ОНЛАЙН-ИССЛЕДОВАНИЯ

О.Ю. Кольцова, К.А. Маслинский
(*Санкт-Петербург*)

ВЫЯВЛЕНИЕ ТЕМАТИЧЕСКОЙ СТРУКТУРЫ РОССИЙСКОЙ БЛОГОСФЕРЫ: АВТОМАТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ТЕКСТОВ[1]

В статье изложены методологические результаты исследования русскоязычных блогов. Произведена адаптация и апробация автоматизированных методов анализа текстов и соответствующего программного обеспечения для решения содержательных задач (выявление тематической структуры блогосферы, описание ее изменений во времени, выявление процесса образования дискуссионных сообществ). Выделяются и описываются два класса методов деления больших массивов текстов на группы – кластерный анализ и тематическое моделирование; из каждой группы выбирается и апробируется программное обеспечение (ПО). Эксперименты проводятся на двух массивах данных в 10^4 постов каждый. Обосновывается выбор в пользу тематического моделирования, представлено описание полной технологической цепочки от сбора до социологического анализа данных.

Ключевые слова: Интернет, блоги, методология социологического исследования, кластеризация текстов, тематическое моделирование, «большие данные».

Олеся Юрьевна Кольцова – кандидат социологических наук, руководитель лаборатории интернет-исследований Высшей школы экономики (Санкт-Петербургский кампус). E-mail: ekoltsova@hse.ru.
Кирилл Александрович Маслинский – научный сотрудник лаборатории социологии образования и науки Высшей школы экономики (Санкт-Петербургский кампус). E-mail: kmaslinsky@hse.ru.

[1] Работа выполнена при поддержке Научного фонда НИУ ВШЭ, грант № 11-04-0006 «Разработка методологии сетевого и семантического анализа блогов для социологических задач». Авторы выражают благодарность С. Кольцову, разработчику ПО сбора данных для данного исследования.

Введение

Только в последние два года 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 новых задач, связанных как с теоретическим осмыслением, так и с поиском методов анализа этих явлений. Казалось бы, данные в сети легкодоступны, но на пути исследования оказываются пока непривычные для социологов проблемы, связанные с этапом как сбора, так и анализа данных. Во-первых, это проблема доступа к генеральной совокупности исходных данных, не опосредованных поисковыми системами с их непрозрачными алгоритмами отбора и сортировки. Во-вторых, это проблема операционализации социологических понятий в терминах интернет-данных, которая ставит перед исследователями целый спектр методологических вопросов, связанных с определением границ объекта и выбором единиц анализа, которыми могут выступать, например, тексты, изображения, гиперссылки, профили пользователей, связи между аккаунтами в социальных сетях. В-третьих, это проблема больших выборок. Есть несколько факторов, которые подталкивают к увеличению размера выборок; среди них большой объем доступных данных, а также, что важно в нашем случае – тематическое разнообразие текстов. Большие объемы данных приносят и новые требования к аналитическим методам, и это четвертый вызов, стоящий перед социальными учеными.

Данное исследование одной из основных задач ставит разработку методологии, позволяющей преодолеть указанные проблемы. Содержательной задачей, на решение которой направлена создаваемая методология, является изучение содержания российской блогосферы как новой формы общественного мнения – мнения интернет-активной части общества – и его изменений во времени.

Исследование сосредоточено на блогплатформе «Живой журнал», которая, по мнению некоторых авторов, вмещает основную часть общественно-политических блогов [5]. Методологические результаты, изложенные в данной статье, носят промежуточный характер.

Статья имеет следующую структуру. Вначале предлагается операционализация понятий блогосферы как площадки для общественно-значимых обсуждений и темы как способа анализа содержания дискуссий. Затем обсуждаются принципы ограничения объекта исследования, методика сбора данных и их структура. После этого анализируются основные группы алгоритмов, которые были задействованы для анализа данных, их возможности и ограничения; описывается, каким образом они применялись к задачам исследования. В заключение описываются методологические результаты исследования.

Основные понятия: блог, блогосфера и тема

Блог – это сайт, представляющий собой дневник, где автор располагает записи в обратном хронологическом порядке. Блог предполагает индивидуальное авторство и, как правило, носит непрофессиональный или неофициальный характер. Значительная доля блогов находится на специальных блог-сервисах или блог-хостингах, предоставляющих простые конструкторы для создания и ведения дневников. Блогосферой называют всю совокупность существующих блогов. Так, в русскоязычной блогосфере насчитывается около 53 млн блогов; из них автономных блогов – чуть менее пяти миллионов [1]. Записи в блогах также называют постами; другие авторы могут оставлять комментарии к каждой записи; на некоторых блог-сервисах комментарии имеют древовидную структуру.

Для русскоязычной блогосферы характерно слияние блог-сервисов и социальных сетей. Ярким примером этого может слу-

жить «Живой журнал» (ЖЖ): 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 блог-хостингов, 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 [5]. Можно предположить, что это свойство скоро перестанет быть специфически российским, так как в связи с миграцией пользователей в социальные сети последние активно впитывают в себя функции блогов.

