АНАЛИЗ ВЗАИМОДЕЙСТВИЯ ПОЛЬЗОВАТЕЛЕЙ СЕТИ TWITTER

А.Н. Исангулова, А.Д. Снида
Научный руководитель: к.ф.-м.н., доцент М. Е. Семенов, к.п.н., доцент Л.А. Сивицкая
Национальный исследовательский Томский политехнический университет,
Россия, г.Томск, пр. Ленина, 30, 634050
E-mail: crazy_157@mail.ru

ANALYSIS OF THE INTERACTION OF TWITTER'S USERS

A.N. Isangulova, A.D. Snida
Scientific Supervisor: PhD, associate prof. M.E. Semenov, PhD, associate prof. L.A. Sivitskaya
Tomsk Polytechnic University, Russia, Tomsk, Lenina str., 30, 634050
E-mail: crazy_157@mail.ru

Abstract. *In our experiment, we present an empirical analysis of influence patterns in the social network Twitter. Using Twitter dataset we compare three different measures of influence: followers, retweets, and mentions. We examine how the three types of influential users performed in spreading popular news topics. The number of followers represents popularity of a user; retweets represent the content value of one's tweets; and mentions represent the name value of a user. Results are providing a better understanding of the different roles users play in social media.*

В связи с развитием социальных сетей – выделился новый класс задач. Эти задачи связаны с обнаружением сообществ и связанных подгрупп, анализом содержания социальных сетей, классификацией вершин в социальных сетях, анализом социального влияния участников социальной сети, прогнозированием формирования связей [1]. Перечисленные задачи могут решаться с использованием алгоритмической теории графов. Среди наиболее известных средств автоматического анализа социальных сетей отметим: NetMiner (http://www.netminer.com/index.php), NetworkX (http://networkx.lanl.gov), SNAP (http://snap.stanford.edu), UCINet (http://www.analytictech.com/ucinet), ORA (http://www.casos.cs.cmu.edu/projects/ora), Cytoscape (http://www.cytoscape.org), NodeXL (http://nodexl.codeplex.com) [2]. Цель работы – определение количественной оценки связи между действиями пользователей, объединенных определенной меткой (хештегом) в социальной сети Twitter.

Для описания отношений следования (направленные дуги) между пользователями (вершины графа) мы использовали социальный граф. На рис. 1 представлена страница из сети Twitter и фрагмент модели – социальный граф, построенный для пользователей, объединенных определенной меткой (хештегом). С применением социального графа описаны следующие отношения следования: а) пользователь *A* подписан на пользователя *B*, б) пользователь *A* поделился сообщением пользователя *B* с пользователями *C* и *D*, в) пользователь *A* прокомментировал сообщение пользователя *B*. Не ограничивая общности рассуждений будем придерживаться терминологии, принятой в Twitter и будем называть отношения следования: а) подписчик (follower), б) ретвит (retweet), в) упоминание (mention) соответственно. Ретвит содержит

ключевой фрагмент *RT@username* или *via@username*, упоминание содержит *@username*, где username – имя пользователя.

Рис.1. а) веб-страница, б) социальный граф (фрагмент)

Наш эксперимент заключается в выявлении линейной зависимости между следующими явлениями: количество подписчиков, ретвитов и упоминаний для конкретного пользователя социальной сети Twitter. Для реализации данного эксперимента были выбраны хештеги, относящиеся к разным темам: #ebola (вирус) и #lol (юмор). Для сбора исходных данных была использована программаNodeXL [2]. Выборка для хештега #lol за 03.02.2016 составила 2376 записей, а для хештега #ebola за период 17-18.10.2015 – 1917 упоминаний и ретвитов. Количество подписчиков зависит от количества пользователей, которые каким-либо образом взаимодействовали с записью, содержащей выбранный хештег.

