ТЕОРИЯ, МЕТОДОЛОГИЯ И МЕТОДЫ

DOI: 10.14515/monitoring.2018.3.05

Правильная ссылка на статью:
Кавеева А. Д., Гурин К. Е. Локальные сети дружбы «ВКонтакте»: восстановление пропущенных данных о городе проживания пользователей // Мониторинг общественного мнения: Экономические и социальные перемены. 2018. № 3. С. 78—90. https://doi.org/10.14515/monitoring.2018.3.05.
For citation:
Kaveeva A. D., Gurin K. E. (2018) VKontakte' local friendship networks: identifying the missed residence of users in profile data. *Monitoring of Public Opinion: Economic and Social Changes.* No. 3. P. 78—90. https://doi.org/10.14515/monitoring.2018.3.05.


А. Д. Кавеева, К. Е. Гурин

ЛОКАЛЬНЫЕ СЕТИ ДРУЖБЫ «ВКОНТАКТЕ»: ВОССТАНОВЛЕНИЕ ПРОПУЩЕННЫХ ДАННЫХ О ГОРОДЕ ПРОЖИВАНИЯ ПОЛЬЗОВАТЕЛЕЙ

ЛОКАЛЬНЫЕ СЕТИ ДРУЖБЫ «ВКОНТАКТЕ»: ВОССТАНОВЛЕНИЕ ПРОПУЩЕННЫХ ДАННЫХ О ГОРОДЕ ПРОЖИВАНИЯ ПОЛЬЗОВАТЕЛЕЙ

КАВЕЕВА Аделя Динаровна — аспирантка Казанского федерального университета, Казань, Россия.
E-MAIL: adele.kaveeva@mail.ru
ORCID: 0000-0002-8689-5532

ГУРИН Константин Евгеньевич — аналитик Корпорация «Центр», Ижевск, Россия.
E-MAIL: rekonchik@mail.ru
ORCID: 0000-0002-1793-2571

VKONTAKTE' LOCAL FRIENDSHIP NETWORKS: IDENTIFYING THE MISSED RESIDENCE OF USERS IN PROFILE DATA

Adelia D. KAVEEVA[1] — Postgraduate Student
E-MAIL: adele.kaveeva@mail.ru
ORCID: 0000-0002-8689-5532

Konstantin E. GURIN[2] — Analyst
E-MAIL: rekonchik@mail.ru
ORCID: 0000-0002-1793-2571

[1] Kazan (Volga region) Federal University, Kazan, Russia
[2] 'The Center' Corporation, Izhevsk, Russia

Аннотация. Социальные онлайн-сети, в частности, самая популярная российская сеть «ВКонтакте», являются источником большого количества доступной информации о пользователях благодаря политике открытости данных. Это дает исследователям возможность изучения топологии сетей взаимодействий, возникающих в онлайн-среде, с применением сетевого подхода (social network analysis). Однако личные данные, которые пользователи сообщают о себе в публичных профилях, зачастую неполны: люди могут по невниманию или умышленно пропускать заполнение тех полей в профиле, которые отражают их пол, возраст, город проживания и другие персональные данные. Эти характеристики играют большую роль при построении социальных сетей в качестве атрибутов «узлов» (то есть пользователей), что позволяет выделять кластеры схожих между собой агентов и их паттерны поведения. Отсутствие некоторых данных может существенно влиять на сетевые метрики (например, размер сети, среднюю длину пути между двумя участниками, распределение числа связей между ними и другие) и искажать полученные результаты. В связи с этим возникает потребность в восполнении пропущенной части данных.

В статье представлен опыт создания и применения классификатора, который позволяет определить, является ли пользователь сети «ВКонтакте», не указавший в профиле место жительства, жителем конкретного города. Классификатор был создан и апробирован на примере сети пользователей из г. Ижевска. Он основан на методе дерева решений, которое поэтапно фильтрует аккаунты через ряд во-

Abstract. Online social networks (e. g. the most popular Russian website 'VKontakte') are a source of available information about users due to the open data policy. Therefore, researchers have great opportunities to study the topology of interaction networks in the online environment using a social network analysis. However, the personal data that users provide in their public profiles are often incomplete: sections on gender, age or city may be missed inadvertently or skipped intentionally. At the same time, these essential characteristics serve as 'nodes' (i. e. users) and help single out clusters of similar agents and their behavior patterns. The absence of some data can significantly affect network metrics (e. g. size of network, average path length between two participants, distribution of the number of connections between them, etc.) and cause distorted results. In this regard, there is a need to fill gaps in data.