Подавляющее большинство записей в блогах и комментариев к ним представляет собой тексты. Так как в нашей работе обсуждается методология анализа содержательной стороны обсуждений, происходящих в блогосфере, можно операционализовать понятие блогосферы как совокупность постов и комментариев к ним, т.е. как текстовую коллекцию.

В качестве категории, позволяющей описывать содержательную общность текстов и оценивать количественные тенденции в текстовой коллекции, естественно использовать тематику. *Тему текста* можно рассматривать как интуитивно ясную категорию, однако наш опыт кодирования как медийных текстов, так и блогов показывает, что тема – характеристика текста, по которой кодировщики демонстрируют невысокую степень согласованности. За внешней очевидностью определения темы скрывается принципиальная проблема выбора уровня обобщения. Впрочем свободу в выборе уровня тематического обобщения можно рассматривать не как методологическую проблему, а как исследовательский инструмент – «микроскоп», сквозь линзы которого рассматривается коллекция текстов. Возможность изучать одну и ту же коллекцию текстов на разных уровнях тематической детализации открывают, в частности, алгоритмические методы анализа текстов.

Тематическую общность текстов можно операционализировать через наличие общих слов в текстах, или, в масштабах всей коллекции, как наличие групп слов, которые имеют тенденцию встречаться совместно в одних и тех же текстах. Такое формальное определение темы позволяет группировать тексты и/или слова в коллекции в тематические группы автоматически, без опоры на знания о языке, на котором написаны тексты, и без предложенной исследователем классификации тем. Формальный подход к операциональному определению темы ставит проблему *интерпретации* полученных автоматически групп слов и текстов. Например, если в автоматически выделенной теме наиболее частотными оказываются слова «Каддафи, Ливия, убить», данную тему можно озаглавить как «война в Ливии и смерть ливийского лидера». Правомерность такой тематической атрибуции должна подтверждаться анализом текстов, отнесенных к этой группе.

Указанное формальное определение тематики позволяет моделировать существующие представления о тематическом членении корпусов текстов, что было показано во многих эмпирических исследованиях (см., например: [2, 3]). Поскольку конкретные определения темы являются продуктом работы алгоритмов разбиения текстов на группы, они будут рассмотрены вместе.

Исходные данные и построение выборки

Естественная единица для построения выборки из блогосферы – это блог. Однако блоги в основном являются политематическими, поэтому продуктивнее рассматривать тему как атрибут отдельного поста, что позволяет группировать сходные по содержанию записи разных блогеров. Выявление тематически близких записей, опубликованных в небольшой промежуток времени, может служить основной для изучения структуры и динамики общественного обсуждения в блогосфере. Таким образом, для нашего исследования конечной единицей наблюдения и отбора является пост, а не блог.

На данном этапе мы решили не включать в тематический анализ комментарии, ограничившись анализом только той тематики, которую выбирают сами авторы блогов.

На момент написания статьи русскоязычная блогосфера без учета микроблогов производит порядка 10^5 постов в день и в несколько раз больше комментариев [1], раскиданных по разным платформам, что остро ставит проблему построения выборки из такой генеральной совокупности**.** Приближением к такой совокупности можно считать публичный рейтинг блогов поисковой системы Яндекс, включающий все русскоязычные блоги, а также микроблоги (*Twitter*), кроме отказавшихся от индексирования (включения в поиск) и тех, которые поисковому роботу Яндекса не удалось найти. Учитывая, что ожидаемое влияние на общественное мнение не проиндексированных блогов минимально, для большинства социологических задач ими можно пренебречь. Однако исчерпывающих списков записей, на основании которых можно было бы строить выборки постов, не существует. В этой ситуации для построения выборки можно прибегнуть к двухступенчатой выборочной процедуре: построить выборку блогов на основании рейтинга и затем отбирать посты из уже выбранных блогов.

Поскольку для автоматизированного сбора блогов требуется разработка специализированного ПО для каждой блог-платформы (которых более ста), мы решили ограничить выборку одной блог-платформой – «Живой журнал». Из предыдущих исследований [5; 6; 7] известно, что социально-политическая тематика обсуждается наиболее активно именно здесь. 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 [1]. ЖЖ публикует собственный рейтинг русскоязычных блогов (он же — их исчерпывающий список, который на момент исследования был методологически прозрачным), мы

использовали именно его для построения сплошной выборки постов за интересующие нас периоды времени.

