Анализ и моделирование пользовательской активности в социальной новостной сети Reddit

А. С. Тамазян
Санкт-Петербургский государственный электротехнический государственный университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

В. С. Пантелеев
Омский государственный университет им. Ф.М. Достоевского

***Abstract*. Analysis of user score distribution of various posts on Reddit social news network was conducted for several subreddits for several years, and Student t-distribution is suggested as model for these scores. Also a comparative analysis of amount of posts with positive, zero or negative rating was done in order to find some similarities in users behavior.**

Keywords: social network; user activity; Reddit

I. Введение

Социальные новостные сети стали одним из главных средств обмена информацией между сетевыми пользователями. В такие сетях пользователи могут делиться информацией в виде отдельных публикаций, а другие пользователи могут голосовать как «за», так и «против» данных публикаций. Исходя из разности между числом голосов «за» и «против» рассчитывается рейтинг публикации, который может быть как положительным, так и отрицательным. В зависимости от данного рейтинга публикация может быть размещена на верхних позициях общего рейтинга публикаций соответствующей тематической секции сайта, или даже может быть помещена на главную страницу ресурса, что повышает ее шанс быть увиденной большим числом пользователей.

Таким образом, при исследовании процессов обмена информацией в данных сетях важной задачей является исследование поведение пользователей. Подобные расследования проводились в работах [1, 2], в том числе для механизмов распространения информации в социальных сетях вроде Twitter и Facebook [4]. Информация в подобных сетях обычно распространяется через «лайки» («мне нравится») или через функцию «поделиться» («ретвит» для Twitter). При этом пользователь копирует исходное сообщение в свой профиль. Также, поведение пользователей изучалось при оценке новостей в сети Reddit [5]. Изучение связи между популярностью и внутренним качеством статьи в Reddit и Hacker News было проведено в [6], в то время как модель интересов для любой социальной сети рассмотрена в [7].

II. Описание данных

Одним из самых популярных агрегаторов социальных новостей является Reddit (http://reddit.com), где зарегистрированные участники могут публиковать различные материалы, такие как текстовые сообщения, изображения и т. д. Сообщения сгруппированы по темам в пользовательские секции под названием «сабреддиты» (subreddits). Каждый пользователь может проголосовать «за» или «против» публикации. Разница между количеством голов «за» и «против» называется рейтингом. Сообщения с более высоким рейтингом доходят до верхних позиций общего рейтинга публикаций, в то время как комментарий с более высоким рейтингом переходит в начало раздела комментария.

Был проведен анализ наборов данных публикаций на Reddit, собранных с 01.01.2011 по 31.12.2016. Мы извлекли данные для сабреддитов AskReddit, IAmA, funny, gaming, science, todayilearned. После этого для были извлечены данные рейтинга публикаций для этих сабреддитов. Тема сабреддита AskReddit – ответы на вопросы пользователей Reddit; в сабреддите IAmA пользователи рассказывают о себе; в funny пользователи рассказывают о разных забавных историях; в gaming пользователи обсуждают различные игры, в частности, видеоигры; в science пользователи Reddit обсуждают различные научные открытия и теории; в todayilearned пользователи рассказывают друг другу различные интересные факты, которые они недавно узнали.

III. Анализ распределения рейтингов

Для анализа распределения положительных рейтингов были оценены оценки плотности вероятности $p(s)$ для положительных рейтингов. Результаты показаны на рис. 1 и 2. На рис. 1 $p(s)$ оценки были сгруппированы по годам, а на рис. 2 они сгруппированы по сабреддитам. Как можем заметить, $p(s)$ показывают экспоненциальное поведение для малых s и имеют тяжелые хвосты при больших значениях s. Разная тематика сабреддитов не оказывает существенного влияния на плотность вероятности положительного рейтинга s.

Далее были предприняты попытки связать результирующие распределения с помощью распределения Стьюдента, которое можно описать как

$$p(s,d,\mu,\lambda) = f\left(\left(s-\mu\right)/\lambda, d\right),$$

Работа выполнена при финансовой поддержке РФФИ, проект №16-37-00374.

