Вопросы доверия данным социальных сетей

© Н. А. Скворцов
Институт проблем информатики ФИЦ ИУ РАН,
Москва
nskv@ipi.ac.ru

Аннотация

Традиционные подходы к сбору социально-ориентированных данных, такие как интервью и анкетирование, часто замещаются на более дешёвый сбор данных, доступных в вебе и онлайновых социальных сетях. Однако по ряду причин анализ данных, полученных из сети, может приводить к другим результатам, нежели даёт анализ опросов. Таким образом, возникает проблема доверия данным социальных сетей. В статье анализируются особенности данных социальных сетей, причины отличий их от данных опросов. Обращается внимание на анализ поведения людей как источника данных социальных сетей и отличие их от данных публичной деятельности. Обсуждаются вопросы доступности данных о поведении пользователей социальных сетей.

1 Введение

Известно, что организация и проведение опросов с целью сбора и анализа данных, отражающих социальные связи и предпочтения людей по тем или иным вопросам, является достаточно затратным делом. Накопление достаточного количества данных для анализа требует привлечения интервьюеров, включает поиск людей, готовых пройти опрос, транспортные и другие издержки, а также занимает много времени.

Данные социальной природы, собираемые в вебе, в онлайновых социальных сетях, становятся при этом желанной альтернативой сбору требуемых данных во взаимодействии с опрашиваемыми людьми. В открытой информационной среде доступны как оперативные данные, отражающие текущие события и настроения, так и данные исторические, хранимые за годы существования определённых веб-проектов и онлайновых социальных сетей. Требуемые данные следует только выделить из всего массива данных, исходя из времени их появления, тематики обсуждений в сети, упоминания тех или иных объектов или событий.

Труды XVII Международной конференции DAMDID/RCDL'2015 «Аналитика и управление данными в областях с интенсивным использованием данных», Обнинск, 13-16 октября 2015

Однако практика показывает, что использование данных онлайновых социальных сетей и веба как источников социальной информации вызывает сложности, так как результаты анализа данных опросов могут сильно отличаться от результатов анализа данных онлайновых социальных сетей. Поэтому при использовании данных, доступных в вебе, неизбежно встаёт вопрос об их качестве и о доверии таким данным. Исследователи данных социальной природы в прикладных областях давно не испытывают эйфории от возможностей, открывающихся благодаря доступности огромных объёмов данных в вебе. Актуальность проблемы доверия данным подтверждается тем, что многие исследователи просто отказываются от использования данных онлайновых соцсетей в пользу традиционных подходов, мотивируя это тем, что открытые данные не отражают их запросов.

Данная статья посвящена проблеме доверия данным социальных сетей. В дальнейшем изложении в разделе 2 обсуждаются способы сбора данных социальной природы и их особенности. Раздел 3 посвящён понятиям оценки качества данных социальной природы. В разделе 4 оцениваются причины несоответствия результатов анализа данных, собираемых с использованием различных подходов. В последнем разделе описываются пути повышения степени доверия данным соцсетей, предлагается уделять большее внимание поведению агентов в сети, нежели той публичной деятельности, которую они ведут.

2 Подходы к сбору данных социальных сетей

Первым делом, обозначим разновидность данных, которые нас интересуют, и возможные их источники. Данные социальных сетей обычно состоят из множества агентов (активных участников сети) с набором описывающих их атрибутов и множества отношений одного или нескольких типов между агентами, также, возможно, снабжённых атрибутами. В некоторых сетях также используются ресурсы (пассивные сущности, в отличие от агентов) и, соответственно, отношения между агентами и ресурсами. Сеть представляется в виде графа, в вершинах которого находятся агенты и ресурсы, а рёбра или дуги представляют отношения. Такие данные могут отражать как достаточно

постоянные связи агентов (такие как родственные связи людей), так и динамику их взаимодействий (телефонные звонки, распространение новостей и другие).

Далее мы будем использовать термины социальная сеть, агент (вершина), отношение (ребро или дуга) в описанном выше смысле, то есть, связанном с графовым представлением данных социальной природы. Мы отличаем эти термины от онлайновых социальных сетей как сервисов для взаимодействия пользователей в интернете. Как данные социальных опросов, так и данные онлайновых соцсетей могут быть представлены в виде графов социальной сети.