Для автоматизированного сбора, хранения, сортировки данных ЖЖ и построения выборок из них в лаборатории интернет-исследований НИУ ВШЭ было разработано программное обеспечение *Koltran Blogminer*. Это ПО осуществляет автоматическую закачку в базу данных на основе MS SQL 2005 следующей информационной структуры: блоггер – посты – комментарии, относящиеся к заданному посту, с учетом древовидной структуры. База не содержит изображений, аудио- и видео файлов и информации о дизайне блогов и не предназначена для визуального анализа. Также ПО позволяет делать простые случайные и систематические вероятностные выборки, выборки по дате, по ключевым словам и др., конвертирует выборки в форматы ряда пакетов для текстового и сетевого анализа. В данной работе использованы две выборки (две «закачки»), включающие все посты топ-2000 блоггеров, но не более 50 последних постов на каждого блогера[1], за периоды: 15 августа – 15 сентября 2011 (контрольный «спокойный» период, для сравнения с периодом активизации политических дискуссий, 24 074 постов) и 27 ноября – 27 декабря 2011 г. (политизированный период вокруг парламентских выборов, 28 252 постов). Периоды выбраны исходя из исследований жизненных циклов новостей в СМИ и в Интернете [8, 9].

Проблемы выбора алгоритмов анализа текстов

Качество алгоритмов, выбор числа тем и лейбелинг

Наша методологическая цель – найти методы разделения текстов на тематические группы, дающие наилучшее качество

[1] Ограничение в 50 последних постов установлено сервером «Живого журнала», выдающим данные по внешним запросам; оно было преодолено в более поздних исследованиях, где главными были не методологические, а содержательные задачи.

при решении социологических задач. Одна из ключевых проблем на пути к этой цели – оценка качества работы алгоритма. Как определить, хорошие, правильные ли получились кластеры либо выявленные темы? Можно условно выделить две группы методов оценки качества работы различных алгоритмов:

– *внешние*: определение доли «правильно» отнесенных единиц через сравнение с образцом. При анализе текстов чаще используются именно такие алгоритмы, 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 стала, в частности, 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;

– *внутренние*: вычисление ряда параметров, таких как соотношение внутрикластерной и межкластерной дисперсии. Эти алгоритмы основаны на формальных допущениях и не имеют эмпирического референта. Когда размеченные коллекции недоступны, внутренние меры – единственный способ оценки качества. Среди их недостатков – тот факт, что, опираясь на заложенную в конкретный алгоритм специфику, они могут быть неприменимы к другим алгоритмам и, следовательно, сравнение принципиально разных алгоритмов на их основе может быть затруднено.

Следует отметить, что методы оценки качества различных алгоритмов анализа находятся в стадии становления [24]. Поэтому социолог сталкивается с проблемой выбора алгоритма из набора средств, надежность которых до конца не установлена. Мы придерживались здесь выбора тех алгоритмов тематической классификации, качество которых уже тестировалось на реальных текстовых коллекциях.

Способы оценки качества классификации также очень важны для определения оптимального количества групп, на которые сле-

дует разделять коллекцию текстов, – будь то кластеры в кластерном анализе или темы в алгоритмах тематического моделирования. При тематическом картировании блогосферы заранее невозможно определить, какое количество групп даст исследователю наиболее удобную и познавательную картину. Один из возможных выходов – выбор между разбиениями на разное количество групп на основе оценки качества каждого из разбиений. Правда, проблема состоит в том, что все известные функции оценки качества, как внешние, так и внутренние, 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. В нашем исследовании мы использовали теорию скачков [10], которая позволяет вычленить такую точку в функции качества кластеризации, после которой прирост качества резко падает и таким образом наращивание числа кластеров можно остановить. Метод основан на использовании анализа функции «искажений» (*distortions*)**,** в которой параметром является число кластеров. В роли функции качества кластеризации можно использовать функцию минимального (максимального или среднеарифметического) значения внутрикластерной дисперсии от числа кластеров. Эта функция не нормированная; наименьшее ее значение соответствует наименьшему разбросу объектов внутри кластера и, следовательно, наилучшему качеству.

В тематическом моделировании широко используется другая мера качества (к которой мы применили теорию скачков – *perplexity* [4; 27, p. 78]), переводимая в зависимости от контекста как мера неопределенности, мера неуверенности или показатель несвязности (далее — перплексивность). Перплексивность изменяется от 0 до 1; наименьшее значение соответствует наилучшему качеству. Эта мера рассчитывается в ряде пакетов для тематического моделирования (например: [21, 22]). Для практической реализации теории скачков нами было разработано специальное программное

обеспечение, которое позволяет определять оптимальное кластерное решение.

Другая серьезная проблема автоматического анализа текстов – автоматизация «лейбелинга» (именования) тематических групп текстов – кластеров или тем. Само по себе получение списка из нескольких сотен групп, в каждой из которых по несколько 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 соавторы [11], качество разделения и качество лейбелинга не могут служить напрямую конкурирующими параметрами, на практике разработчики алгоритмов концентрируются либо на одном, либо на другом.

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 достигнут высокий уровень согласия при классификации) определяются образцы текстов – скажем, «про выборы», «проправительственный», «оппозиционный». 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 в компьютерной лингвистике и машинном обучении принято называть классификацией [12]. В нашем исследовании мы предполагаем, что основной ценностью разрабатываемой методологии может стать возможность находить именно латентные группы, которые могут иметь потенциал не ожидаемых исследователем социальных изменений. Поэтому процедура классификации, предполагающая деление корпуса текстов на заранее известные классы, для нас менее предпочтительна.

