так давно был предложен метод, который стал своего родом прорывом в исследовании сообществ и подтолкнул ученых к развитию этого направления. Речь идет о функции модульности, которую предложили М. Ньюман и М. Гирван, и которая сейчас используется в большинстве алгоритмах выделения сообщество. Согласно этой функции «субграф представляет собой сообщество, если число ребер внутри субграфа превышает ожидаемое число внутренних ребер, которое этот субграф имел бы в нулевой модели» (Fortunato, 86). Модульность является «одновременно глобальным критерием определения сообщества, функцией качества и ключевым ингредиентом наиболее популярных методов кластеризации графа» (Fortunato, 86).

Алгоритмов выделения сообществ, в том числе использующих функцию модульности, в настоящее время разработано много, об этом свидетельствует обширный обзор С. Фортунато (Fortunato, 86). Однако не все они могут быть использованы в социологических исследованиях, поскольку зачастую не интегрированы в доступную к использованию неспециалистом программную среду.

В техническом плане вопрос ставится следующим образом: необходимо подобрать такое программное обеспечение, которое может реализовать методы выделения сообществ на больших сетях такого вида, который имеет сеть блогеров и постов в Живом Журнале, связанных комментариями.

Алгоритмы выделения сообществ в рамках сетевого анализа реализованы в программном пакете igraph, который мы использовали в программной среде R. Данная программная среда позволяет

работать с большими массивами данных, что делает её применение возможным в нашем случае, в отличие от многих видов программного обеспечения, рассчитанного на небольшие массивы.

Всего для igraph (версия 0.5.5-2, вышедшая в сентябре 2011 года) разработано и может быть использовано шесть алгоритмов выделения сообществ, два из них рассчитаны специально на работу с большими массивами. Мы опробовали работу всех шести алгоритмов на графе с известной структурой и оценили их применимость к «большой», полной сети блогеров и постов об исламе в ЖЖ. Критериями для признания алгоритма пригодным для использования в нашем исследовании является, во-первых, возможность получить информацию о принадлежности каждой вершины к определенному сообществу (кластеру, объединению вершин), а во-вторых, адекватность этого приписывания, которую мы оценивали, сопоставляя результаты работы алгоритма с визуальным изображением сети, упорядоченным согласно алгоритму Фруктермана-Рейнголда.

Результаты пробного исследования показали, что из двух алгоритмов, рассчитанных на работу с большими сетями, один показал неудовлетворительные результаты: алгоритм, реализуемый командой label.propagation.community, большинство вершин относит к отдельным сообществам (то есть рассматривает большинство вершин как отдельные сообщества). Возможно, данный алгоритм не пригоден для анализа редких сетей.

Второй из алгоритмов, предназначенный для выделения сообществ в больших сетях - fastgreedy.community, он требует более тщательной, чем в остальных случаях, подготовки данных для анализа, поскольку данный алгоритм более требователен, чем остальные, к структуре входных данных. Вместе с тем, он показал удовлетворительные результаты при разделении сети на сообщества, поэтому его мы считаем наиболее релевантным нашим задачам.

Из оставшихся четырех алгоритмов, возможно, в нашем исследовании может быть применен алгоритм walktrap.community. Он может быть использован в том случае, если размер итоговой сети не превысит несколько сотен тысяч вершин. Данный алгоритм менее требователен к структуре входных данных и показал свою эффективность в применении к тестовой сети.

Выводы

Таким образом, для преодоления концептуальных проблем формирования массива постов об исламе был применен многоступенчатый отбор, который на первом этапе состоял в формировании экспертами списка событий, имеющих отношение к исламу, далее происходил анализ текстов об этих событиях, с выделением лексики, относящейся к теме ислама. Список полученных слов и словосочетаний служит основой для запуска поисковых алгоритмов, а результаты поиска формируется массив данных

Выбор методов выделения сообществ в рамках сетевого анализа сделан в пользу алгоритмов, использующих характеристику модульности, как наиболее эффективно отражающих оптимальные варианты выделения сообществ как групп вершин, связанных между собой сильнее, чем с остальными вершинами сети.

Для преодоления технических проблем были выбраны следующие решения. Для обращения к данным, находящимся в Интернете, было разработано специальное программное обеспечение, позволяющее извлекать (закачивать) содержание Живого Журнала за выбранный период и из выкачанной информации извлекать массив данных для анализа согласно запросу исследователя. Согласно запросу можно получить данные о содержании поста, его маркере, а также маркерах автора и комментаторов, что позволяет решить задачи и семантического, и сетевого анализа.

Для анализа структуры сети – выявления сообществ – был выбран программный пакет igraph, который использовался в программной среде R. Данный программный пакет содержит алгоритмы выделения сообществ, которые могут применяться для анализа больших сетей и дают приемлемые результаты по выявлению сообществ. В результате определен алгоритм, наиболее приемлемый для выделения сообществ в больших сетях – fastgreedy.community.

В данной научной работе использованы результаты, полученные в ходе выполнения проекта «Разработка методологии сетевого и семантического анализа блогов для социологических задач», выполненного в рамках Программы «Научный фонд ГУ-ВШЭ» в 2011 году.

Библиография

1. S. Fortunato. Community detection in graphs. Physics Reports (2010) Volume: 486, Issue: 3-5.