Чем дышит блогосфера? К методологии анализа больших текстовых данных для социологических задач.

Кольцова Олеся
(Государственный Университет — Высшая Школа Экономики)

Значение Интернета и средств мобильной коммуникации для современных обществ трудно переоценить. Только в последний год сначала «арабские революции», а потом и протесты в России убедительно показали влияние Интернета даже в тех обществах, где доля его пользователей не самая большая и где гражданское общество никогда не было самым сильным. В связи с этим перед социологами встает ряд принципиально новых задач, связанных как с теоретическим осмыслением, так и с поиском методов анализа этих явлений. Казалось бы, данные в сети легко доступны для анализа, но на пути социологического анализа оказываются совершенно непривычнее проблемы.

Допустим, исследователю требуется построить репрезентативную выборку российских пользователей социальных сетей или блогов. Для этого ему потребовалось бы загрузить всех пользователей с их текстами, определить параметры репрезентативности (совсем не тождественные офф-лайновым) и автоматизировать построение выборки. Если среди важных параметров для выборки оказывается содержание (тематика) текстов пользователей, требуются математически и вычислительно сложные методы автоматизированного анализа текстов, воспроизводимые разве что в стенах Google или Яндекс. В результате социологический анализ интернета ограничивается, как правило, привычным социологам контент- или дискурс-анализом отдельных сайтов или группы сайтов с неясными критериями выбора.

В данной статье изложены промежуточные результаты большого исследования русскоязычных блогов, проводимого коллекти-

вом Санкт-Петербургской Лаборатории интернет-исследований ВШЭ. Содержательные задачи проекта были следующие: определить тематическую структуру российской блогосферы; выяснить, связано ли образование дискуссионных сообществ (если таковые имеются) с тематикой постов; выяснить, вокруг каких тем существует социальное напряжение. Однако главная задача была методологическая и заключалась в адаптации и апробировании инструментов для решения содержательных задач. Предметом этой статьи являются результаты изучения методов описания тематической структуры блогосферы и тестирования соответствующего программного обеспечения.

Что такое блог

Блог — это сайт, представляющий собой дневник, в котором автор располагает записи в обратном хронологическом порядке, возможно, с картинками, видео- и аудифайлами, ссылками. Отличие блога от новостной ленты — жанровое: блог предполагает индивидуальное авторство и, как правило, носит непрофессиональный или неофициальный характер. Каждый пользователь может самостоятельно создать или заказать сайт для блога, но львиная доля блогов находится на специальных блог-сервисах или блог-хостингах, предоставляющих простые конструкторы для создания блогов. Так, в русскоязычной блогосфере насчитывается около 53 миллионов блогов (что говорит о распространенности и, соответственно, социальной значимости этого явления); из них автономных блогов — чуть менее пяти миллионов [1]. Записи в блогах также называют постами; другие авторы блогов под своими именами (а люди, не ведущие блогов,— анонимно) могут оставлять комментарии к каждой записи; на некоторых блог-сервисах комментарии имеют древовидную структуру (т.е. можно отвечать на конкретный комментарий, а не на сам пост); на других они выстраиваются в линейку. Это определяет и разную структуру дискуссий.

Для русскоязычной блогосферы также характерно слияние блог-сервисов и социальных сетей. Ярким примером этого является Живой журнал: классический блог-хостинг предоставляет не функцию

дружбы, а функцию blog-roll, т.е. ссылок на понравившиеся блоги, зачастую независимо от сервиса, на котором они расположены. Поэтому, например, в США связность блогов зависит не от блогхостингов, на которых они расположены, а в большей степени от социальных факторов (например, общности тематики). В России функции френдования в гораздо большей степени замыкают коммуникацию между блогерами внутри одной блог-платформы. Можно предположить, что это свойство скоро перестанет быть специфически российским, т.к. в связи с миграцией пользователей в социальные сети последние активно впитывают в себя функции блогов.

Что такое тема

Хотя категория «тема» интуитивно понятна, но, по нашему опыту, при ручном кодировании текстов она вызывает большие трудности у кодировщиков. Повседневные тематические классификации дискурса не имеют ни четких оснований, ни общепринятых ad hoc представлений, и их очень трудно операционализировать для исследования однозначным, непересекающимся и исчерпывающимся образом. Темы могут образовываться вокруг событий, социальных проблем, беспроблемных явлений, постоянных и переменных сфер жизни, типов дискурса и систем ценностей. При разной «силе микроскопа», через который рассматривается корпус (коллекция, выборка) текстов (документов), обнаруживаются темы разных масштабов, т.е. возможны более и менее дробные классификации; часть тем вложены одни в другие, часть стоит особняком. В части случаев темы пересекаются и накладываются друг на друга в пределах одного и того же текста, а часть текстов монотематична. Состав тем, особенно в случае блогосферы, быстро меняется во времени, причем темы не только появляются и исчезают, но еще и дробятся, сливаются и мутируют. Наконец, часть текстов вообще не имеет темы и не поддается разумной классификации. Это особенно касается коротких текстов, отсылающих к другим текстам или внетекстовому контексту, напр.: «твою ж мать что вытворяет женя, кмв2 ахах молодец! лапа ахах». Данный текст содержит выраженное позитивное отношение к некому объекту, суть которого не ясна; будем называть такие тексты шумом. Несмотря

на все это, СМИ, поисковые системы и блог-платформы постоянно генерируют тематические классификации своего контента, а миллионы пользователей их успешно употребляют.

