Статистический и тематический профиль
«Живого журнала»[i]

О.Ю.Кольцова, С.Н.Кольцов
Лаборатория Интернет - Исследований, НИУ Высшая Школа Экономики
ekoltsova@hse.ru; skoltsov@hse.ru

Аннотация
Цель этого исследования - выработка системы основных статистических показателей, характеризующих русскоязычную блогосферу с точки зрения задач исследований в области социологии, маркетинга, лингвистики и ряда других отраслей научного и практического знания, а также изучение основных взаимосвязей между ними. Необходимость сбора такой базовой статистики обусловлена ее отсутствием в публичном доступе – в отличие, например, от социо-демографических показателей населения, без которых не возможны социологические и маркетинговые исследования, или различных словарей русского языка, без которых были бы затруднены исследования в области языкознания. В данном исследовании не ставится задача проверки конкретных гипотез и получения законченных содержательных результатов; вместо этого, исследуются возможности решения различных исследовательских задач с помощью ряда предлагаемых показателей и на основании знаний о связи между ними. Рассматриваются такие показатели, как дата и время поста, длина, тематика и комментируемость поста, строятся временные циклы активности блогеров и их тематические профили. Тематика постов извлекается методами автоматического тематического моделирования, применение которых в сочетании с анализом других данных представляет собой основную новизну работы.

1. Данные

Исследование основано на данных российской блог-платформы «Живой журнал» - лидера по активности, измеряемой в постах в день, и одной из шести лидирующих блог-платформ, которые вместе аккумулируют около 90% блог-аккаунтов в русскоязычном пространстве Интернета. В анализ вошли четыре временных периода, включающие все посты 2000 блоггеров, которые на момент сбора данных занимали первые 2000 мест в рейтинге популярности Живого журнала, измеряемого как функция от количества друзей блоггера, с учетом читательской активности этих друзей. Фокусирование данного исследования на топовых блогерах обусловлено их повышенным влиянием на публичное пространство, однако в дальнейшем планируется их сравнение с «обычными» блогерами с целью выявления общих тенденции блогосферы и факторов, приводящих блогеров в топ рейтинга.

Для осуществления планомерных загрузок было разработано специализированное программное обеспечение "БлогМайнер" по клиент - серверной технологии на основе MS SQL и API, которое предоставляется платформой ЖЖ. С помощью этого программного обеспечения были осуществлены множественные закачки, из которых были сформированы четыре месячных выгрузки (выборки), чьи характеристики представлены в таблице 1. При изучении этих выгрузок оказалось, что в среднем чуть меньше трети блоггеров в каждом периоде не имеют постов за закачиваемый период.

Таблица 1. Основные параметры выборок.

Период, за который сделана выгрузка из базы	Количество постов в выгрузке	Количество блоггеров, имеющих посты в выгрузке	Кол-во постов, вошедших в тематический анализ	Кол-во блоггеров с постами, взятыми в анализ	Количество комментариев в выгрузке
10.08.2011-10.09.2011	54182	1822	49811	1798	291209
27.11.2011-27.12.2011	57282	1482	53210	1476	445819
03.02.2012-03.03.2012	61868	1604	57177	1593	655437
12.03.2012-12.04.2012	62067	1563	57835	1551	730980

1. Методы: тематическое моделирование.

В рамках данного исследования тематик постов был использован подход тематического моделирования, с использованием алгоритма латентного размещения Дирихле (Latent Dirichlet Allocation (LDA)) [2], суть которого заключается в следующем. Каждый документ в корпусе текстов рассматривается как наблюдаемая случайная независимая выборка слов (мешок слов), порождённая некоторым, скрытым (латентным) множеством тем. То есть, каждый документ в

коллекции представлен смесью распределения латентных тем, а каждая тема определяется вероятностным распределением на множестве слов. В методе LDA в качестве распределения используется функция Дирихле.

По заданным исходным данным (по заданной коллекции документов и заданной величине предполагаемых тем) требуется восстановить вероятностные распределения всех тем в корпусе и определить, каким именно подмножеством тем порождён каждый документ. Соответственно, результатом расчета в методе LDA являются функция распределения документов по темам и функция распределения слов по темам. Распределения представлены в виде двух матриц: во-первых, матрица распределения вероятностей каждого уникального слова по темам, во-вторых, матрица распределения всех документов по темам.

Соответственно, выбирая наиболее вероятностные слова (например, десять наиболее вероятностных слов), можно определить характер каждой темы. Выбирая документы с наибольшими вероятностями (например, первые десять документов), исследователь получает более подробное представление о теме. Необходимо отметить, что уровень вероятности отсечения текстов исследователь должен определять самостоятельно. Кроме того, в качестве одного из нескольких входных параметров, исследователь должен определить количество тем, на которые LDA будет распределять документы.