где $f(x,d)=\dfrac{\Gamma\left((d+1)/2\right)}{\sqrt{\pi d}\,\Gamma\left(d/2\right)\left(1+x^2/d\right)^{(d+1)/2}}$,

λ –параметр масштаба, μ – параметр смещения, d – степень параметра свободы.

Аппроксимация для разных сабреддитов показана на рис. 1. Стоит отметить, что значение d = 0,8 для всех рассмотренных сабреддитов, что подразумевает сходство распределения рейтинга, несмотря на различную тематику сабреддитов.

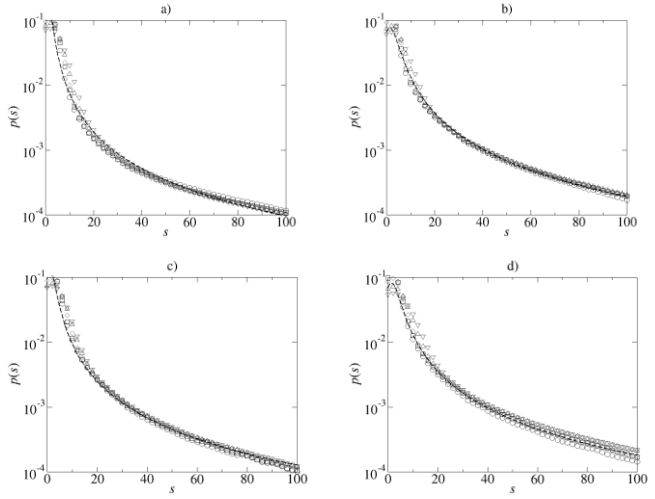

Рис. 1. Оценка плотности вероятности постов за 2012 (точки), 2013 (квадраты), 2014 (ромбы), 2015 (треугольники вверх) и 2016 (треугольники вниз), сгруппированный по сабреддитам: AskReddit (a), funny (b), science (c) и todayilearned (d). Аппроксимация распределением Стьюдента показана черной пунктирной линией

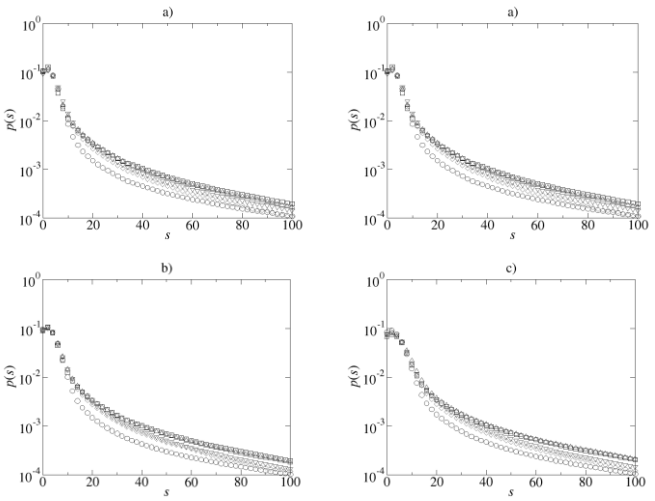

Рис. 2. Оценка плотности вероятности оценок постов, тематических секций AskReddit (точки), funny (квадраты), gaming (ромбы), science (треугольники вверх) и todayilearned (треугольники вниз), сгруппированные по годам: 2013 (a), 2014 (b), 2015 (c) 2016 (d).

IV. Анализ доли различных рейтингов

Для оценки доли публикаций с отрицательным рейтингом было рассчитана величина отношения общего количества публикаций с отрицательным рейтингом к общему количеству публикаций с положительным рейтингом для разных сабреддитов за разные года с шагом в месяц. Данное отношение может быть представлено как

$$\beta = N_d / N_u$$

где N_d – общее количество публикаций с отрицательными оценками в сабреддите, N_u – общее количество публикаций с положительным рейтингом в том же сабреддите. Результаты показаны на рис. 3.

Аналогичным образом мы вычисляем отношение общего количества публикаций с нулевым рейтингом в сабреддите к общему количеству публикаций в том же сабреддите. Данное отношение может быть представлено как

$$\beta_0 = N_0 / N$$

где N_0 – общее количество публикаций с нулевым рейтингом, N – общее количество публикаций. Результаты показаны на рис. 4.

