

УДК 004.77

Е. В. Митягина, Т. В. Ашихмина, Е. В. Котельников

МЕТОДЫ АНАЛИЗА СОЦИАЛЬНЫХ МЕДИА (НА ПРИМЕРЕ СОЦИАЛЬНОЙ СЕТИ ВКОНТАКТЕ)

В статье представлен анализ социальной сети ВКонтакте как одного из наиболее популярных социальных медиа среди жителей г. Кирова. Описаны проблемы, которые возникли при анализе данной социальной сети, такие как проблема получения информации, связанная с программным интерфейсом, проблема актуальности полученной информации и проблема пропущенных данных. В качестве решения указанных проблем предложен постоянный мониторинг обновлений, производимых пользователями, совершенствование программ обработки данных, а также анализ только заполненных пользователями полей.

Проанализированы социально-демографические характеристики пользователей с социальной сети, их семейное и профессиональное положение, отношение к вредным привычкам. Особое внимание уделено жизненной позиции кировчан, ключевыми категориями в которой являются семья и дети и саморазвитие. Показано, как при помощи количественных статистических методов, таких как «хи-квадрат» и номинальный регрессионный анализ, можно анализировать данные социальных медиа.

Ключевые слова: социальные медиа, социальная сеть ВКонтакте, методы социологического анализа социальных медиа, анализ связи между номинальными признаками.

Среди новых методов социологического анализа общественных и бизнес-процессов особое место занимает исследование социальных медиа [2]. Социальные медиа – это совокупность интернет-площадок, которые на основе сетевых технологий предоставляют пользователям возможность устанавливать коммуникацию друг с другом: социальные сети, блоги, форумы и т. д. [1]. Список предметных областей, которые можно подвергнуть анализу, достаточно обширный, это: отзывы о товарах (например, о фотокамерах, ноутбуках, мобильных телефо-

нах, автомобилях), о произведениях искусства (например, о книгах, фильмах, спектаклях), об организациях (например, о ресторанах, отелях, банках) и т.п.

В настоящей работе для пилотажного социологического исследования и апробации программных средств была выбрана социальная сеть ВКонтакте в связи с высокой её популярностью среди жителей г. Кирова и возможностью получения доступа к открытой информации, то есть к данным, которые заполняются пользователем самостоятельно, и сообщениям с его «стены». На 10 февраля 2017 г. в социальной сети ВКонтакте было зарегистрировано 518 710 пользователей, указавших в качестве места жительства г. Киров. При этом по данным Росстата на 1 января 2016 г. численность населения г. Кирова составляет 523 119 человек [4].

Был осуществлен доступ к следующей открытой информации о пользователях:

- id (идентификатор пользователя),
- first_name и last_name (имя и фамилия),
- bdate (дата рождения),
- followers_count (количество подписчиков пользователя),
- has_mobile (информация о том, известен ли номер мобильного телефона пользователя),
- has_photo (наличие фотографии для профиля),
- last_seen – time (время последнего посещения),
- last_seen – platform (тип платформы, через которую был осуществлен последний выход),
- occupation – type (информация о текущем роде занятия пользователя),
- occupation – name (название школы, вуза или места работы),
- personal – political (политические предпочтения),
- personal – religion (религиозные взгляды, мировоззрение),
- personal – inspired_by (источник вдохновения),
- personal – people_main (главное в людях),
- personal – life_main (главное в жизни),
- personal – smoking (отношение к курению),

- personal – alcohol (отношение к алкоголю),
- relation (семейное положение),
- sex (пол).

При анализе социальных медиа возникает ряд проблем:

- 1) проблема получения информации, связанная с программным интерфейсом социальной сети;
- 2) проблема актуальности полученной информации;
- 3) проблема пропущенных данных.

Рассмотрим указанные проблемы подробнее.

ВКонтакте имеет интерфейс прикладного программирования (Application Programming Interface, API), включающий более 350 методов и позволяющий программно использовать функционал социальной сети и получать из неё разнообразную информацию [5]. Однако при запросах имеется ограничение на количество возвращаемых результатов, как правило, не более 1000. Таким образом, напрямую получить список всех пользователей из Кирова невозможно. Для составления такого списка был использован следующий алгоритм: на первом шаге был получен список из тысячи наиболее популярных кировских групп и отобраны только кировчане, состоящие в этих группах; на втором шаге для всех кировчан из предыдущего шага просмотрены их друзья и также отобраны кировчане; второй шаг повторялся рекурсивно до тех пор, пока список не перестал пополняться.

