ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 62 за 2014 г.

Чекмышев О.А., Яшунский А.Д.

Извлечение и использование данных из электронных социальных сетей

Рекомендуемая форма библиографической ссылки: Чекмышев О.А., Яшунский А.Д. Извлечение и использование данных из электронных социальных сетей // Препринты ИПМ им. М.В.Келдыша. 2014. № 62. 16 с. URL: http://library.keldysh.ru/preprint.asp?id=2014-62

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

О. А. Чекмышев, А. Д. Яшунский

Извлечение и использование данных из электронных социальных сетей

Москва — 2014

О. А. Чекмышев, А. Д. Яшунский

Извлечение и использование данных из электронных социальных сетей

В работе обобщен накопленный авторами опыт в области сбора и обработки данных из электронных социальных сетей в приложении к социально-географическим исследованиям. Описаны основные аспекты, проблемы и перспективы развития используемых методов.

Ключевые слова: социальные сети, сбор данных, автоматизация, Интернет, алгоритмы на графах, социология, демография, география

Oleg Alexandrovich Chekmyshev, Alexey Dmitrievich Yashunsky

Extraction and usage of online social network data

The paper summarizes the authors' experience in collecting and processing data from online social networks for use in social geographic research. The general aspects, problems and development perspectives are outlined.

Key words: social networks, data collection, automation, Internet, graph algorithms, sociology, demography, geography

Оглавление

Введение

Вряд ли кто-то станет отрицать повсеместное проникновение инфокоммуникационных устройств в жизнь современных людей. Происходящее можно оценивать положительно или отрицательно, но нельзя отрицать того, что всеобщая «телефонизация» и «интернетизация» создает как новые угрозы, так и новые возможности в различных областях человеческой жизни.

В частности, «цифровой след», который оставляют люди, пользуясь электронными средствами связи, открывает новые возможности для наук о человеке: социологии, демографии, социально-экономической географии и т.д.

Несомненно, наибольшим потенциалом обладают данные, которые доступны операторам сетей сотовой связи. Во многих странах число действующих контрактов на услуги сотовой связи уже превысило население [15], а данные о подключении аппарата к базовой станции позволяют достаточно точно определить местоположение абонента в каждый момент. Эти два фактора в совокупности создают предпосылки для круглосуточного отслеживания перемещений почти всех жителей страны. Естественно, информация о перемещениях конкретного человека защищена законодательно, однако уже есть прецеденты использования агрегированных данных от операторов сотовой связи для анализа внутригородских перемещений жителей [6].

Детальному анализу изменения местоположения абонентов сотовой связи препятствуют не только юридические аспекты, но и вычислительная трудоемкость. Подобные массивы информации — яркий пример «больших данных» (англ. «big data»), привлекающих внимание специалистов в области информационных технологий в последние годы.

Вместе с тем, помимо весьма детальных и объемных данных, которые могли бы быть получены из сетей мобильной связи, имеются и другие проявления «электронного следа» человека, менее детальные и полные (как следствие, менее трудоемкие в обработке), и размещенные в открытом доступе с согласия индивидуума. Речь идет об анкетах пользователей в электронных социальных сетях. Термин «социальная сеть» возник в социологии еще до появления современной вычислительной техники и обозначал модель взаимодействия индивидов (акторов) в обществе (см., например, [9]). Электронные социальные сети, появившиеся в начале 2000-х годов, можно считать своеобразным «отражением» или «проекцией» социальных сетей, как их понимают в социологии. В дальнейшем мы будем использовать термин *социальная сеть* для обозначения именно электронной социальной сети — фактически, базы данных о пользователях и связях между ними.

Существующие в настоящее время в Интернете социальные сети весьма разнообразны по своим целям, месту проживания и возрастному составу пользователей. Вместе с тем, сама природа социальной сети диктует наличие

элементов, присутствующих во всех сетях: а именно — страниц (анкет) пользователей и связей (как минимум, между пользователями, а зачастую еще и между другими объектами, выступающими в качестве узлов сети).

Данные в анкете заполняются пользователем самостоятельно, добровольно и размещаются в открытом доступе с его ведома. Таким образом, доступ к этим данным с правовой точки зрения ничем не отличается от доступа к любым другим данным из сети Интернет. Массовое извлечение и обработка этих данных требует определенных технических средств, однако это несоизмеримо доступнее, чем данные из сетей мобильной связи.

В данной работе освещаются различные вопросы, связанные с извлечением и обработкой данных из социальных сетей: потенциальные области применения, степень достоверности данных, технические аспекты сбора данных, алгоритмические аспекты обработки данных, примеры использования.