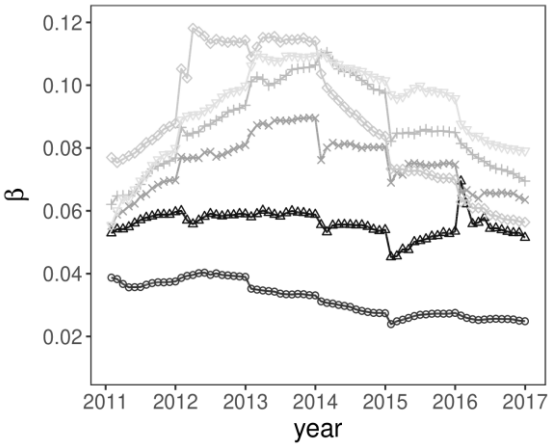

Рис. 3. Отношение общего количества публикаций с отрицательным рейтингом к общему количеству публикаций с положительным рейтингом для пяти сабреддитов: AskReddit (точки), IAmA (треугольники вверх), funny (плюсы), gaming (кресты), science (ромбы), todayilearned (треугольники вниз).

Как видно из рис. 4, отношение количества публикаций с нулевым рейтингом к общему количеству публикаций не так сильно меняется с годами, как отношения общего количества публикаций с отрицательным рейтингом к общему количеству публикаций с положительным рейтингом, как показано на рис. 3.

Кроме того, следует отметить, что характер зависимости β_0 схож для разных сабреддитов. Также отметим, что в 2013-2015 годах наблюдался некоторый рост доли публикаций с нулевым рейтингом для всех.

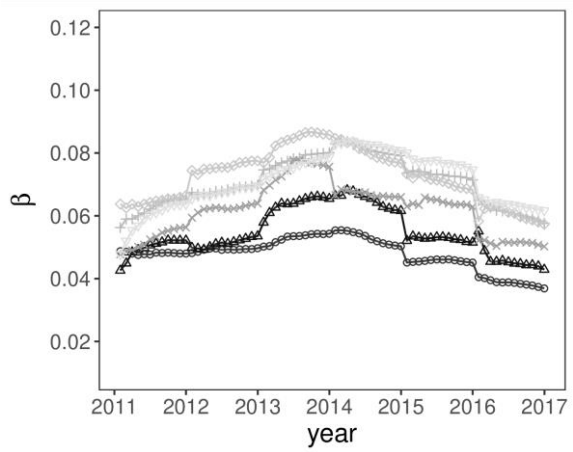

Рис. 4. Отношение общего количества публикаций с нулевым рейтингом в сабреддите к общему количеству публикаций: AskReddit (точки), IAmA (треугольники вверх), funny (плюсы), gaming (кресты), science (ромбы), todayilearned (треугольники вниз)

Кроме того, было показано, что отношение количества публикаций с нулевым рейтингом к общему количеству публикаций аналогично для разных сабреддитов, тогда как отношения общего количества публикаций с отрицательным рейтингом к общему количеству публикаций с положительным рейтингом отличается для сабреддитов AskReddit и IAmA.

Список литературы

[1] Rybski D. et al. Scaling laws of human interaction activity //Proceedings of the National Academy of Sciences. 2009. Т. 106. №. 31. С. 12640-12645.

[2] Rybski D. et al. Communication activity in a social network: relation between long-term correlations and inter-event clustering //Scientific reports. 2012. Т. 2. С. 560.

[3] Kawamoto T. A stochastic model of tweet diffusion on the Twitter network //Physica A: Statistical Mechanics and its Applications. 2013. Т. 392. №. 16. С. 3470-3475.

[4] Kawamoto T., Hatano N. Viral spreading of daily information in online social networks //Physica A: Statistical Mechanics and its Applications. 2014. Т. 406. С. 34-41.

[5] Van Mieghem P. Human psychology of common appraisal: The Reddit score //IEEE Transactions on Multimedia. 2011. Т. 13. №. 6. С. 1404-1406.

[6] Stoddard G. Popularity Dynamics and Intrinsic Quality in Reddit and Hacker News //ICWSM. 2015. С. 416-425.

[7] Olson R. S., Neal Z. P. Navigating the massive world of reddit: Using backbone networks to map user interests in social media //PeerJ Computer Science. 2015. Т. 1. С. e4.