The paper presents a case study on the design and applications of a classifier which would determine whether a VKontakte user whose location was not specified in the profile is a resident of a particular city. The classifier was created and tested for the Izhevsk city user network. It is based on the decision tree method which gradually filters the accounts by a series of questions. The paper explains the choice of the main indicators helping the classifier to determine the user's city, describes the algorithm and shows how the network topology changes as the missing data on user's location are added.

просов, а затем принимает решение, считать ли данный аккаунт профилем ижевчанина или нет. В статье объяснен выбор основных показателей, которые помогают классификатору определить город пользователя; описан алгоритм работы классификатора и показано, как изменяется топология сети, когда в нее добавляются пропущенные данные о городе проживания пользователей.

Ключевые слова: анализ социальных сетей, онлайн-сообщества, ВКонтакте, топология сетей, большие данные, анализ данных в R, сетевая гомофилия, пропущенные данные

Keywords: social network analysis, online communities, VKontakte, network topology, big data, using R for data analysis, network homophily, missing data

Благодарность: Авторы статьи выражают благодарность Д. Сорокину (Университет ИТМО), разработавшему библиотеку «VKR» для языка программирования R, с помощью которой были собраны данные для проведения исследования. Работа выполнена при поддержке Программы повышения конкурентоспособности Казанского (Приволжского) федерального университета.

Acknowledgement. The authors thank Dmitry Sorokin from ITMO University who developed 'VKR' package for R programming language. This research was financially supported by the Russian Government Program of Competitive Growth of Kazan Federal University.

Введение, или проблема неполных данных о пользователях социальных онлайн-сетей

Профили пользователей составляют основу любых социальных медиа [boyd, Ellison, 2008], и особенно отчетливо это проявляется для такой их разновидности, как социальные онлайн-сети, или сайты социальных сетей. В зависимости от конкретного приложения варианты идентификации пользователя в нем могут сильно различаться, однако практически все включают в себя возможность создания личного профиля с именем, контактной информацией и фотографией пользователя. Необходимость создания профиля обусловлена тем, что иначе было бы невозможно создание социальной сети[1] и взаимосвязей между пользователями, невозможен поиск и подключение к другим участникам. Большинство функций

[1] Под «социальной сетью» авторами понимается социальное образование, состоящее из акторов (людей или групп) и связей между ними, необязательно опосредованных виртуально. Онлайн-платформы и веб-сайты, предназначенные для создания онлайн-профилей и онлайн-взаимодействий, мы будем называть социальными онлайн-сетями или сайтами социальных сетей.

социальных медиа имеют смысл только при идентификации пользователя (например, рейтинги или обмен «лайками»).

Одним из способов создания социальной сети на базе социальных медиа является создание списка пользователей, с которыми вы хотели бы взаимодействовать — например, формирование списка друзей во «ВКонтакте». После того как такой список создан, пользователи могут взаимодействовать в этой сети и видоизменять ее посредством включения или исключения пользователей из своей сети и действий в их отношении (комментирование, демонстрация симпатии и пр.).

Характеристики пользователей, которые они размещают в своих публичных профилях на сайтах социальных сетей, представляют собой доступную и значимую социологическую информацию. Применение подхода и набора методов, известного как сетевой анализ, позволяет выявлять структуру и характеристики сетей, в которые объединяются пользователи. Сетевые метрики, рассчитываемые для сетей, могут многое сказать о сплоченности сообщества, подгруппах, которые оно содержит, наиболее значимых узлах сети, за счет которых происходит ее рост и эффективное распространение информации. К таким сетевым метрикам относятся, например:

— размер сети — число «узлов», или «вершин» (пользователей);
— средняя длина пути между двумя любыми участниками сети;
— распределение степеней (т. е. числа связей, которыми пользователь связан с другими);
— ассортативность — тенденция узлов с одинаковой степенью образовывать связи друг с другом; ассортативность означает, что пользователи объединены связями с теми, у кого схожее с ними количество друзей;
— транзитивность — доля закрытых «триад», где все трое связаны между собой («друг моего друга — мой друг»);
— модулярность — свойство, характеризующее степень кластеризации узлов, когда внутри кластера плотность сети высокая, а между кластерами — низкая; и другие.

