

Прикладная эконометрика, 2017, т. 47, с. 74–99.
Applied Econometrics, 2017, v. 47, pp. 74–99.

Е. С. Котырло¹

Социальные медиа: о чем и кому пишут их пользователи? Некоторые подходы к анализу данных

Изучение поведения социальных групп и отдельных индивидов, сегментация пользователей на основе проявляемых интересов и характера внутрисетевого взаимодействия — это те возможности, которые предоставляют социальные медиа. В статье обобщаются подходы к анализу социальных медиа на базе анализа текстов и анализа социальных сетей, сложившиеся к настоящему времени. Имеющийся инструментарий дополняется использованием показателей концентрации и расширенными возможностями визуализации результатов сетевого анализа. Проведенный на модельном примере анализ позволяет сделать следующие выводы: 1) пользователи могут быть успешно сегментированы по темам публикуемых сообщений для проведения маркетинговой кампании; 2) индекс Джини демонстрирует высокую степень неравенства пользователей по количеству публикуемых сообщений, фактически 10% пользователей, объединенных интересом к теме и связями друг с другом, публикуют 70–80% контента; 3) пользователи, общающиеся друг с другом через социальные медиа, проживают в относительной близости; 4) активность и центральность пользователя тесно связаны между собой.

Ключевые слова: анализ текстов; социально-сетевой анализ; социальные медиа; регрессионный анализ; индекс Джини.

JEL classification: M39; C18.

1. Введение

Доступность Интернета (табл. 1) и рост популярности социальных медиа позволяют рассматривать их как новый источник данных в маркетинге, политических исследованиях и других сферах деятельности. Методология анализа социальных медиа находится в стадии формирования, как и исследование пределов его возможностей. С одной стороны, пользователи социальных медиа выступают участниками социальной сети, подчиняясь законам сетевой динамики. С другой стороны, со временем накапливается огромный массив различных видов активности пользователей: размещение сообщений, графической и видео информации, репостов, комментариев и лайков, показывающих отношение к информации, размещенной другими авторами. Эта информация позволяет довольно детально реконструировать предпочтения пользователя и фокусировать маркетинговые или идеологические акции

¹ Котырло Елена Станиславовна — Национальный исследовательский университет «Высшая школа экономики», Москва; ekotyrl@hse.ru.

с учетом этих предпочтений. В то же время, знание статуса пользователя в социальной сети позволяет еще больше выиграть от точечной адресации рекламы, которая достоверно будет растиражирована ключевыми акторами. Изучение пользователей и сообществ, сегментация акторов на основе проявляемых интересов и характера внутрисетевого взаимодействия — это те уникальные возможности, которые предоставляют социальные медиа.

Таблица 1. Пользователи Интернета и социальных медиа в России

	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Доля пользователей Интернета во взрослом населении*	18	24.7	26.8	29	43	49.0	63.8	68.0	70.5	70.1
Число пользователей социальных медиа (млн чел.) **	—	—	—	—	—	—	—	60.5	65.5	69.2

Источник:

* <http://data.worldbank.org/indicator/IT.NET.USER.ZS?end=2015&locations=RU&start=2002&view=chart>.

** <http://www.statista.com/statistics/278410/number-of-social-network-users-in-russia/>.

В статье обобщаются подходы к анализу социальных медиа, сложившиеся к настоящему времени. Имеющийся инструментарий дополняется использованием показателей концентрации, расширенными возможностями визуализации результатов сетевого анализа. В работе используется совмещение анализа текстов и социально-сетевого анализа, структурирование результатов семантического анализа, обсуждаются проблемы и перспективы анализа социальных медиа в решении коммерческих и управленческих задач.

Анализ данных социальных медиа проиллюстрирован модельным примером. Статья подготовлена на основе серии проведенных исследований социальных медиа сетей. В качестве предмета анализа может выступать распространение в сети информации о бренде, политике, компании, товаре и т. д. В целях соблюдения коммерческой тайны предмет анализа, представленный в статье, изменен. Вместе с тем, характер распределения сообщений, так же как и метрики взаимодействия пользователей, сохраняют особенности, присущие реальным выборкам. Проведенное модельное исследование охватывает 800 участников одной из популярных в России социальных медиа, проживающих в 150 городах, и охватывает 2010–2015 гг. Пользователи, являясь членами различных групп, получали рассылку от 50 тыс. сообществ, имели 150 тыс. друзей и разместили на своих страницах почти 3 млн сообщений. Исследование выполнено в статистическом пакете R, находящемся в открытом доступе.

В качестве результатов исследования можно отметить несколько характерных моментов того, о чем и кому пишут пользователи социальных медиа.

1. Интерес к теме проявляют пользователи, имеющие набор определенных предпочтений при обсуждении в социальных сетях. Таким образом, пользователи могут быть успешно сегментированы по этим предпочтениям для проведения маркетинговой кампании.

2. Индекс Джини показывает, что пользователи существенно разнятся по своей активности в социальных медиа в отношении конкретной обсуждаемой темы. Фактически 10% пользователей, объединенных интересом к теме и связями друг с другом, публикуют 70–80% контента. Это может указывать на то, что часть пользователей преследует коммерческие интересы, публикуя сообщения по теме. Доминирование одного пользователя и низкая плотность сети свидетельствуют об использовании сети скорее как механизма направленной передачи информации, чем в качестве форума, где пользователи равноправны.

3. Пользователи, общающиеся через социальные медиа друг с другом, проживают в носительной близости. Это знание также полезно для планирования маркетинговой или управленческой акции.

4. Активность пользователей и центральность пользователя тесно связаны между собой. Это подтверждает результаты, полученные Kumar, Mirchandani (2012) и Kim, Han (2009).

2. Обзор литературы

Анализ социальных медиа может пониматься в двух смыслах. «Social media analytics» подразумевает анализ контента: внутреннего, относящегося к действиям или реакции на действия конкретной компании, которая анализирует эффективность собственного маркетинга (social media marketing); или внешнего — анализа сообщений пользователей социальных медиа с целью исследования предпочтений целевой или потенциально-целевой группы маркетинга (Actionable social analytics..., 2012; Social media marketing..., 2013). «Social network analysis», не обязательно анализ виртуальных социальных сетей, подразумевает анализ взаимодействия участников сети. Таким образом, направления «анализ текстов» и «анализ социальных взаимодействий» развиваются независимо и привлекают к себе внимание многих исследователей. С одной стороны, эти исследования имеют коммерческий интерес (Lewis et al., 2008). С другой стороны, развитие вычислительных мощностей в последние десятилетия позволило сделать применение этих методов масштабными, а не элитными, как ранее. Данные социальных медиа хорошо структурированы в той части, которая касается профиля пользователя, что упрощает анализ (Lewis et al., 2008). Однажды разработанный инструментарий может использоваться многократно.

Профессиональные пакеты для медиа аналитики и маркетинга в какой-то мере служат образцом того, какие инструменты служат для анализа социальных медиа. Например, в (Social media marketing..., 2013) сравниваются около 20 решений ведущих производителей программного обеспечения. Позиции сравнения указывают на то, что единицей анализа является либо сообщение и его содержание, либо пользователь, как производитель медиа контента. В частности, для анализа доступны число подписок и подписчиков. При этом связи между пользователями, как возможность направленного обмена сообщениями, не выступают объектом анализа. Таким образом, коммерческий анализ социальных медиа в большей мере фокусируется на контенте и тех сетевых характеристиках пользователя, которые не связывают его с другими конкретными пользователями.

Анализ активности и роли участников сетей непосредственно по данным социальных медиа ограничивается доступом к информации о пользователях. Наиболее открытая и поэтому чаще представляемая в результатах исследований сеть — это социальная сеть Twitter. Данные этой сети являются основой большинства исследований в этой области (см. например, Cha et al., 2010; Cheng et al., 2010; Cheong, Cheong, 2011; Huberman et al., 2009; Kwak et al., 2010; Morstatter et al., 2013; Suh et al., 2010; Wakamiya et al., 2011). В Twitter исследователям доступны данные о конкретных подписках и подписчиках пользователей, что позволяет применить к анализу этой сети методы анализа социальных сетей.

Несмотря на то что маркетинг — не единственная причина исследований социальных медиа, есть мнение, что практически каждый проект, созданный во «ВКонтакте», на Facebook или в Instagram, является коммерческим. И в открытом доступе можно найти расценки на рекламу

в популярных группах социальных медиа, цена которой варьируется от нескольких сотен до сотен тысяч рублей в зависимости от числа подписчиков группы (<https://vk.com/public97197153/>).

2.1. Анализ текстов пользователей социальных медиа

Анализ текстов (text mining) представлен трудами (Bouchet-Valat, 2013; Feinerer, 2008; Feinerer et al., 2008; Feinerer, 2017; Lebart et al., 1998; Lang, 2004; Williams, 2016). Единицей такого рода анализа служит сообщение. Это наиболее доступная часть для анализа социальных медиа, что подтверждается многочисленными программными решениями, предлагающими стандартные показатели того, о чем пишут пользователи (Social media marketing..., 2013). Практические материалы предлагают следующие индикаторы анализа социальных медиа для достижения коммерческих целей (см., например, (Actionable social analytics..., 2012)): доля сообщений с ключевыми упоминаниями в числе релевантных теме сообщений, число подписчиков по ключевой теме (достижимость аудитории), отношение числа комментариев и других пользовательских реакций к общему числу тематических сообщений (вовлеченность), темпы роста этих показателей во времени и в период проведения маркетинговой акции, доля конверсий в продажи через социальные сети от общего числа продаж. При этом авторы отмечают неэффективность таких индикаторов как слишком общих и на самом деле не дающих достоверной картины ни о степени вовлечения аудитории, ни о потенциальном влиянии бренда, ни об отношении к нему (Actionable social analytics..., 2012; Ruhi, 2014, Social media marketing..., 2013). Таким образом, поиск индикаторов и методик, характеризующих социальные медиа, остается весьма популярной и коммерчески востребованной темой.