Источники и подходы к сбору таких данных могут быть различны. К традиционным подходам, широко используемым в социологии, политологии, маркетинге, относятся анкетирование и интервьюирование. Первый предполагает самостоятельное заполнение форм опроса респондентом, второй – работу интервьюера, заполняющего формы опросов в процессе или по результатам интервью с респондентом, проводимому по заранее заготовленному сценарию. Этот сценарий, впрочем, в некоторых случаях может корректироваться в зависимости от промежуточных результатов интервьюирования некоторого количества людей. Таким образом, данные сети формируются из фрагментов, соответствующих личным (эгоцентрическим) сетям респондентов.

Ещё одним подходом является наблюдение. Оно предполагает фиксацию взаимодействий в определённой группе людей сторонним наблюдателем без вербального контакта с ними. Этот подход часто оказывается более точным, нежели другие, так как основан на фиксации фактов взаимодействия, а не на прошлых взаимодействиях, воспоминания о которых могли спутаться в памяти или забыться. Среди других традиционных разновидностей сбора данных можно отметить проведение экспериментов (как наблюдение за взаимодействиями в контролируемых условиях), анализ архивов, включающих печатные публикации, протоколы и другие документы.

Цифровые способы сбора данных, в первую очередь, повторяют перечисленные выше подходы, переносимые в цифровой вид вручную, но отличаются тем, что готовятся в цифровом виде изначально. Формы приложений и веб-сайтов реализуют опросы. Записи баз данных и логи заменяют наблюдение и работу с архивами. Структуры этих источников данных часто изначально рассчитаны на сбор данных о тех или иных типах взаимодействия агентов. Разновидностей источников данных социальных сетей здесь может быть великое множество, включая корпоративные базы данных о действиях клиентов или сотрудников, транзакциях, коммуникациях, перемещениях.

Наиболее знаменательным изменением в составе данных социальных сетей в последнее время стали общедоступные данные веба и деятельность пользователей онлайновых сетей и форумов. Радикальным отличием таких данных от данных традиционных источников является то, что доступные данные часто не имеют прямого отношения к исследуемым вопросам даже в тематическом плане и содержат помимо информации, полезной в решаемых задачах, большой объём сторонних данных. Коэффициент полезности таких данных слабый, на выделение полезной части данных и их очистку приходится тратить дополнительные усилия. Но за счёт колоссальных объёмов доступных данных они имеют большую ценность для исследований.

В вебе, форумах и онлайновых социальных сетях можно почерпнуть социальные данные различного характера.

Частично доступна личная информация о пользователях, предоставляемая ими о себе в сетях. Эта информация может представляться как значения атрибутов вершин формируемой социальной сети, описывающих её агентов.

Одним из наиболее используемых видов данных является публичная деятельность пользователей сети, включающая содержимое публикаций: сообщения, заметки, ответы на сообщения других пользователей. Они наиболее ценны в качестве источников данных об отношении пользователей сети к тем или иным объектам или событиям.

Помимо формирования содержимого самими пользователями сети, большинство онлайновых сетей и форумов предоставляют возможность обмена содержимым, например, перепубликации и отметки, говорящие о публичных предпочтениях пользователей.

Информация об использовании сетей, такая как длительность сессий использования ресурса, количество просмотров страниц определённой тематики и другие данные о поведении участников сети весьма и весьма важны для оценки непубличных предпочтений людей в сети.

Графы социальных сетей могут формироваться на основе статических отношений, например, неориентированные графы социальных связей отражают связи друзей в онлайновой социальной сети, ориентированные графы подписки – связей следования.

Графы на основе динамических отношений агентов могут отмечать либо видимые взаимодействия, такие как публикации на стене других пользователей сети и перепубликации сообщений других пользователей, либо скрытые взаимодействия, такие как посещения страниц других пользователей.

Затруднение сбора данных онлайновых соцсетей заключается в том, что в них часто присутствует разграничение доступа к тем или иным видам данных. Для сбора данных часто используется

доступ через API конкретных онлайновых сетей, либо если таковые не предоставляются, то через HTTP. Информация о поведении пользователей в сети гораздо реже бывает доступной. Для сбора данных о поведении могут использоваться специальные приложения в онлайновых сетях, на которые подписываются пользователи. Также и сбор анкет в онлайновых сетях может быть организован при помощи специализированных приложений. Предложение заполнить анкету, в том числе, может быть мотивировано непосредственно предшествующей этому деятельностью человека в сети [1].