Существует две разновидности метода LDA. Первая разновидность - вариационная модель LDA [2], чья численная схема основана на принципе максимизации функции правдоподобия. Расчет максимальной величины правдоподобия реализован на EM-алгоритме. В рамках данной модели реализовано предположение о том, что одна функция Дирихле описывает лишь одно распределение (одного слова по темам или одного документа по темам); соответственно, поиск распределения каждого слова и каждого документа по темам приводит к работе с огромными матрицами. Таким образом, размерность матриц существенно зависит от размера словаря, поэтому качественный препроцессинг документов играет важную роль в тематическом моделировании. Кроме того, наличие произведения большого числа функций приводит к множеству локальных максимумов в функции правдоподобия [3]. Таким образом, метод максимального правдоподобия может приводить не к оптимальным результатам, так как этот метод лишь дает гарантию попадания в один из локальных максимумов, но не позволяет находить наибольший максимум среди множества локальных экстремальных точек.

Второй разновидностью метода LDA является метод сэмплирования Гиббса – статистический алгоритм на основе методов Монте-Карло, в котором строится марковская цепь, сходящаяся к апостериорному распределению тем, по которым далее строятся оценки параметров [1]. Такой подход позволяет эффективно находить скрытые темы в больших корпусах текстов. В большинстве случаев, сэмплирование Гиббса оказывается более эффективным, чем вариационные методы [5].

Прежде всего, в методе сэмплирования Гиббса используются симметричные функции Дирихле, что позволяет существенно упростить вычисление. Это означает, что все распределения описываются одинаковыми функциями; соответственно, размерность матриц оказывается существенно меньшей. В итоговых формулах в методе сэмплирования функции Дирихле уже не фигурируют, и весь алгоритм основан на подсчете числа слов, которые попадают в тот или иной документ и на подсчете числа слов, которые отнесены к разным темам. По сути дела, алгоритм сэмплирования вычисляет три величины:

1. матрица того, сколько раз слово w было отнесено к теме t,

2. матрица того, сколько раз слово документа d было отнесено к теме t,

3. значение того, сколько слов было отнесено к теме t,

Вычисление матриц распределений документов и слов по темам происходит после окончания сэмплирования.

В силу того, что сэмплирование Гиббса дает более качественный результат, для исследования тематик в постах ЖЖ было использовано именно оно. В качестве программного обеспечения был использован пакет Stanford Topic Modeling Toolbox [*11*], [12]. Данный метод и данное программное обеспечение было успешно применено в других исследованиях авторов [7], [10].

2. Посты: активность и тематика аккаунтов.

В блогосфере возможны следующие типы (мета)текстов: блог, пост, комментарий и их комбинации: пост с всеми комментариями и блог со всеми комментариями. Нетекстовая информация постов находится за пределами данного исследования. Сразу отметим, что у всех этих единиц, помимо наблюдаемых характеристик, есть, очевидно, важные, но скрытые параметры, доступа к которым мы не имеем, а именно связанные с их аудиторией (количество просмотров, количество уникальных посетителей и некоторые другие).

Дата и время. Помимо того, что данный параметр важен для изучения динамики блогосферы и связи ее активности с оффлайновой реальностью, важно отметить, что эта активность имеет определенную цикличность, которую необходимо учитывать.

На рисунке 1 дан пример распределения количества постов по дням недели; это количество

значимо отличается в выходные от будней (при зависимой «день недели» коэффициент Эта =0,61 в данном периоде, а в другие исследуемые периоды он достигал 0,82). Голдер и соавторы [4] получили сходные данные по активности американских студентов в Фэйсбуке, хотя в их исследовании суббота всегда значимо ниже, чем воскресенье. Соответственно, при расчете изменения освещения той или иной темы по дням следует вводить поправочный коэффициент, который лучше рассчитывать для каждого периода отдельно, но в перспективе при накоплении больших массивов данных можно учитывать средний коэффициент. Так, например, в указанном на рисунке периоде (март-апрель 2012) количество постов 4 марта (1950) и 11 марта (1895) примерно равно среднему по будним дням (1951) и было бы не заметно в общем распределении, однако оно существенно больше среднего по выходным (1488). Нетрудно догадаться, что 4 марта – день президентских выборов, и видно, что наибольший всплеск активности происходит на следующий день. Всплеск 11 марта, на первый взгляд, менее очевиден; такого рода неочевидные всплески – повод для исследования скрытой повестки дня блогосферы. 11 марта – это день, следующий за митингом на тему «Итоги выборов», сопровождавшегося задержаниями активистов. Связь взлета активности с этим событием может быть проверена сравнением тематики данного дня с другими выходными.

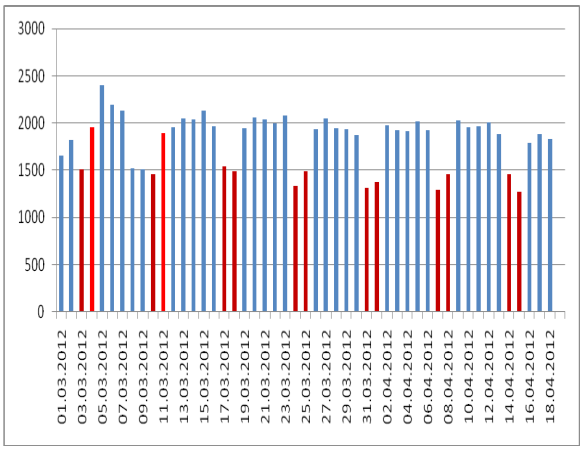

Рис. 1. Распределение количества постов по дням недели, март-апрель 2012.