Suh et al. (2010) находят, что источник распространения информации (URL) и ее тема (hashtag) наиболее важны в распространении информации через социальные медиа (через репосты). Число подписчиков и длительность использования аккаунта также положительно связаны с потенциальной аудиторией охвата, в то время как активность пользователя (число авторских постов) не влияет на характер распространения новой новости и ожидаемый охват аудитории (там же).

В качестве индикаторов более сложного уровня используются отношение к бренду (sentiment analysis) (Asur, Huberman, 2010; Liviu, 2011; O'Connor et al., 2010; Zhang, Skiena, 2009), размер аудитории (influence), сегментация пользователей по социально-демографическим характеристикам, если такие данные доступны (Actionable social analytics..., 2012; Ruhi, 2014; Social media marketing..., 2013). Для задач маркетинга наиболее интересными являются как раз показатели влияния на принятие решения, например доля покупок в общем числе пользователей, проявляющих интерес к продукту или бренду (коэффициент конверсии в продажи). Однако расчет такого рода индикаторов затруднен тем, что покупка совершается, как правило, на иной платформе, и не наблюдается непосредственно в социальных медиа (Ruhi, 2014). Хотя появляются решения, совмещающие функции социальных медиа и интернет-магазинов, например китайский мессенджер WeChat (Луганская, 2017). Обобщая, можно сказать, что любой реальный выбор — голосование, покупки и т. д. — не может быть напрямую ассоциирован с тем, что пользователи об этом пишут. Однако можно ожидать некий агрегированный эффект проявленного интереса на распознавание бренда или результаты выборов. В качестве примеров эффективного использования такого рода анализа приводят предсказания распределения премий Оскар (Yu, Kak, 2012) или результаты выборов (Ceron et al., 2014).

2.2. Онлайн активность пользователей социальных медиа

Исследователи оперируют понятием онлайн-активности пользователя (Asur, Huberman, 2010; Ceron et al., 2014; Corley et al., 2010; Lewis et al., 2008), которая позволяет добавить пользовательские характеристики к анализу контента. В качестве индикаторов активности выступают частота посещений определенной социальной сети или период с момента последнего обновления (Lewis et al., 2008). Анализ частоты сообщений во времени оказывается полезным для предсказания кассовых сборов от кинопремьеры (Asur, Huberman, 2010). Corley et al. (2010) используют этот показатель в анализе упоминаний о гриппе, а Ceron et al. (2014) — для измерения рейтинга претендентов на пост премьер-министра Италии.

Ряд исследований посвящен анализу причин обращения пользователей к социальным медиа. Исследователи находят, что активность в социальных медиа дополняет оффлайн-жизнь, усиливая имеющиеся социальные связи (социальный капитал), что, в частности, может быть верифицировано географией общения (Смирнов и др., 2016). Онлайн-активность также дополняет оффлайн теми возможностями, которые иначе вряд ли имели бы место, например простотой общения с незнакомцами (Ellison et al., 2007, 2011; Hampton, 2002; Hampton, Wellman, 2003; Lampe et al., 2006; Wellman et al., 1996; Смирнов и др., 2016). Хотя есть исследования, подтверждающие не дополнение, а замещение реальной социальной жизни общением в социальных медиа, и негативное влияние такого замещения на психоэмоциональное состояние индивида (Bessiere et al., 2008).

Burke et al. (2011) выделяют: 1) прямое общение с друзьями, которое включает сообщения, посты, чат; 2) пассивное наблюдение за социальной жизнью; 3) вещание, когда социальные медиа используются именно как медиа, т. е. для распространения информации, не имеющей конкретного адресата. Они отмечают различие между типологизацией по мотиву использования социальных медиа и по характеру их использования, сложность наблюдения как мотивов, так и поведения, и отсутствие корреляции между ними. Ellison et al. (2007) типологизируют активность в сети следующим образом: поиск социальной информации (social information seeking), т. е. информации о людях, которых пользователь встретил в реальной жизни; эмоциональный обмен информацией (expressive information sharing). К имеющимся типологиям Papacharissi, Mendelson (2008) добавляют привычное времяпровождение, конкретную мотивацию которого сложно выявить. Следовательно, такое поведение вряд ли можно связать с какой-либо ожидаемой результативностью, например, побуждением к покупке, формированием симпатии к бренду или персоне. Следует отметить, что такого рода анализ выходит за рамки анализа самих социальных медиа и требует использования дополнительных источников, например опросов.

2.3. Анализ социальных медиа средствами социально-сетевого анализа

Анализ взаимодействия пользователей сети опирается на более зрелую теорию в сравнении с анализом текстов (Burt, 1992; Butts, 2014; Correa et al., 2012; Freeman, 1977, 1979; Goodreau et al., 2008; Sabidussi, 1966; Wasserman, Faust, 1994; Winship, Mandel, 1983). Однако в практическом плане данные сетей сложнее поддаются анализу и накоплению информации. С одной стороны, большинство социальных медиа не позволяют использовать персональную информацию о пользователях, в том числе, персонифицировать входящие и исходящие

контакты. С другой стороны, алгоритмическая сложность решаемых задач быстрее чем линейно растет с увеличением числа участников сети (Butts, 2000). В то же время, именно статус пользователя в сети, т.е. его участие в передаче информации, играет важную роль в оценке потенциально охватываемой аудитории, эффективности охвата во времени и, соответственно, в плане достижения лояльного отношения к бренду (продукту, политике и т.д.). Описанию доступных на социальных медиа техник анализа социальных сетей посвящены работы (Borgatti et al., 2009; Catanese et al., 2011; Lewis et al., 2008).

2.4. Совмещенный анализ контента и социально-сетевой анализ

Анализ социальных медиа сетей, расширяющий представление о взаимодействии акторов сети и анализа публикуемого контента, получил свое развитие в работах последнего десятилетия (Bhatt et al., 2010; Cha et al., 2010; Cheong, Cheong, 2011; Cheng et al., 2010; Kim, Han, 2009; Kumar, Mirchandani, 2012; Kwak et al., 2010; Morstatter et al., 2013; Ruhi, 2014; Wakamiya et al., 2011; Yu, Kak, 2012). Концептуально подход продолжает складываться, поэтому эта область утилизирует и тестирует те возможности, которые разработаны в смежных областях знаний, например, статистики и маркетинга (Ruhi, 2014).

Предсказание поведения аудитории в зависимости от содержания, частоты сообщений и взаимосвязей между акторами выступает важной задачей как в целях маркетинга, так и других управленческих целях. Yu, Kak (2012) рассматривают две ключевые с этой точки зрения метрики — эмоциональный окрас (sentiment index) и частоту сообщений в единицу времени (postrate). Huberman et al. (2009) находят, что число подписчиков у того или иного пользователя (in-degree) растет вместе с его активностью в сети (число постов) до определенного уровня. Дальнейший рост подписчиков после достижения этого уровня уже не зависит от активности. Для сети Twitter таким порогом выступает 300 подписчиков (там же). Однако число подписок и подписчиков (out-degree и in-degree) не позволяют судить о реальном влиянии пользователя в сети, т.е. выделить потенциально эффективных распространителей информации (influencers) (Bhatt et al., 2010; Cha et al., 2010; Kumar, Mirchandani, 2012).

Характеристики сети в целом, например плотность сети (density), также позволяют судить об активности пользователей в обмене информацией, а потому интересны для анализа. Исследователи отмечают, что, несмотря на довольно высокую плотность сетей в социальных медиа, только каждый пятый контакт предполагает взаимный, а не односторонний обмен сообщениями (Kwak et al., 2010). Важную роль в выборе пользователей сети, рекрутируемых для проведения «вирусной» маркетинговой кампании (viral marketing), играют их индивидуальные показатели эффективности сетевых позиций — собственный вектор эффективности сетевых позиций (eigenvector centrality scores), близости (closeness), посредничества (betweenness), централизации (centralization) — как характеристики роли пользователя в обеспечении обмена информацией между другими участниками и эффективности сети (Burt, 1992; Cheong, Cheong, 2011; Freeman, 1977, 1979; Hinds, McGrath, 2006; Kim, Han, 2009; Morstatter et al., 2013). Morstatter et al. (2013) рассчитывают достижимость аудитории как число акторов сети, соединенных друг с другом менее чем через заданное число посредников.