Как ни странно, наилучшее приближение к этим повседневным классификациям дают типологии текстов, построенные на основе понятия темы, определенного статистически, при помощи анализа частот слов и их совместной встречаемости в текстах. В таких подходах темы в общем виде понимаются как неявные единства слов, которые наиболее часто встречаются друг с другом в одних текстах. Например, если слова «Каддафи, Ливия, убить», часто употребляются вместе в разных комбинациях в одни и тех же текстах, то они формируют основу темы, которую можно озаглавить как «война в Ливии и смерть ливийского лидера». Такое понимание темы позволяет не только хорошо моделировать существующие представления о тематическом членении корпусов текстов (что экспериментально проверялось в разных исследованиях [2; 3], но и генерировать новые классификации, отвечающие исследовательским потребностям и поддающиеся разумной социологической интерпретации. Т.к. при таком подходе определения темы являются продуктом работы алгоритма разбиения текстов на группы, они неотделимы от этого алгоритма и будут рассмотрены вместе с соответствующими алгоритмами.

Хотя в данной статье мы не рассматриваем выявление дискуссионных сообществ вокруг тем, сами темы ищутся нами в контексте именно этой задачи, поэтому коротко поясним, что мы видим дискуссионное сообщество там, где группа постов со сходными характеристиками (например, на одну тему) комментируется примерно одним и тем же составом блогеров. Операционализируется это понятие так же, как и тема, алгоритмически, но с применением инструментов сетевого анализа (network analysis), где сообщество представляется как субграф в бимодальном графе постов и комментаторов, имеющий плотность большую, чем случайная.

Получение исходных данных и построение выборок

Для построения корректных выборок, как известно, нужно иметь представление о распределении оцениваемых параметров в генераль-

ной совокупности и технический доступ к выбранным единицам, что и в «офф-лайновой» социологии, как мы знаем, случается далеко не всегда. В области блогов есть свои особенности. Исчерпывающего списка блогов нет почти ни в одной стране или языковой зоне. России в этом смысле повезло: здесь существует публичный рейтинг блогов поисковой системы Яндекс, включающий все русскоязычные блоги, кроме отказавшихся от индексирования (включения в поиск) и тех, которые краулеру Яндекса не удалось найти. Учитывая, что влияние таких блогов на общественное мнение минимально, для большинства социологических задач ими можно пренебречь. Таким образом, в России теоретически возможно формировать случайные выборки пригодных для ручного анализа размеров.

Однако есть проблемы с составом и распределением параметров, участвующих в определении размеров выборок. Пол и возраст блогеров, во-первых, указаны далеко не всегда; во-вторых, если указаны, не ясно, что они на самом деле отражают; в-третьих, на самом деле пол и возраст авторов не являются характеристиками блогов. Ими являются количество и длина постов, частота обновления, возраст блога, количество друзей и комментариев и т.д. Их распределения изучены не до конца, хотя достаточно, чтобы в целом понимать, что большинство семантических и сетевых параметров распределено не нормально, а, в основном, по т.н. «безмасштабным распределениям» (Ципфа, Парето и др.)

Главный вопрос в том, какие задачи можно решать на основе таких «ручных» выборок. Оценка формальных количественных параметров, таких как перечислены выше, осуществляется самими поисковыми системами и отдельными блог-платформами на основе полных данных. По таким выборкам также можно оценить жанровые и тематические склонности авторов, но по ним нельзя судить о тематической структуре блогосферы и о связях и сообществах. Блоги в основном политематичны. Посты монотематичны или, во всяком случае, в существенной их части, в которую не входит шум, они представляют собой смесь очень ограниченного количества тем. Таким образом, поскольку тема — атрибут поста, единицей анализа тематического членения блогосферы является пост, и выборку следует делать из постов. А это уже гораздо более сложная задача.

На момент написания статьи русскоязычная блогосфера производит порядка 10^5 постов в день (с микроблогами — 10^6) и в несколько раз больше комментариев. Исчерпывающих списков записей, тем более списков за период времени, из которых можно было бы делать выборки, не существует. Формированию случайных выборок мешает, во-первых, отсутствие знаний о составе значимых параметров постов. Во-вторых — исследования, в которых было выяснено, что в больших текстовых коллекциях темы достаточно мелки, а их количество очень велико (на коллекциях 10^5 оно может измеряться десятками и сотнями [4]). Из этого следует необходимость большого размера выборок, что исключает не только ручной анализ, но и даже ручную закачку данных. А это, в свою очередь, означает необходимость создания специальных программных средств.

Казалось бы, простой альтернативой является ограничение выборки интересующей исследователя тематикой, определенной через ключевые слова. Такой подход используется не только в повседневной жизни, но и в маркетинге, где обычно требуется составлять выборки текстов о простых конкретных сущностях (например, марках товаров). Мы провели ряд экспериментов в отношении социальных тем. Экспертами составлялись списки событий, относящихся к широкой социально значимой теме, а кодировщики собирали посты об этих событиях и составляли списки ключевых слов, читая эти посты; в конечный список попадали слова с наибольшим коэффициентом согласия между кодировщиками. Результаты эксперимента отрицательные. От построения выборки поиском постов по получившимся ключевым словам в поисковой машине сразу пришлось отказаться, по следующим причинам: (а) ни один поисковик не воспринимает длинных списков, неизбежных при широких социальных темах, а при делении списков выдает дублирующиеся результаты, которые невозможно вычистить вручную при больших данных; (б) поисковики не выдают более 1000 страниц поиска; (в) главное — поисковики ранжируют результаты по непрозрачному алгоритму, в состав которого входит недавность публикации, ее популярность на момент поиска и другие не публикуемые поисковиками критерии релевантности. Таким образом, такая выборка не является не только репрезентативной, но и прозрачной по методике формирования, так что не ясно, какие выводы можно

делать на ее основе. Поэтому мы искали посты в нашей базе данных, созданной на основе сплошной закачки (об этом см. ниже), которая не ранжирует и не отсекает никаких результатов. В конечной выборке документов оказалось множество текстов, не имеющих никакого отношения к теме (например, содержащих омонимы к ключевым словам), в то время как, по мнению тех же кодировщиков, большая часть релевантных текстов не попала в выборку.