Второй подход – это формальное вычисление сходства. В кластерном анализе текстов используется вариант, основанный на представлении текста в векторной форме (см., например: [12]). 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 чаще всего используется косинусная мера – вычисление косинуса многомерного угла между каждой парой векторов [26, с. 296]. Эта мера обращает внимание на сходство/различие в соотношении частот слов в сравниваемых текстах, чем на сходство/различие в абсолютной частоте слов, и таким образом позволяет улавливать сходство между лексически близкими текстами разной длины. Вычисленные расстояния между векторами записываются в матрицу расстояний, или различий.

Одна из проблем векторного и других частотных подходов – так называемое проклятие размерности [25]. Подавляющее большин-

ство слов в любом корпусе встречается в ничтожно малой доле текстов, а еще заметная часть встречаются 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 исключение ста самых частотных слов и всех слов, встречающихся менее чем в пяти текстах, существенно сжимает матрицу, при этом часть текстов оказываются пустыми; вопрос, каково значение этого эффекта для качества получаемых решений, требует дальнейшего изучения.

Инструментарий: алгоритмы и ПО для тематической классификации

Кластерный анализ

Социологам хорошо знакомы основные виды «классического» кластерного анализа, при реализации которого объекты сравниваются между собой напрямую на основе заданной меры близости и по результатам этого сравнения размещаются в заданное число групп, поэтому детальное описание этих алгоритмов не входит в нашу задачу. Считается, что все виды «классической» кластеризации имеют ряд достоинств и недостатков: так, известный алгоритм *k*-средних и производные алгоритмы плоской кластеризации зависимы от выбора начальных объектов и поэтому могут останавливаться на субоптимальных решениях (см., например: [12]). Однако на практике используются не виды кластеризации, а конкретные итеративные алгоритмы, действительное качество и быстродействие которых зависит от многих деталей. При кластеризации текстов, в частности, важно: 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).

Существуют также десятки алгоритмов, совершенствующих основные виды кластеризации и предлагающих новые (обзор см., например: [11]). Назовем две основные новые группы.

Первая – генеративные алгоритмы; они основаны на предположении о том, что каждый кластер описывается одинаковыми функциями распределения документов внутри него (например, Гауссово распределение или распределение фон Мизеса-Фишера). Каждый кластер характеризуется своей величиной математического ожидания и дисперсией, а совокупность кластеров определяется ковариационной матрицей. Процедура кластеризации представляет собой итеративный процесс, в рамках которого производится оценка вероятности принадлежности документа тому или иному кластеру на основе математического ожидания, дисперсии и ковариационной матрицы. Цель итеративного процесса – максимизация величины суммы всех вероятностей по всем документам.

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 ко-кластеризовать одновременно и документы, и слова (см., например: [13]), то получится алгоритм, очень схожий с алгоритмом *LSA*, описанным далее. Кроме того, матрица может быть представлена в виде полного графа, где тексты – вершины, 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 он работает именно на изучаемом массиве или близких к нему тестовых массивах. В нашем случае ПО должно позволять тестировать качество алгоритмов, работать 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 рассчитана на большие объемы данных. Единственный известный нам пакет кластеризации, работающий с большими объемами, это *gCLUTO* (*George Karypis Lab*, университет Миннесота) [14]. Данный пакет не осуществляет препроцессинга и с трудом поддался настройке на кириллицу. Среди его преимущество – поддержка четырех разных алгоритмов (*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 некоторые другие) и двух внешних функций оценки качества разбиения – энтропия и чистота (понимаемая как средняя доля доминирующего класса в каждом кластере), которые можно применять для выбора параметров алгоритмов, если есть образцовая коллекция. По *gCLUTO* авторами проведено множество тестов, в том числе на данных высокой размерности (текстах), подробно описанных в [15; 16].

Латентно-семантический анализ и тематическое моделирование

Другой тип алгоритмов выявления тематических групп представляет тематическое моделирование. Если кластерный анализ развивался как статистическая процедура для группирования

разных объектов в разных дисциплинах, то тематическое моделирование возникло в сфере автоматического анализа текстов. Основные подходы в порядке появления один из другого – латентно-семантический анализ (*LSA*) [18], вероятностный латентно-семантический анализ (*pLSA*) [19] и латентное размещение Дирихле (*LDA*) [4]. Каждый представлен целым рядом алгоритмов с различными усовершенствованиями.

Все это направление условно можно считать развитием логики факторного анализа [19, с. 8]. Эти 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. Иными словами, если в тексте *t* присутствует тема, к которой относится слово *w*, текст *t* и слово *w* «встретятся».

Для *LSA*, как и для описанных в предыдущем разделе алгоритмов кластеризации, исходными данными служит матрица терминов-документов. Латентные факторы вычисляются за счет ряда операций по сокращению размерности матрицы терминов-документов. Конечным продуктом LSA являются матрицы сходств между текстами, между словами и между текстами и словами.