Характеристики пользователей играют важную роль и при выборе ими контента, а также его производстве. Например, в исследовании М. Косински и соавторов, где были проанализированы «лайки», которые оставляют друг другу пользователи Facebook, показано, что социально-демографические характеристики участников (пол, возраст, уровень образования, этническая, религиозная, расовая принадлежность, политические взгляды, потребление алкоголя и табака) сильно взаимосвязаны с выбираемым ими контентом. Ученые смогли построить модель, которая способна предсказывать характеристики участников сети по «лайкам», которые они поставили [Kosinski et al., 2013: 5802—5805].

Однако, как отмечают авторы отчета AAPOR, «большие данные зачастую выборочны, неполны и содержат ошибки» [Джапек и др., 2015: 28], и эта проблема, безусловно, затрагивает и данные из социальных онлайн-сетей. Наиболее значительны такие возможные искажения данных из профилей пользователей, как их неполнота и недостоверность. В настоящей статье мы фокусируемся на первом типе искажений — неполноте данных о пользователе, представленных в его профиле.

По сравнению с блогами и микроблогами, сайты социальных сетей располагают более развернутыми анкетами, где пользователи могут указать имя, пол, дату

рождения, место проживания, работы или учебы, также эти анкеты содержат информацию о «друзьях». Указание полных сведений о себе упрощает установление и поддержание контактов, поскольку облегчает нахождение пользователя его офлайн-знакомыми, коллегами, одноклассниками. Однако заполнение этих данных пользователем абсолютно добровольно, в связи с чем и возникает неполнота сведений, которые можно собрать о пользователе [Чекмышев, Яшунский, 2014: 4].

При этом данные характеристики весьма значимы при построении адекватных социальных сетей пользователей, поскольку играют роль атрибутов «узлов» (т. е. пользователей). Отсутствие некоторых данных может существенно влиять на сетевые метрики (например, размер сети, среднюю длину пути между двумя участниками, распределение числа связей между ними и другие) и искажать полученные результаты. В связи с этим возникает потребность в восполнении пропущенной части данных. Так, например, исследователями Института системного программирования РАН был разработан метод определения демографических атрибутов пользователей сети Twitter по текстам их сообщений [Коршунов, Белобородов, Гомзин и др., 2013].

Наш исследовательский интерес сосредоточен на сайтах социальных сетей, а именно самой популярной российской социальной сети «ВКонтакте»[2], и на таком виде пропущенных данных из профиля пользователя, как город его проживания. В статье описан способ определения принадлежности пользователя к конкретному городу, на основании других открытых данных из его профиля. Это позволяет восстановить полную локальную сеть пользователей, проживающих в этом городе, добавив в нее участников, проигнорировавших заполнение графы о месте жительства в анкете.

Определение города проживания пользователя: выбор характеристик для классификатора

Задачей предлагаемого нами классификатора является определение города проживания для тех пользователей сайта «ВКонтакте», которые не указали его в своем профиле. Территория проживания является одним из значимых атрибутов узлов в сети, позволяя выяснить, насколько реальная географическая близость участников влияет на формирование онлайн-связей между ними. Кроме того, особую значимость этот атрибут имеет, если нас интересуют жители конкретного города. В таком случае построенная нами сеть жителей будет неполной, поскольку она не будет содержать тех участников, которые не указали информацию о месте проживания в анкете.

Мы предполагаем, что дополнение сети пользователей информацией о месте проживания (т. е. добавление в сеть пользователей, с высокой вероятностью являющихся жителями определенного города, но не отметивших в профиле место жительства) повлияет на наблюдаемые сетевые метрики (транзитивность, модулярность, среднюю длину пути, распределение степеней, размер сети, ассортативность по числу друзей).

В качестве примера для создания и апробации классификатора была выбрана сеть жителей города Ижевска. Этот выбор обусловлен проживанием одного из авторов статьи в этом городе, а также его относительно небольшим размером, что

[2] Согласно рейтингу Top Websites Ranking [Электронный ресурс]. URL: https://www.similarweb.com/top-websites (дата обращения: 1.08.2017).