Гораздо более корректным является получение сплошной выборки постов блогосферы за определенный период с последующим автоматизированным делением на темы. Такие выборки можно получать по договоренности с поисковыми фирмами, а для самостоятельного формирования требуется написание отдельного ПО на каждую из блог-платформ (которых более ста) и сведение результатов в единую базу, что фактически означает создание мини-Гугла в домашних условиях, и большинству социологов, включая нас, недоступно. Поэтому мы решили ограничиться одной блог-платформой, а дополнительным аргументом стало то, что комментирование в российской блогосфере не выходит за пределы отдельных блог-платформ [5]. Таким образом, изучение множества платформ для другой нашей цели — поиска дискуссионно-комментовых сообществ — не только не полезно, но и даже вредно, т. к. влияние платформы на структуру сети будет гораздо сильнее, чем влияние всех социальных факторов.

Мы остановились на Живом журнале. Из предыдущих исследований [6; 7] известно, что социально-политическая тематика обсуждается наиболее активно именно в нем. По своему размеру ЖЖ замыкает четверку платформ-лидеров, имеющих свыше миллиона аккаунтов и составляющих вместе около пятой части русскоязычной блогосферы по числу аккаунтов [1]. При этом по активности ЖЖ абсолютный лидер, примерно на треть опережающий ближайшего конкурента. Еще одним аргументом было то, что рейтинг пользователей ЖЖ (он же — их исчерпывающий список) имел прозрачную методику и был основан только на количестве друзей. С сожалением можем констатировать, что это полностью не избавило нас от методологической непрозрачности, так как в ходе исследования ЖЖ изменил свой рейтинг в пользу учета активности, посещаемости и других неизвестных параметров, отчего рейтинг ЖЖ стал не менее непрозрачным, чем рейтинги Ян-

декса. Однако рейтинг пользователей — меньшее методологическое зло по сравнению с рейтингом постов. Во-первых, имея полный список аккаунтов, рейтинга в принципе можно избежать, закачивая либо всех пользователей, либо большую случайную выборку, и эти функции сейчас реализуются. Во-вторых, полная закачка всех постов даже только топовых блогеров, пусть непрозрачно отобранных, позволяет получать достаточно большие массивы постов, доступные для автоматизированного выделения тем и создания тематических карт пусть не всей блогосферы, но, во всяком случае, топа ЖЖ, которые можно экстраполировать на посты всех блогеров ЖЖ за период времени с помощью некоторых методов машинного обучения. После этого возможно корректное использование сетевого анализа для выявления сообществ.

Итак, нами было разработано ПО, которое закачивает в базу данных имена (ники) блоггеров (авторов постов), тексты и URL их постов с датами и временем и относящиеся к ним тексты комментариев с датами, временем и никами комментаторов. База не содержит картинок, аудио- и видео файлов и информации о дизайне блогов и не предназначена для визуального анализа. В базе данных реализован полнотекстовый поиск, который выдает все данные (а не только первую тысячу) и не рейтингует их. База также позволяет делать случайные и пошаговые выборки, выборки по дате и др., конвертирует выборки в форматы ряда пакетов для текстового и сетевого анализа, т.е. приспособлена для социологических задач. На данный момент нами сформировано три выборки, включающие все посты и комментарии топ-2000 блоггеров за периоды: 15 августа — 15 сентября 2011 (спокойный период), 27 ноября — 27 декабря 2011 (вокруг парламентских выборов) и 4 февраля — 4 марта 2012 (перед президентскими выборами; период после них в стадии формирования). Периоды выбраны исходя из исследований жизненных циклов новостей в СМИ и в Интернете [8;9]. Основные эксперименты проводились на декабрьской выборке в 28252 постов и августовской выборке в 24074 постов.

Основные проблемы алгоритмов анализа текстов.

Наша методологическая цель — найти методы разделения текстов на тематические группы, дающие наилучшее качество при ре-

шении наших социологических задач. Общими проблемами всех алгоритмов анализа — будь то сетевой или текстовый — является соотношение качества и вычислительной сложности (O). Последняя оценивается приблизительно, как функция от количества данных, а в простейшем случае — коэффициент, на который надо умножить количество данных, чтобы получить количество условных шагов работы алгоритма и таким образом оценить время его работы. O — не просто вопрос того, сколько времени будет работать компьютер, но также вопрос того, сможет ли он работать вообще (например, хватит ли у него оперативной памяти). Особенно критичным это является для анализа текстов; так, по данным наших экспериментов, популярный у социологов R не справляется с выборками наших размеров.

Вторая проблема — оценка качества анализа. Как определить, хорошие, правильные ли получились кластеры либо выявленные сообщества? Существуют две основные группы методов оценки качества работы различных алгоритмов: (а) внешние — определение доли «правильно» отнесенных единиц через сравнение с образцом, и (б) внутренние — вычисление ряда параметров, таких как соотношение внутрикластерной и межкластерной дисперсии и десятки других функций. Для методов анализа текстов ведущими методами являются внешние, основанные на сравнении с образцовым корпусом, разделенным на группы вручную с помощью кодировщиков (например, чистота, точность, F-мера, энтропия и их модификации). Проблемой этого подхода является распространенность некритичного отношения к результатам кодирования и проблематичность экстраполяции результатов, полученных на одних типах образцовых корпусов, на другие типы (например, другой тематики), а также невозможность их применения на больших гетерогенных коллекциях, где требуется ручная обработка десятков тысяч текстов. Следует отметить, что методы оценки качества различных алгоритмов анализа, как и сами алгоритмы широко дебатируются в математическом сообществе. Поэтому социолог сталкивается с проблемой выбора алгоритма из набора средств, надежность которых до конца не установлена. В нашей работе мы придерживались выбора тех алгоритмов, по которым проводилось хоть какое-то тестирование.