Kumar, Mirchandani (2012) и Kim, Han (2009), опираясь на анализ контента и социально-сетевого анализа, предлагают индекс «липкости» (stickiness index), который выделяет

пользователей — не только потенциально эффективных распространителей информации по индексу центральности, но и заинтересованных в продвигаемом продукте или торговой марке. Этот интерес определяется как число постов по данной теме среди общего числа сообщений отобранных пользователей. Включение таких пользователей в маркетинговую кампанию способно удвоить распознавание торговых марок и доходов компании. Примером служит маркетинговая кампания производителя мороженого премиум-класса в Индии Hokey Pokey, реализованная через отобранных пользователей социальных сетей (Kumar, Mirchandani, 2012).

Morstatter et al. (2013) развивают концепцию сетевого взаимодействия в Twitter и предлагают в качестве показателей потенциального влияния: 1) подсети пользователей, публикующих взаимные репосты (User $\times$ User retweet networks); 2) подсети, образованные по принципу отбора пользователей, публикующих по данной теме (User $\times$ Hashtag content networks); 3) темы, одновременно встречающиеся в сети (Hashtag $\times$ Hashtag co-occurrence networks). Однако в своей работе они анализируют только первый вариант подсетей.

Поиск общих характеристик между участниками, формирование профиля типичного эффективного распространителя информации — важная маркетинговая задача (Kumar, Mirchandani, 2012). Yu, Kak (2012) выделяют регрессионный, кластерный анализ, анализ искусственных нейронных сетей и дерева решений в качестве методов, позволяющих описывать поведение аудитории и связь с реальными событиями. Они отмечают, что на данном этапе использование этих методов для изучения социальных медиа незначительно, но определенно открывает новые возможности. Следует отметить работу Polonski, Hogan (2015). Авторы анализируют сообщества в Facebook, используя многоуровневый регрессионный анализ, и находят, что метрики центральности пользователя значимо влияют на его активность в сети. Эта активность измеряется в вероятности начала общения, получении отклика на размещенный контент и реагировании на сообщения других. Регрессионный анализ используется Asur, Huberman (2010) для предсказания кассовых сборов от проката кинофильмов.

Lomi et al. (2011) обращают внимание на одновременное формирование стратегии поведения пользователя в сети, принадлежности к той или иной социальной группе (сети) и статуса в ней, как результата личных достижений. Можно сказать, что оффлайн- и онлайн-поведение пользователя постоянно меняются под влиянием друг друга, меняет пользовательские предпочтения, выбор друзей и социальный статус как в его реальной, так и виртуальной жизни. Это усложняет объяснение причинно-следственной связи между ними, и, соответственно, достоверность результатов регрессионного анализа.

Для изучения характера числа упоминаний интересующей темы можно успешно применить такие популярные в экономике измерения, как индекс Джини и кривую Лоренца, служащие для оценки уровня неравенства по доходам и его концентрации. По аналогии с этим подходом можно рассмотреть неравенство по числу сообщений, публикуемых одним пользователем, понять, насколько равноправны пользователи в распространении информации. Возможность применения показателей концентрации для социальных сетей обоснована в работах (Kelly et al., 2014; Kunegis, 2017; Lopes et al., 2012; van Mierlo et al., 2016). Показатели концентрации представляются автору интересными для анализа социальных медиа, хотя они пока не получили широкого признания у аналитиков. Например, можно исследовать, в какой мере высокие показатели концентрации связаны с длительностью существования темы или подсети.

2.5. Проблема репрезентативности выборки

Вопрос о репрезентативности выборки, построенной на отборе пользователей социальных медиа или их сообщений, не является праздным. Во-первых, при оперировании одновременно массивом сообщений и выборкой индивидов возникает вопрос, что должно быть репрезентативным — пользователь или сообщение? Во-вторых, относительно чего должна оцениваться репрезентативность пользователей? Относительно среднего пользователя социальных медиа, некоторой социальной группы, населения, пользующегося или не пользующегося социальными медиа? В отношении репрезентативности сообщений, большое число сообщений может гарантировать репрезентативность. Это дает возможность относиться к выборке из сотен тысяч и миллионов сообщений как к репрезентативной. Репрезентативность пользователей как участников сети более проблематична. Во-первых, сеть, которую можно проанализировать современными техническими средствами без временных потерь, включает не так уж много пользователей, в лучшем случае несколько тысяч. Во-вторых, отбор пользователей социальных медиа по принципу включенности в ту или иную социальную сеть или социальную группу неизбежно ведет к нерепрезентативности выборки (Lewis et al., 2008).

Проблеме отбора выборки, сохраняющей характеристики социальной сети в целом, посвящены работы (Borgatti et al., 2006; Costenbader, Valente, 2003; Granovetter, 1976), которые демонстрируют, что в плотных сетях показатели центральности устойчивы к случайному добавлению новых акторов и ребер. Lewis et al. (2008) отмечают, что характеристики сети зависят не только от социально-демографических характеристик пользователей, но и от их активности.

Таким образом, проблема репрезентативности действительно есть и является полем для дальнейших научных исследований. В то же время, данные социальных медиа имеют неоспоримое достоинство, поскольку являются результатом своего рода натурального эксперимента и позволяют избежать искажений, которые неизбежно возникают при ответах на вопросы интервьюера (Lewis et al., 2008; Marsden, 2003), несовершенства в отражении событий и чувств (Brewer, Webster, 1999; Brewer, 2000), и других искажений, возникающих при проведении социологических опросов (Bernard et al., 1984; Butts, 2003; Feld, Carter, 2002; Marsden, 1990).

3. Эмпирический анализ

3.1. Основные гипотезы. Данные

Несмотря на то что статья скорее представляет собой обзор возможного инструментария для анализа социальных медиа, можно сформулировать гипотезы, нашедшие подтверждение по итогам проведенного анализа.

Гипотеза 1. Ключевой интерес к теме проявляют пользователи, имеющие набор определенных предпочтений или тем, упоминаемых в социальных сетях.

Гипотеза 2. Пользователи существенно разнятся по своей активности в социальных медиа в отношении конкретной обсуждаемой темы. Это может указывать на то, что часть пользователей преследует коммерческие интересы, публикуя сообщения по определенной теме.

Гипотеза 3. Пользователи, общающиеся через социальные медиа друг с другом, проживают в относительной близости.

Гипотеза 4. Активность пользователей и центральность пользователя тесно связаны между собой.

В работе используются методы, разработанные для статистической программы R, хотя подобные средства разработаны и для других статистических пакетов: SPSS, Stata, Python и др. Структура данных представлена в табл. 2.

Выборка строилась следующим образом. Сначала были отобраны пользователи, в чьих сообщениях упоминалась ключевая для исследования тема (предмет анализа). Для этих пользователей был полностью выгружен контент сообщений за доступный период, находящийся в публичном доступе. В соответствии с методологией Morstatter et al. (2013), в анализе присутствует подсеть пользователей, публикующихся по данной теме (User $\times$ Hashtag content networks). По этому массиву частотный анализ позволил выделить другие интересы отобранных пользователей (Hashtag $\times$ Hashtag co-occurrence networks). Для отобранных пользователей были составлены списки групп, на которые они подписаны, и друзей, с которыми они находились в контакте. Сообщения пользователей датированы, что позволяет проанализировать динамику интереса пользователей к теме. Длительный период позволяет отнести к одним темам, постоянно упоминаемым на протяжении длительного времени, как к предикторам по отношению к другим, относительно новым, например предмету анализа. Число подписок и подписчиков доступно только на конец исследования. Поэтому динамика социальных сетей в данном исследовании не обсуждается.

Таблица 2. Структура данных

Выгрузка с сообщениями пользователей социальной сети	Выгрузка с сообществами пользователя и друзьями
Дата	ID пользователя
ID сообщения	Url пользователя
Заголовок	ID группы
Url сообщения	Url группы
ID автора	Название группы
Url автора	ID пользователя
Страна	Url пользователя
Регион	ID друга
Город	Url друга
Текст	

Проведенное модельное исследование охватывает 800 участников одной из популярных в России социальных медиа, проживающих в 150 городах России, и период 2010–2015 гг. В данном случае география пользователей сети следует за проводимой маркетинговой кампанией, т. е. включает крупные города по всем федеральным округам, и те средние города, которые были затронуты акцией. Пользователи, являясь членами различных групп, получали рассылку от 50 тыс. сообществ, имели 150 тыс. друзей и разместили на своих страницах почти 3 млн сообщений.

3.2. Анализ текстов

Контент-анализ позволяет выявить темы, интересующие пользователей из сотен и тысяч страниц текста. В анализе социальных медиа как социальных сетей использованы работы Butts (2014), Correa et al. (2012), Goodreau et al. (2008). Для представления результатов использовался пакет графического анализа gplots (Warnes et al., 2016). Будем считать, что темой исследования был водный туризм с целью выявления пользователей, через которых можно распространять информацию о маркетинговой кампании. Ключевые слова для отбора тех, кто увлекается данной темой: байдарка, сплав, рафтинг, катамаран и др. Видно, что попытка объединить пользователей по этим словам неизбежно приведет к определенному зашумлению выборки, поскольку одно и то же слово может использоваться в разных значениях. Сплав может быть спортивным, а может быть транспортировкой леса по воде или сплавом металлов и т. д. К сожалению, зашумление выборки и возможное смещение оценок — это неотъемлемый негативный эффект автоматической обработки больших массивов неструктурированных данных. Масштабы зашумления зависят от грамотного составления словаря, соответствующего интересу к определенной теме. Таким образом, интерес к теме формируется не одним ключевым словом, а целым словарем, составление которого — ответственная и кропотливая задача.