LDA и *pLSA*, несмотря на сходство названия последнего с *LSA*, относятся к другому классу – генеративных вероятностных моделей (ср. генеративные алгоритмы кластеризации). Эти подходы рассматривают каждый текст как смесь нескольких 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 – это смесь 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 вероятност-

ные модели, используемые в *LDA*, считаются более точными, т.е. лучше моделирующими реальные данные [20, 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 можно, например, взять в каждую группу тексты с вероятностью принадлежности к ней выше определенного порога или просто «топ-n» текстов. Кроме того, сумма вероятностей всех слов в теме служит количественным показателем ее важности, «веса» в коллекции.

В наших экспериментах мы использовали ПО *Stanford Topic Modeling Toolbox* (*TMT*) [21]. Этот пакет, в отличие от большинства других, написан специально для социальных исследователей; хотя он не прост в освоении, зато имеет открытый код и поддается настройке. Он без больших проблем воспринимает кириллицу, имеет многие встроенные функции препроцессинга (кроме лемматизации), встроенную функцию внутренней оценки качества получаемого решения – перплексивность, 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 тем *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. Одна из проблем этого ПО – недостаток инструкций с точным описанием того, как работают алгоритмы, но это отчасти компенсируется открытостью кода; существует короткий обзор ПО для социальных исследователей [22] и доклады с результатами экспериментов по применению алгоритма (см., например: [23]).

Эксперименты

Кластерный анализ

Для тестирования *gCLUTO* на русскоязычных данных, нашими кодировщиками была вручную составлена выборка из трехсот русскоязычных постов, принадлежащих к трем сильно отличающимся темам. На основании публикаций по теме [15; 16; 17] мы выбрали для тестирования два алгоритма – аггломеративный (*agglomerative*) и повторного разбиения пополам (*repeated bisection*), косинусную меру близости и две критериальные функции, называемые авторами тестов *I*2 и *H*2 и показавшие наилучшие результаты на их данных. Наилучшие результаты при кластеризации нашего массива данных показывает *repeated bisection* в сочетании с *H*2 (в частности, энтропия 0,14 по сравнению с 0,47-0,6 у других сочетаний).

На основе этой комбинации нами было получено по шесть кластерных решений для каждой из коллекций (август-сентябрь 2011 г., 24 074 поста, и декабрь, 28 252 поста), с числом кластеров от 50 до 300, т.е. с шагом 50, а также с числом кластеров от 90 до 140 (с шагом 10). С помощью нашего ПО мы нашли скачки в функции качества, называемой ISim (среднее расстояние между всеми парами объектов внутри всех кластеров данного решения). Оптимальным решением для сентября оказалось число кластеров, равное 120; для декабря – 130.

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, и *gCLUTO* – это не пакет с открытым кодом, который можно было бы подправить. Попытка назначать лейблы кластерам на основе четырех слов и случайных текстов привела к очень большому числу случаев, когда кодировщики затруднялись интерпретировать значение кластеров. Поэтому мы отказались от дальнейшего использования *gCLUTO* и

от трудоемкой работы по поиску средств сравнения его качества с качеством алгоритма *LDA*, на котором и сосредоточили свои усилия.

Тематическое моделирование (LDA)

По аналогии с кластерным анализом нами был получен ряд решений *LDA* сначала с шагом 50, а затем 9 решений на более узком промежутке от 80 до 160 тем с шагом 10. Разработанное нами ПО позволило рассчитать скачки функции перплексивности, которые оказались чрезвычайно малы; разница между первым и вторым по величине скачками – в седьмом знаке после запятой. Наибольший скачок в сентябре 2011 г. наблюдается на 130 темах, в декабре – на 140 темах; вторые по величине скачки в обоих периодах наблюдаются на 100 темах. Поскольку одной из содержательных задач исследования было сравнение периодов, для дальнейшего анализа нами были выбраны решения, соответствующие вторым по величине скачкам, так как они содержат одинаковое количество тем и более пригодны для сравнительного анализа.

При поверхностном просмотре топ-слов (*табл.* 1) и топ-текстов (*табл.* 2), входящих в темы с наибольшей вероятностью, складывается впечатление общей осмысленности результатов. Так, большинству тем легко приписать названия на основании топ-20 слов. Например, тема, наиболее вероятные слова в которой: *суд, судья, адвокат, уголовный, судебный, Барановский, прокурор*, легко обозначается как «судебные разбирательства». Тексты, с наибольшей вероятностью отнесенные к этой теме представляют собой официальную и неофициальную судебную хронику, публикуемую в блогах. Другие темы охватывают здоровье, домашних животных, фотографию, туризм, школу и т. п. В декабре 2011 г. по сравнению с августом возрастает доля и вес тем, связанных с выборами и протестами; в части из них среди топ-20 слов фигурируют упоминания конкретных персонажей и событий. Впрочем ядро некоторых тем составляет общая лексика, не позволяющая отнести данную тему к той или иной предметной области, например: *вопрос, более, именно, иметь, проблема, сторона, важный, решение*.