С помощью рассмотренного алгоритма удалось сформировать базу данных о 421 468 пользователей из г. Кирова. Остальные примерно 97 000 пользовательских профилей, по нашей оценке, представляют собой, в основном, либо неактуальные, либо фейковые аккаунты: у них нет ни одного друга из собранных нами профилей кировчан и они не состоят в популярных кировских группах.

Следующей важной проблемой при анализе социальных сетей является получение полной и актуальной информации. Ежедневно пользователи обновляют профили, добавляют новые комментарии и фотографии, ставят «лайки». Например, в период сбора данных с 28.12.2016 г. по 11.01.2017 г. название родного го-

рода поменяли 160 человек. Эта проблема может быть решена только путём непрерывного мониторинга информационных потоков в социальных медиа, то есть постоянного наблюдения за активностью и сообщениями пользователей, содержащими интересующую исследователя информацию.

Еще одной проблемой, требующей решения при анализе данных социальных сетей, является значительное количество незаполненных полей. Самой полной информацией оказались данные о *гендерной принадлежности*¹. Не указавших пол в базе лишь 13 человек. Но эта проблема была решена путем соотнесения данной переменной с фамилией и именем. Интересно, что все пользователи, не указавшие пол, оказались мужчинами.

Поле *семейное положение* не заполнили 72% пользователей, *отношение к алкоголю и к курению* – 91,7% и 91,4% соответственно. Информации о *текущем роде занятий* нет у 81,4%, о *политических предпочтениях* – у 90,7%, о *религиозных взглядах (мировоззрении)* – 91,3%.

Наиболее высокую сложность представляет корректная обработка поля *дата рождения*. Помимо пропущенных значений (57,4%), сложностью является то, что разработчики социальной сети ВКонтакте самым ранним годом рождения позволяют ставить 1901 год. Значительное количество жителей города, не желая делиться информацией о возрасте, поставило начало XX века временем своего рождения. Так, 1901 г.р. указали 1272 человека. После анализа профилей кировчан было принято решение в качестве начальной учитываемой даты рождения использовать 1950 год. Поставившие более ранние даты рождения рассматривались как не указавшие возраст. Последними датами рождения, указанными в профилях пользователей, стали годы с 2000 по 2003 гг. Для упрощения анализа пользователи были поделены на четыре возрастные группы: юность (13–18 лет), молодость (19–35 лет), зрелость (36–60 лет) и старость (старше 60 лет).

Проблема незаполненных полей является одной из наиболее значимых при анализе социальных медиа. Частично она может быть решена за счет воз-

¹ Здесь и далее все анализируемые переменные будут указаны курсивом.

возможности анализа только заполненных полей, так как выборка является значительной и репрезентативной. Например, семейное положение отметили 114 757 человек, что, несомненно, является достаточным для социологического анализа.

Основными методами анализа полученной количественной информации явились:

1) методы частотного анализа для номинальных переменных. Для определения центральной тенденции использовались мода и медиана (в ранжированном списке) [3, с. 164–178];

2) методы выявления статистически значимых связей между номинальными переменными: коэффициенты связи, основанные на критерии «хи-квадрат» (коэффициенты Пирсона, Чупрова и Крамера), номинальный регрессионный анализ [5].

Из всех зарегистрированных пользователей 54,2% – это женщины и 45,8% – мужчины. Среди отметивших возраст наиболее распространенными являются 1980–1989 г.р. (36,3%) и 1990–1999 г.р. (32,2%)². Семейное положение чаще всего определяется как *женат/замужем* (47,6%) и *не женат/не замужем* (19%). В активном поиске находятся 16% зарегистрированных в социальной сети. Интересно, что женщины значительно чаще, чем мужчины, отмечают свое семейное положение как состоящие в браке (55,1% против 39,9%).

У большинства пользователей указан *номер телефона* (84,6%), есть *фотография* (93%). У каждого второго пользователя *типом платформы, через которую был осуществлен последний выход*, является полная версия сайта (48,3%). Через приложение для Android выход осуществили 26,8%, мобильной версией сайта воспользовались 16,7%, приложением для iPhone – 6,4%.

Социальная сеть предлагает посетителям поразмышлять над такими вопросами как *главное в людях, главное в жизни, отношение к курению и алкоголю, политические предпочтения*. Вопросы являются закрытыми, предполагая несколько

² Здесь и далее проценты даются от числа заполнивших данное поле.

вариантов ответов. Наиболее часто пользователи указывают, что *главным в людях* являются доброта и честность (58,4%), а также юмор и жизнелюбие (18,2%). *Отношение к курению* чаще всего отмечается как негативное (30,8%) и резко негативное (32,8%). О нейтральной позиции по отношению к этой вредной привычке заявили 21,1%, о компромиссной – 8,2%. *Отношение к алкоголю* более терпимое, чем к курению: резко негативное (25,8%) и негативное отношение (24,6%) более тесно соседствует с нейтральным (26,8%) и компромиссным (16,6%).