Еще одна проблема — как отличить пользователей-профессионалов от любителей. В пользовательских группах по любой тематике присутствуют и те и другие. Задачей первых является формирование лояльности к бренду. В частности, ими постоянно упоминается продвигаемая торговая марка или продукт. Любители проявляют интерес от случая к случаю, могут резко менять предмет обсуждения, поэтому поиск по одному ключевому слову не позволит собрать полную картину интереса к теме. Они «стремятся» не к финансовой выгоде, а к получению психо-эмоционального комфорта за счет общения, к экономической поддержке имеющихся контактов up-to-date или получению информации, поддерживающей социальный статус (endorsement). Смешение пользователей, проявляющих спонтанный интерес, и тех, кто целенаправленно участвует в продвижении некоторого продукта, приводит к искажению результатов анализа — смещению оцениваемого профиля пользователя и, соответственно, к недостоверному прогнозу результативности маркетинговой кампании. Выделение профессионалов из общего числа пользователей, обсуждающих конкретную тему, представляется интересной и полезной исследовательской задачей, но в данной статье она не решается.

Для исследования корреляций между словами можно использовать функцию FindAssocс пакета tm, которая задает параметры близости в соответствии с особенностями анализа текстовой информации. При этом ассоциации будут устанавливаться между конкретными словами. Для их синонимов будут посчитаны новые метрики. Но такой путь является слишком длинным. Поэтому дальнейший подход к анализу сообщений строится не на возможностях, которые разработаны для неструктурированной информации, а на переходе к привычному структурированному формату данных.

Полученные частотные характеристики вхождения слов в сообщения позволяют выявить темы, популярные у пользователей, а затем поставить сообщениям пользователей в соответствие маркеры. Водный туризм может быть связан с темами: рыбалка, бардовская песня, коллекционирование бабочек, кулинария, фотография и многими другими хобби и профессиональными занятиями. Эти маркеры далее служат атрибутами сообщений: 1 — есть

слово или целая тема, скомпонованная по ключевым словам, 0 — нет. Здесь следует обратить внимание на то, что более достоверно отбирать именно некоторый набор ключевых слов по теме, чтобы отметить интерес к ней маркером. Проведенное с помощью маркеров структурирование сообщений позволяет понять, как сообщения по интересующей теме распределены по пользователям, и далее перейти к традиционным методам статистической обработки данных — кластеризации, корреляционному и регрессионному анализу. Благодаря этому можно установить профили пользователей как совокупности устойчивых интересов и социально-демографических характеристик. Таким образом, далее фокусируемся не на исследовании содержания сообщений в целом, а на изучении отдельных пользователей: того, что они пишут, и кто потенциально это читает.

То, что время и дата публикации сообщения регистрируются, позволяет понять, какова динамика интереса к той или иной теме (рис. 1). По динамике упоминания ключевых слов можно установить, проявлялся ли интерес к объекту в течение долгого времени или появился внезапно, ситуативно, переключился ли интерес с какого-то другого бренда, оценить регулярность интереса к объекту по частоте сообщений в единицу времени и т. д. Можно наложить на динамику сообщений временной ряд проходящих PR-акций и оценить их эффективность.

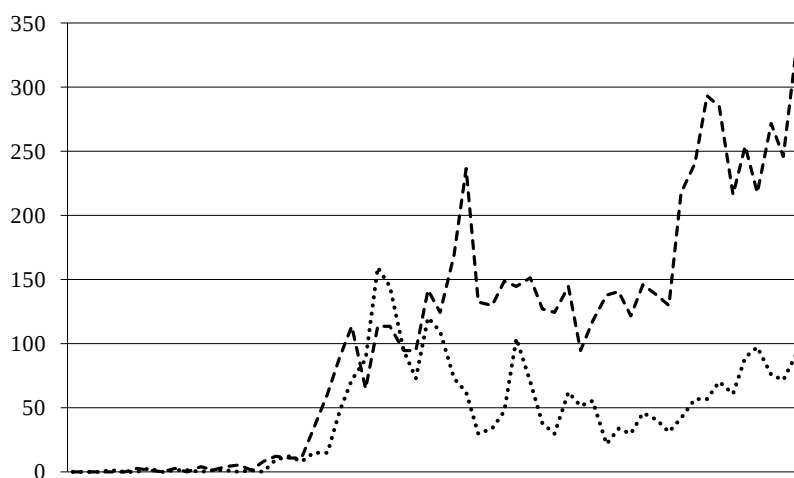


Рис. 1. Динамика числа сообщений (условных единиц) с упоминаниями двух конкурирующих брендов (за условный период)

Независимо от предмета исследования, была выделена следующая особенность создания сообщений в социальных сетях: существенное различие медианы и среднего арифметического. Из таблицы 3 видно, что есть единичные пользователи (или боты?), которые создали десятки тысяч сообщений, тогда как половина пользователей написала лишь несколько или несколько десятков сообщений, в зависимости от темы.

Результаты расчета индекса концентрации по анализируемым темам показывают высокие коэффициенты Джини (более 0.7), что, возможно, свидетельствует о незрелости сети. Интерпретация коэффициента следующая: «небольшое число участников соцсетей (10%) играет ключевую роль в вещании по теме, передавая 70–80% сообщений». Максимальное число сообщений одного участника по условной теме «Интерес пользователя» достигает

Таблица 3. Распределение числа сообщений по темам (сообщения 800 отобранных участников)

Тема	Медиана			Среднее	Максимум на одного пользователя	Всего сообщений
	Все	Женщины	Мужчины			
Всего сообщений	700	900	210	3297	116330	2637599
Тема 1. Интерес пользователя	85	165	8	1435	80183	1148038
Тема 2	2	4	0	15	750	12260
Тема 3	1	2	0	2	390	1960
Тема 4	3	3	4	51	11571	40535
Тема 5	1	1	1	30	7546	23922
Тема 6	40	40	14	208	6452	166337
...						

80 тыс. (табл. 3). Следует отметить, что большое число публикуемых сообщений вовсе не свидетельствует о большой аудитории у автора. Для того чтобы оценить потенциальную аудиторию, перейдем к анализу взаимосвязей между пользователями.

Нередко частотный анализ выступает главной и завершающей процедурой. Между тем, для изучения пользователей как группы полезно понять, каковы социально-демографические характеристики исследуемой группы, и какие еще темы интересуют пользователей группы. Наличие этих характеристик пользователей дает возможность рассмотреть интерес к теме с точки зрения половозрастного распределения, географии пользователей или уровня их образования. Хотя следует оговориться, что пользователи нередко указывают вымышленные данные или опускают эти сведения при регистрации в соцсетях (Смирнов и др., 2016).

3.3. Взаимодействие пользователей социальных медиа

Пользователи социальных медиа имеют друзей и подписаны на тематические группы. Анализ структуры социальных медиа дает возможность понять, как передается информация. Он предполагает исследование взаимодействия пользователей между собой. В статистической программе R такой анализ реализуется в пакетах социально-сетевых анализа: sna, statnet, network и др. В частности, можно сделать качественные и количественные выводы об эффективности передачи и распространения информации в социальных сообществах, сравнить их между собой, как информационные каналы. Метрики, определяющие статус в сети отдельных пользователей, могут быть включены в качестве характеристик для их кластеризации. Таким образом, к социально-демографическим характеристикам анализ соцсетей позволяет добавить потенциал пользователя в продвижении идей, товаров, брендов.

Как было сказано выше, пользователи, являясь членами различных групп, получали рассылку от 50 тыс. сообществ и имели 150 тыс. друзей. Анализ такого количества взаимосвязей, во-первых, требует больших вычислительных мощностей, а во-вторых, визуализация результатов проблематична. Поскольку в фокусе исследования находились 800 пользователей, было решено ограничиться теми из них, кто знаком хотя бы с одним пользователем из выборки. Это позволило сократить рассматриваемую группу до 380 пользователей. Такая

подсеть репрезентативна по отношению к теме анализа и пользователям, общающимся друг с другом и имеющим общий интерес к теме.

Характеристики этой подсети следующие. Из потенциально возможных 72 тыс. прямых связей между пользователями существует только 1165 контактов, что отражается в относительно низкой плотности сети (0.016) и ее несвязности, т. е. не все пользователи потенциально могут получать информацию друг от друга. Однако 99% могут получать информацию через одного посредника (транзитивность). Предполагается, что сеть не ориентирована, все пользователи как принимают, так и передают информацию друг другу. Следовательно, она не иерархична. Центральность сети по степени контактов и возможностям посредничества — 0.3. В среднем каналы передачи взаимозаменяемы с показателем 7.7 возможных путей для передачи информации между двумя пользователями.

Обычно социальный граф представляет собой набор вершин, соединенных ребрами. Рисунок 2² иллюстрирует, как можно комплексно визуализировать и оценить характеристики пользователя соцсети. Вершины — это пользователи, ребра — контакты (друг, подписчик группы). Дополнительно цветом можно показать пол пользователя (возраст, географию, образование). Число сообщений данного пользователя по исследуемой теме можно отобразить размером узла, а центральность — многоугольником, соответствующим вершине: чем больше углов, тем выше значение индекса центральности. Таким образом, можно не только визуализировать взаимосвязи между участниками сети, но и отметить различия акторов по их центральности, активности, социально-демографическим характеристикам.