То же можно сказать и о распределении во времени: при отслеживании динамики по часам ночной провал может оказаться важным. Существуют также сезонные колебания, но для их расчета пока недостаточно данных. Кроме того, все эти временные циклы могут отличаться у разных блоггеров или, что важнее, категорий / групп блоггеров.

Рисунок 2 иллюстрируют разные «временные профили» различных блоггеров из числа наиболее активных в декабре 2011. Временные циклы генерирования сообщений, в частности, могут помогать выявлять и удалять из анализа ботов, коллективные или коммерческие блоги, а также тех, кто наиболее / наименее активно отреагировал на какое-либо событие.

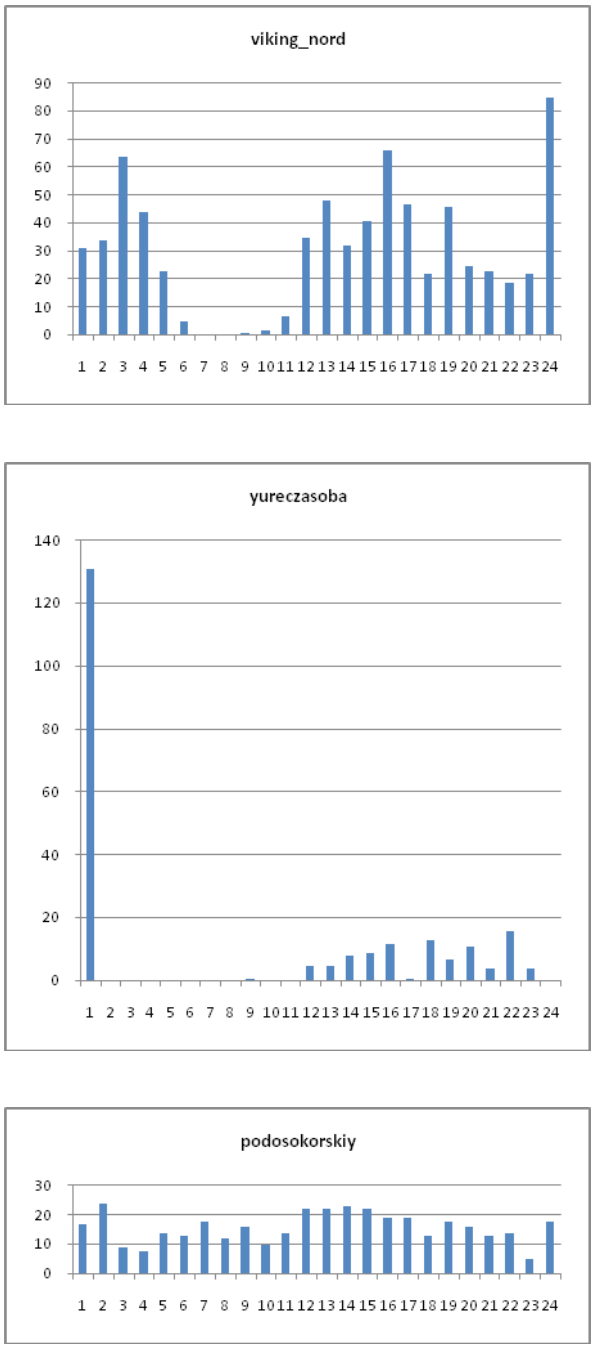

Рис. 2. Временные циклы активности отдельных блоггеров.

Длина текста. Исследование показывает, что средняя длина постов в разных периодах колеблется незначительно (например, сентябрь 2011 – 1761

символ; декабрь 2011 – 1724; март 2012 – 1884), а распределение практически идентично. Оно дает скачок от 100 к 200 символам, а затем резко падает по степенной функции. Зная эти распределения, можно адекватно выявлять необычно длинные и необычно короткие тексты либо для того, чтобы ими пренебречь, либо для того, чтобы сосредоточить на них свое внимание в лингвистических, социо-лингвистических и других целях. Например, из предварительного ручного анализа видно, что более длинные тексты имеют больше вероятности быть перепостами новостей, а более короткие – собственными комментариями, более эмоционально нагруженными. Методы автоматического анализа тональности текстов (sentiment analysis) различны для разных длин и жанров, поэтому для этих задач классификация по длине может оказаться очень важна. Кроме того, длина постов некоторым образом связана с их авторством; средняя длина поста от автора к автору сильно разнится; пример распределения дан на рисунке 3. Из него видно, что можно выделить небольшое количество блоггеров, склонных писать очень длинные или очень короткие посты. Средняя длина постов блоггеров не связана с их активностью, измеренной в количестве постов; коэффициент корреляции близок к нулю. С чем связано это распределение и связано ли, может стать предметом отдельного исследования.