облегчает сбор сетевых данных (по данным на 1 июля 2017 г., на сайте «ВКонтакте» зарегистрировано 507 748 ижевских аккаунтов).

Объектом исследования стала локальная сеть дружеских связей в социальной онлайн-сети «ВКонтакте» на примере г. Ижевска, а предметом — дополнение этой сети пользователями, не указавшими город проживания в своем профиле. Таким образом, цель данного исследования — определение изменений характеристик сети при учете пользователей, не указавших город проживания. Для достижения исследовательской цели были поставлены и реализованы следующие задачи:

— определить характеристики, по которым можно определить город проживания пользователя;
— создать классификатор на основе отобранных характеристик для установления, является ли пользователь жителем выбранного города или нет;
— оценить влияние на локальную сеть и ее характеристики добавления в нее пользователей, не указавших город.

Для создания классификатора необходимо было в первую очередь выделить среди доступных из профиля пользователя характеристик те, через которые можно определить город его проживания. К числу потенциально значимых в контексте данной задачи были отнесены следующие характеристики:

1. *Доля жителей города среди друзей пользователя.* Это важнейшая характеристика, поскольку многие зарубежные и российские сетевые исследования онлайн-сообществ выявляют гомофилию по территории проживания, несмотря на отсутствие в интернете географических барьеров для коммуникации (например, [Ugander, Karrer, Backstrom et al., 2011; Takhteyev, Gruzd, Wellman, 2012; Гурин, 2016: 69]). Под гомофилией понимается склонность людей формировать связи с другими на основании общих черт (пола, расы или, как в нашем случае, места проживания), что структурирует и дифференцирует сеть.
2. *Членство пользователя в популярных среди жителей города онлайн-сообществах* (в тех, где ощутима доля представителей рассматриваемого города).
3. Принадлежность *вуза,* указанного пользователем в профиле, к изучаемому городу;
4. Принадлежность *школы,* указанной пользователем в профиле, к изучаемому городу;
5. Наличие среди указанных в профиле *родственников и романтического партнера* жителей изучаемого города.

Сбор данных о пользователях был осуществлен в мае 2017 г. при помощи языка программирования R и библиотеки *VKR*, разработанной Д. Сорокиным и находящейся в открытом доступе[3]. Пакет *VKR* содержит набор команд, обращающихся к социальной сети «ВКонтакте» после получения специального ключа приложения, позволяет выгружать и анализировать данные о пользователях и сообществах. В качестве сети жителей Ижевска рассматривалась гигантская связная компонента графа (большая часть сети, с которой узлы соединены хотя бы одной связью). Сбор происходил следующим образом: от случайного набора ижевчан собиралась сеть их друзей, из этой сети собирались ижевчане, и поиск друзей

[3] AccesstoVK (Vkontakte) API via R. [Электронный ресурс]. URL: https://github.com/Dementiy/vkR (дата обращения: 1.08.2017).

велся уже по ним. Итерации продолжались до полного сбора данных. Всего в гигантской связной компоненте оказалось 477 тыс. пользователей (94% всех ижевчан). Вместе с этим собирались связи между участниками и жителями других городов и пользователями без отметки о городе проживания. Так, в окрестностях сети Ижевска находится 4766 тыс. пользователей других городов, то есть в десять раз больше самой сети. Всего же в окрестности сети города Ижевска 2603 тыс. пользователей не указали город проживания. Многие из них оказались заблокированными или удаленными аккаунтами.

После сбора сети ижевчан для каждого пользователя из окрестностей сети Ижевска собиралась сеть друзей и рассчитывалась доля ижевчан в окрестности пользователя, отметивших город. Таким образом, «вторая окрестность» сети Ижевска составила более 126 млн пользователей. Несмотря на сложности, возникающие в сборе и агрегации доли друзей жителей Ижевска для иногородней окрестности сети жителей Ижевска, этот показатель оказался одним из самых информативных и значимых.

Также была собрана информация о местах обучения, которые отметили пользователи (школах и университетах). Если пользователь не оставлял информации об образовании, классификатор принимал на вход информацию о том, что нам неизвестно, учился конкретный пользователь в Ижевске или нет. Аналогичным образом обрабатывалась информация о романтическом партнере и наличии в Ижевске родственников.