Таким образом, подходы, описанные выше, годятся для сбора социальных данных разных видов, объёма, изменчивости и качества, что, несомненно, сказывается на их применимости для анализа и решения определённых задач.

3 Оценки качества данных

При оценке качества сетевых данных обычно используется несколько понятий [2].

Во-первых, на основе сравнения социальной сети, полученной в результате сбора данных, с реальной сетью определяется точность (accuracy) измерений. Например, точность измерения можно оценить, сравнивая результаты опросов с результатами наблюдения, так как наблюдение часто даёт гораздо более отражающую реальность картину. Измерения в когнитивных сетях, в которых отношения формируются на основе мнения или оценочного восприятия респондентов об отношениях других людей, гораздо менее точны, нежели измерения в эгоцентрических сетях.

Во-вторых, оценивается надёжность (reliability) измерений, определяемая разбросом результатов при повторных измерениях. Повтор измерений во времени в социальных сетях имеет смысл только для достаточно статичных отношений. В динамически меняющейся социальной сети надёжность измерений может быть оценена только для данных, полученных в одно время из разных источников или разными методами.

В-третьих, важной характеристикой качества сетевых исследований является достоверность (validity) измерений. Она учитывает корректность метода: постановки вопросов при сборе данных, анализа данных для решения определённой задачи, теоретических прогнозов в модели.

В то время как точность и надёжность измерений в социальных сетях отвечают за случайную погрешность измерений, достоверность измерений относится к оценке систематической погрешности. Безусловно, все три характеристики существенны при оценке доверия данным.

4 Различия характеристик данных онлайновых социальных сетей и опросов

Данные социальных сетей, полученные различными способами, имеют свои особенности и различные характеристики, которые могут быть принципиально важны для исследователей с точки зрения пригодности данных для тех или иных целей исследования. Понимание особенностей может влиять на применяемые методы исследования.

Причины различий в подходах к измерению на данных онлайновых социальных сетей, при сравнении их с измерениями на данных опросов, достаточно разнородны.

4.1 Ограничения применимости данных социальных сетей

Данные онлайновых социальных сетей обычно бывают ограничены в нескольких направлениях. Во-первых, они не могут охватить часть населения, не использующую социальные сети. Это существенное ограничение, невосполнимое без использования традиционных подходов к сбору данных. Соответственно, невозможно обобщить результаты исследований на всех жителей территории. Однако многие исследования рассчитаны на изучение измерений, связанных только с людьми, пользующимися социальными сетями.

Во-вторых, доступ к данным бывает ограничен правилами конкретной онлайновой социальной сети, принимаемыми её пользователями, либо установленным ими уровнем конфиденциальности данных. Публикации в онлайновых социальных сетях либо доступны для всех, либо могут быть доступны при условии установления связей с их автором.

В-третьих, необходимо учитывать, что на вопрос анкеты человек как-то ответит в любом случае, если того требуют правила опроса, а пользователь социальной сети просто не будет писать о том, чем не хотел бы делиться со своей аудиторией в сети. Например, известно, что люди редко пишут о собственных проблемах со здоровьем.

4.2 Сложности интерпретации данных социальных сетей для исследований

При интерпретации данных онлайновых социальных сетей возникает много неоднозначностей. Безусловно, фактов, которые могли бы быть полезны для использования в определённых задачах исследования в сетях гораздо больше, чем можно накопить с помощью опросов. Однако, в отличие от опросов, они не отвечают на чётко сформулированные вопросы. Текстовые публикации в онлайновых социальных сетях, несмотря на присутствие упоминаний сущностей, задействованных в исследовании, могут совершенно не относиться к теме исследований. Текст может

содержать цитирования других людей, суждения в иронической форме, в переносном значении.

В опросах одна заполненная анкета соответствует мнению одного респондента, учитываемому в исследовании. Количество же упоминаний в сети исследуемых сущностей и отношений не означает количество учтённых мнений. Но необходимо учитывать, что один человек в диалоге может не один раз обозначить свою позицию, выявлять первоначальный источник при перепубликациях или цитированиях содержимого, а также идентифицировать одних и тех же людей в случае использования данных из нескольких источников. Практически всегда приходится учитывать также временной и географический контекст. В некоторых исследованиях целесообразно ослаблять вес сообщений от более активных агентов или агентов с большим количеством отношений.