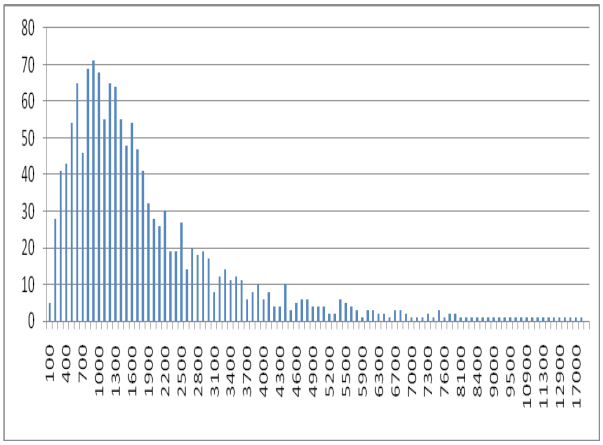

Рис. 3. Распределение средней длины поста у блоггера по количеству блоггеров, декабрь 2011.

Тематика поста. Важнее было бы обнаружить связь между длиной поста и его тематикой как ключевым показателем для большинства исследований, однако пока такой связи не выявляется, зато выявляется связь между длиной текста и количеством тем, к которым он относится с ненулевой вероятностью. Здесь следует отметить, что тематика есть не «данная» его характеристика, а результат работы определенного алгоритма в сочетании с различными выборами и интерпретациями исследователя. Все тексты в данном подходе рассматриваются как политематические, включающие все искомые темы, но большинство из этих тем входят в каждый текст с нулевой вероятностью, и только несколько – с ненулевой; они и составляют тематический «профиль» данного текста. Среднее количество тем, к которым используемый нами алгоритм относит тексты с ненулевой вероятностью, является константой на всех массивах данных, (в нашем случае – 6,3-6,5 тем на пост), т.к. мы использовали настройки, которые показали удовлетворительную применимость на разнообразных данных [13, p. 432] Постоянны и распределения количества тем на пост.

Количество тем, приписываемых посту, коррелирует с его длиной (коэффициент корреляции Пирсона ≈ 0,7): чем больше длина, тем больше лексическое разнообразие текста и, соответственно, тем больше у него вероятность попасть в большее количество тем. Поэтому мы наблюдаем и сходную корреляцию по авторам: среднее количество тем на пост у автора коррелирует со средней длиной поста у автора. Кроме того, существуют тексты, относительно равномерно «размазанные» по множеству тем; и существуют тексты, где одна тема доминирует. Также, существуют авторы, склонные писать «размазанные» тексты, и авторы, склонные писать тематически «сконцетрированные» тексты. Высокое значение доли домнирующей темы среди всех тем текста поможет быстро выявлять монотематические тексты, если, например, есть задача исключать из анализа тексты «обо всем». Этот же высокий показатель у автора (то есть насколько велика в среднем доля доминирующей темы в его постах) не свидетельствует о том, что автор всегда пишет об одном и том же; он свидетельствует о том, что автор склонен посвящать посты одной теме; таких авторов меньшинство. Тем не менее, для ряда задач может оказаться важным работать только с «сконцетрированными» блогами – например, если есть задача вычленять спектры мнений о темах, тексты «обо всем» могут сильно затруднять анализ.

Если суммировать все вероятности, с которыми тексты данного блоггера приписаны к данной теме, и сделать это для каждой темы, мы получим тематический профиль блоггера. В данном исследовании темы закодированы вручную – следует оговориться, закодированы предварительно, кодирование будет уточняться. Затем они разделены на 11 групп: группы 1-4 – социально-политические темы; группы 5-8 – культура и приватная сфера; группы 9 -11 – шум (подробнее в таблице 2). Распределение интереса блогеров к групам тем крайне неравномерно. Так, пользователь navalny, как и следовало ожидать, представляет собой пример профиля с акцентом на социально-политические темы (группы 1-4). Пользователь _arlekin_ в исследуемом периоде имел акцент на темы культуры (группа 5), профиль блоггера prosto_telo демонстрировал пики на рекреативной и потребительской темах (группы 6 и 8). Юзер shorec – оказался типичным монотематиеским профилем; автор

в данном периоде писал почти исключительно об акциях протеста. Такая монотематичность встречается у блоггеров с низкой или умеренной активностью. Были также аккаунты с высокой активностью с относительно большим, хотя и неравномерным вниманием, к большинству тем и, наоборот, малоактивные, избирательные, но при этом относительно разнообразные аккаунты. В таких профилях зачастую высока доля «мусорных» тем (т.е. тем, интерпретация которых затруднена, группа 11). Следует подчеркнуть, что профиль с акцентом, скажем, на экономику, не обязательно означает, что у автора большинство постов посвящено экономике; он означает, что тема экономики присутствует в блоге автора сильнее других тем; она может быть как распределена по множеству постов, так и сильно сконцентрирована в некоторой их части. Также следует помнить, что тематический профиль отражает состояние блога за определенный период времени.

Таблица 2. Классификация групп тем, выработанная для февраля 2012 года.