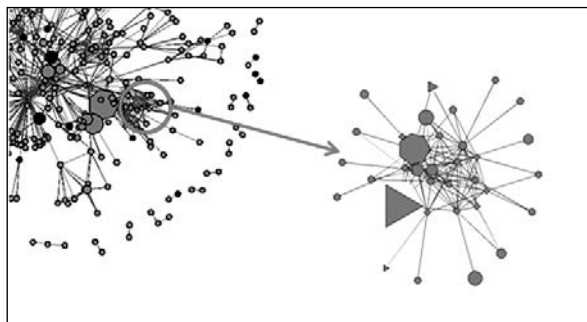


Рис. 2. Взаимодействие пользователей социальных медиа.

Размер точки — число созданных сообщений, цвет — пол пользователя, число углов у вершины — центральность пользователя в сети

Рисунок 3³ отражает распределение участников по территории России. На нем хорошо видно, что, несмотря на возможности поддерживать контакты через Интернет со всем миром, большая часть общающихся друг с другом через социальные медиа пользователей проживает в относительной близости друг к другу (в одном федеральном округе) и, скорее всего, локализована общим местом работы или обучения (Смирнов и др., 2016).

² В статье приведен лишь фрагмент, рисунок в полном формате можно получить по запросу у автора, ekotyrllo@hse.ru.

³ См. предыдущую сноску.

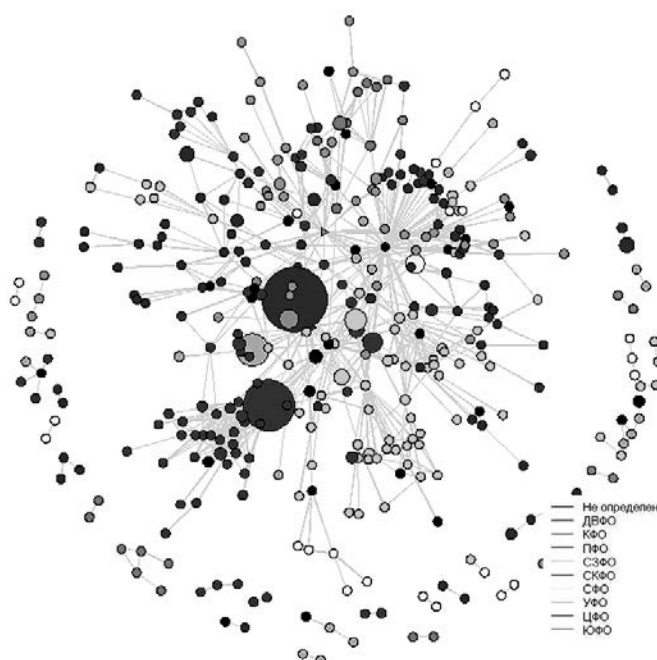


Рис. 3. География пользователей. Размер точки — число сообщений по исследуемой теме данного актора. Цвет — регистрация пользователя в определенном федеральном округе

Таблица 4. Характеристики центральности наиболее активных пользователей (сообщения 800 отобранных участников)

	Идентификатор пользователя					Максимум	Медиана
	id1	id2	id3	id4	...		
Число прямых контактов	35	6	15	3		121	3
Посредничество	4716	588	3883	0		44996	0
Индекс центральности (evcent)*	0.30	0.02	0.03	0.00		0.30	0.00
Число контактов через одного посредника	51	9	21	4		171	4
Число контактов через двух посредников	417	8	49	3		501	0
Число контактов через трех посредников	4600	34	327	5		5749	0
Число контактов через пять посредников	47552	297	1926	7		59437	0
Кластер по близости друг к другу	1	1	1	1		35	1
Число сообщений по ключевой теме	504	54	227	27		504	0

Примечание. * evcent — собственный вектор эффективности сетевых позиций (eigenvector centrality scores).

По данным, представленным в табл. 4, видно, что пользователи, наиболее часто упоминающие тему, выступающую предметом анализа, характеризуются высокими индексами центральности, т. е. имеют возможность распространения информации. Они также

объединены в общий кластер по признаку близости, т. е. обмениваются между собой информацией при минимальном числе посредников.

3.4. Корреляционный, кластерный и регрессионный анализ

Корреляционный, кластерный и регрессионный анализ позволяют выявить профили пользователей и разбить пользователей на сегменты, которые помогут провести сфокусированное продвижение товара или бренда. Очевидно, что социально-демографические характеристики инвариантны, тогда как интерес к той или иной теме — явление временное. С этой точки зрения более правомерно использовать множественный корреляционный анализ или кластеризацию, которые не устанавливают причинно-следственные связи между атрибутами сообщений (темами), а просто дают возможность выявить устойчивые взаимосвязи. Но в ряде задач, по мнению автора, правомерно применение регрессионного анализа. Например, если анализируется эффективность конкретной PR-кампании, которая опиралась на определенные стабильные потребительские предпочтения, проявляющиеся в обсуждаемых темах. В этом случае тема, соответствующая контенту PR-кампании, может рассматриваться как эндогенная по отношению к другим пользовательским интересам. Такой эндогенной темой может быть интерес к акции распродажи инвентаря для водного спорта некоторой крупной сетью спортивных магазинов в конкретном сезоне.

Следует отметить, что используемые регрессоры выступают в качестве предикторов. Построенная модель использует данные в статике, т. е. исследуется скорее ассоциативная связь между темами, социально-демографическими характеристиками и статусом в социальных медиа. Речь не идет о причинно-следственной связи между темами или центральностью и интересом к другим темам из-за очевидной эндогенности. Решением проблемы эндогенности могло бы стать применение инструментов. Однако выбор инструментов из имеющихся данных затруднен. В качестве наиболее доступных кандидатов можно рассмотреть пол, возраст, место жительства пользователя. Однако эти характеристики слишком общие и будут плохими инструментами. Существуют факторы, которые могли бы послужить хорошими инструментами, но они недоступны для наблюдения при анализе только социальных медиа. Например, исследователю социальных медиа неизвестно, является ли активность в сети дополнением или замещением оффлайн-активности. Хотя очевидно, что социальный статус пользователя в реальной жизни — важный предиктор как онлайн-активности, так и онлайн-статуса. Другим инструментом могла бы послужить маркировка профессионалов и любителей.

По значимости оценок параметров регрессии (табл. 5) можно установить, является ли тот или иной интерес ассоциированным с этой распродажей и насколько такая взаимосвязь существенна, принимая во внимание широкий спектр возможных интересов. Среди сформированных двадцати тем, наиболее часто обсуждаемых пользователями выборки, только пять оказались связаны с «Интересом пользователя», причем одна из них — обратно пропорционально. В нашем конкретном примере это может быть интерес-конкурент (скажем, пляжный отдых за рубежом) или обременение (например, наличие маленьких детей).

Индивидуальные показатели эффективности сетевых позиций или центральности могут рассматриваться как характеристики потенциала пользователя в распространении информации. Регрессионный анализ показывает, что этот потенциал (в данном случае в качестве показателя сетевых позиций взят собственный вектор, *evscent*) положительно связан с числом

сообщений по ключевой теме или активностью пользователя. Коэффициенты модели являются коэффициентами эластичности и показывают, как в среднем меняется число сообщений по теме «Интерес пользователя» при изменении значения показателя на 1% (см. комментарий к табл. 5). В частности, интерпретировать взаимосвязь между центральностью и активностью можно как рост числа сообщений на 2.77% в ответ на 1 п.п. прироста индекса центральности.

Таблица 5. Регрессионный анализ взаимосвязи между интересом к ключевой теме для исследования и другими темами с учетом пола, региона и статуса в сети (сообщения 800 отобранных участников)

Показатель	Коэффициенты	Влияет на интерес к ключевой теме	Характер влияния	Доверительная вероятность
Пол	-0.017 (0.111)	Нет		
Индекс центральности (evcent)	2.772*** (0.795)	Да	Положительно	99%
Количество сообщений по теме 1	0.223*** (0.033)	Да	Положительно	99%
Количество сообщений по теме 2	1.013*** (0.069)	Да	Положительно	99%
Количество сообщений по теме 3	0.140** (0.067)	Да	Положительно	95%
Количество сообщений по теме 4	-0.131** (0.056)	Да	Отрицательно	95%
Количество сообщений по теме 5	0.007 (0.088)	Нет		
...				
Количество сообщений по теме 20	-0.014 (0.071)	Нет		
Регион 1	-0.162 (0.219)	Нет		
...				
Регион 10	0.149 (0.161)	Нет		

Примечание. Анализ включает характеристики 380 пользователей. $R^2 = 0.68$, $R^2_{adj} = 0.65$.

Статистическая значимость: ** — $p < 0.05$, *** — $p < 0.01$. В скобках указаны стандартные отклонения коэффициентов. Число сообщений по теме представлено в модели в форме $\log(x + 1)$.

Индекс центральности используется в модели после логарифмического преобразования.

Регион 1,..., Регион 10 — фиксированные эффекты по федеральным округам (ФО), ФО не указан — референс (не все пользователи указывают свои данные).