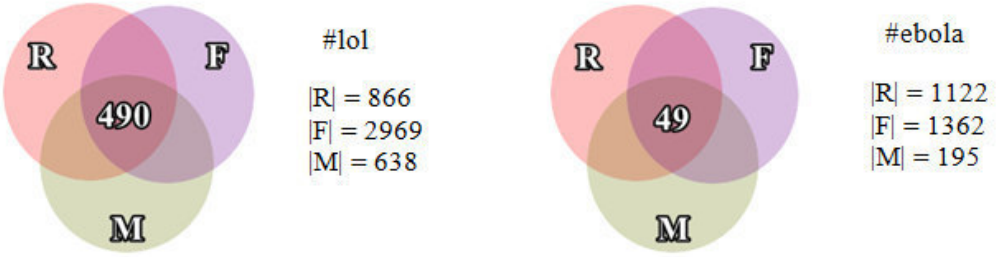

Рис. 2. Диаграммы вариационных рядов

На основе собранных данных сформированы три вариационных ряда: подписчики (F), ретвиты (R) и упоминания (M), мощности этих множеств приведены на рис. 1. Общее количество пользователей, находящихся во всех трех рядах для хештегов #lol и #ebola составляет |R∩F∩M|=490 и 49 соответственно (рис. 1). Фрагменты этих рядов, отсортированные по алфавиту имен пользователей, приведены в табл. 1. Для количественной оценки связи между явлениями использован коэффициент ранговой корреляции Спирмена [3]: $\rho = 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n}(x_i - y_i)^2$, где x_i и y_i – ранги пользователей, а n – объем выборки. Каждому элементу вариационного ряда присвоен соответствующий ранг, при совпадении значений признаков ранг вычисляется как среднее арифметическое. В нашем эксперименте ранг Спирмена для большинства случаев положительный. Это говорит том о линейной, но достаточно слабой связи. Самая высокая корреляция наблюдается между количеством ретвитов и упоминаний (табл. 2). Для проверки уровня

значимости коэффициента была вычислена статистика [3]: $T_{кр} = t(\alpha, k)\sqrt{\left(1-\rho^2\right)/(n-2)}$, $t(\alpha, k)$ – критическая точка двусторонней критической области. Если $|\rho|<T_{кр}$ – ранговая корреляционная связь между качественными признаками незначима, в противном случае существует значимая ранговая корреляционная связь.

Таблица 1

Вариационные ряды (фрагмент)

Retweets			**Followers**			**Mentions**		
Username	**R**	**Rank**	**Username**	**F**	**Rank**	**Username**	**M**	**Rank**
chrisblack10	1	271	chrisblack10	337	335	chrisblack10	1	319,5
damien_fern	1	271	damien_fern	43	465	damien_fern	2	92,5
ebitinnn	1	271	ebitinnn	155	459	ebitinnn	3	92,5
fabmaichard	1	271	fabmaichard	150	236	fabmaichard	1	319,5
gelinalaurente	2	32	gelinalaurente	293	348	gelinalaurente	2	92,5
gobiglexis	4	3	gobiglexis	298	346	gobiglexis	1	319,5
hailieem	1	271	hailieem	3828	93	hailieem	1	319,5
i_m_alive_	1	271	i_m_alive_	492	288	i_m_alive_	1	319,5
newday	3	9	newday	166348	26	newday	1	319,5
obeyfemmes	1	271	obeyfemmes	25871	52	obeyfemmes	2	92,5
. . .								
zlishhhh	1	271	zlishhhh	835	224	zlishhhh	1	319,5

Таблица 2

Коэффициент корреляции Спирмена ρ *при доверительной вероятности* α=*0,95*

Вариационные ряды	**#lol**			**#ebola**		
	ρ	$T_{кр}$	**связь**	**ρ**	$T_{кр}$	**связь**
Retweets, **Followers**	0,111	0,20	незначима	–0,12	0,29	незначима
Retweets, **Mentions**	0,538	0,17	значима	0,368	0,27	значима
Followers, **Mentions**	0,286	0,19	значима	0,131	0,29	незначима

Для достижения поставленной цели построен социальный граф, сформированы списки пользователей, использовавших определенный хештег в своих записях (упоминаниях и ретвитах). На основе полученных данных построены три вариационных ряда, вычислен коэффициент корреляции Спирмена, получена оценка значимости коэффициента корреляции Спирмена.

Установлено, что реакция аудитории на новости юмористического характера (хештег #lol) статически более значима, так как для вариационных рядов (Retweets, Mentions) наблюдается наиболее высокая корреляция ρ=0,538.

СПИСОК ЛИТЕРАТУРЫ

1. Батура Т.В. Методы анализа компьютерных социальных сетей // Вестник НГУ. Серия: Информационные технологии. 2012. – Том 10. – Выпуск 4. – с. 13-28.
2. Hansen D.L., Shneiderman B., Smith M. A. Analyzing Social Media Networks with NodeXL – Insights from a connected world, Elsevier Inc. – 2011, 284 p.
3. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. М.: Физматлит. – 2006, 816 с.