Номер группы	Содержание группы тем
1	Внутренняя политика России, в особенности выборы, протесты и права человека
2	Другая политика (включая внешнюю, региональное управление и историю СССР)
3	Социальные темы
4	Экономика
5	Культура
6	Рекреативная деятельность (отдых, туризм, творчество)
7	Приватная сфера (семья, дети, межличностные отношения, секс)
8	Потребление
9	Посты на других языках (украинский, английский)
10	Посты, объединенные по специфической лексике (напр, сниженной)
11	Не поддающиеся интерпретации темы.

Тематические профили блогов могут служить для разных задач. С помощью них можно выявлять аккаунты, акцентирующие определенные социально значимые или социально опасные темы, либо потребительские темы, акцент на которых может указывать на авторитет блогера в потреблении определенной товарной категории и может использоваться для рекламных и маркетинговых задач. Можно проводить тематическую кластеризацию аккаунтов на основе их профилей и, таким образом, автоматически выявлять целые группы тематически сходных блогеров, которые могут быть основой целевой аудитории для контекстной рекламы. Это невозможно при использовании целых аккаунтов как единых текстов, так как они оказываются настолько политематичными, что кластеры не выявляются.

Количество постов в аккаунте за период. Сумма всех вероятностей по всем темам у блоггера равна количеству его постов, так как сумма всех вероятностей одного поста равна 1. Активность аккаунтов, измеренная в количестве постов, сильно разнится между авторами и распределена по степенному закону. Можно предположить, что в летние месяцы активность снижается, однако для этого на данный момент недостаточно данных.

Активность аккаунтов является важнейшим статистическим показателем, так как большинство аккаунтов не активны или настолько слабоактивны, что включение их в анализ для определенных задач бессмысленно. Например, если осуществляется поиск связи между тематикой постов и характеристиками аккаунта, количество постов ниже определенного порога в исследуемом периоде может оказаться отсекающим критерием.

Для некоторых задач – например, связанных с изучением структуры дискуссий – активность аккаунтов может быть также измерена в количестве комментариев, оставленных автором. Однако сбор такой информации через сбор по выборке блогов не возможен, так как комментарии хранятся в тех блогах, к которым относятся. В профилях блоггеров есть количество оставленных ими комментариев за весь период существования аккаунта, что не дает представления о текущей комментовой активности блоггера. Эту проблему можно решить только периодической закачкой профилей аккаунтов.

1. Комментарии: популярность и комментовая активность аккаунтов.

Переходя к комментариям, следует сказать, что анализ их содержания пока не реализован. Тематическое моделирование комментариев осложнено их малым размером и широким использованием общей лексики, не позволяющим эффективно кластеризовать их на основе их лексического состава. Применение сентимент-анализа осложнено отсутствием качественных словарей эмоциональной лексики для русского языка. Ли и Цанг [9] кластеризовали посты вместе с текстами комментариев к ним, присоединенных к телу каждого поста, и получили более высокое качество на основе внешних мер качества по сравнению с кластеризацией постов отдельно. Проблема измерения качества тематического моделирования и кластеризации текстов при этом сохраняется. Важнее, что при такой кластеризации получается содержательно иной результат: если тематическая структура постов может свидетельствовать о повестке дня, сознательно задаваемой авторами, то тематическая структура постов

вместе с их комментариями отражает повестку дня дискуссий. Такая кластеризация – тема дальнейших исследований.

Количество комментариев на пост. В выборке представлены блоггеры, имеющие наибольшее количество друзей, то есть находящиеся ближе к верхней границе любого из рейтингов «Живого журнала», поэтому количество комментариев на пост у них, очевидно, намного выше среднего по ЖЖ и колеблется от 5 в августе-сентябре 2011 до 11 в феврале 2012. Однако оно сильно неравномерно распределено как по постам, так и по блоггерам. В табл. 3 дан пример самых комментируемых постов за февраль 2012. Следует отметить, что в этом и других исследуемых периодах тематика блогов ЖЖ находилась под сильным влиянием электорального цикла 2011-2012; в другие периоды, изученные в других исследованиях , доля общественно-политической тематики ниже. Далее опишем основные характеристики структуры комментирования

Таблица 3. Десять самых комментируемых постов, период: февраль 2012.

Кол-во комментов	Автор поста	Название поста	Содержание поста
1015	Navalny	Правильный конкурс плаката	объявление о конкурсе плаката против Путина
891	Tema	В трех словах	«Отвечаю на любой вопрос про дизайн, компанию, деловые отношения, отношения людей в трех словах»
828	tipaa-etaa	Поднимите это говно в топ, пожалуйста	призыв поднять пост с изображением кала в топ путем обильного комментирования
826	Navalny	не врать и не воровать	призыв прочитать статью автора в "Ведомостях"
793	md-prokhorov	Нет	комментарий к программе автора по политике в области здравоохранения
778	Navalny	это мы неплохо сегодня прошлись	репортаж с митингов 2 февраля
728	Zyalt	3 вопроса	о журналистской концепции блога автора
673	sobchak-xenia	Ответ	Ответ на статью Виктора Шендеровича «отскок от борта»
671	sobchak-xenia	Нет	рассуждения о выборах и этике
661	shemka	Конкурс от InLoveWithFashion!)	объявление о конкурсе с призом в виде модного платья

также на основе февральской выборки. Из 68686 постов более 17 тысяч вообще не имеют комментариев; более 10 комментариев имеют около 15 тысяч постов (рисунок 4).