О характере доступных данных

В зависимости от социальной сети, доступные о пользователях данные, конечно, варьируются. Если причислять к социальным сетям популярные сервисы блогов, микроблогов или фото-блогов (например, livejournal.com, twitter.com и instagram.com), то, из «анкетной» информации о пользователе гарантированно можно рассчитывать только на какое-то имя, возможно вымышленное.

Если же ограничиться «классическими» социальными сетями (vk.com, facebook.com, ok.ru), то анкеты пользователей становятся более развернутыми — пользователям предлагается указать свое имя, дату рождения и пол, а также сведения о местах проживания, учебы и работы. Конечно, заполнение всех этих полей добровольное. Однако, учитывая, что основная цель социальной сети как сервиса — способствовать общению и установлению контактов, указание достаточно подробных сведений о себе представляется разумным с точки зрения пользователей, которые хотят, чтобы их было легче найти их одноклассникам, сокурсникам, коллегам по работе.

Помимо более или менее развернутых анкетных данных, страница пользователя содержит информацию о «друзьях» — связях с другими пользователями социальной сети, и сообществах (группах по интересам), в которых состоит пользователь. Все эти сведения, даже при том, что часть их может отсутствовать, носят достаточно регламентированный характер: администраторы социальных сетей стараются следить за тем, чтобы упоминаемые в анкетах сущности (города, школы, вузы) не отличались на страницах разных пользователей, хоть и не всегда успешно.

Кроме регламентированных анкетных данных, страницы пользователей обычно содержат нерегламентированные пользовательские данные — фотографии, текстовые заметки, аудио- и видео-файлы, прочие документы.

Нерегламентированное содержание также характерно для упомянутых выше сервисов блогов, только в их случае оно практически не дополняется анкетными данными. Анализ нерегламентированных данных также (по крайней мере, теоретически) возможен, но эта задача намного сложнее простой обработки анкетных данных. Далее мы ограничимся вопросами извлечения и анализа анкетных данных.

В социологических и демографических исследованиях источником данных могут быть статистические сведения, собранные уполномоченной государственной структурой (в рамках переписи населения или иной процедуры), или результаты опроса, проведенного в исследуемой группе людей. В первом случае исследователь ограничен в выборе показателей для которых имеются сведения. Во втором случае достоверность получаемых результатов напрямую зависит от качества выборки, обеспечение которого может оказаться не только трудоемким, но и дорогостоящим.

В отличие от опросного листа, анкета пользователя социальной сети имеет заранее определенный список полей (он задан создателями конкретной социальной сети) и исследователь не может на него влиять. По своей природе данные из социальных сетей ближе к данным переписи населения, чем к результатам опроса. Подобного рода источники информации использовались в исследованиях и в «докомпьютерную» эпоху. Так, например, в Советском Союзе для исследования миграций на микро-уровне географы использовали записи из домовых книг и учетных карточек отделов кадров предприятий (см. подробное описание в [4, с. 20–30]).

Вообще говоря, можно использовать социальные сети как инструмент для проведения опросов, но социологи относятся к подобной практике достаточно настороженно, поскольку обеспечить и доказать репрезентативность выборке при таком «электронном» опросе гораздо сложнее, чем при проведении классических «поквартирного» или «телефонного» опросов. Тем не менее, имеются прецеденты использования данных из социальных сетей для формирования выборки в социологических исследованиях [10]. И все же социальные сети в социологических исследованиях обычно не столько источник информации, сколько объект исследования. Подобные исследования имеют в первую очередь прикладное значение для владельцев социальных сетей, поскольку способствуют повышению качества «таргетирования» — демонстрации рекламной информации согласно предпочтениям пользователя. С точки зрения этой прикладной задачи вполне естественными выглядят исследования поло-возрастной, социальной и территориальной структуры различных социальных сетей (например, [13]). Примечательно, что в этих исследованиях используются панельные данные, а не обработка всего массива информации о пользователях той или иной социальной сети. Статистическая обработка всех анкет пользователей социальной сети вполне технически осуществима [14], но почему-то не используется в подобных исследованиях.

Потенциальная нерепрезентативность — главный недостаток данных из

социальных сетей и в том случае, если их рассматривать как альтернативу статистической информации, полученной из официальных источников. Как будет показано далее, этот недостаток может сойти на нет при выборе генеральной совокупности «хорошо» представленной в социальной сети. К достоинствам данных из социальных сетей относятся:

1. Наличие сведений, не предоставляемых статистическими службами.
2. Детальность: данные собираются изначально в неагрегированном виде, могут быть агрегированы любым требуемым способом.
3. Ретроспективность: анкеты пользователей зачастую содержат не только информацию о текущем статусе пользователя, но и биографическую информацию, позволяющую формировать сведения о пользователе на произвольный прошедший момент времени.