Среди сообществ по интересам был выставлен порог для сбора по следующему правилу: сообщество должно включать не менее 500 участников-ижевчан. В случае сбора сообществ с меньшим числом ижевчан количество групп для обработки росло по экспоненте, что вызывало затруднения в получении данных. Всего была собрана информация о 8610 группах, в крупнейшей из них состояло 56 тыс. подписчиков из Ижевска. В ходе определения числа групп и их характеристик для максимальной информативности было установлено, что оптимальный порог доли ижевчан в сообществе — 5%. Более «жесткие» или «мягкие» условия отбора делали этот критерий менее эффективным при ответе на вопрос, живет ли пользователь в Ижевске. В конечном итоге была использована информация о 1513 группах.

Создание классификатора для определения города проживания пользователя

Для создания классификатора треть наблюдений, выбранных случайным образом, была определена в тестовую выборку, чтобы проверять на переобучение модели, построенные на обучающей выборке. После ряда тестов авторы статьи остановились на модели классификатора, основанной на дереве решений, отвечающего на вопрос: проживает ли пользователь в Ижевске или нет.

Деревья решений — популярный алгоритм классификации, в котором решающие правила извлекаются непосредственно из исходных данных в процессе обучения. Они представляют собой иерархическую последовательность правил вида «Если…, то…». Этот метод классификации отличается хорошей интерпретацией и визуализацией. Также этот метод лежит в основе множества других современных подходов к построению моделей (таких как случайные леса, GBM, XGBoost). Классификатор представлен на рисунке 1.