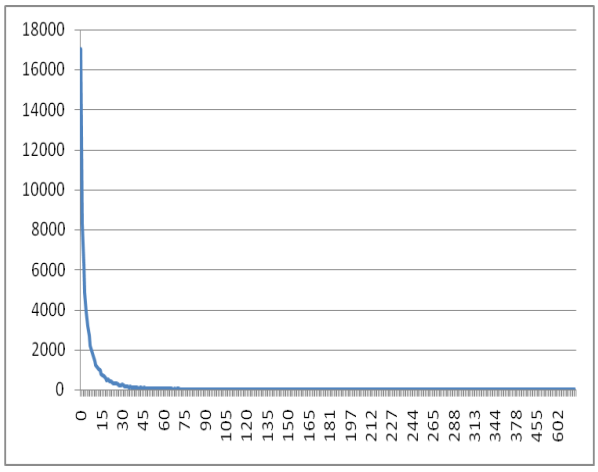

Рис. 4. Распределение количества комментариев на пост по количеству постов, февраль 2012 г.

Обнаружение связи между комментируемостью постов и доминирующей в них темой может стать альтернативным или дополнительным способом определения важности тем в общественном сознании блог-сообщества. Если количество постов, содержащих тему, или вес темы в коллекции постов в целом показывают важность темы для тех, кто пишет посты о ней, то комментируемость постов, содержащих тему, может указывать на ее важность для более широкого круга блоггеров – тех, кто сам о ней не пишет постов или вообще не пишет постов (например, владельцев связанных аккаунтов, предназначенных только для комментирования). Связь между комментируемостью поста и его длиной, временем, количеством тем может дать «портрет» поста, оптимального с точки зрения сбора наибольшего количества комментариев, что может быть полезным для маркетинговых задач.

Количество комментариев на автора постов. Количество комментариев, полученных блоггером, также неравномерно и может служить индикатором популярности блоггера, в особенности в отсутствие данных о количестве просмотров и уникальных посетителей его блога. Количество полученных комментариев не очень сильно коррелирует с количеством написанных постов (коэффициент

корреляции Пирсона =0,4). Это значит, есть блоггеры, которые при небольшом количестве постов могут собирать большое количество комментариев; среднее количество комментариев на пост автора может быть одним из индексов коммуникативной эффективности автора.

Распределение «эффективности» по авторам (рисунок 5) качественно отличается от распределений количества постов по авторам и полученных комментариев по авторам. Последние два распределения односторонние и демонстрируют падение по степенному закону (как на рис. 4), а вот график эффективности все-таки имеет пик на значение 1. Это значит, что авторов, у которых среднее количество полученных комментариев на пост равно 0, все-таки меньше, чем тех, у кого оно равно единице. Самые активные, самые комментируемые и самые «эффективные» блоггеры – это тоже не обязательно одни и те же люди (таблица 4).

Характеристики комментаторов. На данный момент работа с этими данными не завершена, так как сделанные выгрузки не позволяют точно оценить эти характеристики. В среднем комментаторы комментируют по выборке чуть более шести постов. Распределение количества оставленных комментариев по комментаторам имеет еще более

Таблица 4. Лидеры по трем показателям: активность, комментируемость, эффективность (февраль 2012).

Самые активные	
Ник блоггера	Кол-во постов
Carabaas	967
lvoropaeva	790
prosto-telo	762
Ibigdan	734
Uzoranet	700
viking-nord	690
Dgz	624
prikolizmus	624
bloggmaster	611
Pryf	584

Самые комментируемые	
Ник блоггера	Кол-во получ. Коммен-тариев
Navalny	18532
teh-nomad	17304
Ibigdan	16408
Tema	13072
putnik1	8759
Avmalgin	8315
Drugoi	8215
Zyalt	7563
md-prokhorov	6376
Dolboeb	6309

Самые эффективные		
Ник блоггера	Кол-во постов	Кол-во коммен-тариев на пост
sobchak-xenia	6	398
Dolgachov	1	372
pesen-net	1	308
navalny	65	285
zhgun	1	258
prostitutka-ket	13	256
md-prokhorov	26	245
fritzmorgen	30	189
drugoi	44	187
naganoff	11	175

резкий спад, чем распределение постов по авторам (от 9830 до 1), но важно понимать, что оно не отражает их полной комментовой активности, а представляет только их «интерес» к комментированию топовых блоггеров. По предыдущим исследованиям можно сказать, что тематика постов, комментируемых одним комментатором, обычно разнообразна [6], хотя слабая тенденция комментировать сходные темы все же прослеживается [8].