Способы оценки качества также очень важны для определения наилучшего количества групп, на которое следует разделять коллекцию текстов — будь то кластеры в классическом кластерном анализе или темы в алгоритмах тематического моделирования. При тематическом картировании блогосферы заранее невозможно определить, сколько там «на самом деле» групп (если придерживаться позитивистских подходов) или же сколько групп даст исследователю наиболее удобную и познавательную картину (если придерживаться более конструктивистских подходов). Один из возможных выходов — выбор между разбивками на разное количество групп на основе оценки качества каждой из разбивок. Правда, проблема состоит в том, что все известные функции оценки качества как внешней, так и внутренней, монотонно изменяются с ростом числа групп. Поэтому очень непросто определить точку скачка функции, после которого прирост качества резко уменьшается, что могло бы служить сигналом для прекращения наращивания числа групп. Существует ряд математических подходов к решению этой проблемы, например, обзор и результаты сравнения множества функций остановки иерархической кластеризации даны в известной статье Миллигана и Купер [10]. Две функции, которые по итогам этого исследования оказались наиболее удачными — Calínski & Harabasz pseudo-*F* index и Duda & Hart Je (2)/Je (1) index — присутствуют в нескольких стандартных статистических пакетах, таких как STATA или SAS, но, к сожалению, нам не удалось найти алгоритмов со встроенной функцией определения количества кластеров, которые были бы реализованы в каком-либо ПО, способном работать с большими массивами текстовых данных. И, конечно, эти классические функции на таких массивах не тестировались. Нами был взят и запрограммирован в виде отдельного кода один из современных подходов, позволяющих находить скрытые скачки в функциях качества кластеризации [11]. Сейчас его работа проверяется на размеченных вручную коллекциях.

Наконец, еще одна серьезная проблема автоматического анализа текстов — автоматизация «лейбелинга» групп. На первый взгляд она кажется побочной. Однако само по себе получение списка из нескольких сотен групп, в каждой из которых по нескольку тысяч текстов, ничего не прибавляет к знанию исследователя о коллекции текстов и о ее

тематике, даже когда алгоритм работает качественно и быстро. Если для определения тематики каждой группы требуется вручную перечитать все тексты, автоматизированный анализ обесценивается. Можно назвать несколько видов «подсказок» исследователю, которые алгоритмы в принципе способны генерировать: списки наиболее частотных слов или фраз, информация о центроиде («главном» тексте группы) и о расстояниях от других текстов до него, или о вероятности принадлежности текста группе, что позволяет строить списки топ-текстов и читать только их. К сожалению, как отмечают Карпинето и соавт. [12], хотя качество разделения и качество лейбелинга не являются напрямую конкурирующими параметрами, на практике разработчики алгоритмов концентрируются либо на одном, либо на другом. Причем академические разработчики нацеливаются на большие объемы данных и качество, а коммерческие — на лейбелинг и скорость в ущерб больших объемов, и выход из этой ситуации не прост.

Вычисление сходства между текстами

Главная задача тематического картирования — сформировать группы текстов, сходных по тематике, и затем изучить отношения между ними, но что такое более или менее похожие тексты? Здесь возможны два основных подхода. Первый подход заключается в том, что экспертами (между которыми достигнута высокая надежность интеркодирования) определяются образцы текстов — скажем, «про выборы», «проправительственный», «оппозиционный» и т.д. Затем алгоритм анализирует частотно-лексические характеристики этих текстов и экстраполирует получившиеся наборы признаков на новые тексты, раскладывая их по группам, к которым каждый текст находится ближе всего. Эту операцию принято называть классификацией, т.к. она не предполагает поиска латентных групп, а лишь делит корпус на заранее известные. Так же как и кластеризация, она может быть полной или неполной, четкой или нечеткой. В нашем исследовании мы предполагаем, что основной ценностью разрабатываемой методологии может стать возможность находить именно латентные группы, которые могут иметь потенциал неожидаемых социальных изменений. Поэтому классификация для нас является менее предпочтительной процедурой.

Второй подход — это формальное вычисление сходства. В классическом кластерном анализе используется вариант, основанный на представлении текста в векторной форме (описание см. напр. в [13]). При обработке больших массивов тексты представляются в виде «мешка» слов, точнее, их лемм (корней) или начальных форм, частоты которых подсчитываются в каждом тексте и располагаются в таблице, называемой матрицей терминов-документов. Далее в векторном подходе каждая лемма представляется в виде измерения в N-мерном пространстве, где N — общее количество уникальных лемм, встречающихся в корпусе. Каждый текст представляется в виде вектора в этом пространстве; частоты лемм в данном тексте соответствуют длине проекции вектора на ось соответствующего данной лемме измерения. Такие вектора становятся сравнимыми. Есть несколько способов вычисления расстояния между ними, однако в анализе текстов принята косинусная мера — вычисление косинуса многомерного угла между каждой парой векторов. Эта мера привилегирует разницу в угле перед разницей в длине, т.е. обращает большее внимание на наличие/отсутствие общих слов, чем на сходство/различие в частоте общих слов. Вычисленные расстояния между векторами записываются в матрицу расстояний, или различий.

