Картирование комментовых сообществ в Живом Журнале

О. Ю. Кольцова, Ю. Г. Рыков, С. Н. Кольцов

Национальный исследовательский университет «Высшая школа экономики»
(НИУ ВШЭ — Санкт-Петербург)
olessia.koltsova@gmail.com, rykyur@gmail.com, kol-sergei@yandex.ru

Аннотация

Общая цель проекта — выявить закономерности образования дискуссионных сообществ в социальных сетях для того, чтобы иметь возможность обнаруживать «точки» возникновения социального напряжения и мобилизации. До сих пор объектом исследования становились только сообщества, образуемые сетями дружбы или подписки, и не существует исследований сообществ, в которых люди объединены комментированием одних и тех же текстов. Задача излагаемой части проекта — выявить, образуются ли сообщества комментирования в блогах вокруг тематики комментируемых постов или вокруг авторов (групп авторов).

1. Введение

Выявление сообществ (community detection) - одна из задач анализа социальных сетей (Social Network Analysis). Изучение виртуальной социальной реальности, представленной Интернет-блогами и социальными сетями, представляет особый социологический интерес, так как в последние годы их роль и влияние на современное общество значительно возросли. Особенно заметно влияние виртуальных социальных сетей на изменение политических практик, на политическую мобилизацию.

Таким образом, выявление и картирование сообществ предполагает построение сетевой модели Живого Журнала на базе данных о комментировании пользователями различных постов в определенные периоды времени. В полученной сети можно выделить группы постов, которые были прокомментированы одними и теми же пользователями. В результате анализа такой сети мы можем узнать на каком основании группируются пользователи, комментирующие одни и те же посты (вокруг блоггера или вокруг темы), и возникают ли устойчивые комментовые сообщества в Живом Журнале. Мы полагаем, что комментовые сообщества ("сетевые уплотнения") являются структурными элементами сети, поэтому получение карты таких сообществ представляется важной задачей при описании социальной структуры Живого Журнала[2].

2. Дизайн исследования

2.1. Модель сети

Дадим операциональные определения ключевым понятиям данного исследования. Основным является понятие сообщества, под которым в теории графов понимается сегмент графа (субграф), где фактическое количество ребер превышает ожидаемое количество ребер в случайном графе таких же размеров. Другими словами, сообщество - это подмножество вершин, связанных между собой большим количеством ребер, чем с "внешними" вершинами. Визуально сообщества выглядят как " уплотнения" в графе всей сети.

Комментовое сообщество (comment-based community) существует, когда примерно один тот же круг постов комментируется примерно одной и той же группой блоггеров. Таким образом, если комментовое сообщество устойчиво существует, мы могли бы определить на основе чего оно интегрировано: на основе общей темы или общего авторства постов.

Единицами анализа в нашем исследовании являются зарегистрированные пользователи Живого Журнала (блоггеры), их посты и факты комментирования. Посты являются вершинами (узлами) сети. Связь между парой постов возникает в том случае, если один тот же пользователь оставил хотя бы по одному комментарию к каждому из этих постов. Таким образом, ребро в сети - это наличие общего комментатора. Чем больше общих комментаторов, тем больше ребер соединяет пару постов, и тем сильнее связь между постами. В результате таких исходных данных мы получаем сеть постов, связанных общими комментаторами.

Нулевая гипотеза данного исследовании, выглядит следующим образом: мы предполагаем, что комментовые сообщества в Живом Журнале возникают вокруг общих комментируемых тем.

2.2. Данные и метод

Данные из Живого Журнала были загружены за два трехдневных периода при помощи специализированного краулера "Blogminer"[1]: 21-23 и 24-26

Труды XV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2012), Санкт-Петербург, Россия, 2012.

декабря 2011 года. Это время выбрано не случайно и соответствует фазе ожидаемой реакции со стороны "населения" российской блогосферы на выборы в Государственную Думу, состоявшиеся 4 декабря. Период с 21 по 26 декабря был разбит на две части, которые анализировались по отдельности, в силу нехватки вычислительных мощностей используемого программного обеспечения. Для анализа данных использовалась программа NodeXL [3].

Сеть за период 21-23 декабря состоит из 4220 постов, написанных в общей сложности 898 блоггерами. За период 24-26 декабря было загружено 3721 постов, принадлежащих 831 блоггеру. Все закаченные посты принадлежат топовым блогерам, входящим в первую тысячу блогеров рейтинга Живого Журнала.