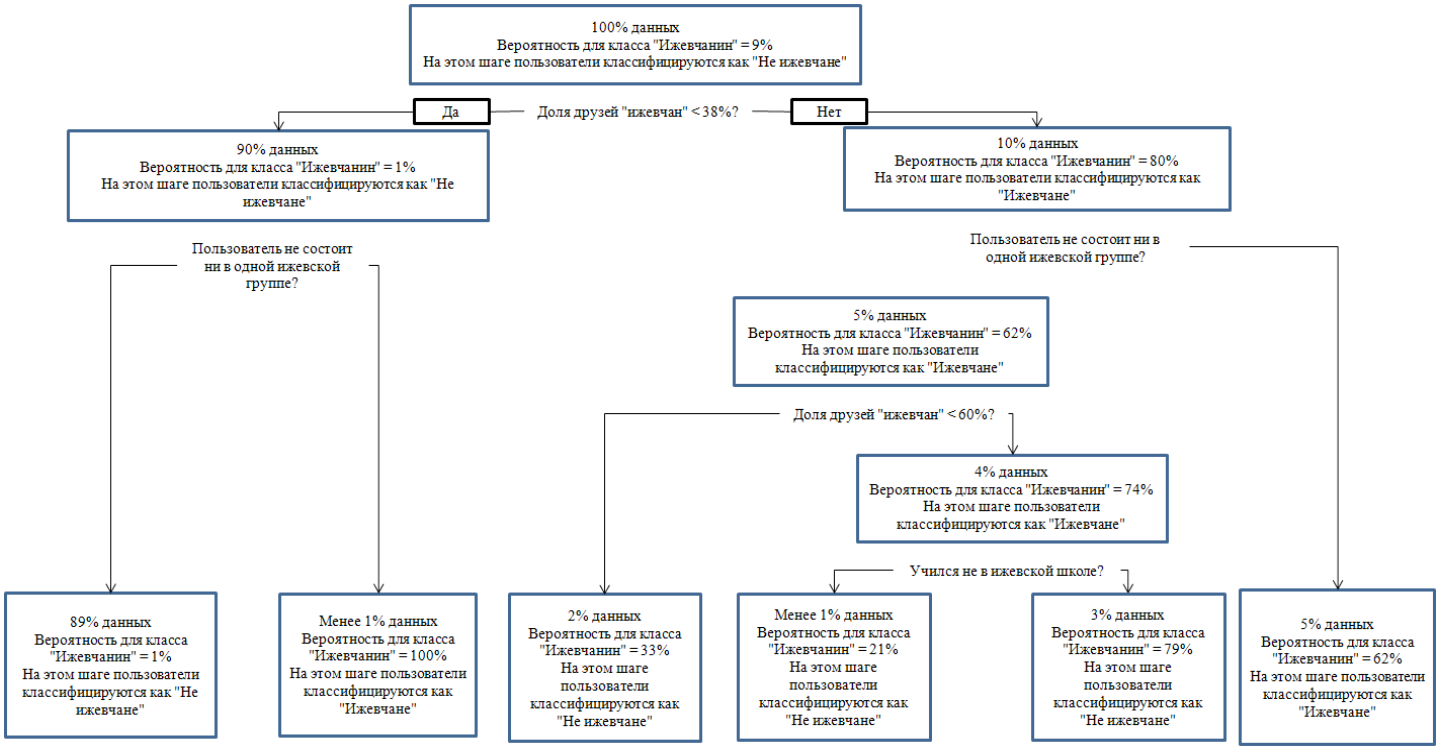

Рисунок 1. Классификатор для определения принадлежности пользователя, не указавшего место проживания, к рассматриваемому городу

Как видно из рисунка, классификатор создал решающее дерево, которое поэтапно фильтрует аккаунты через ряд вопросов, а затем принимает решение, считать ли аккаунт принадлежащим ижевчанину или нет. Например, если у пользователя 80 % друзей проживают в Ижевске, но пользователь учился в школе вне Ижевска, классификатор не причислит его к ижевчанам. Несмотря на простоту дерева, оно ошибается в 1,8 % случаев. Относительно ижевчан классификатор неверно приписывает 8,1 % аккаунтов к ижевчанам от общего числа проклассифицированных как жителей Ижевска, однако 12,2 % реальных ижевчан маркируются как иногородние.

Не все используемые признаки аккаунтов попали в классификатор. Так, на фоне информации о доле друзей-ижевчан в окрестности пользователя информация о наличии родственников или романтического партнера, проживающих в Ижевске, дублируется. Информация об обучении в университете также теряет актуальность, когда комбинируются признаки, попавшие в классификатор.

Если рассматривать, жителей каких городов классификатор часто причисляет к жителям Ижевска, то прослеживаются две наиболее сильные тенденции. Во-первых, классификатор часто приписывает проживание в Ижевске жителям крупных городов — Москве (20,6 % от иногородних аккаунтов, приписанных к Ижевску), Санкт-Петербурге (6,8 %), Казани (3,7 %), Екатеринбурге (1,9 %). Во-вторых, к ижевчанам причисляются жители многих близлежащих населенных пунктов и городов (Сарапул — 1,7 %, Завьялово — 1,4 %, Воткинск — 1,3 %, Можга — 1,3 %).

Было установлено, что уровень ошибок классификатора зависит также от включенности пользователя в социальную сеть. Чем сильнее пользователь интегрирован в социальную сеть (хотябы по числу связей дружбы), тем с меньшей ошибкой классификатор будет определять его класс (ижевчанин или нет). Например, если рассматривать пользователей с более чем одной дружеской связью, то доля неверно отнесенным к ижевчанам снижается с 8,1 % до 6,2 %, а неправильный класс приписывается лишь 10,2 % жителей Ижевска (вместо 12,2 %). При этом, несмотря на потерю общего размера тестовой выборки на 60 %, доля причисленных к ижевчанам сократилась всего на 7 % от первоначального количества.

Изменения сетевых метрик после дополнения сети недостающими данными

После создания алгоритма определения ижевчан классификатор был применен к аккаунтам, в списке друзей которых есть хотя бы один ижевчанин. Первоначально были отобраны 2 603 905 аккаунтов, однако более половины из них оказались удалены или заблокированы. После отсева недействующих профилей осталось 1 294 758 пользователей, 142 978 из которых классификатор промаркировал как ижевчан.

Для сети до добавления аккаунтов без отметки города (далее — сеть № 1) и сети после (сеть № 2) были произведены замеры сетевых характеристик, представленные в таблице 1.

Как видно из таблицы 1, несмотря на рост размера сети, ряд структурных характеристик, таких как транзитивность, ассортативность и средняя длина пути,

не изменяются. Однако сплоченность в сети уменьшается, подгруппы становятся более выраженными. (Разбиение на подгруппы производилось при помощи алгоритма, предложенного в [Blondel et al., 2008], при вычислениях была использована функция *cluster_louvain* библиотеки *igraph* для языка R.) Также подгруппы из первой сети распадаются на более мелкие кластеры при добавлении аккаунтов, число таких подгрупп вырастает с 273 до 330, из добавленных аккаунтов образуются еще четыре малых кластера. Таким образом, включение в сеть аккаунтов пользователей, не указавших город проживания, увеличивает размер сети и ее дифференцированность.