4. Достоинства и ограничения анализа социальных медиа

В статье обобщаются подходы к анализу социальных медиа. По мнению автора, поток информации (постов), который накапливается в социальных сетях, дополняет классический сетевой анализ. И наоборот, социальные медиа служат информационным источником не только для

анализа текстов, но и для анализа социальных сетей. Вместе два этих ресурса дают хороший потенциал для использования в коммерческих и политических целях. Автор дополняет имеющийся инструментарий дискуссией о преимуществах анализа социальных медиа методами анализа текстов и анализа социальных сетей, а также применением показателей концентрации (коэффициент Джини) и расширенными возможностями визуализации результатов анализа.

Модельный пример, рассмотренный в статье, позволяет сделать следующие выводы.

1. Интерес к теме проявляют пользователи, имеющие набор предпочтительных тем для обсуждения в социальных сетях.

2. Индекс Джини демонстрирует высокую неравномерность публикации сообщений авторами: 10% пользователей, объединенных интересом к теме и связями друг с другом, публикуют 70–80% контента.

3. Пользователи, общающиеся друг с другом через социальные медиа, находятся, как правило, в географической близости.

4. Активность пользователей, измеряемая в числе публикуемых сообщений, и центральность пользователя тесно связаны между собой, что соответствует работам других авторов (Kumar, Mirchandani, 2012, Kim, Han, 2009).

Несмотря на широкие возможности использования информации о том, что и кому пишут пользователи социальных медиа, существуют ограничения, которые необходимо учитывать в выводах исследования. Во-первых, отобранная выборка пользователей не является репрезентативной к населению в целом. Во-вторых, в отличие от опросов, располагаемые данные регистрируют не ответ пользователя на поставленный вопрос, а некоторое поведение в социальной медиа. При этом вряд ли можно установить причинно-следственную связь, почему было написано сообщение по некоторой теме: продиктовано ли оно реально имеющимися у пользователя потребностями или просто желанием включиться в обсуждение. В-третьих, нельзя достоверно сказать, было ли сообщение прочитано другими пользователями. Просмотр страницы с сообщением и лайки могут лишь приблизительно показать интерес к нему. В-четвертых, троллинг и тиражирование сообщений (боты), не имеющих прямого отношения к теме обсуждения, могут занимать существенную долю контента и не быть обнаруженными. В-пятых, необходимо отметить очень важный момент — это отбор тем и соответствующих им слов и словосочетаний, т. е. словарей ключевых понятий, по результатам контент-анализа. Любой язык, в том числе русский, богат синонимами и омонимами, к тому же продолжает непрерывно меняться. Поэтому ошибочная ассоциация слова с той или иной темой или эмоцией может существенно исказить интерпретацию исследуемых взаимосвязей. Результаты анализа могут быть также искажены вымышленными социально-демографическими характеристиками пользователей.

Еще раз отметим, что регулярное присутствие в сети может быть связано с профессиональными интересами, например, проведением маркетинговой кампании. При этом один и тот же пользователь может выполнять две принципиально разных роли: профессионал, преследующий рост объемов продаж своей компании, и любитель-индивид. Исследование профессионального присутствия в сравнении российских и зарубежных социальных медиа, возможно, позволит выявить отечественные особенности формирования социальных сетей и использования социальных медиа в российском бизнесе, где доля серого бизнеса относительно высока.

Вместе с тем, анализ контента и взаимодействия пользователей социальных медиа может оказаться очень ценным при подготовке к проведению социологического исследования,

для выявления проблематики данной социальной группы, ее предполагаемого профиля. При этом такой анализ сравнительно недорог, поскольку не требует затрат на организацию работы интервьюеров, обработку анкет и т. д.

Благодарности. Автор благодарит участников сессии «Новые медиа и социальные отношения» XVII Апрельской международной научной конференции по проблемам развития экономики и общества (НИУ ВШЭ, 2016) за возможность обсуждения результатов, представленных в статье. Автор благодарит анонимных рецензентов, чьи замечания и советы позволили значительно улучшить структуру статьи и изложение результатов.

Список литературы

- Луганская Д. (2017). Мессенджер для всего. *Газета РБК*, 57 (2554), 12–13.
- Смирнов И. Б., Сивак Е. В., Козьмина Я. Я. (2016). В поисках утраченных профилей: достоверность данных «ВКонтакте» и их значение для исследований образования. *Вопросы образования*, 4, 106–123.
- Actionable social analytics: From social media metrics to business insights (2012). <http://www.cbpp.uaa.alaska.edu/afef/Actionable-Social-Analytics.pdf>.
- Asur S., Huberman B. A. (2010). Predicting the future with social media. In: *Web Intelligence and Intelligent Agent Technology (WI-IAT)*. *IEEE/WIC/ACM International Conference*, 1 (6), 492–499.
- Bernard H. R., Killworth P., Kronenfeld D., Sailer L. (1984). The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology*, 13, 495–517.
- Bessiere K., Kiesler S., Kraut R., Boneva B. (2008). Effects of Internet use and social resources on changes in depression. *Information, Communication and Society*, 11, 47–70.
- Bhatt R., Chaoji V., Parekh R. (2010). Predicting product adoption in large-scale social networks. *Proceedings of the 19th ACM international conference on Information and knowledge management — CIKM'10*, 83 (6), 1039.
- Borgatti S. P., Carley K. M., Krackhardt D. (2006). On the robustness of centrality measures under conditions of imperfect data. *Social Networks*, 28 (2), 124–136.
- Borgatti S. P., Mehra A., Brass D. J., Labianca G. (2009). Network analysis in the social sciences. *Science*, 323 (5916), 892–895.
- Bouchet-Valat M. (2013). Package ‘SnowballC’. <https://r-forge.r-project.org/projects/r-temis/>.
- Brewer D. D. (2000). Forgetting in the recall-based elicitation of personal and social networks. *Social Networks*, 22, 29–43.
- Brewer D. D., Webster C. M. (1999). Forgetting of friends and its effects on measuring friendship networks. *Social Networks*, 21, 361–373.
- Burke M., Kraut R., Marlow C. (2011). Social capital on Facebook: Differentiating uses and users. *CHI Session: Facebook*, May 7–12, 2011. Vancouver, BC, Canada.
- Burt R. S. (1992). *Structural holes: The social structure of competition*. Harvard University Press, Cambridge, MA.
- Butts C. T. (2000). An axiomatic approach to network complexity. *Journal of Mathematical Sociology*, 24 (4), 273–301.
- Butts C. T. (2003). Network inference, error, and informant (in)accuracy: A Bayesian approach. *Social Networks*, 25, 103–140.

- Butts C. T. (2014). Tools for social network analysis package 'sna' Version 2.3-2. <http://www.statnet.org>.
- Catanese S. A., De Meo P., Ferrara E., Fiumara G., Provetti A. (2011). Crawling Facebook for social network analysis purposes. *WIMS'11, May 25–27, 2011 Sogndal, Norway*. <https://core.ac.uk/download/pdf/107692.pdf>.
- Ceron A., Curini L., Miacus S., Porro G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New media and society*, 16 (2), 340–358.
- Cha M., Haddadi H., Benevenuto F., Gummadi K. P. (2010). Measuring user influence in Twitter: The million follower fallacy. *4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 14 (1), 8.
- Cheng Z., Caverlee J., Lee K. (2010). You are where you tweet: A content-based approach to geo-locating Twitter users. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 759–768. Toronto, Canada.
- Cheong F., Cheong C. (2011). Social media data mining: A social network analysis of tweets during the 2010–2011 Australian Floods. *PACIS 2011 Proceedings*. 46. <http://aisel.aisnet.org/pacis2011/46>.
- Corley C. D., Cook D. J., Mikler A. R., Singh K. P. (2010). Text and structural data mining of influenza mentions in Web and social media. *International Journal of Environmental Resources and Public Health*, 7, 596–615.
- Correa C. D., Crnovrsanin T., Ma K.-L. (2012). Visual reasoning about social networks using centrality sensitivity. *IEEE Transactions on Visualisation and Computer Graphics*, 18 (1), 106–120.
- Costenbader E., Valente T. W. (2003). The stability of centrality measures when networks are sampled. *Social networks*, 25 (4), 283–307.
- Ellison N., Steinfield C., Lampe C. (2011). Connection strategies: Social capital implications of Facebook-enabled communication practices. *New Media and Society*, 13 (6), 873–892.
- Ellison N. B., Steinfield C., Lampe C. (2007). The benefits of Facebook «Friends»: Social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, 12, 1143–1168.
- Feinerer I. (2008). An introduction to text mining in R. *R News*, 8/2, 19–22. https://cran.r-project.org/doc/Rnews/Rnews_2008-2.pdf.
- Feinerer I. (2017). Introduction to the tm package text mining in R. <https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf>.
- Feinerer I., Hornik K., Meyer D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, 25 (5), 1–54. <http://www.jstatsoft.org/v25/i05>.
- Feld S. L., Carter W. C. (2002). Detecting measurement bias in respondent reports of personal networks. *Social Networks*, 24, 365–383.
- Freeman L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40 (1), 35.
- Freeman L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, 1 (3), 215–239.
- Goodreau S. M., Handcock M. S., Hunter D. R., Butts C. T., Morris M. (2008). A statnet tutorial. *Journal of Statistical Software*, 24 (9). <https://www.jstatsoft.org/article/view/v024i09/v24i09.pdf>.
- Granovetter M. (1976). Network sampling: Some first steps. *American Journal of Sociology*, 81 (6), 1287–1303.
- Hampton K. (2002). Place-based and IT mediated «community». *Planning Theory and Practice*, 3 (2), 228–231.