В то время как в анкетах ответы зависят от памяти о событиях, что само по себе вносит неточность данных, в соцсетях неопределённость ещё возрастает из-за того, что сообщения могут описывать как текущее состояние, так и воспоминания о взаимодействии. Время описываемых событий выбирается пользователей сети произвольно, а не задаётся, как в анкетах.

Социальные опросы основаны на заранее разработанных формулировках, и респондентам приходится действовать в рамках контролируемого словаря опроса. Пользователи онлайновых соцсетей не связаны какими-либо вопросами в выражении своих переживаний и при этом формулируют мнение об исследуемых сущностях на привычном для них языке. Задачей исследователей становится учёт языка, используемого в предметной области исследования.

Существенно влияют на содержимое возможности, предоставляемые теми или иными онлайновыми социальными сетями. Например, функциональности, связанные с перепубликацией сообщений, публикацией фотографий и видео, или их отсутствие существенно меняют поведение людей в сети.

4.3 Психологические факторы отличия данных соцсетей от данных опросов

Наиболее сложным для учёта в исследованиях остаётся человеческий фактор, так как при опросах и в онлайновых социальных сетях люди оказываются в существенно различающихся ситуациях [3], что может влиять на позицию, которую люди занимают по тем или иным вопросам, являющимся темой исследований.

В интервью и анкетировании респонденты выходят из обычной для себя обстановки специально, чтобы принять участие в опросе. За ограниченное время, которое респонденты специально выделяют на опрос и стремятся минимизировать, они дают ответы на темы, о которых могли не задумываться заранее.

Общение же в онлайновых социальных сетях является повседневным делом, на которое человек в комфортной для него обстановке по собственной инициативе выделяет время для создания сообщений. При публикации сообщений пользователи соцсетей могут быть ограничены во времени только вследствие ведения диалога или собственных представлений о срочности и актуальности сообщений, однако чаще всего пользователи уделяют достаточно времени для вдумчивой разработки сообщений. Они вольны задаваться интересными либо актуальными для них темами, затрагивать обсуждаемые другими темы, констатировать свои недавние взаимодействия или вспоминать прошлое. В соцсетях и форумах люди стараются публиковать собственный анализ и обдуманное отношение к сущностям и событиям окружающего мира, но в то же время могут высказывать импульсивные и поверхностные оценки. Таким образом, подготовленность к высказываниям на определённые темы и возможность сосредоточения на них в соцсетях выше.

Другим ракурсом, от которого оказывается зависимой специфика социальных данных, является разная аудитория, на которую рассчитывает человек при формировании мнения. При опросах её составляют интервьюер и исследователи, а при формировании содержимого в онлайновых социальных сетях – сообщество любых пользователей этой сети или знакомые люди, с которыми установлены связи в сети. Соответственно, человек, отвечающий на вопросы, будет отвечать так, чтобы качество ответов понравилось интервьюеру или исследователю. Человек, пишущий в онлайновой соцсети, будет стараться создать посредством сообщений определённый имидж себе перед аудиторией соцсети. Подрывающим доверие к данным социальных сетей оказывается вопрос несоответствия убеждений и обстоятельств человека тому, как он себя публично представляет в соцсетях.

В анкетах респонденты могут давать неверные ответы с целью манипулирования результатом исследования. И если в каждом конкретном случае это влияние ничтожно, то в анкетах, заполняемых в интернете и социальных сетях, результаты часто бывают существенно изменены из-за целенаправленного массового манипулирования результатами, организуемого участниками некоторых заинтересованных сообществ.

Среди пользователей интернета раньше была принята высокая степень анонимности представления себя в форумах. Эта ситуация заметно изменилась в сторону открытости с распространением больших онлайновых социальных сетей, в которых люди устанавливают отношения друг с другом, повторяя те или отношения в реальной жизни, и имеют возможность

заполнять профиль реальными данными о себе. Однако после раскрытия информации о всеобъемлющем слежении ряда правительств за информационными коммуникациями граждан многих государств по всему миру всё же не следует ожидать улучшения точности в личных данных пользователей социальных сетей. Таким образом, при представлении себя и своего мнения в сети в предполагаемую аудиторию люди стали включать и контролирующие или даже потенциально недружественные институты власти. Это влияет и на представление своей позиции по некоторым вопросам, и на реальность личных данных, представляемых людьми. Часто данные профилей не оставляют пустыми, но заполняют заведомо неверными сведениями, что существенно спутывает оценку социально-демографических характеристик на основе данных профилей.