Опыт использования данных из социальных сетей в исследованиях показывает, что можно реализовать эти преимущества (в особенности первые два) и избежать недостатка репрезентативности, если подавляющее большинство исследуемых представлено в социальной сети. Вполне естественно, что доля людей зарегистрированных в социальных сетях существенно выше среди более молодых возрастных групп.

По нашим наблюдениям, при отборе среди пользователей социальной сети людей в возрасте 20–29 лет (т. е. родившихся между 1984 и 1994 годами), проживающих в определенном населенном пункте, получаемое число анкет сопоставимо с количеством жителей соответствующих возрастов в населенном пункте. Использовавшийся способ поиска пользователей (как когда-либо учившихся в определенных школах) обеспечивает исключение из выборки части анкет несуществующих (вымышленных) пользователей, поскольку анкеты таких пользователей редко бывают подробно заполненными (с указанием школы). Конечно, гарантировать полное исключение подобных анкет невозможно, и задача фильтрации обрабатываемых данных заслуживает особого внимания, но «в первом приближении» можно считать, что для указанной возрастной группы по репрезентативности данные из социальных сетей приближаются к данным переписи населения.

С точки зрения перспектив использования данных социальных сетей в социологических и демографических исследованиях несомненный интерес представляет масштабный мониторинг анкет пользователей различных социальных сетей (например, в пределах Российской Федерации) для выявления территориальных и возрастных групп, в которых проникновение социальных сетей близко к 100%. Помимо выявления таких групп людей, подобный мониторинг также создаст возможность применять веса [1] для собранных данных и за счет этого, рассматривая данные уже как выборку, переходить к другим (более широким) изучаемым генеральным совокупностям. Вместе с тем, вопрос случайности такой выборки требует отдельного исследования, которое тем важнее, чем больше отличается выборка от 100% генеральной совокупности.

Все вопросы о репрезентативности данных, естественно, снимаются, если в качестве генеральной совокупности выступает множество пользователей конкретной социальной сети.

Рассмотрим теперь, как соотносятся анкеты пользователей (на примере социальной сети ВКонтакте) с анкетными листами переписи населения Российской Федерации 2010 г. [12] (см. таблицу).

Таблица

Поле в анкете переписи	**Наличие данных в анкете социальной сети**
Номер по порядку лица в домохозяйстве	Отсутствует
Отношения (родственные) к первому лицу в домохозяйстве	Присутствует возможность указывать родственников
Пол	Присутствует
Дата рождения	Присутствует
Место рождения	Присутствует
Состояние в браке	Присутствует, но варианты ответов отличаются от переписной анкеты
Гражданство	Отсутствует
Национальная принадлежность	Отсутствует
Образование и обучение	Присутствует в более детальном варианте
Владение языками	Присутствует
Источники средств к существованию	Отсутствует
Занятость и безработица	Присутствует в менее детальном варианте
Миграция	Присутствует в менее детальном варианте
Количество детей	Присутствует

Помимо перечисленных выше анкетных данных, анкета пользователя социальной сети содержит и другую информацию (например, пользователь имеет возможность перечислить места своего досуга), которая может быть использована для отслеживания перемещений пользователя по территории страны, тем самым существенно дополняя данные о миграциях пользователей.

Конечно, наличие полей в анкете не гарантирует, что они будут заполнены, и, тем более, что они будут содержать достоверные сведения. По опыту обработки данных из сети ВКонтакте можно сказать, что поля,

относящиеся к полу, возрасту и образованию, чаще всего заполнены, и, скорее всего, содержат достаточно достоверные сведения. В анкете социальной сети предусмотрена возможность указать свое состояние в браке, родственников и детей, но часто пользователи (особенно молодые) игнорируют эти поля или заполняют их вымышленной информацией.

Выработка и алгоритмизация критериев, позволяющих судить о достоверности заполнения тех или иных полей в анкете — еще одна задача, решение которой необходимо для масштабного использования данных социальных сетей в демографических и социологических исследованиях.

Технические аспекты извлечения данных

С технической точки зрения массовое извлечение данных из социальных сетей осуществимо достаточно просто, но все-таки находится за пределами возможностей рядового пользователя персонального компьютера. В задаче извлечения данных можно выделить три этапа:

1. Формирование списка необходимых анкет.
2. Выгрузка анкет из социальной сети.
3. Обработка анкет для извлечения данных.

Для проведения широкомасштабных исследований может иметь смысл создание «зеркала» социальной сети, т. е. сначала выгрузить все доступные анкеты, а затем уже выбирать среди них нужные, тем самым выполнив сначала пп. 2 и 3, а затем уже многократно выполнять п. 1. Если же исследуемая генеральная совокупность мала по сравнению с общим массивом анкет (в настоящее время крупнейшие социальные сети насчитывают десятки миллионов пользователей), то разумнее выгружать и обрабатывать только требуемые в исследовании анкеты.