Остановимся на поле *главное в жизни*, так как в приведенных ниже примерах оно будет проанализировано более подробно. Условно эта переменная была обозначена как *жизненная позиция*. Модальной категорией при определении жизненной позиции является *семья и дети* (59,4%). В этой же строке будет находиться и медиана, определяемая для номинальных переменных в ранжированном списке [3, с. 172]. Поэтому с уверенностью можно сказать, что семья и дети являются приоритетом в определении жизненной позиции. Если рассматривать центральную тенденцию как бимодальную, то можно говорить о включении в список позиций категории *саморазвитие* (17,9%). Но, несомненно, она будет гораздо менее репрезентативна. Абсолютно незначимыми для жителей г. Кирова являются *наука и исследования* (1,4%), а также *слава и влияние* (1,2%), может быть в силу того, что данный регион представляет мало возможностей для развития данных позиций (табл. 1).

Таблица 1

Жизненная позиция пользователей социальной сети ВКонтакте,
жителей г. Кирова, %

Жизненная позиция	Частота	%
1. Семья и дети	21 036	59,4
2. Саморазвитие	6 341	17,9
3. Развлечения и отдых	2 505	7,1
4. Совершенствование мира	1 785	5,0
5. Карьера и деньги	1 488	4,2
6. Красота и искусство	1 328	3,8
7. Наука и исследования	492	1,4
8. Слава и влияние	434	1,2
Итого	35 409	100,0

Исследование показало ряд сильных и слабых зависимостей между отдельными позициями полей. Для их анализа были использованы коэффициенты связи, основанные на критерии «хи-квадрат» (коэффициенты Пирсона, Чупрова и Крамера), а также номинальный регрессионный анализ [5].

Приведем пример расчётов силы связи с использованием коэффициентов связи, основанных на критерии «хи-квадрат», для полей *жизненная позиция* Y, принимающего 8 значений и *возрастная группа* X, принимающего 4 значения. Для поиска силы связи необходимо проверить статистическую гипотезу (H_0) о независимости признаков X и Y, применяя статистический критерий «хи-квадрат».

Расчет критерия (его эмпирического, или выборочного значения $\chi^2_{\text{выб}}$) производится по формуле:

$$\chi^2 = \sum_{i,j} \left[\frac{(n_{ij}^{\text{теор}} - n_{ij}^{\text{эмн}})^2}{n_{ij}^{\text{теор}}} \right], \quad (1)$$

где $n_{ij}^{\text{эмн}}$ – наблюдаемая частота, расположенная на пересечении i -й строки и j -го столбца таблицы сопряженности (эмпирическая частота), $n_{ij}^{\text{теор}}$ – частота, которая располагалась бы в той же ячейке таблицы сопряженности, если бы переменные были статистически независимы (теоретическая частота).

Теоретическая частота находится по формуле:

$$n_{ij}^{\text{теор}} = \frac{\sum_{j=1}^c n_{ij}^{\text{эмн}} \cdot \sum_{i=1}^r n_{ij}^{\text{эмн}}}{n}, \quad (2)$$

$$n = \sum_{i=1}^r \sum_{j=1}^c n_{ij}^{\text{эмн}}. \quad (3)$$

Приведем пример расчёта силы связи для полей *жизненная позиция* и *возрастная группа*. Рассмотрим эмпирическую частоту ($n_{ij}^{\text{эмн}}$) на пересечении полей *жизненная позиция* и *возрастная группа* (табл. 2).

Таблица 2

Эмпирическая частота пересечения полей жизненная позиция и возрастная группа

	1	2	3	4	5	6	7	8	Сумма
13–18 лет	1 464	119	320	59	116	324	130	37	2 569
19–35 лет	6 858	579	817	143	501	2 137	343	149	11 527
36–60 лет	2 543	82	91	25	146	552	60	32	3 531
Старше 61	281	9	19	3	28	74	14	2	430
Сумма	11 146	789	1 247	230	791	3 087	547	220	18 057

Примечание: 1 – семья и дети; 2 – карьера и деньги; 3 – развлечения и отдых; 4 – наука и исследования; 5 – совершенствование мира; 6 – саморазвитие; 7 – красота и искусство; 8 – слава и влияние.