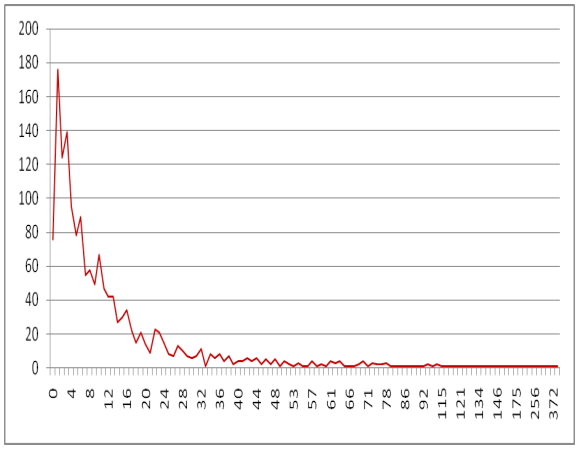

Рис. 5. Распределение среднего количества комментариев на пост автора по количеству авторов (мин. = 0; макс. = 398), февраль 2012

Всего в выгрузке для анализа количества комментариев за февраль 2012 года – 115503 блоггера; из них только 1603 являются авторами постов. Отсюда видно, что подавляющее большинство комментариев топовым блоггерам оставляют те, кто сами в топ не входят. Насколько велика комментовая активность самих топовых блоггеров, пока не ясно.

5.Заключение и перспективы дальнейших исследований.

В настоящем исследовании было показано, что открытые данные блогов в сочетании с методами тематического моделирования дают множество показателей и индексов, важных для маркетинговых и социологических задач и которые составляют статистический и тематический профиль ЖЖ. В работе было описано, каким образом можно собирать такие данные, как автоматически извлекать из них тематику и анализировать ее связь с другими доступными статистическими показателями. В частности, было показано, что построение тематических профилей блогеров информативно и может быть использовано для размещения контекстной рекламы как в блогах, так и в социальных сетях, в особенности при наложении кластерного анализа на тематическое моделирование. Также была показана информативность индекса эффективности блогеров; польза этого индекса может вырасти при изучении предикторов его роста.

В целом, следует отметить, что проделанная работа лишь отмечает начало более обширного направления исследования, которое имеет перспективы развития по множеству путей. Выделим здесь пути, открывающиеся при дополнении имеющихся данных мета-данными аккаунтов (работа над которыми ведется в данный момент):

Изучение показателей активности блоггера, таких как возраст аккаунта (дата создания), количество

постов, количество оставленных комментариев в сочетании друг с другом позволяет решать несколько разных задач. В частности, соотношение числа постов и оставленных комментариев блоггера может подразделить блоггеров на преимущественно «монологичных» и преимущественно «дискутирующих». Не исключено, что роль «дискутирующих» блоггеров в определении повестки дня, во влиянии на общественное мнение или поведение (электоральное, потребительское) не ниже, чем роль «монологичных», но остается латентной, и это может стать предметом исследования.

Важным представляется изучение показателей популярности или успешности блоггера и их предикторов. К первым можно отнести соотношение количества полученных комментариев и количества постов (упоминалось выше); количество блоггеров, зафрендовавших блоггера; соотношение зафрендовавших и зафреднованных; социальный капитал, рейтинг по социальному капиталу и рейтинг по просмотрам. Количество зафрендовавших является прозрачным показателем популярности, однако его недостаток в его статичности: количество друзей имеет тенденцию замирать, однажды образовавшись, и не отражает текущее состояние аккаунта. Социальный капитал, наоборот, учитывает только тех зафрендовавших, которые реально читали исследуемый аккаунт в последнее время, а также сами имеют большое количество активных друзей. Недостаток этого показателя в том, что он является коммерческой тайной компании «Суп», непрозрачен, и его методика время от времени меняется разработчиками. В этом смысле более прозрачен рейтинг по просмотрам, но его значение меняется ежесуточно. Среднемесячные рейтинги не доступны, а ежедневная закачка рейтингов технически сложна. Несмотря на это, вероятно, комбинация количества просмотров с количеством зафрендовавших может дать сводный прозрачный индекс популярности. Предикторами популярности могут выступать, среди прочего, тематический профиль блогера, соотношение тематики аккаунта с популярными темами в СМИ, упоминаемость самого блогера в СМИ, объем визуального материала, а также показатели социальности блогера (количество зафрендованных блоггеров, количество сообществ и интересов, количество указанных аккаунтов в других интернет-сервисах (Facebook, Twitter, OpenID, VK, Skype, Google+ и другие).

Аккаунты могут быть кластеризованы не только по тематическим профилям блоггеров, вытекающим из тематического анализа, но и по тематической самоидентификации блогера (сообщества, интересы и «о себе»). Это может дать не только более объемное представление об изучаемом объекте; если исследовательские ресурсы ограничены, выявление степени соотношения самоидентификации с «реальной» тематической структурой поможет определить реальные тематические наклонности блогера только по его самоидентификации, без более трудоемкого тематического моделирования. Принадлежность к сообществам может быть использована для кластеризации так же, как и перечень интересов, но также может стать основой сетевого анализа. Если сети дружбы могут не отражать текущего состояния, или отражать декларированные отношения со значимыми другими, то принадлежность к сообществам может отражать представления о референтных группах или же реальную тематизированную активность. Очевидно, отражение реальной активности будет тем ниже, чем больше сообществ декларирует автор, а порог, после которого группы перестают отражать активность, можно установить эмпирически.