Одна из проблем векторного и других частотных подходов — т. н. «проклятие многомерности». Подавляющее большинство слов в любом корпусе встречается в ничтожно малой доле текстов, а еще заметная часть встречается везде; ни те, ни другие не имеют дискриминационной силы, а лишь увеличивают бесполезный размер матрицы, утяжеляют вычисления и ухудшают его результаты. Есть разные способы уменьшения размерности матриц как математические, так и «механические», через «отрезание» редких и частых слов. Использованное нами «отрезание» ста самых частотных слов и всех слов, встречающихся менее, чем в пяти текстах, существенно сжимает матрицу, при этом часть текстов оказываются пустыми; каково значение этого эффекта, требует дальнейшего изучения.

Классический кластерный анализ

Поскольку социологам хорошо знакомы основные виды кластерного анализа (плоская, восходящая и нисходящая), мы не бу-

дем их описывать; вместо этого скажем несколько слов о современных алгоритмах. Считается, что все виды кластеризации имеют постоянный ряд достоинств и недостатков. Так, известный алгоритм k-means и производные зависимы от выбора начальных точек и поэтому не дают стабильных результатов, могут останавливаться на субоптимальных решениях и вычислительно сложны (см. напр [13]). Однако на практике используются не виды кластеризации, а конкретные итеративные алгоритмы, действительное качество и быстродействие которых зависит от многих деталей. Так, при кластеризации текстов важно следующее: какая мера близости текстов используется (косинусная, эвклидова, другая); как при плоской кластеризации или на каждом шаге иерархической кластеризации рассчитываются расстояния между кластерами, как оптимизируются и оптимизируются ли какие-либо шаги, распределяются ли объекты по кластерам однозначно или с коэффициентами принадлежности к нескольким кластерам (нечеткая кластеризация) и т.д. Существуют десятки алгоритмов, совершенствующих основные виды кластеризации и предлагающих новые (обзор см. [12]). Назовем две основные новые группы.

Первую называют генеративными алгоритмами, или алгоритмами, основанными на распределениях, или алгоритмами, основанными на вероятностных моделях. В основе этих алгоритмов лежат какие-либо предположения о распределении параметров данных массива, которые можно представить в виде суммы распределений параметров в субмассивах. Задача таких алгоримтов — отнести каждый текст к субмассиву (возможно, нечетко) на основе сравнения распределений параметров внутри текста с распределениями их же в смоделированных субмассивах [14].

Во вторую группу можно объединить алгоритмы, основанные на анализе матриц и графов. Так, математические способы уменьшения размерности матриц можно использовать не только для ее «чистки» от шума, но и как средство кластерного анализа (это называют спектральной кластеризацией). Если таким способом кокластеризовать одновременно и документы, и слова [15], то получится алгоритм, очень схожий с алгоритмом следующего поколения — LSA (см. в следующем разделе). Кроме того, матрица может быть пред-

ставлена в виде полного графа, где тексты — вершины, а расстояния — взвешенные ребра, а к графу применимы как алгоритмы спектрального деления графов, так и не связанные с матричными вычислениями алгоритмы выявления сообществ, понимаемых как кластеры.

Социологу во множестве этих алгоритмов легко потеряться; часть из них не тестировалась совсем, а часть на разных массивах данных давала совершенно разные результаты, поэтому выбор алгоритма в конечном итоге должен определяться тем, как данный алгоритм работает именно на изучаемом массиве или близких к нему тестовых массивах. Таким образом, оказалось, что для наших задач необходимо ПО, которое позволяет тестировать качество алгоритмов, работает с большими текстовыми данными (10^4–10^5 текстов) на кириллице, осуществляя их самостоятельную закачку и препроцессинг (чистку, лемматизацию, векторизацию и др.). Среди более чем сорока изученных пакетов такого ПО найти не удалось; большая часть ПО не содержит информации о своих алгоритмах и не рассчитано на большие объемы данных. Единственным известным нам пакетом кластеризации, работающим с большими объемами, является gCLUTO (George Karypis Lab, университет Миннесота) [16]. Он не осуществляет препроцессинга и с трудом поддался настройке на кириллицу, однако в нем реализовано четыре разных алгоритма (direct — вариант плоской кластеризации, agglomerative, repeated bisection и graph), несколько мер близости текстов, несколько функций расчета расстояний между кластерами, оптимизируемых в иерархической кластеризации (criterion functions); опция вычисления нескольких внутренних функций качества (внутри- и межкластернация дисперсия и т.д.) и двух внешних функций качества — энтропии и чистоты, которые можно применять для выбора параметров алгоритмов, если есть образцовая коллекция. По gCLUTO авторами проведено множество тестов, в т.ч. на данных высокой размерности (текстах), подробно описанных в публикациях [17; 18].

Нашими кодировщиками была вручную составлена выборка из трехсот русскоязычных постов, принадлежащих к трем сильно отличающимся темам, которую мы ввели в gCLUTO. На основании опубликованных авторами тестов [17; 18; 19] мы выбрали для тестирования два алгоритма — agglomerative и repeated bisection, косинус-

ную меру близости и две критериальные функции, называемые авторами *I2* и *H2* и показавшие наилучшие результаты в их тестах. На наших данных лидирует repeated bisection в сочетании с *H2* (энтропия 0,14 по сравнению с 0,47–0,6 у других сочетаний; чистота 0,94 по сравнению с 0,62–0,75 у других), а использование методов автоматического определения количества кластеров позволяет надеяться на хорошее качество и при работе с большими коллекциями. Однако именно при работе с большими коллекциями gCLUTO сталкивается с практически неразрешимой проблемой интерпретации кластеров. В качестве «подсказки» gCLUTO выдает только четыре наиболее частотных слова, по которым не удается определить тематику кластера. Тексты внутри кластеров не ранжированы, информации о центроидах нет. Поэтому следующую серию экспериментов мы провели на ПО, представляющем иной тип алгоритмов деления текстов на группы.