Hampton K., Wellman B. (2003). Neighboring in Netville: How the Internet supports community and social capital in a wired suburb. *City and Community*, 2 (4), 277–311.

Hinds P., McGrath C. (2006). Structures that work: Social structure, work structure and coordination ease in geographically distributed teams. In: *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, 1 (3), 343–352.

Huberman B. A., Romero D. M., Wu F. (2009). Social networks that matter: Twitter under the microscope. *First Monday*, 14 (1), 8.

Kelly G., Cooper A., Pinkerton E. (2014). Social network analysis, Markov chains and input-output models: Combining tools to map and measure the circulation of currency in small economies. *Journal of Rural and Community Development*, 9 (3), 118–141.

Kim S., Han S. (2009). An analytical way to find influencers on social networks and validate their effects in disseminating social games. In: *2009 International Conference on Advances in Social Network Analysis and Mining*, 20–22 July 2009, 41–46. DOI: 10.1109/ASONAM.2009.59.

Kumar V., Mirchandani R. (2012). Increasing the ROI of social media marketing. *MIT Sloan Management Review*, 54 (1).

Kunegis J. (2017). Handbook of network analysis KONECT — the Koblenz network collection. *arXiv: 1402.5500v3 [cs.SI]*. <https://arxiv.org/pdf/1402.5500.pdf>.

Kwak H., Lee C., Park H. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th international conference on World wide web*, 591–600.

Lampe C., Ellison N., Steinfield C. (2006). A Facebook in the crowd: Social searching vs. social browsing. Community and social capital in a wired suburb. *City and Community*, 2 (4), 277–311.

Lang D. T. (2004). Word stemming in R. <http://www.omegahat.net/Rstem/stemming.pdf>.

Lebart L., Salem A., Berry L. (1998). *Exploring textual data*. Kluwer Academic Press, Dordrecht/Boston.

Lewis K., Kaufman J., Gonzalez M., Wimmer A., Christakis N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, 30, 330–342.

Liviu L. (2011). Predicting product performance with social media. *Informatics in education*, 15 (2), 46–56.

Lomi A., Snijders T. A. B., Steglich C. E. G., Torlód V. J. (2011). Why are some more peer than others? Evidence from a longitudinal study of social networks and individual academic performance. *Social Science Research*, 40 (6), 1506–1520.

Lopes G. R., da Silva R., Moro M. M., de Oliveira J. P. M. (2012). Scientific collaboration in research networks: A quantification method by using Gini coefficient. *International Journal of Computer Science and Applications*, 9 (2), 15–31.

Marsden P. V. (1990). Network data and measurement. *Annual Review of Sociology*, 16, 435–463.

Marsden P. V. (2003). Interviewer effects in measuring network size using a single name generator. *Social Networks*, 25, 1–16.

Morstatter F., Pfeffer J., Liu H., Carley K. M. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. *arXiv: 1306.5204v1 [cs.SI]*. <https://arxiv.org/pdf/1306.5204.pdf>.

O'Connor B., Balasubramanyan R., Routledge B. R., Smith N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 122–129.

Papacharissi Z., Mendelson A. (2008). Toward a new(er) sociability: Uses, gratifications, and social capital on Facebook. *Media perspectives for the 21st century*, 212.

Polonski V. W., Hogan B. (2015). Assessing the structural correlates between friendship networks and conversational agency in Facebook groups. *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, 674–677.

Ruhi U. (2014). Social media analytics as a business intelligence practice: Current landscape and future prospects. *Journal of Internet Social Networking and Virtual Communities*, Vol. 2014, Article ID 920553, DOI: 10.5171/2014.920553.

Sabidussi G. (1966). The centrality index of a graph. *Psychometrika*, 31 (4), 581–603.

Social media marketing and analytics (2013). *Market landscape report*. The Enterprise Strategy Group. <http://www.oracle.com/us/corporate/analystreports/es-social-landscape-1936813.pdf>.

Suh B., Hong L., Piroli P. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. *IEEE International Conference on Social Computing / IEEE International Conference on Privacy, Security, Risk and Trust*, 177–184, DOI: 10.1109/SocialCom.2010.33.

Van Mierlo T., Hyatt D., Ching A. T. (2016). Employing the Gini coefficient to measure participation inequality in treatment-focused digital health social networks. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 5–32.

Wakamiya S., Lee R., Sumiya K. (2011). Crowd-based urban characterization: extracting crowd behavioral patterns in urban areas from twitter. In: *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN'11*, 77–84. New York: ACM.

Warnes G. R., Bolker B., Bonebakker L., Gentleman R., Liaw W. H. A., Lumley Th., Maechler M., Magnusson A., Moeller S., Schwartz M., Venables B. (2016). Various R programming tools for plotting data. <https://cran.r-project.org/web/packages/gplots/index.html>.

Wasserman S., Faust K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

Wellman B., Salaff J., Dimitrova D., Garton L., Gulia M., Haythornthwaite C. (1996). Computer networks as social networks: Collaborative work, telework, and virtual community. *Annual Review of Sociology*, 22, 213–238.

Williams G. (2016). Hands-on data science with R text mining. <http://handsondatascience.com/Text-MiningO.pdf>.

Winship C., Mandel M. (1983). Roles and positions: A critique and extension of the blockmodeling approach. In: Leinhardt S. (ed.), *Sociological Methodology 1983–1984*. Jossey-Bass, San Francisco, 314–344.

Yu S., Kak S. (2012). A survey of prediction using social media. *arXiv*: 1203.1647v1 [cs.SI]. <https://arxiv.org/ftp/arxiv/papers/1203/1203.1647.pdf>.

Zhang W., Skiena S. (2009). Improving movie gross prediction through news analysis. *IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 30 (2), 301–304.

Поступила в редакцию 09.01.2017;
принята в печать 08.06.2017.

Kotyrlo E. S. Social network sites: What users post and to whom they address. Some approaches to the study. *Applied Econometrics*, 2017, v. 47, pp. 74–99.

Elena Kotyrlo

National Research University Higher School of Economics (NRU HSE), Moscow,
Russian Federation; ekotyrlo@hse.ru

Social network sites: What users post and to whom they address. Some approaches to the study

Study of users and their segmentation, based on users' preferred topics of discussion and their networking, is the unique opportunity offered by social networks. Variety of approaches to social media analysis based on social network analysis and text mining is summarized in the paper. It is extended by concentration index application and visualizing of the results of social network analysis.

The study of a model set exhibits that: 1) users can be successfully segmented on the base of their most mentioned topics, which is useful for a product placement and other commercial purposes; 2) distribution of number of posts by authors is highly uneven regardless to the topic of discussion; 3) users connected on-line typically live in the same geographical area; 4) users' number of posts and centrality indices are correlated.

Keywords: text mining; social network analysis; social network sites; regression analysis; Gini coefficient.

JEL classification: M39; C18.

References

- Luganskaja D. (2017). Messendzher dlja vsego. *Gazeta RBK*, 57 (2554), 12–13 (in Russian).
- Smirnov I., Sivak E., Kozmina Y. (2016). In Search of Lost Profiles: The Reliability of VKontakte Data and Its Importance for Educational Research. *Educational Studies*, 4, 106–123 (in Russian).
- Actionable social analytics: From social media metrics to business insights (2012). <http://www.cbpp.uaa.alaska.edu/afef/Actionable-Social-Analytics.pdf>.
- Asur S., Huberman B. A. (2010). Predicting the future with social media. In: *Web Intelligence and Intelligent Agent Technology (WI-IAT)*. *IEEE/WIC/ACM International Conference*, 1 (6), 492–499.
- Bernard H. R., Killworth P., Kronenfeld D., Sailer L. (1984). The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology*, 13, 495–517.
- Bessiere K., Kiesler S., Kraut R., Boneva B. (2008). Effects of Internet use and social resources on changes in depression. *Information, Communication and Society*, 11, 47–70.
- Bhatt R., Chaoji V., Parekh R. (2010). Predicting product adoption in large-scale social networks. *Proceedings of the 19th ACM international conference on Information and knowledge management — CIKM'10*, 83 (6), 1039.
- Borgatti S. P., Carley K. M., Krackhardt D. (2006). On the robustness of centrality measures under conditions of imperfect data. *Social Networks*, 28 (2), 124–136.