Рассчитав теоретические частоты ($n_{ij}^{теор}$) по формулам (2) и (3), вычислим выборочное значение «хи-квадрат» ($\chi^2_{выб}$) по формуле (1):

$$\chi^2_{выб} = 492,739.$$

Определим число степеней свободы по формуле:

$$df = (r-1)(c-1), \quad (4)$$

где r и c – количество значений рассматриваемых признаков X и Y соответственно: $r = 4$, $c = 8$. Зададим уровень значимости $\alpha = 0,05$. Найдем по таблице распределения χ^2 критическое значение критерия $\chi^2_{крит}$, для которого выполняется неравенство:

$$P(\xi \geq \chi^2_{крит}) = \alpha, \quad (5)$$

где ξ – обозначение случайной величины, имеющей распределение χ^2 с рассматриваемым числом степеней свободы.

Для указанных условий значение $\chi^2_{крит} = 32,671$.

Таким образом, получили $\chi^2_{выб} \geq \chi^2_{крит}$ ($492,739 \geq 32,671$). Следовательно, с вероятностью 95% ($1 - \alpha$) мы можем говорить о наличии зависимости между возрастом респондентов и тем, что они считают главным в жизни, то есть их жизненными ценностями.

Строго говоря, значения рассматриваемого критерия не совсем пригодны для оценки связи между признаками, поскольку они зависят от объема выборки и

других обстоятельств, носящих случайный характер по отношению к силе измеряемой связи. Возникает необходимость определенной нормировки найденного значения критерия, то есть такого его преобразования, которое устранил описанную зависимость от случайных (для оценки связи) факторов.

Имеются разные подходы к требующейся нормировке. Наиболее известными являются такие, которые превращают критерий «хи-квадрат» в известные коэффициенты, называемые обычно по именам впервые предложивших их авторов – Пирсона, Чупрова, Крамера:

$$\text{коэффициент Пирсона:} \quad P = \sqrt{\frac{\chi^2}{\chi^2 + n}}, \quad (6)$$

$$\text{коэффициент Чупрова:} \quad T = \sqrt{\frac{\chi^2}{n \sqrt{(c-1)(r-1)}}}, \quad (7)$$

$$\text{коэффициент Крамера:} \quad K(\text{или } C) = \sqrt{\frac{\chi^2}{n \times \min(c-1, r-1)}}. \quad (8)$$

Все коэффициенты изменяются от 0 до 1 и равны нулю в случае полной независимости признаков. Как и критерий «хи-квадрат», эти показатели являются симметричными относительно признаков: с их помощью нельзя выделить зависимую и независимую переменную, на основе их анализа нельзя говорить о том, какая переменная на какую «влияет».

Для рассматриваемого примера коэффициенты принимают следующие значения: 0,1630 (коэффициент Пирсона), 0,0772 (коэффициент Чупрова), 0,0954 (Коэффициент Крамера). Полученные данные говорят о слабой степени зависимости между *возрастом респондентов* и их *жизненными ценностями*.

Следующим методом анализа переменных стал номинальный регрессионный анализ. Для примера снова рассмотрим зависимый номинальный признак *жизненная позиция* Y , принимающий 8 значений, и независимые номинальные признаки X^1 (*пол*, 2 значения), X^2 (*возрастная группа*, 4 значения), X^3 (*семейное положение*, 7 значений).

Выполним дихотомизацию номинальных данных, в результате чего независимый признак превратится в 8 дихотомических признаков Y_1 – Y_8 , а каждый из трех зависимых признаков в соответствующие дихотомические признаки: X_1^1 (женщины), X_2^1 (мужчины), X_1^2 (юность), X_2^2 (молодость), X_3^2 (взрослость), X_4^2 (старость), а также X_1^3 – X_7^3 , соответствующие различным состояниям семейного положения респондента.

Коэффициенты уравнений регрессии определяются по правилам классического регрессионного анализа, например, для Y_1 уравнение регрессии будет иметь следующий вид:

$$Y_1 = 0,025 X_1^1 + 0,013 X_2^1 + 0,025 X_1^2 + 0,013 X_2^2 - 0,004 X_3^2 - 0,005 X_4^2 + 0,083 X_1^3 + 0,109 X_2^3 + 0,132 X_3^3 + 0,091 X_4^3 + 0,090 X_5^3 + 0,105 X_6^3 + 0,143 X_7^3.$$

Проанализировав полученные 8 линейных регрессионных уравнений, можно сделать следующие выводы.