Судя по отнесенным к этой и аналогичным темам текстам, они характеризует тип дискурса (в данном случае проблематизирующие рассуждения), а не тематику в обыденном понимании. Наконец, в части тем объединяется лексика, которая совместно встречается в текстах по каким-то формальным причинам. Таковы темы, в топ-20 слов которых встречаются названия дней недели (сюда попадают тексты, включающие календари), имена, числа, текстовые элементы интерфейсов (*кнопка, оглавление, щелчок, ссылка)*. Для части таких тем не удается определить, какие формальные свойства текстов в коллекции привели к выделению данной темы.

Таблица 1

ПРИМЕР НАИБОЛЕЕ ВЕРОЯТНЫХ СЛОВ ДВУХ ТЕМ
(декабрь 2011 г.)

Тема 1	25397	Тема 2	29420
украина	1003	немцов	905
президент	611	путин	477
российский	532	навальный	368
украинский	440	собчак	286
лукашенко	234	оппозиция	252
глава	229	разговор	209
министр	211	болотный	204
виктор	180	божен	200
союз	178	народ	197
янукович	175	выходить	196
государство	170	сахаров	195
тимошенко	166	борис	184
владимир	164	выступать	171
заявлять	163	кургинян	166
беларусь	163	проспект	163
совет	161	приходить	161
ядерный	161	ксения	158
безопасность	158	лидер	152
москва	147	освистывать	152
белорусский	147	называть	139

Таблица 2

ПРИМЕР ТОП-ТЕКСТОВ ОДНОЙ ИЗ ТЕМ
(декабрь 2011 г.)

Текст	Вероятность появления текста в теме
Из опубликованных ФСБ телефонных разговоров Немцова с Яшиным, Пархоменко, Пономаревым, Панюшкиным и другими я понял только то, что у Бориса Ефимовича какие-то недопонимания с Евгенией Чириковой и что он реально не хотел подставлять людей под омоновские дубинки	1,0
Из чужого твиттера: один чувак отнес Навальному LJUSERnavalny в изолятор мандарины, чурчхелы, сулугуни и лаваш. И сопроводил это запиской: «от Кавказа который хватит кормить»	0,94
А чо, теперь самые главные революционеры – это собчак, канделака и боженарынска? Суркофф – замечательный разводчик всё-таки!	0,88
Белые ленточки. Путин. Выборы. Беспредел. Митинг на Болотной. Посмотрим. Набигут ли спамеры	0,87
После того как выложили записи Немцова – надо точно идти 24-го на митинг. Пусть все знают, что я трусливое офисное хомячье! Ура, товарищи!	0,83

В ходе этой предварительной интерпретации результатов стало понятно, что планировавшаяся ранее оценка качества алгоритма по тому, насколько он правильно относит конкретные тексты к темам, не оптимальна. Основным достоинством алгоритма оказалось его способность выявлять «яркие» и ясные темы-факторы и определять их вес в коллекции, т.е. определять «повестку дня» текстовой коллекции, а не имитировать то, как люди распределили бы тексты по группам. Поэтому и меры оценки качества из кластерного анализа здесь не подошли бы – что, однако, сделало

невозможным прямое сравнение кластерного анализа и тематического моделирования. Для оценки ясности тем на данный момент предложен только трудоемкий метод экспериментов с участием людей [27], который мы заменили простым кодированием легкости интерпретации тем.

Таким образом, мы провели ручной лэйбелинг ста тем декабрьской и августовской выборки на основании топ-30 текстов, кодирование простоты лейбелинга, а также исследовали некоторые статистические свойства соотношения текстов и тем в декабрьском и августовском массивах. Около трети (24, на которые приходится 37,6% веса в августе, и 28 с 33,7% веса в декабре 2011 г.) тем содержит общую лексику в топ-словах, разнородные или бессмысленные тексты и не поддается лейбелингу, тогда как топ-20 слов остальных тем представляют собой группы лексики, достаточно однозначно связанные с определенной предметной областью (как в приведенном выше примере темы «суд»). Примечательно, что среди тем, соотнесенных с некоторой предметной областью, есть темы, которые касаются острых социальных вопросов, но при этом не видны при менее дробном тематическом членении (мы также проводили разбиение на 30 и на 50 тем). Тем не менее увеличение числа тем нельзя считать безусловно предпочтительным, так как при дробном членении возрастает и количество неинтерпретируемых тем.