Описываемые ниже технологии работы с социальными сетями основаны на опыте по извлечению данных из сетей ВКонтакте и Facebook, однако, по-видимому, они могут быть, с соответствующими изменениями применены и для других социальных сетей.

Средством формирования списка необходимых анкет выступает обычный механизм поиска пользователей, имеющийся в социальных сетях. Обычно такие механизмы позволяют устанавливать фильтры (ограничивая отбираемые анкеты по полу, возрасту, месту проживания, местам учебы), после чего выдается список анкет пользователей, удовлетворяющих критериям запроса. В сети ВКонтакте фильтры представляют собой набор заполняемых полей, а в сети Facebook изначально пользователю предлагается ввести поисковый запрос на естественном языке (например, «люди, которые учились в МГУ»), который затем может быть дополнен фильтрами, заданным с помощью заполняемых полей. В обеих сетях список результатов имеет фиксированный размер (порядка тысячи анкет), и если число результатов превышает этот порог, какие-то анкеты не будут показаны. В связи с этим необходимо формировать

поисковые запросы так, чтобы количество подпадающих под них анкет было заведомо небольшим.

При использовании данных ВКонтакте хорошей практикой оказалась выборка одновременно по возрасту и школе обучения — для каждого года рождения в конкретной школе учится лишь несколько десятков человек. При экспериментах с сетью Facebook, однако, использовать такие запросы оказалось гораздо менее удобно, потому что пользователи Facebook слишком часто не указывают свою дату рождения в анкете.

Примечательно, что поисковая программа сети ВКонтакте имеет доступ в том числе и к данным, которые пользователь не делал публичными: в выборку двадцатилетних попадают и те анкеты, где дата рождения скрыта.

На уровне браузера поисковые запросы транслируются в HTTP-запросы с достаточно простой структурой, что создает предпосылки для частичной автоматизации формирования таких запросов, и, как следствие, облегчает процесс формирования списка анкет.

После того, как получен список анкет пользователей, можно переходить к выгрузке анкетных данных. Это можно осуществить, как минимум, двумя способами. Во-первых, каждая анкета — это просто веб-страница, которая может быть сохранена в файл. Во-вторых, социальные сети часто предоставляют программные интерфейсы для доступа создания приложений, работающих в рамках социальной сети (API — Application Programming Interface), которые также можно использовать для получения необходимых данных (см, например, [11]). Поскольку основное назначение этих интерфейсов иное, этот способ часто приходится комбинировать с предыдущим, однако, он может быть предпочтительнее, так как облегчает последующий разбор анкет.

Оба способа технически несложно реализуются и сводятся к отправке на сайт социальной сети последовательности HTTP-запросов. Единственный нюанс, который надо учитывать при автоматизации — это ограничение на число запросов в единицу времени. При слишком частых обращениях к сайту, действия могут быть восприняты как атака на сайт, что приведет к блокировке доступа.

Данные, полученные через API социальной сети, заведомо структурированы, что существенно упрощает их обработку, хотя и в этом случае она может понадобиться. Так, например, населенные пункты из анкеты могут быть представлены числовыми идентификаторами, которые потребуют дополнительной «расшифровки». Для данных, полученных в виде HTML-страниц, неизбежно потребуется программная обработка, извлекающая из содержимого страницы значения интересующих полей анкеты (или другую имеющуюся там информацию). При работе с данными ВКонтакте для этого использовались стандартные средства POSIX-среды: программа текстового поиска и фильтрации grep, потоковый редактор sed и язык awk [5].

Результаты извлечения данных из анкет собираются в таблицы, в которых каждая анкета представлена строчкой, а поля анкеты — столбцами. К

таким таблицам легко применять агрегирующие операторы и делать выборки (например, средствами SQL). При исследовании взаимосвязей пользователей социальной сети предпочтительной может оказаться графовая модель хранения, как более соответствующая внутренней структуре данных.

Математические модели и алгоритмы

В простейшем варианте, когда рассматриваются только анкетные данные пользователей, но не рассматриваются связи между ними, обработка заключается, фактически, в агрегировании данных по каким-либо показателям. В проведенных нами исследованиях данные социальных сетей использовались для отслеживания миграций [2, 3, 6], в связи с чем особый интерес представляли данные, дающие территориально-временну́ю привязку пользователя.

В предположении, что проникновение социальной сети ВКонтакте в возрастной группе 20–29 лет близко к 100%, были проанализированы данные о текущем месте проживания (согласно анкете ВКонтакте) всех пользователей соответствующей возрастной группы, когда-либо учившихся в школе конкретного города (например, Норильска). Целью исследования было выявление мест скопления мигрантов из Норильска.