Альтернативы классической кластеризации

Другое направление алгоритмов выявления тематических групп представляет тематическое моделирование. Если кластерный анализ развивался как статистическая процедура для группирования разных объектов в разных дисциплинах, то тематическое моделирование, как не трудно догадаться из названия, возникло в сфере автоматического анализа текстов. Оно было предназначено не только для разбивки корпуса текстов на группы, однако с успехом применяется и для этой задачи. Основные подходы в порядке появления один из другого — латентно-семантический анализ (LSA) [20], вероятностный латентно-семантический анализ (pLSA) [21] и латентное размещение Дирихле (LDA) [4], каждый из которых представлен целым рядом алгоритмов с различными усовершенствованиями.

Все это направление можно считать развитием логики факторного анализа; по крайней мере, наиболее типичная для LSA процедура уменьшения размерности матрицы является математической генерализацией факторного анализа [21, с. 8]. При этом, как говорилось выше, LSA сходен со спектральной ко-кластеризацией, которая также кластеризует одновременно тексты и слова через уменьшение размерности матрицы.

Все перечисленные подходы основываются на предположении о том, что совместная встречаемость текста t и слова w (проще — появление слова w в тексте t) объясняется латентными переменными, похожими на факторы, которые в применении к анализу текстов можно считать темами. Т.е. если текст t и слово w принадлежат к одной теме, они «встретятся». Исходными данными для всех тематических подходов является матрица терминов-документов. Информация о сходствах и различиях между текстами и между словами является результатом работы этих алгоритмов, с которым можно поступить по-разному, например, кластеризовать. LSA получает эту информацию через ряд операций по уменьшению размерности матрицы терминов-документов. Конечным продуктом LSA являются матрицы сходств между текстами, между словами и между текстами и словами; при решении задачи разбивки текстов на группы последующая кластеризация неизбежна.

LDA и pLSA, несмотря на сходство названия последнего с LSA, относятся к другому классу — классу генеративных вероятностных моделей (sf генеративные алгоритмы кластеризации). Эти подходы рассматривают каждый текст как смесь латентных переменных (тем), к каждой из которых текст принадлежит с разной вероятностью. Также смесью тем являются и слова, каждое из которых тоже принадлежит к каждой теме с разной вероятностью. Таким образом, тема является смесью слов, принадлежащих к ней с разной вероятностью, и «фактором», в отношении которого оценивается вероятность того, что именно он «породил» данный текст. pLSA и LDA отличаются в основном предположениями о распределениях указанных вероятностей, причем вероятностные модели, используемые в LDA, считаются более точными, т.е. лучше моделирующими реальные данные, и, кроме того, они отличаются меньшей вычислительной сложностью [22, с.11]. Конечным продуктом LDA являются матрица вероятностей принадлежности слов к темам и матрица вероятностей принадлежности текстов к темам. Для задачи разбиения на тематические группы последнюю матрицу можно кластеризовать, но если нет задачи безостаточного распределения текстов по группам (как у нас) можно, например, взять в каждую группу тексты с вероятностью принадлежности к ней выше определенного порога или просто топ n текстов.

В наших экспериментах мы использовали ПО Stanford Topic Modeling Toolbox (TMT) [23]. Этот пакет, в отличие от большинства других, написан специально для социальных исследователей; хотя он не очень прост в освоении, зато имеет открытый код и поддается настройке. Он без проблем воспринимает кириллицу, имеет многие встроенные функции препроцессинга (кроме лемматизации), встроенную функцию внутренней оценки качества получаемого решения — меру неопределенности (perplexity), возможность использования части коллекции как обучающей, на основании которой затем производится оценка другой части коллекции, а также функцию анализа изменений тематической структуры во времени. В качестве лейбелинга TMT выдает список топ-20 слов с их весами принадлежности к теме (вес — функция от вероятности) и вес «значимости» самой темы, являющийся суммой весов всех слов по теме. Кроме того, поскольку TMT выдает полные матрицы вероятностей текстов в темах и весов слов в темах, легко самостоятельно составлять списки топ-слов и топ-текстов любой длины, необходимой для анализа. Одной из проблем этого ПО является недостаток инструкций с точным описанием того, как работают алгоритмы, но это отчасти компенсируется открытостью кода. Существует короткий обзор ПО для социальных исследователей [24] и доклады с результатами экспериментов по применению алгоритма [25].

В TMT нами вводилась та же коллекция, в то и в gCLUTO (август-сентябрь), и декабрьская коллекция. Обе анализировались с разным количеством тем, и строился график изменения меры неопределенности, которая, что неудивительно, изменяется довольно монотонно. Д. Блеем, автором LDA, разработан метод оценки количества тем [26] и код для него, однако код рассчитан на программистов. Поэтому мы опирались в своем выборе на визуальный анализ функции (так же, как в [9]) и на собственное ручное кодирование осмысленности тем. При поверхностном просмотре топ-слов и топ-текстов, входящих в темы с наибольшей вероятностью, складывается впечатление общей осмысленности результатов. Так, большинству тем легко приписать названия на основании топ-20 слов, а топ-тексты чаще всего соответствуют этим названиям; в декабре по сравнению с августом возрастает доля и вес тем, связанных с выборами и протестами; часть из них ясно привязана к конкретным персонажам и событиям.