Проявляя свою *жизненную позицию* в качестве главного в жизни пользователя ВКонтакте чаще всего выбирают:

- семью и детей – влюбленные (0,14) и помолвленные (0,13);
- карьеру и деньги – респонденты, находящиеся в активном поиске (0,013) и со сложностями с семейным положением (0,011);
- развлечения и отдых – респонденты влюбленные (0,021) и находящиеся в активном поиске (0,023);
- науку и исследования – респонденты со сложностями с семейным положением (0,006) и имеющие друга/подругу (0,003);
- совершенствование мира – респонденты со сложностями с семейным положением (0,014) и неженатые/незамужние (0,009);
- саморазвитие – помолвленные (0,044) и неженатые/незамужние (0,038);
- красоту и искусство – влюбленные (0,0098) и имеющие друга/подругу (0,0095);
- славу и влияние – респонденты со сложностями с семейным положением (0,007) и помолвленные (0,005).

Сравнивая числовые значения коэффициентов, расположенные в регрессионных уравнениях перед переменными X_1^1 (женщины) и X_2^1 (мужчины), можно сделать выводы о том, что главное в жизни выбирают чаще всего мужчины, а что – женщины:

- семью и детей, красоту и искусство – женщины;
- карьеру и деньги, развлечения и отдых, науку и исследования, совершенствование мира, саморазвитие, славу и влияние – мужчины;

Сравнивая числовые значения коэффициентов, соответствующих в регрессионных уравнениях переменным, характеризующим возрастную группу респондентов X_1^2 – X_4^2 , можно сделать выводы о том, респонденты какого возраста чаще всего выбирают каждую из предложенных ценностей жизни:

- семью и детей – 13-18 лет;
- карьеру и деньги – 19-35 лет;
- развлечения и отдых – 13-18 лет;
- науку и исследования – 13-18 лет;
- совершенствование мира – от 61 года;
- саморазвитие – 19-35 лет;
- красоту и искусство – 13-18 лет;
- славу и влияние – 13-18 лет.

Исследование информации, доступной в социальных сетях, является одним из наиболее перспективных методов социологического анализа, который, однако, содержит в себе ряд проблем. Проблемы в первую очередь связаны с получением актуальной и достоверной информации. Профили социальных сетей постоянно обновляются, их сложно верифицировать, пользователи оставляют много «пропущенных» полей. Решением указанных проблем может стать постоянный мониторинг обновлений, совершенствование программ обработки данных, а также анализ только заполненных полей.

Анализ социально-демографических характеристик пользователей, зарегистрированных ВКонтакте, показал, что данная социальная сеть пользуется попу-

лярностью у жителей. Приоритетное место в жизненной позиции кировчан занимают *семья и дети*, которые являются важными для всех категорий пользователей. Но значительно чаще её выбирают девушки в возрасте 13-18 лет, определяющие свое семейное положение как влюбленные и помолвленные.

Работа выполнена при финансовой поддержке Министерства образования и науки РФ, государственное задание ВятГУ № 34.2092.2017/ПЧ, проект «Разработка и исследование словарей оценочной лексики для анализа тональности текстов» (2017–2019 гг.).

Список литературы

1. Дукин Р. А. Феномен социальных медиа: проблемы социологического осмысления // Вестник Нижегородского университета им. Н. И. Лобачевского. Сер.: Социальные науки. – 2015. – № 4(40). – С. 122–126.
2. Митягина Е. В., Котельников Е. В., Окулов С. М. Применение автоматических технологий анализа социальных медиа в социологических исследованиях // Перспективы развития информационных технологий. – 2016. – № 33. – С. 56–60.
3. Ненашев М. И. Методы исследования в количественной социологии : учеб. пособие для студ. спец. «Социология». – Киров, 2016. – 267 с.
4. Численность населения Российской Федерации по муниципальным образованиям на 01.01.2016 г. – URL: http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/population/demography/ (дата обращения 10.02.2017).
5. Толстова Ю. Н. Анализ социологических данных. Методология, дескриптивная статистика, изучение связей между номинальными признаками. – М., 2000. – 352 с.
6. API ВКонтакте. – URL: https://vk.com/dev/first_guide (дата обращения 10.02.2017).

МИТЯГИНА Екатерина Владимировна – доктор социологических наук, профессор кафедры социологии и социальной психологии, Вятский государственный университет. 610000, г. Киров, ул. Московская, 36.

E-mail: mityagina2004@yandex.ru

АШИХМИНА Татьяна Викторовна – старший преподаватель кафедры фундаментальной информатики и прикладной математики, Вятский государственный университет. 610000, г. Киров, ул. Московская, 36.

E-mail: tv_ashihmina@vyatsu.ru

КОТЕЛЬНИКОВ Евгений Вячеславович – кандидат технических наук, доцент кафедры фундаментальной информатики и прикладной математики, Вятский государственный университет. 610000, г. Киров, ул. Московская, 36.

E-mail: kotelnikov.ev@gmail.com