Одна из наиболее распространенных моделей миграционного взаимодействия (в частности, в эконометрике) — гравитационная. В рамках этой модели предполагается, что величина миграционного потока между двумя населенными пунктами определяется соотношением, сходным с законом всемирного тяготения; при этом вместо масс в зависимости участвуют людности населенных пунктов. Обозначая число мигрантов M, людности населенных пунктов через P_1 и P_2, а расстояние через D имеем:

$$M = P_1 \cdot P_2/D^2.$$

Дальнейшее развитие гравитационных моделей (см., например, [8]) заключалось в добавлении в эту формулу множителей, призванных учесть какие-то еще дополнительные факторы, а также возможность вхождения фактора в произвольной степени. Сами эти степени характеризуют важность фактора для объяснения миграции. В частности, показатель с которым входит в формулу расстояние называется «пространственное трение».

В итоге, формулы гравитационных моделей записываются в виде:

$$M = K_1^{a1} \cdot K_2^{a2} \cdot \ldots \cdot K_m^{am},$$

где K_i— значения факторов, а a_i — показатели степеней. Показатели a_i могут быть как положительными, так и отрицательными.

Переход в этом соотношении к логарифмам величин (например,

десятичным), превращает соотношение в линейное:

$$\lg M = a_1 \lg K_1 + a_2 \lg K_2 + \ldots + a_m \lg K_m.$$

Тогда, если имеется достаточное количество миграционных потоков, подчиняющихся одним и тем же законам, то при известных значениях миграционных потоков M и факторов $K_1,...,K_m$, коэффициенты $a_1,...,a_m$ могут быть определены линейной регрессией.

Допускается, что некоторые факторы выражены фиктивными переменными: для них в уравнение регрессии вместо слагаемого $a_i \lg K_i$ входит просто слагаемое a_i, если фактор фактор равен 1 (и не входит ничего, если он равен 0).

Эта модель была применена к данным о миграциях молодежи (20-29 лет), полученным из анкет пользователей сети ВКонтакте. В качестве факторов миграции использовались характеристики целевого населенного пункта: численность населения (нас.), расстояние до Норильска (расст.), признак расположения в Красноярском крае (рег., фиктивная переменная, равная 1 для населенных пунктов Красноярского края), признак «столичности» (столица, фиктивная переменная, равная 1 только для Москвы), признак центра региона (центр, фиктивная переменная, равная 1 для центров субъектов РФ), стоимость квадратного метра жилья в регионе (цена м²).

Полученные в результате линейной регрессии показатели a_i для каждого из факторов приведены в заголовке графика, изображенного на рисунке. Из полученных таким образом показателей было найдено «прогнозное» значение для числа мигрантов в каждый из городов-целей. На графике (см. рис. 1) в логарифмических осях показано соотношение «прогнозного» (по регрессии) и «реального» (по данным сети ВКонтакте) числа мигрантов.

Целью такого анализа было выявление городов, в которых число мигрантов не объясняется перечисленными выше факторами — т. е. таких, которые выбиваются из общей картины данных (для таких данных часто используется английский термин «outlier»).

В качестве порогового значения для определения «нестандартных» целей была выбрана разница в 0,5 между логарифмами прогнозной и реальной величины миграционных потоков. То есть, если потоки различались в $10^{0,5} \approx 3$ раза, то считалось, что такой город-цель не укладывается в модель. Исходя из графика, можно выделить как города, в которые поток оказался больше, чем предполагалось моделью, так и города, в которые поток оказался меньше.

Превышение над прогнозным значением демонстрируют: Санкт-Петербург (13-кратное), Белгород (9-кратное), Новосибирск (9-кратное), Кедровый (7-кратное), Киров и Красноярск (6-кратное), Абакан, Краснодар, Нижний Новгород и Старый Оскол (4-кратное), Обнинск (3-кратное). Список городов, в которые поехало меньше мигрантов, чем ожидалось, также обширен: в Астрахани, Владивостоке, Вологде, Петропавловске-Камчатском и Хабаровске

поток в 5 раз меньше прогнозного, в Йошкар-Оле, Махачкале, Нижнем Тагиле, Сургуте и Чите — в 4 раза, в Дзержинске и Мурманске — в 3 раза.