В целом степень интерпретируемости темы можно рассматривать как градуированную величину: от тем, не поддающихся интерпретации, которые приходится рассматривать как информационный шум, к «дискурсивным» темам и темам, характеризующим отдельные предметные области («цельные» темы). Провести четкую границу между этими типами тем во многих случаях затруднительно, и, с нашей точки зрения, не очень корректно. В среднем, при имеющихся настройках в обоих массивах данных алгоритм относит с ненулевой вероятностью к каждой теме по 6,8% текстов (по 1 921 на массиве 28 253 текста и по 1 680 на массиве 24 074 текстов). Распределение «размеров» тем показано на *рис*. 1. Больше по-

ловины 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. *рис*. 2). Случаев отнесения к какой-либо теме с вероятностью больше 0,5 – всего 5%. Наблюдается довольно четкая связь между размером темы и простотой ее интерпретации. Почти все большие темы (более трех тысяч текстов) – сформированы в основном общей лексикой и должны быть отнесены к классу «дискурсивных». Для таких тем также характерна малая (меньше средней) доля отнесений с высокой степенью вероятности, хотя здесь связь более слабая. Наоборот, в темах, связанных с конкретной предметной областью, такая доля выше. Самыми «цельными» оказались темы, собирающие тексты на украинском языке (у них минимальное количество общих с другими 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). Одна из политических тем очень четко соотносится с обсуждением конкретного события – 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 однозначности интерпретации располагаются в середине списка. И по числу таких тем, и по количеству текстов, к ним относимых, они занимают около трети тематического пространства.

В большей части тем, соотносимых исследователями с некоторой предметной областью, прослеживается упоминание двух или более событий, объединенных в одной теме благодаря общей лексике, характерной для данной предметной области. Например, рассказы о совершенно разных, не связанных друг с другом пре-

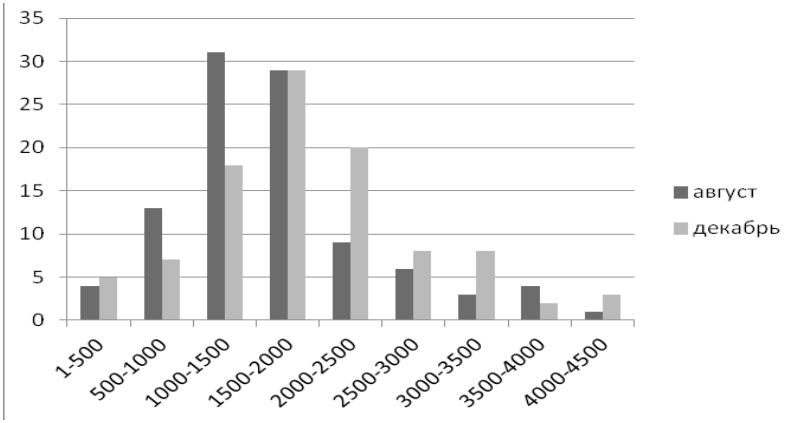

Рис. 1. Распределение размеров тем

Примечание:
Ось X: количество текстов, отнесенных к теме с ненулевой вероятностью
Ось Y: число тем с таким количеством текстов

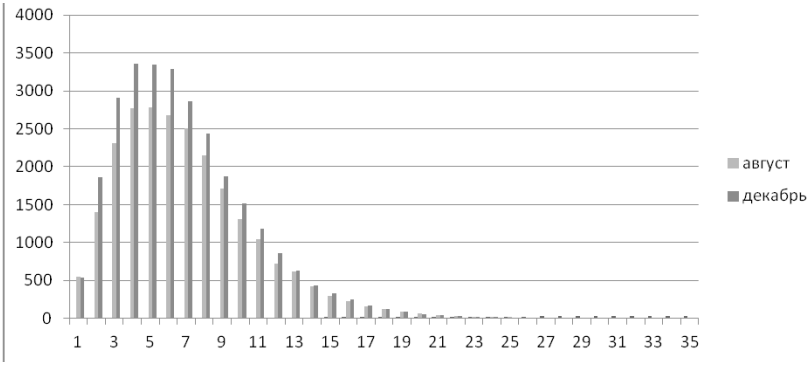

Рис. 2 . Распределение текстов по темам

Примечание:
Ось X: число тем, к которому одновременно отнесен один текст
Ось Y: количество текстов, отнесенных к данному числу тем

ступлениях притягиваются друг к другу 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 позволяет предсказать содержание топ-текстов, так как в топ-20 слов могла попасть лексика из одной подтемы, а в топ-30 текстов – посты из другой. Тексты, соответствующие топ-словам, могут находиться во второй (третьей, четвертой) двадцатке, равно как и слова, соответствующие топ-текстам, могут находиться ниже в списке. При этом подтемы легко вычленимы, что – наряду с несоответствием топ-20 слов и топ-20 текстов – может служить косвенным признаком «склеенной» темы.

Заключение

Алгоритмы тематического моделирования решают задачу определения тематической структуры коллекции текстов более явным образом, чем алгоритмы кластеризации, формируя темы-факторы как списки слов и подбирая к ним наиболее типичные тексты. Задача безостаточного и точного распределения текстов по группам для определения общей повестки дня оказывается излишней, а задача определения соотношения тем по важности кластерным анализом напрямую не решается вообще. С этой точки зрения тематическое моделирование предпочтительнее кластерного анализа. К недостаткам тематического моделирования стоит отнести неразвитость (по сравнению с кластерным анализом) внешних мер качества и неясность критериев их развития. И кластерный анализ, и тематическое моделирование могут быть с успехом использованы для быстрого вычленения

поддающихся ручному анализу выборок из больших массивов данных: например, в нашем случае среди 100 тем в декабре 2011 г. 15 относилось к выборам и протестам, так что чтение всего нескольких сотен топовых текстов дало представление о характере освещения данной тематики в блогах «Живого журнала». При оценке перспектив использования тематического моделирования для анализа больших текстовых массивов следует принимать во внимание то обстоятельство, что различные модификации метода стремительно развиваются. Так, уже предложены алгоритмы, отслеживающие рост и упадок тем во времени, а также сочетающие выявление тем с анализом их эмоциональной окрашенности [28]. Дальнейшая настройка и адаптация таких алгоритмов исследователями для решения своих задач представляется важным направлением развития социологических методов работы с большими текстовыми данными.