Кроме того, перспективным направлением представляется исследование наиболее комментируемых тем. Это можно сделать, просуммировав все вероятности, с которыми посты приписаны данной теме, взвесив их на количество комментариев, полученных постами. Такая методика, насколько нам известно, не применялась; ближайшим по задачам исследованием является работа Тсагкиса и соавторов [14], в которой они вырабатывают индексы комментируемости новостей онлайновых СМИ, но применительно к цельным текстам, а не к темам, распределенным по текстам. Наконец, отметим, что все описанные индексы и показатели применимы не только к блогам, чей рост в последнее время замедлился, но и к активно растущим социальным сетям, вбирающим в себя функции блогов и производящим все больше контента.

Литература

[1] Andrieu C., Freitas N.D., Doucet A., Jordan M. An introduction to MCMC for machine learning. Journal of Machine Learning. // Machine Learning, 2003. Mol. 50 № 1. P. 5–43.

[2] Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet Allocation // Journal of Machine Learning Research, 2003. Vol. 3. P. 993–1022.

[3] Daud A., Li J., Zhou L., Muhammad F. Knowledge discovery through directed probabilistic topic models: a survey. // Journal of Frontiers of Computer Science in China (FCS), 2010. Vol. 4, № 2. P. 280-301.

[4] Golder S.A., Wilkinson D.M., Hubennan B.A. Rhythms of Social Interaction: Messaging Within a Massive Online Network / Communities and Technologies, Conference Proceedings, 2007. P. 41-66.

[5] Griffiths T.L., Steyvers M. Finding scientific topics. // Proceedings of the National Academy of Sciences, 2004. № 101. P. 5228–5235.

[6] Jamali S., Rangwala H. Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis. // International Conference on Web Information Systems and Mining, Conference Proceedings, 2009. P. 32-38.
[7] Koltsova O., Koltcov S. Mapping the Public Agenda with Topic Modeling: The Case of the Russian LiveJournal // Policy & Internet, 2013.Vol. 5, № 2. P. 70-89
[8] Koltsova, Koltcov. Comment-based communities in the Russian Livejournal and their topical coherence. / XXXIII Sunbelt Social Networks Conference of the International Network for Social Network Analysis (INSNA), 21-26 May 2013, Hamburg, Germany.
[9] Li B., Xu Sh., Zhang J. Enhancing Clustering Blog Documents by Utilizing Author/Reader Comments. / ACM-SE 45 Proceedings of the 45th annual southeast regional conference, 2007. P. 94-99.
[10] Maslinsky K., Koltsova O., Koltcov S. Changes in the Topical Structure of Russian Language Livejournal: The Impact of Elections 2011 / Серия препринтов «Sociology», Высшая школа экономики, 01/2013. -21 с. http://publications.hse.ru/preprints/72804584
[11] *Ramage D., Dumais S., Liebling D*. Characterising Microblogs with Topic Models. / ICWSM 2010. Association for the Advancement of Artificial Intelligence, Conference Proceedings, 2010. URL: http://www.stanford.edu/dramage/papers/twitter-icwsm10.pdf (дата обращения: 19.04.2012).
[12] *Ramage D., Rosen E., Chuang J., Manning C.D., McFarland* D.A. Topic Modeling for the Social Sciences. / Workshop on Applications for Topic Models. NIPS, 2009. URL: http://vis.stanford.edu/files/2009-TopicModels-NIPS-Workshop.pdf.
[13] Steyvers, M., Griffiths T. Probabilistic Topic Models. / Landauer T., McNamara D., Dennis S, Kintsch W. (eds). Handbook of Latent Semantic Analysis. Hillsdale, NJ. 2007.
[14] Tsagkias M., Weerkamp W., de Rijke M. News Comments: Exploring, Modeling, and Online Prediction // Advances in Information Retrieval. Lecture Notes in Computer Science, 2010. Vol. 5993, 201. P. 191-203.

A STATISICAL AND TOPICAL PORTRAIT OF LIVEJOURNAL

Olessia Koltsova, Sergei Koltcov

The purpose of this work is to explore the basic statistical properties of the Russian blogosphere for the goals of their future application in sociology, marketing, linguistics and other fields of scientific and practical knowledge. The need for such descriptive statistics is driven by its being unavailable in the public domain – unlike socio-demographic properties of population widely used in sociological and marketing research, or various Russian language dictionaries and lexicons facilitating linguistic research. This explorative study does not aim at hypotheses testing; instead, it sketches possible solutions for various research tasks basing them on the knowledge about relations between the properties under scrutiny. The paper considers relations between posts' dates, lengths, topic composition, number of comments and some others; it also makes sense of cycles of bloggers' activity and builds their thematic profiles. Thematic structure is extracted with automatic topic modeling, which, combined with the analysis of other variables, presents the main novelty of the paper.

i Работа выполнена при поддержке Центра Фундаментальных Исследований НИУ «Высшая школа экономики», ТЗ-51, 2012.