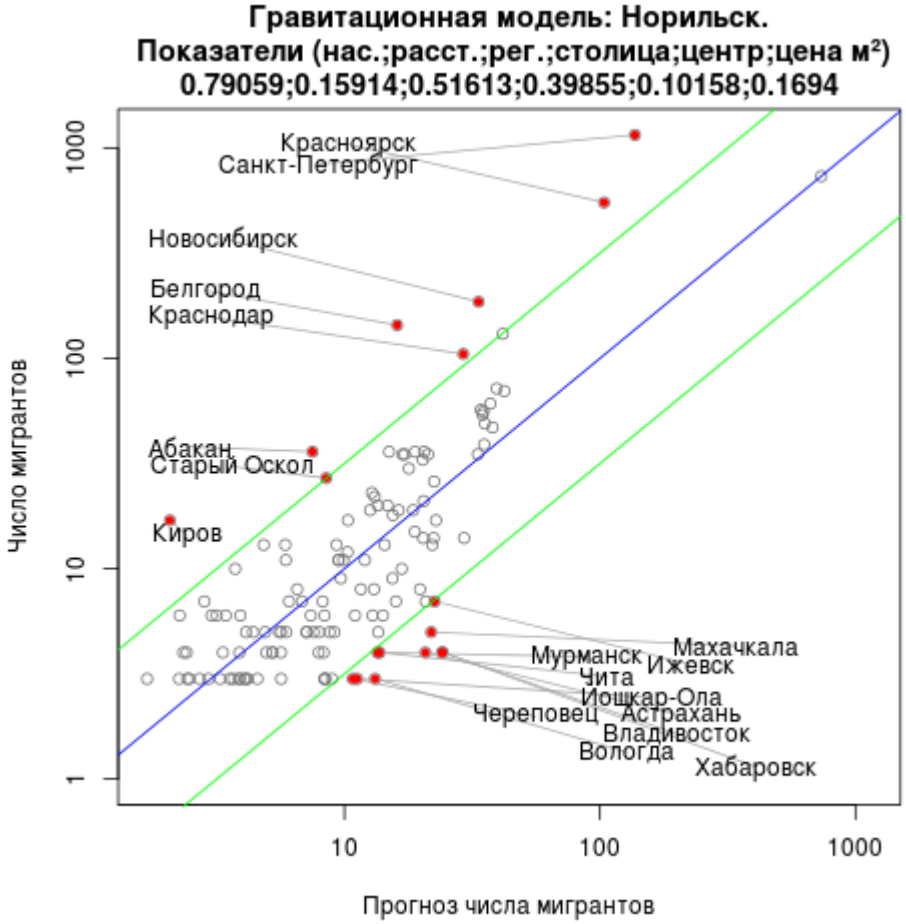

Рис. 1. Соотношение числа мигрантов по данным сети ВКонтакте и «прогнозных» значений, полученных из линейной регрессии

Эти расхождения означают, что для указанных городов-целей величина миграционного потока не объясняется уже рассмотренными факторами, а требует введения в рассмотрение еще каких-либо факторов.

Отметим, что подобное исследование невозможно было бы провести на основе данных переписи, так как миграции в анкете переписи учитываются только с точностью до субъекта федерации.

В описанном исследовании отслеживалось только две точки в траектории каждого человека. Более детальный анализ требует формализации того, как перемещения человека в пространстве отражаются в его анкете в социальной сети. Обобщая, можно говорить о *карьерной траектории* человека [3], рассматривая не только его места жительства, но и места учебы и работы. Анализировать совпадения карьерных траекторий или их фрагментов становится удобно, если рассматривать двудольный граф, в котором вершины одной доли соответствуют пользователям, а вершины другой доли — местам

жительства, учебы, работы. Тогда совпадениям (хотя бы фрагментарным) карьерных траекторий соответствуют полные подграфы этого двудольного графа.

Эксперименты по выявлению совпадений в карьерных траекториях были проведены нами на данных о жителях севера Финляндии. Было выгружено около 18 000 анкет пользователей сети Facebook, отобранных по запросу «когда-либо проживавшие» в одном из четырех городов: Оулу (Oulu), Кайани (Kajaani), Рованиеми (Rovaniemi), Раахе (Raahe). Из анкет пользователей извлекалась вся доступная информация, имеющая отношение к карьерной траектории и перемещениям пользователя. По полученным данным был сформирован двудольный граф (в одной доле — пользователи, в другой — точки карьерных траекторий), после чего использовалась модификация алгоритма Брона — Кербоша [7] для поиска полных двудольных подграфов.