Пакет NodeXL содержит три общепризнанных алгоритма по кластеризации сообществ: Ньюмана-Гривана, Клозэ-Ньюмана-Мура [4] и Вакита-Цуруми [5]. Для выявления сообществ мы использовали алгоритм Вакита-Цуруми, так как он наиболее пригоден для анализа крупных сетей, и в качестве контрольного, алгоритм Клозэ-Ньюмана-Мура.

После операции по выявлению сообществ, которая разделила полную сеть постов на отдельные подмножества, мы отобрали несколько групп постов для качественного анализа. Его целью было установить, связаны ли посты, входящую в одну группу, по смыслу (тематически) или каким-либо другим образом (принадлежат перу одного или нескольких авторов).

Напомним, что каждый кластер постов интегрирован группой общих комментаторов, то есть за каждой группой потов стоит определенное комментовое сообщество.

3. Результаты

Мы анализировали всю совокупность данных в виде двух отдельных массивов и строили две отдельные сети, однако, так как были получены близкие результаты, мы приводим иллюстрации из обоих кейсов.

Приведем метрику полных графов для I-го (21-23 дек.) и II-го (24-26 дек.) периодов в форме таблицы.

Как видно из Таблицы 1, сети обоих периодов различаются между собой несущественно: наблюдаемые различия обусловлены изначальным неравным размером сетей.

Далее приведем результаты операции по выявлению кластеров в сети. Стоит отметить, что алгоритмы Вакита-Цуруми и Клозэ-Ньюмана-Мура разбили граф на практически равное количество сообществ (57 и 53 для данных I периода). На основе этого мы полагаем, что количество полученных сообществ является объективной характеристикой графа. Граф I периода был разбит на 57 кластеров, II периода - на 52 кластера. Эти различия также обуславливаются неравными размерами графов.

Таблица 1. Параметры графов за оба периода.

Параметры графа	**Значение**	
	I период	**II период**
Количество вершин	4220	3721
Диаметр (максимальная геодезическая дистанция)	7	8
Средняя геодезическая дистанция	2,685	2.762
Средняя степень (degree)	66,494	58,522
Медиана степени (degree)	31	28
Плотность графа	0,01576	0,01573

Отметим, что каждый пост входит только в один кластер, тогда как блоггер может принадлежать сразу к нескольким группам (в силу того, что посты одного блоггера могут входить в разные группы).

В двух сетях плотность сообществ обратно пропорциональна их размерам (были получены достаточно высокие отрицательные значения корреляции плотности: -0,61 с количеством вершин/постов и -0,41 с количеством ребер/общих комментаторов), что говорит о существенной рыхлости сетей. Отсюда, предпочтительнее рассматривать небольшие группы. Также бессмысленно рассматривать совсем маленькие группы (менее 10 постов) — полученные результаты не будут достаточно надежны. Таким образом, для качественного анализа оказались пригодны лишь 10 групп, содержащие от 10 до 61 постов. Такая же картина наблюдается с графом за II период: всего лишь 10 групп (от 10 до 71 постов и от 7 до 29 блоггеров) удовлетворяют условиям. Однако настолько высокая одинаковость результатов для I и II периодов (одинаковость распределения постов и блоггеров по группам) заставляет нас сомневаться, что именно такое разбиение на группы отражает некие объективные характеристики сети. И в том и в другом случае только ровно по 10 групп (из 50-ти) претендуют на осмысленную связанность.

Качественное изучение постов из отобранных групп не выявило значимой связи между темами постов одного сообщества. Вот некоторые темы потов одного кластера: медицина, семейная жизнь, путешествия, подарки, политика РФ, пропаганда митинга, IT. В некоторых группах прослеживается влияние политической тематики (выборы, митинги), но нельзя утверждать, что она доминирует. Кроме того, многие темы повторяются от группы к группе, например тема католического Рождества.

Итак, наша гипотеза не подтвердилась: мы не можем утверждать, что комментовые сообщества интегрированы общими темами в Живом Журнале. Вероятно, механизмы образования комментовых сообществ более сложные, и в их генезисе задействовано большее количество факторов.