Однако мы столкнулись с проблемой более точной оценки качества работы этого алгоритма на наших данных (и, соответственно, с проблемой сравнения его с методами кластерного анализа). LDA не работает с маленькими коллекциями, поэтому у нас не было возможности проверить его на размеченных нами трехстах текстах и пока не было возможности закодировать несколько тысяч текстов. Кроме того, даже если бы такая возможность была, LDA все равно относит каждый текст к каждой теме, так что процедура прямого сравнения ручного и машинного отнесения к группе невозможна.

Мы провели ручной лэйбелинг ста тем декабрьской и августовской выборки на основании сначала топ-20, а потом топ-30 текстов, кодирование простоты лейбелинга, а также исследовали некоторые статистические свойства соотношения текстов и тем в декабрьском и августовском массивах. Около 15% тем содержит общую лексику в топ-словах, разнородные или бессмысленные тексты и не поддается лейбелингу, тогда как некоторые выглядят очень цельными. Наличие бессмысленных тем могло бы говорить о слишком большом заданном числе тем, однако среди цельных тем есть такие, которые касаются острых социальных вопросов, но при этом не видны при членении на меньшее количество (мы также проводили разбиение на 30 и на 50 тем). Т.к. при еще более дробном членении количество неинтерпретируемых тем сильно возрастает, предварительно можно сказать, что данное количество является разумным компромиссом.

«Бессмысленные» и «цельные» темы имеют разные статистические свойства. В среднем, в обоих массивах данных алгоритм относит с ненулевой вероятностью к каждой теме по 6.8% текстов (по 1921 на массиве 28253 текста и по 1680 на массиве 24074). Распределение «размеров» тем показано на гистограмме 1. Больше половины случаев отнесения этих текстов к темам имеет вероятность менее 0,1 (поскольку отнесение множественное, общее количество отнесений больше количества текстов; в среднем каждый текст относится к 7 темам; распределение см. гистограмма 2). Случаев отнесения к какой-либо теме с вероятностью больше 0,5 всего 5%. Есть очень четкая связь между размером и осмысленностью темы. Почти все большие темы (более трех тысяч текстов) неинтерпретируемы;

и наоборот, среди неинтерпретируемых большая доля больших тем. Также для «бессмысленных» тем характерна малая (менее средней) доля отнесений с высокой степенью вероятности, хотя здесь связь более слабая. Наоборот, в наиболее «цельных» темах такая доля выше. Самыми «цельными» темами являются темы, собирающие тексты на украинском языке (у них минимальное количество общих с другим текстами слов), на английском или русско-английской смеси, на «компьютерно-английской» смеси, а также календарь, кулинарные рецепты и темы, содержащие много перепостов одного и того же текста (например, спам). Из политических тем очень цельной является тема ареста Удальцова. Большинство текстов в топе этой темы рассказывают именно об этом событии или комментируют его, а меньшинство посвящено арестам Навального и Яшина. Все три персонажа — политические активисты, арестованные за участие в митингах за честные выборы. В целом, социально-политические темы по количеству отнесенных к ним текстов и по цельности прижимаются к середине списка. И по числу таких тем, и по количеству текстов, к ним относимых, они занимают около трети тематического пространства.

Важно, что кроме цельных и неинтерпретируемых тем есть большой класс того, что можно было бы назвать «склеенными» темами. В них прослеживаются две или более темы, притянувшиеся друг к другу на основе общей лексики. Например, рассказы о совершенно разных, не связанных друг с другом преступлениях, притягиваются на основании наличия общих слов, типичных для криминальной хроники. Бывают и более отдаленные «склейки». В частности, в декабрьской выборке есть тема, объединяющая дело коммерсанта Барановского, обвиняемого в финансовых преступлениях, и разнородные события из исламских регионов и стран на основании того, что Барановский — ветеран-афганец. Такие темы нельзя назвать неинтерпретируемыми, но они требуют большей ручной работы. В них часто список топ-слов не совпадает с содержанием топ-текстов, т. к. в топ-20 слов могла попасть лексика из одной подтемы, а в топ-20 или топ-30 текстов — посты из другой. Тексты, соответствующие топ-словам, могут находиться по второй (третьей, четвертой) двадцатке, равно как и слова, соответствующие топ-текстам, могут находиться ниже в списке. При этом подтемы легко вычленимы, так

что эта вычленимость вкупе с резким несоответствием топ-20 слов и топ-20 текстов является признаком «склеенной» темы. В то же время в неинтерпретируемых темах определить тематическую область как топ-текстов, так и топ-слов не удается вообще.

Таким образом, рассмотренные тематические членения содержат как хорошо, так и плохо интерпретируемые группы текстов, однако первые превалируют. В целом на данный момент при помощи TMT удалось получить более интерпретируемые данные, чем при помощи gCLUTO.

Заключительные замечания

Главное значение LDA и сходных подходов для социальных исследователей состоит в том, что они позволяют быстро разбивать большие, не поддающиеся ручному чтению массивы текстов на легко интерпретируемые темы и выделять для анализа только то, что отвечает задачам исследования, таким образом сократив объем текстов для чтения на один-два порядка. Так, если мы ставим задачу определения наиболее «горячих» социально-политических тем в блогах через сравнение разных периодов, мы выделяем темы, специфические для данного времени (для декабря 2011 — это протесты и выборы) и получаем весь спектр текстов в виде небольшой «выжимки» в несколько сотен постов. Их легко проанализировать качественными методами на предмет выделения подтем, жанров, социально-коммуникативных функций, эмоциональной заряженности и др. В целом же автоматизированный анализ больших текстовых данных настолько молодая отрасль, что отладка процедур таких исследований в социальных науках потребует еще не один год и усилий множества людей.