Анализ полученных клик, к сожалению, не выявил неожиданных закономерностей. Крупнейшие клики содержали по две точки карьерной траектории и порядка 500 пользователей. Своим появлением они были обязаны чаще всего весьма естественному совпадению у множества пользователей пары «место проживания, учебное заведение». Наиболее крупные клики из тех, в которых точки карьерной траектории были территориально удалены, соответствовали основным миграционным потокам, например, из Рованиеми в Оулу — крупный образовательный центр. Совпадение трех и более точек в карьерной траектории происходило существенно реже — такие клики содержали лишь несколько десятков пользователей, и чаще всего в них территориально удалены были лишь две точки. Отсутствие значительного количества совпадений сложных фрагментов карьерных траекторий не означает, что подобные совпадения невозможны, а лишь говорит, что они относительно редки и при обнаружении заслуживают пристального внимания. Фрагменты карьерных траекторий, полученных в данном исследовании, отображены на карте на рис. 2. Они позволяют сделать неочевидный вывод о преобладании миграционных потоков между городами севера над оттоком в столицу — Хельсинки (Helsinki).

Помимо искусственно сформированного графа, содержащего отображение карьерных траекторий, можно анализировать естественным образом заданный социальной сетью граф взаимосвязей (дружбы) между пользователями — *социального графа*. В этой области имеется, например, исследование «дружбы между странами» [16], в котором вершины социального графа объединены по территориальному (страновому) признаку. В результате, каждой стране соответствует вершина, а кратность ребер между странами соответствует интенсивности дружбы. Исследование неожиданно продемонстрировало тесные связи Казахстана с Центральной Африканской Республикой, что, впрочем, вполне может объясняться ошибкой выборки. К сожалению, методология этого исследования не прописана детально, что затрудняет анализ полученных результатов.