Таблица 1. ***Характеристики сети до и после добавления пользователей, проклассифицированных как ижевчане (сеть № 1 и сеть № 2)***

Характеристики сети	Сеть № 1	Сеть № 2
Число узлов в сети	477 057	620 035
Число ребер	17 742 662	22 790 631
Транзитивность в сети	0,090	0,088
Ассортативность по числу друзей	0,162	0,166
Средняя длина пути в сети	3,590	3,679
Модулярность	0,377	0,402
Число подгрупп	273	334

Для каждого участника сети № 1 также были замерены показатели кластеризации, центральности по числу друзей и центральности по собственному значению векторов (*eigenvector centrality*), характеризующие значимость профиля в сети. Распределения показателей мер центральности до добавления новых участников и после не изменяются (корреляции Спирмена равны 1). В то же время кластеризация окрестностей узлов становится уже менее стабильной (корреляция Спирмена равна 0,95). Соответственно, исключение из сети города жителей, не указавших места проживания, может исказить и локальную структуру сети по конкретному пользователю.

Также интуитивно понятно, что добавление новых аккаунтов в сеть увеличивает число дружеских контактов старых аккаунтов. Однако в связи с тем, что сети формируются в значительной мере по принципу предпочтительного присоединения [Barabasi, Albert, 1999: 509—512], прирост числа контактов будет неравномерным для разных пользователей. Этот прирост для участников сети № 1 можно описать уравнением:

$$Log(\text{Число друзей в сети № 2}) = 0{,}1151 + 1{,}0047(Log(\text{Число друзей в сети № 1}))$$

Это означает, что при росте числа друзей в сети № 1 у пользователя на 1 % число его друзей в сети № 2 увеличится на 1,0047. Например, если у пользователя до добавления новых аккаунтов было 120 друзей, то в сети № 2 у этого пользова-

теля будет в среднем 137,7 друзей. Для случая с 50 друзьями в первой сети после добавления аккаунтов станет 57 друзей. Таким образом, различия в уровне числа друзей будут усиливаться с добавлением новых профилей.

Заключение и области применения

«Цифровые следы», которые оставляют пользователи сайтов социальных сетей, открывают новые возможности для исследователей в области социальных наук, а также в сфере маркетинга и политике работы с молодежью, составляющей большинство активных онлайн-пользователей. В частности, сайт «ВКонтакте» предоставляет широкий набор методов для сбора открытой информации о пользователях. Однако важным фактором, который необходимо учитывать при сборе и анализе данных о пользователях, является возможная неполнота сведений, которые содержатся в их анкетах. Отсутствие информации о гендерной принадлежности, возрасте и месте проживания может приводить к искажениям получаемой сети и препятствовать ее правильной интерпретации.

Предложенный в статье классификатор был создан с целью определения города проживания для тех пользователей сайта «ВКонтакте», которые не указали его в своем профиле. Место проживания зачастую представляет собой один из самых значимых атрибутов узлов в сети, особенно в ситуациях, когда необходимо проанализировать полную сеть пользователей из конкретного города. Классификатор основан на дереве решений, которое отвечает на вопрос, является ли пользователь, не указавший в профиле город проживания, ижевчанином или нет. Дерево решений извлекает правила из других данных о пользователе, которые имеются в его профиле. Так, было установлено, что наиболее эффективна для определения города такая характеристика, как наличие и доля ижевчан в дружеской сети пользователя, а также членство пользователя в онлайн-сообществах с аудиторией, большую долю которых составляют жители данного города, вуз или школа, относящиеся к городу, и проживание в нем родственников или романтического партнера пользователя.

Сравнение сети жителей Ижевска до и после добавления в нее пользователей, проклассифицированных как ижевчане, показало следующее. Включение в сеть таких аккаунтов увеличивает размер сети и ее дифференцированность. Также добавление новых аккаунтов в сеть увеличивает число дружеских контактов между пользователями (то есть число ребер в сети). Несмотря на рост размера сети, ряд структурных характеристик (транзитивность, ассортативность, средняя длина пути) остаются неизменными. Однако при добавлении аккаунтов сплоченность в сети уменьшается, подгруппы в сети становятся более выраженными.