ЛИТЕРАТУРА

1. URL: http://blogs.yeandex.ru.
2. *Biro I.* Document Classification with Latent Dirichlet Allocation. PhD thesis. Eötvös Loránd University.— Budapest, 2009.
3. *Zhao Y., Karypis G.* Evaluation of Hierarchical Clustering Algorithms for Document Datasets. CIKM '02 Proceedings of the eleventh international conference on information and knowledge management. ACM New York.— N.Y., USA, 2002.
4. *David M. Blei, Andrew Y. Ng, Michael I. Jordan.* Latent Dirichlet allocation//Journal of Machine Learning Research 3 (Jan). 2003.— P. 993–1022.

5. *Этлинг Б., Алексанян К., Келли Дж., Палфри Дж., Гассер У.* Публичный дискурс в российской блогосфере: анализ публичной политики и мобилизации// Исследования центра Беркмана. 19 октября 2010, № 2010–11.— URL: http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/Public_Discourse_in_the_Russian_Blogosphere-RUSSIAN.pdf. English language original: http://cyber.law.harvard.edu/publications/2010/Public_Discourse_Russian_Blogosphere.
6. *Alexanyan K., Koltsova O.* Blogging in Russia is not Russian blogging/Russel A. Echchaibi N. (eds.) International Blogging: Identity, Politics and Networked Publics. Peter Lang, 2009.
7. *Gorny E.* Russian LiveJournal: National specifics in the Development of a Virtual Community. Version 1.0 of 13 May 2004. Russian-cyberspace.org. –
URL: http://www.ruhr-uni-bochum.de/russ-cyb/library/texts/en/gorny_rlj.pdf.
8. *Koltsova O.* Coverage of Social Problems in St.Petersburg Press/Cecilia von Feilitzen & Peter Petrov (eds). Usa and Views of Media in Sweden & Russia, 2011.
9. *Wu S., Hofman J. M., Mason W., Watts D. J.* Who Says What to Whom on Twitter//International WWW Conference 2011, March 28 — April 1.— Hyderabad, India, 2011. Copyright 2011 ACM 978–1-4503–0637–9/11/03.
10. *Milligan G. W., Cooper M. C.* An Examination of Procedures of Determining the Number of Clusters in Data Set. Psychometrika. June 1985. Vol. 50. №. 2.— C. 59–179.
11. *Sugar C., James G.* Finding the Number of Clusters in a Data Set: An Information Theoretic Approach//Journal of the American Statistical Association. 2003, № 98.— P. 750–763.
12. *Carpineto C., Osiński S., Romano G., Weiss D.* A Survey of Web Clustering Engines. ACM Computing Surveys (CSUR). 2009. Vol. 41. Issue 3. Article № 17.
13. *Andrews N.O, Fox E. A.* Recent Developments in Document Clustering, October 16, 2007.— URL: http://eprints.cs.vt.edu/archive/00001000/01/docclust.pdf.
14. *Ahlquist J. S., Breunig C.* Model-Based Clustering and Typologies in the Social Sciences. Political Analysis. 2011. Vol. 20. Issue 1.— P. 92–112.
15. *Kummamuru K., Dhawale A., Krishnapuram R.* Fuzzy Co-clustering of Documents and Keywords. FUZZ '03: 12th IEEE international conference on fuzzy systems.— 2003. P. 772–777.
16. *George Karypis* Lab, страница gCLUTO с обзором, публикациями и ПО.— URL: http://glaros.dtc.umn.edu/gkhome/cluto/gcluto/overview.
17. *Rasmussen M., Karypis G.* gCLUTO: An Interactive Clustering, Visualization, and Analysis System. UMN-CS TR-04–021, 2004.
18. *Zhao Y., Karypis G.* Emprical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering. Machine Learning, 55. 2004.— P.311–331.
19. *Zhao Y., Karypis G.* Hierarchical Clustering Algorithms for Document Clustering. Data Mining and Knowledge Discovery. 2005. Vol. 10. № 2.— P. 141–168.
20. *Landauer, T. K., Foltz, P. W., & Laham, D.* Introduction to Latent Semantic Analysis. Discourse Processes. 25. 1998.— P. 259–284.
21. *Hoffman T.* Probabilistic Latent Semantic Analysis. Uncertainty in Artificial Intelligence. UAI'99.— Stockholm. 1999.
22. *Daud A., Li J., Zhou L., Muhammad F.* Knowledge Discovery Through Directed Probabilistic Topic Models: a Survey. Frontiers of Computer Science in China. 2010. Vol. 4. Issue 2.— P. 280–301/пер. К.В. Воронцов, А.В. Темлянцев и др.— URL: www.machine-learning.ru/wiki/images/9/90/Daud2009survey-rus.pdf.
23. The Stanford Natural Language Processing Group, страница TMT с кодом

и инструкцией. — URL: http://nlp.stanford.edu/software/tmt/tmt-04/.
24. *Ramage D., Rosen E., Chuang J., Manning C. D., McFarland D. A.* Topic Modeling for the Social Sciences. NIPS 2009 Workshop on Applications for Topic Models.
25. *Ramage D., Dumais S., Liebling D.* Characterising Microblogs with Topic Models. ICWSM 2010. — URL: http://www.stanford.edu/~dramage/papers/twitter-icwsm10.pdf.
26. Teh Y. W., Jordan M. I., Beal M. J., Blei D. M. Hierarchical Dirichlet processes//Journal of the American Statistical Association. 2004. Vol. 101, N. 476. P. 1566–1581.