4.4 Влияние рекламных технологий и вредоносной деятельности в сети

При анализе данных социальных сетей приходится учитывать последствия рекламных технологий и вредоносной деятельности в онлайновых социальных сетях.

В качестве рекламы сервисы онлайновых социальных сетей искусственно могут навязывать обсуждения определённых тематик, регулируя появление рекламируемых сообщений в лентах новостей и оповещениях пользователей. Алгоритмы такого регулирования могут быть различны и меняться со временем без какого-либо их открытого описания.

Фиктивные учётные записи, созданные для спама, формирования общественного мнения, доступа к данным профилей пользователей сети и других целей, необходимо выявлять и исключать при сборе данных для анализа. Реклама и фиктивные пользователи могут весьма существенно искажать результаты измерений на данных социальных сетей.

Соглашения, принимаемые при регистрации в онлайновых социальных сетях, обычно включают согласие на обработку данных для исследований самой сетью и, возможно, третьей стороной. Однако пользователи сетей не узнают о реальном использовании их данных. По этому поводу ведутся дискуссии об этической стороне использования данных в исследованиях. Использование скрытых данных онлайновых социальных сетей, например, через создание учётных записей, используемых для доступа к конфиденциальным личным данным, не только противоречит этике социальных исследований, но и усложняет работу другим исследователям, которым приходится выявлять подобные учётные записи для исключения из своих графов социальных сетей.

5 Направления для повышения доверия данным социальных сетей

Сбор данных онлайновых социальных сетей дёшев, но анализ данных обычно требует привлечения квалифицированных кадров. В онлайновых социальных сетях сбор данных должен начинаться не с одной вершины, а с некоторого множества доступных вершин. Это позволяет выбрать более представительные данные, а также рассредоточить сбор данных на многих вычислительных узлах. При невозможности охватить все данные большой социальной сети прибегают к техникам выборочного сбора данных (sampling), которые позволяют сформировать некоторый подграф данных социальной сети и при этом постараться восстановить на подграфе характеристики, подобные полной сети [4]. Это важно для того, чтобы по небольшой части данных онлайновой социальной сети можно было с какой-то степенью уверенности судить обо всём множестве данных сети. Для разных социальных сетей и специфики исследуемых данных (текстовые сообщения или отношения в сети) эффективные алгоритмы сбора данных могут отличаться.

Как видно из рассуждений предыдущей главы, доверие к данным социальных сетей зависит от множества факторов. Эта проблема сложная и комплексная. Но для повышения степени доверия к данным вполне можно предпринимать определённые меры, разрабатывая принципы их сбора и обработки и следя за качеством обрабатываемых данных.

5.1 Совместное использование данных соцсетей с данными опросов

Для возможности оценки качества данных необходимо сравнение данных, полученных разными способами или из разных источников (раздел 3). Рекомендуется совмещать данные, полученные разными способами, в одном исследовании [1]. Это позволяет, во-первых, сравнивать данные из разных источников, оценивая их точность и надёжность и принимая меры к их повышению, во-вторых, обогащать данные, полученные одним способом, данными, полученными с помощью других подходов.

Данные из онлайновых социальных сетей целесообразно сопровождать данными анкет, связанных с деятельностью в сети её пользователей. То есть, если в сети состоялось реальное социальное взаимодействие, детали и мотивы этого взаимодействия можно выяснить, например, предложив пользователю сети заполнить анкету. Анкетирование целесообразно проводить незамедлительно после факта взаимодействия, так как через некоторое время человек забывает детали тех или иных взаимодействий в сети. Сопоставляя описания пользователем взаимодействий с данными, основанными на регистрации

взаимодействий в сети, можно оценить точность измерения.