Данный классификатор будет полезен для решения как исследовательских, так и бизнес-задач. Так, определение социально-демографических атрибутов пользователя представляет большую ценность для таргетированного продвижения товаров и услуг среди целевой аудитории. В частности, сбор более полных сетей о жителях города эффективно используется в рамках CRM-системы (системы управления взаимоотношениями с клиентами, которая предназначена для автоматизации стратегий взаимодействия с ними). В свою очередь, исследовательские центры могут использовать полные локальные сети пользователей для анализа

и моделирования различных социокультурных, политических, экономических процессов.

Список литературы (References)

Гурин К. Е. Структурирование сетей дружбы в онлайн-сообществах СМИ // Дискуссия. 2016. № 6 (69). С. 64—71.
Gurin K. E. (2016) Friendship networks structuring of mass media online communities. *Discussion*. No. 6 (69). P. 64—71. (In Russ.)

Джапек Л., Крейтер Ф., Берг М. и др. Отчет AAPOR о больших данных: 12 февраля 2015 [Электронный ресурс] / Американская ассоциация исследователей общественного мнения; Пер. с англ. Д. Рогозина, А. Ипатовой, Е. Вьюговской ; предисловие Д. Рогозина. М., 2015. Систем. требования: Adobe Acrobat Reader. URL: https://wciom.ru/fileadmin/file/nauka/grusha2015/AAPOR_big_data.pdf (дата обращения: 1.08.2017).
Japec L., Kreuter F., Berg M. et al. (2015) AAPOR Report: BIG DATA. February 12, 2015. American Association for Public Opinion Research. Transl. from the English by Rogozin D., Ipatova A., Vyugova E. URL: https://wciom.ru/fileadmin/file/nauka/grusha2015/AAPOR_big_data.pdf (accessed: 1.08.2017). (In Russ.)

Коршунов А., Белобородов И., Гомзин А. и др. Определение демографических атрибутов пользователей микроблогов // Труды Института системного программирования РАН. 2013. Т. 25. С. 179—194. https://doi.org/10.15514/ISPRAS-2013-25-10.
Korshunov A., Beloborodov I., Gomzin A. et al. (2013) Detection of demographic attributes of microblog users. In: *Proceedings of ISP RAS*. Vol. 25. P. 179—194. https://doi.org/10.15514/ISPRAS-2013-25-10. (In Russ.)

Чекмышев О. А., Яшунский А. Д. Извлечение и использование данных из электронных социальных сетей // Препринты ИПМ им. М. В. Келдыша. 2014. № 62. [Электронный ресурс]. Систем. требования: Adobe Acrobat Reader. URL: http://www.keldysh.ru/papers/2014/prep2014_62.pdf (дата обращения: 1.08.2017).
Chekmyshev O. A., Yashunsky A. D. (2014) Extraction and usage of online social network data. Preprints of Keldysh Institute of Applied Mathematics. No. 62. URL: http://www.keldysh.ru/papers/2014/prep2014_62.pdf (accessed: 1.08.2017). (In Russ.)

Barabasi A. L., Albert R. (1999) Emergence of scaling in random networks. *Science*. Vol. 286., No. 5439. P. 509—512. https://doi.org/10.1126/science.286.5439.509.

Blondel V. D., Guillaume J.-L., Lambiotte R. et al. (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*. P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.

boyd d. m., Ellison N. B. (2008) Social network sites: Definition, history, and scholarship. *Journal of Computer Mediated Communication*. Vol. 13. No. 1. P. 210—230. https://doi.org/10.1111/j.1083-6101.2007.00393.x.

Kosinski M., Graepel T., Stillwell D. (2013) Private traits and attributes are predictable from digital records of human behavior. In: Proceedings of the National Academy of Sciences. P. 5802—5805. https://doi.org/10.1073/pnas.1218772110.

Takhteyev Y., Gruzd A., Wellman B. (2012) Geography of Twitter networks. *Social Networks*. Vol. 34. No. 1. P. 73—81.

Ugander J., Karrer B., Backstrom L. et al. (2011) The anatomy of the Facebook social graph. *Cornell University Library*. URL: https://arxiv.org/abs/1111.4503 (accessed: 1.08.2017).