ЛИТЕРАТУРА

1. Яндекс-блоги. URL: http://blogs.yandex.ru (дата обращения 05.04.2012).

2. *Biro I.* Document Classification with Latent Dirichlet Allocation. PhD thesis. Budapest: Eötvös Loránd University, 2009.

3. *Zha, 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, 2002.

4. *Blei D.M., Ng A.Y., Jordan M.I., Lafferty J.* Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. No. 3. 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 No 2010-11, 19 октября 2010 г. URL: http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/Public_Discourse_in_the_Russian_Blogosphere-RUSSIAN.pdf (дата обращения 17.04.2012).

6. *Alexanyan K., Koltsova O.* Blogging in Russia is not Russian blogging // International Blogging: Identity, Politics and Networked Publics / Ed. A. Russel, N. Echchaibi. N.Y.: 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 (дата обращения 05.04.2012).

8. *Koltsova O.* Coverage of Social Problems in St.Petersburg Press // Use and Views of Media in Sweden & Russia / Ed. C. von Feilitzen, P. Petrov Stockholm: Sodertorn University, 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, 2011, Hyderabad, India.

10. *Sugar C., James G.* Finding the Number of Clusters in a Data Set: An Information Theoretic Approach // Journal of the American Statistical Association. 2003. No. 98. P. 750–763.

11. *Carpineto C., Osiński S., Romano G., Weiss D.* A Survey of Web Clustering Engines // ACM Computing Surveys (CSUR). 2009. Vol. 41. Iss. 3. No. 17.

12. *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 (дата обращения 17.04.2012).

13. *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.

14. gCLUTO – Graphical Clustering Toolkit. URL: http://glaros.dtc.umn.edu/gkhome/cluto/gcluto/overview (дата обращения 19.04.2012).

15. *Rasmussen M., Karypis G.* gCLUTO: An Interactive Clustering, Visualization, and Analysis System // UMN-CS TR-04-021, 2004.

16. *Zhao Y., Karypis G.* Emperical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering // Machine Learning. 2004. Vol. 55. P. 311–331.

17. *Zhao Y., Karypis G.* Hierarchical Clustering Algorithms for Document Clustering // Data Mining and Knowledge Discovery. 2005. Vol. 10. No. 2. P. 141–168.

18. *Landauer T.K., Foltz P.W., Laham D.* Introduction to Latent Semantic Analysis // Discourse Processes. 1998. Vol. 25. P. 259–284.

19. *Hoffman T.* Probabilistic Latent Semantic Analysis // Uncertainty in Artificial Intelligence, UAI'99. Stockholm, 1999.

20. Обзор по вероятностным тематическим моделям / Пер. с англ. К.В. Воронцова, А.В. Темлянцева и др. URL: http://www.machinelearning.ru/wiki/images/9/90/Daud2009survey-rus.pdf (дата обращения 19.02.2012).

21. Stanford Topic Modeling Toolbox // The Stanford Natural Language Processing Group. URL: http://nlp.stanford.edu/software/tmt/tmt-0.4/ (дата обращения 19.04.2012).

22. *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. URL: http://vis.stanford.edu/papers/topic-modeling-social-sciences (дата обращения 19.04.2012).

23. *Ramage D., Dumais S., Liebling D.* Characterising Microblogs with Topic Models // ICWSM. 2010. URL: http://www.stanford.edu/~dramage/papers/twitter-icwsm10.pdf (дата обращения 19.04.2012).

24. *Wallach H., Murray I., Salakhutdinov R. & Mimno D.* Evaluation Methods for Topic Models // Proceedings of the 26th International Conference on Machine Learning. Montreal, 2009.

25. *Bellman R.E.* Dynamic Programming. Princeton, NJ: Princeton University Press, 1957.

26. *Manning C., Schutze H.* Foundations of Natural Language Processing. Cambridge: The MIT Press, 1999.

27. *Chang J., Boyd-Graber J., Wang C., Gerrish S., Blei D.M.* Reading Tea Leaves: How Humans Interpret Topic Models // Neural Information Processing Systems, 2009. Vol. 22. P. 288–296.

28. *Li F., Huang M., Zhu X.* Sentiment Analysis with Global Topics and Local Dependency // Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10). 2010. Atlanta, USA, July 11–15, 2010. P. 1371-1376.