С другой стороны, если исследование основано на анкетировании или интервьюировании, в которых респонденты описывают своих социальные взаимодействия, целесообразно сопровождать их в исследовании данными наблюдений и измерений, тем самым проверяя объективность описания взаимодействия агентами сети.

В целом, наблюдаемое поведение пользователей сети чаще представляет объективные факты о взаимодействиях, а описание собственных взаимодействий пользователями сети – субъективное отношение к ним. Поэтому важным становится не только анализ данных, которые сознательно предоставляются аудитории пользователями онлайновых социальных сетей, но и анализ их поведения в сети.

5.2 Наблюдаемые и скрытые взаимодействия пользователей сети

К наблюдаемым взаимодействиям можно отнести перепубликации, отметки, сообщения на стенах друзей и другую подобную деятельность в онлайновых социальных сетях, данные о которых открыты для внешних наблюдателей и для исследования. Пользователи сети устанавливают социальные связи друг с другом, однако взаимодействуют только с небольшой частью связанных друзей [5]. На основании реальных наблюдаемых взаимодействий можно, например, существенно корректировать граф социальных связей [6].

Скрытые взаимодействия представляют собой действия пользователей сети по отношению к другим, не имеющие видимого следа в соответствии с правилами сети или выбранным уровнем конфиденциальности. Количество скрытых взаимодействий оказывается значительно больше, чем видимых [7].

С помощью регистрации трафика запросов пользователей онлайновых социальных через HTTP возможен сбор данных о сессиях. Такой способ доступа к данным позволяет собирать как собственно данные сессий (об их частоте и продолжительности), так и данные практически обо всех типах взаимодействий и идентифицируемых пользователях. Он может использоваться на закрытых данных для мониторинга социальных сетей, либо опираться на пользователей сети, устанавливающих специальное приложение.

Помимо взаимодействий, анализу подлежит множество других событий, составляющих динамические процессы в сети: присоединение новых участников, наведение новых и разрыв существующих связей, присоединение к сообществам и другие изменения.

Если в социально-ориентированных науках доступ к данным поведения составляет большую проблему, то для анализа социальных данных в бизнесе обычно бывают доступны данные поведения в корпоративных и коммерческих онлайновых сетях, генерируемые действиями людей на ресурсах, принадлежащих компаниям, заинтересованным в анализе данных. Если же компаниям необходимо анализировать данные открытых соцсетей, они сталкиваются с той же проблемой доступа к данным, что и социально-ориентированные науки.

На анализе поведения участников социальных сетей завязан широкий круг задач, небезынтересных деловым кругам, таких как анализ интересов и намерений, динамика использования ресурсов, анализ факторов влияния на выбор, поиск конкурентов или соратников, определение оптимальных каналов какой-либо деятельности и тому подобных.

При анализе поведения агентов сети возникают проблемы скорости сбора и обработки данных. Для решения этих проблем немаловажны алгоритмы выборки подграфов социальной сети, обеспечивающие распределённое хранение с учётом локализации и временной принадлежности данных.

5.3 Происхождение данных

Отдельными проблемами в вебе и онлайновых соцсетях становятся идентификация пользователей, принадлежность данных, их актуальность во времени и другие подобные вопросы. Для повышения качества измерений и доверия данным социальных сетей при сборе необходимо сопровождать их аннотациями о происхождении данных [8]. Данные без указания источника и идентификации авторства имеют существенно меньшую степень доверия даже при возможно высоком качестве их отбора. Метаданные происхождения могут включать, социально-демографические характеристики агентов сети, данные об изначальных источниках данных, периоды действительности данных, географические метки, пути миграции данных, способы получения вычислимых параметров и другие важные характеристики. При очистке данных на всех уровнях их использования метаданные об их происхождении оказываются незаменимыми для установления достоверности, подлинности, актуальности, точности данных и снятия множества вопросов, подрывающих доверие к ним.

5.4 Обнаружение вредоносной деятельности

Как на этапе сбора, так и на этапе анализа данных важно прикладывать усилия к очистке данных для того, чтобы фильтровать возможные фиктивные данные или навязанную деятельность.

С целью изменения характеристик сети, распространения спама, раскрытия конфиденциальных данных пользователей сети и

других вредоносных целей в онлайновых социальных сетях злоумышленники могут создавать множество поддельных учётных записей, компрометировать существующие записи реальных пользователей, распространять вредоносные приложения для сбора личных данных и рассылки спама [9].