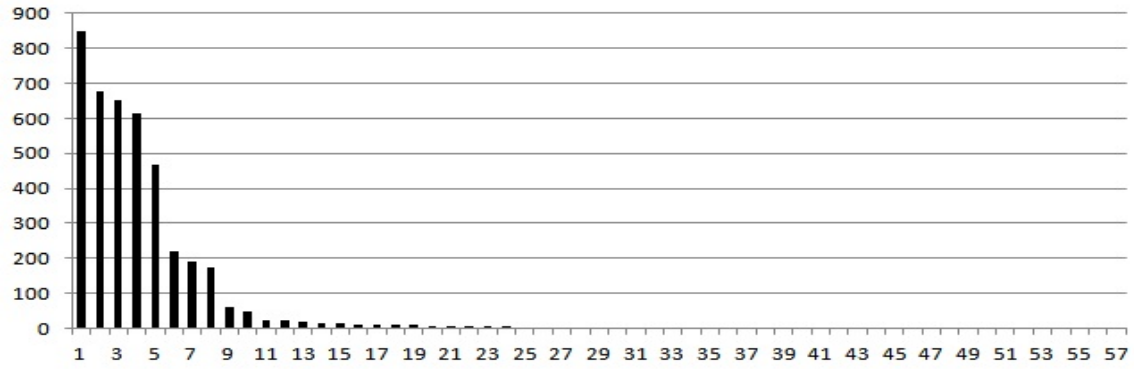

Рис. 1. Распределение постов по группам для I периода (21-23 декабря)

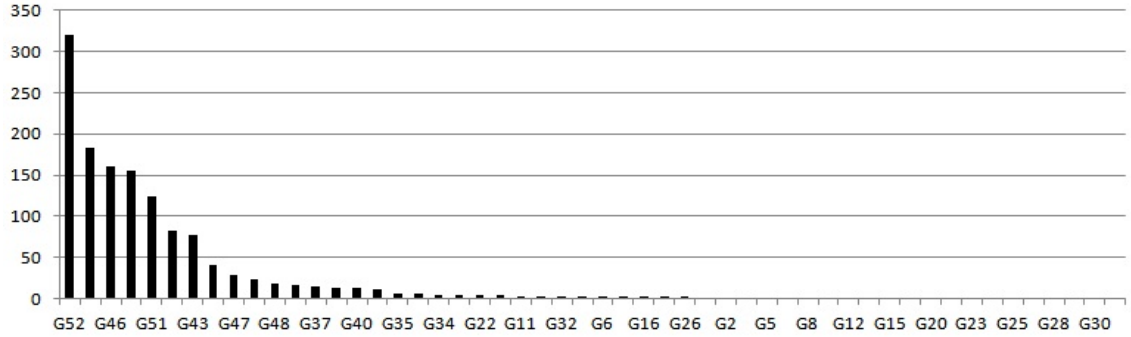

Рис. 2. Распределение блоггеров по группам для II периода (24-26 декабря)

4. Заключение: проблемы и перспективы

Разница между параметрами сетей за два различных периода не существенна, поэтому мы считаем, что сеть постов определенных размеров, построенная на основе общих комментаторов, имеет некие постоянные величины. Однако, пока неизвестно какова природа этих постоянных: или они обусловлены каким-то социальными факторами, или же это побочный эффект использования данного математического инструментария и ПО.

Построенная в исследовании сеть является не совсем социальной, так как узлами сети являются не пользователи, а посты (коммуникации). Однако данный подход позволяет оценить связь между комментовым сообществом и блогерам.

Перспективным, с нашей точки зрения, является дальнейший анализ сети, в которой в качестве узлов берутся пользователи комментирующие друг друга. В подобной сети кластеры будут являться сообществами в более строгом социологическом смысле: как сообщество пользователей, связанные густой сетью коммуникаций.

Литература

[1] Кольцов С.Н. ПО ‘Blogminer’, ООО ‘колтран-Лабс’, 2012, http://www.koltran-labs.ru/.

[2] Кольцова О.Ю., К проблеме применения алгоритмов выявления сообществ в больших сетях для социологических задач // Бизнес информатика, статья подана 19.05.2012.

[3] Clauset A, Newman M. E. J., Moore C. Finding community structure in very large networks, Physical Review E. Vol. 70. No. 6. (30 Dec 2004)

[4] Hansen D., Shneiderman B., Smith M. Analyzing Socail Media Networks with NodeXL: insights from a connected world. Elsevier, London, 2011.

[5] Wakita K, Toshiyuki Tsurumi, Finding Community Structure in Mega-scale Social Networks,URL http://arxiv.org/pdf/cs.CY/0702048.pdf.

Mapping comment-based communities in Life Journal

Olessia Koltsova, Yurii Rykov, Sergei Koltsov

This study aims to describe the social structure of comment-based communities of posts by the most popular Russian LiveJournal bloggers. This research investigates links between comment-based communities and topics of posts which include in the same cluster.