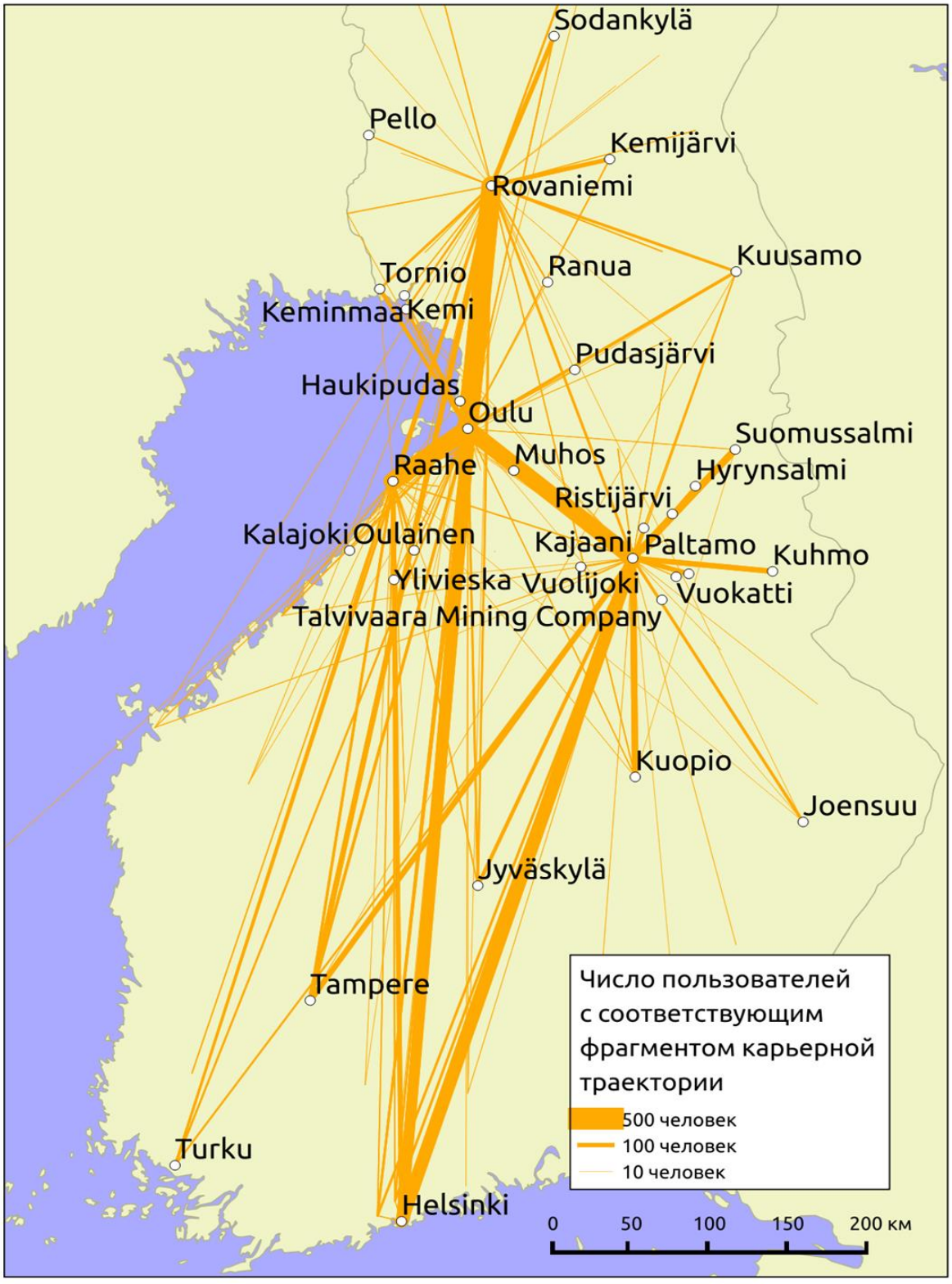

Рис. 2. Фрагменты карьерных траекторий для пользователей сети Facebook, когда-либо проживавших в городах Кайани, Оулу, Раахе, Рованиеми.

Одновременный анализ структуры социального графа и анкетной информации, которой нагружены его вершины, открывает пути для верификации этой информации. Например, можно предполагать, что учащийся

некоторого учебного заведения должен иметь среди друзей в социальной сети определенное количество учащихся того же учебного заведения. В противном случае, эта информация может быть поставлена под сомнение. Подобные соображения позволят алгоритмизировать критерии отбора достоверных анкет, тем самым способствуя повышению качества данных ценой сокращения выборки.

Заключение

Изложенные выше соображения дают представление как о характере задач, возникающих при работе с данными из социальных сетей, так и о характере результатов, которые можно ожидать при обработке этих данных. Использование данных из социальных сетей позволило получить нетривиальные результаты при изучении миграций молодежи. Дальнейший более глубокий анализ данных из социальных сетей, с одной стороны, ставит ряд задач, интересных с точки зрения информатики и алгоритмов, а, с другой стороны, — открывает перспективы качественно новых подходов к сбору данных в социологических, демографических, географических исследованиях.

Список литературы

1. Батыгин Г. С. Лекции по методологии социологических исследований. Учебник для высших учебных заведений. — М.: Изд-во РУДН, 2007.
2. Замятина Н. Ю. Метод изучения миграций молодежи по данным социальных Интернет-сетей: Томский государственный университет как «центр производства и распределения» человеческого капитала (по данным социальной Интернет-сети «ВКонтакте») // Региональные исследования. — 2012, №2. — С. 15–28.
3. Замятина Н.Ю., Яшунский А.Д. Межрегиональные центры образования // Отечественные записки. — 2012. — № 3(48). — С. 74–84.
4. Маергойз И.М. Географическое учение о городах. — М.: Наука, 1987.
5. Яшунский А.Д., Замятина Н. Ю. Севера как зона роста российской провинции // Отечественные записки. — 2012. — № 5(50). — С. 227–239.
6. Bogorov V., Novikov A., Serova E. Self-exploration of the City URL: http://issuu.com/mosurbanforum/docs/eng-uf_354-373_data__tm
7. Bron C., Kerbosh J. Algorithm 457 — Finding all cliques of an undirected graph // Comm. of ACM. — 1973. — V. 16. — P. 575—577.
8. Etzo I. Internal migration: a review of the literature // MPRA Paper No. 8783, University Library of Munich, Germany, 2008. — Pp. 1–27.
9. Granovetter M. S. The Strength of Weak Ties // The American Journal of Sociology. — 1973. — 78 (6). — Pp. 1360–1380.
10. Wadhwa V., Saxenian A., Freeman R., Gereffi G., Salkever A. America’s Loss is the World’s Gain. Part IV. America’s New Immigrant Entrepreneurs. (March 2,

2009). Available at SSRN: http://ssrn.com/abstract=1348616 or doi:10.2139/ssrn.1348616
11. Запросы к API ВКонтакте. URL: https://vk.com/dev/api_requests
12. Переписные листы переписи 2010 г. Лицевая сторона, URL: http://www.gks.ru/free_doc/new_site/perepis2010/croc/Documents/Vol1/l1.pdf Оборотная сторона, URL: http://www.gks.ru/free_doc/new_site/perepis2010/croc/Documents/Vol1/l2.pdf
13. Социальные сети в России. Дата публикации: 24.03.2014. URL: http://corp.mail.ru/media/files/issledovanie-auditorij-sotcialnykh-setej.pdf
14. Статистика по профилям пользователей Вконтакте. Дата публикации: 11.07.2011. URL: http://habrahabr.ru/post/123856/
15. Число контрактов на мобильную телефонную связь (на 100 чел.), данные International Telecommunications Union. Mobile cellular subscriptions (per 100 people). URL: http://data.worldbank.org/indicator/IT.CEL.SETS.P2/countries
16. Interactive: Mapping the World's Friendships. URL: http://www.facebookstories.com/stories/1574/interactive-mapping-the-world-s-friendships