Помимо прочего, результатом вредоносной деятельности в социальных сетях является понижение качества данных и повышение сложности их анализа. Из-за поддельных учётных записей и их связей меняется конфигурация сети и её характеристики. Фиктивные пользователи встраиваются в сеть таким образом, чтобы максимизировать свою центральность. Они генерируют содержимое для имитации реальных участников, а также содержат вредоносное содержимое, рекламу, ссылки на заражённые узлы. Контроль над множеством учётных записей реальных людей захватывается злоумышленниками с теми же целями.

Поддельные учётные записи для повышения центральности образуют друг с другом достаточно плотные подграфы. Более редкими являются связи, которые им удаётся установить с реальными людьми. Поэтому естественным решением для исключения таких подграфов является подход, основанный на методах выявления сообществ. Сообщества, содержащие проверенные вершины, рассматриваются как реальные, а не содержащие проверенных вершин – как поддельные [10].

6 Заключение

Проведённый обзор представляет видение проблемы доверия данным, полученным из веба и онлайновых социальных сетей. На основании него можно сделать вывод о том, что данные социологических опросов и данные онлайновых соцсетей имеют множество факторов отличия. Их комплексный учёт представляется затруднительным, в том числе, потому что формирование содержимого разными способами создаёт для участвующих людей разные ситуации психологического характера. Для повышения доверия к данным онлайновых социальных сетей рекомендуется совместно использовать данные опросов, данные соцсетей, данные наблюдения, оценивать их качество, предпринимать усилия по их очистке, снабжать метаданными происхождения.

Литература

[1] F. Abdesslem, I. Parris, T. Henderson. Reliable online social network data collection // Computational Social Networks. – Springer London, 2012. – P. 183-210.

[2] S. Wasserman, K. Faust. Social Network Analysis: Methods and Applications. – Cambridge University Press, 1994. – 825 p. – ISBN: 9780521387071.

[3] F. Conrad, M. Schober, C. Lampe, J. Pasek, L. Guggenheim. Can analyses of social media ever replace survey estimates? – 5th sociological conference named by B. A. Grushin “Big sociology: dataspace extension”. – 2015.

[4] M. Gjoka, M. Kurant, C. Butts, A. Markopoulou. Practical recommendations on crawling online social networks // Selected Areas in Communications, IEEE Journal on. – 2011. – V. 29. No. 9. – P. 1872-1892.

[5] Исследовательский проект «Вес Рунета». – URL: http://ifnm.ru/weight/

[6] C. Wilson, B. Boe, A. Sala, K. Puttaswamy, B. Zhao. User interactions in social networks and their implications // The 4th ACM European conference on Computer systems. – ACM, 2009. – P. 205-218.

[7] J. Jiang, C. Wilson, X. Wang, W. Sha, P. Huang, Y. Dai, B. Zhao. Understanding latent interactions in online social networks // ACM Transactions on the Web (TWEB). – 2013. – Vol. 7, No. 4. – P. 18:1-39.

[8] G. Barbier, Z. Feng, P. Gundecha, H. Liu. Provenance data in social media // Synthesis Lectures on Data Mining and Knowledge Discovery. – 2013. – Vol. 4, No. 1. – P. 1-84.

[9] H. Gao, J. Hu, T. Huang, J. Wang, Y. Chen. Security issues in online social networks // Internet Computing, IEEE. – 2011. – Vol. 15, No. 4. – P. 56-63.

[10] Q. Cao, M. Sirivianos, X. Yang, T. Pregueiro. Aiding the detection of fake accounts in large-scale social online services // The 9th USENIX conference on Networked Systems Design and Implementation. – USENIX Association, 2012. – P. 197-210.

Social Network Data Trust Issues

Nikolay A. Skvortsov

Traditional approaches to social-oriented data acquisition such as interviewing and questionnaires are often replaced with a cheaper web-based and social network data acquisition. However, for a number of reasons analysis of data obtained from the internet may lead to different results than using traditional approaches. Thus the problem of social network data reliability appears. In the paper, specificity of social network data and factors of its dissimilarity from questionnaire data are analyzed. Attention drawn to person behavior analysis as a social network data source and to its difference from public activity. Accessibility issues of social network user behavior are discussed.