- Borgatti S. P., Mehra A., Brass D. J., Labianca G. (2009). Network analysis in the social sciences. *Science*, 323 (5916), 892–895.
- Bouchet-Valat M. (2013). Package ‘SnowballC’. <https://r-forge.r-project.org/projects/r-temis/>.
- Brewer D. D. (2000). Forgetting in the recall-based elicitation of personal and social networks. *Social Networks*, 22, 29–43.
- Brewer D. D., Webster C. M. (1999). Forgetting of friends and its effects on measuring friendship networks. *Social Networks*, 21, 361–373.
- Burke M., Kraut R., Marlow C. (2011). Social capital on Facebook: Differentiating uses and users. *CHI Session: Facebook*, May 7–12, 2011. Vancouver, BC, Canada.
- Burt R. S. (1992). *Structural holes: The social structure of competition*. Harvard University Press, Cambridge, MA.
- Butts C. T. (2000). An axiomatic approach to network complexity. *Journal of Mathematical Sociology*, 24 (4), 273–301.
- Butts C. T. (2003). Network inference, error, and informant (in)accuracy: A Bayesian approach. *Social Networks*, 25, 103–140.
- Butts C. T. (2014). Tools for social network analysis package ‘sna’ Version 2.3-2. <http://www.statnet.org>.
- Catanese S. A., De Meo P., Ferrara E., Fiumara G., Provetti A. (2011). Crawling Facebook for social network analysis purposes. *WIMS'11, May 25–27, 2011 Sogndal, Norway*. <https://core.ac.uk/download/pdf/107692.pdf>.
- Ceron A., Curini L., Miacus S., Porro G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens’ political preferences with an application to Italy and France. *New media and society*, 16 (2), 340–358.
- Cha M., Haddadi H., Benevenuto F., Gummadi K. P. (2010). Measuring user influence in Twitter: The million follower fallacy. *4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 14 (1), 8.
- Cheng Z., Caverlee J., Lee K. (2010). You are where you tweet: A content-based approach to geo-locating Twitter users. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 759–768. Toronto, Canada.
- Cheong F., Cheong C. (2011). Social media data mining: A social network analysis of tweets during the 2010–2011 Australian Floods. *PACIS 2011 Proceedings*. 46. <http://aisel.aisnet.org/pacis2011/46>.
- Corley C. D., Cook D. J., Mikler A. R., Singh K. P. (2010). Text and structural data mining of influenza mentions in Web and social media. *International Journal of Environmental Resources and Public Health*, 7, 596–615.
- Correa C. D., Crnovrsanin T., Ma K.-L. (2012). Visual reasoning about social networks using centrality sensitivity. *IEEE Transactions on Visualisation and Computer Graphics*, 18 (1), 106–120.
- Costenbader E., Valente T. W. (2003). The stability of centrality measures when networks are sampled. *Social networks*, 25 (4), 283–307.
- Ellison N., Steinfield C., Lampe C. (2011). Connection strategies: Social capital implications of Facebook-enabled communication practices. *New Media and Society*, 13 (6), 873–892.
- Ellison N. B., Steinfield C., Lampe C. (2007). The benefits of Facebook «Friends»: Social capital and college students’ use of online social network sites. *Journal of Computer-Mediated Communication*, 12, 1143–1168.
- Feinerer I. (2008). An introduction to text mining in R. *R News*, 8/2, 19–22. https://cran.r-project.org/doc/Rnews/Rnews_2008-2.pdf.

- Feinerer I. (2017). Introduction to the tm package text mining in R. <https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf>.
- Feinerer I., Hornik K., Meyer D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, 25 (5), 1–54. <http://www.jstatsoft.org/v25/i05>.
- Feld S. L., Carter W. C. (2002). Detecting measurement bias in respondent reports of personal networks. *Social Networks*, 24, 365–383.
- Freeman L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40 (1), 35.
- Freeman L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, 1 (3), 215–239.
- Goodreau S. M., Handcock M. S., Hunter D. R., Butts C. T., Morris M. (2008). A statnet tutorial. *Journal of Statistical Software*, 24 (9). <https://www.jstatsoft.org/article/view/v024i09/v24i09.pdf>.
- Granovetter M. (1976). Network sampling: Some first steps. *American Journal of Sociology*, 81 (6), 1287–1303.
- Hampton K. (2002). Place-based and IT mediated «community». *Planning Theory and Practice*, 3 (2), 228–231.
- Hampton K., Wellman B. (2003). Neighboring in Netville: How the Internet supports community and social capital in a wired suburb. *City and Community*, 2 (4), 277–311.
- Hinds P., McGrath C. (2006). Structures that work: Social structure, work structure and coordination ease in geographically distributed teams. In: *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, 1 (3), 343–352.
- Huberman B. A., Romero D. M., Wu F. (2009). Social networks that matter: Twitter under the microscope. *First Monday*, 14 (1), 8.
- Kelly G., Cooper A., Pinkerton E. (2014). Social network analysis, Markov chains and input-output models: Combining tools to map and measure the circulation of currency in small economies. *Journal of Rural and Community Development*, 9 (3), 118–141.
- Kim S., Han S. (2009). An analytical way to find influencers on social networks and validate their effects in disseminating social games. In: *2009 International Conference on Advances in Social Network Analysis and Mining, 20–22 July 2009*, 41–46. DOI: 10.1109/ASONAM.2009.59.
- Kumar V., Mirchandani R. (2012). Increasing the ROI of social media marketing. *MIT Sloan Management Review*, 54 (1).
- Kunegis J. (2017). Handbook of network analysis KONECT — the Koblenz network collection. *arXiv*: 1402.5500v3 [cs.SI]. <https://arxiv.org/pdf/1402.5500.pdf>.
- Kwak H., Lee C., Park H. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th international conference on World wide web*, 591–600.
- Lampe C., Ellison N., Steinfield C. (2006). A Facebook in the crowd: Social searching vs. social browsing. Community and social capital in a wired suburb. *City and Community*, 2 (4), 277–311.
- Lang D. T. (2004). Word stemming in R. <http://www.omegahat.net/Rstem/stemming.pdf>.
- Lebart L., Salem A., Berry L. (1998). *Exploring textual data*. Kluwer Academic Press, Dordrecht/Boston.
- Lewis K., Kaufman J., Gonzalez M., Wimmer A., Christakis N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, 30, 330–342.
- Liviu L. (2011). Predicting product performance with social media. *Informatics in education*, 15 (2), 46–56.

Lomi A., Snijders T. A. B., Steglich C. E. G., Torlód V. J. (2011). Why are some more peer than others? Evidence from a longitudinal study of social networks and individual academic performance. *Social Science Research*, 40 (6), 1506–1520.

Lopes G. R., da Silva R., Moro M. M., de Oliveira J. P. M. (2012). Scientific collaboration in research networks: A quantification method by using Gini coefficient. *International Journal of Computer Science and Applications*, 9 (2), 15–31.

Marsden P. V. (1990). Network data and measurement. *Annual Review of Sociology*, 16, 435–463.

Marsden P. V. (2003). Interviewer effects in measuring network size using a single name generator. *Social Networks*, 25, 1–16.

Morstatter F., Pfeiffer J., Liu H., Carley K. M. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. *arXiv*: 1306.5204v1 [cs.SI]. <https://arxiv.org/pdf/1306.5204.pdf>.

O'Connor B., Balasubramanyan R., Routledge B. R., Smith N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 122–129.

Papacharissi Z., Mendelson A. (2008). Toward a new(er) sociability: Uses, gratifications, and social capital on Facebook. *Media perspectives for the 21st century*, 212.

Polonski V. W., Hogan B. (2015). Assessing the structural correlates between friendship networks and conversational agency in Facebook groups. *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, 674–677.

Ruhi U. (2014). Social media analytics as a business intelligence practice: Current landscape and future prospects. *Journal of Internet Social Networking and Virtual Communities*, Vol. 2014, Article ID 920553, DOI: 10.5171/2014.920553.

Sabidussi G. (1966). The centrality index of a graph. *Psychometrika*, 31 (4), 581–603.

Social media marketing and analytics (2013). *Market landscape report*. The Enterprise Strategy Group. <http://www.oracle.com/us/corporate/analystreports/es-social-landscape-1936813.pdf>.

Suh B., Hong L., Piroli P. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. *IEEE International Conference on Social Computing / IEEE International Conference on Privacy, Security, Risk and Trust*, 177–184, DOI: 10.1109/SocialCom.2010.33.

Van Mierlo T., Hyatt D., Ching A. T. (2016). Employing the Gini coefficient to measure participation inequality in treatment-focused digital health social networks. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 5–32.

Wakamiya S., Lee R., Sumiya K. (2011). Crowd-based urban characterization: extracting crowd behavioral patterns in urban areas from twitter. In: *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN'11*, 77–84. New York: ACM.

Warnes G. R., Bolker B., Bonebakker L., Gentleman R., Liaw W. H. A., Lumley Th., Maechler M., Magnusson A., Moeller S., Schwartz M., Venables B. (2016). Various R programming tools for plotting data. <https://cran.r-project.org/web/packages/gplots/index.html>.

Wasserman S., Faust K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

Wellman B., Salaff J., Dimitrova D., Garton L., Gulia M., Haythornthwaite C. (1996). Computer networks as social networks: Collaborative work, telework, and virtual community. *Annual Review of Sociology*, 22, 213–238.

Williams G. (2016). Hands-on data science with R text mining. <http://handsondatascience.com/Text-MiningO.pdf>.

Winship C., Mandel M. (1983). Roles and positions: A critique and extension of the blockmodeling approach. In: Leinhardt S. (ed.), *Sociological Methodology 1983–1984*. Jossey-Bass, San Francisco, 314–344.

Yu S., Kak S. (2012). A survey of prediction using social media. *arXiv*: 1203.1647v1 [cs.SI]. <https://arxiv.org/ftp/arxiv/papers/1203/1203.1647.pdf>.

Zhang W., Skiena S. (2009). Improving movie gross prediction through news analysis. *IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 30 (2), 301–304.

Received 09.01.2017; accepted 08.06.2017.