Задачи и методы определения атрибутов пользователей социальных сетей[1]

© Антон Коршунов
Институт системного программирования РАН, Москва
korshunov@ispras.ru

Аннотация

Рост популярности онлайновых сервисов социальных сетей — основных источников персональных данных о пользователях Интернета — открывает беспрецедентные возможности для решения исследовательских и бизнес-задач, а также создания вспомогательных сервисов и приложений для пользователей социальных сетей. Определение неизвестных атрибутов пользователей является одной из фундаментальных проблем анализа социальных данных. В представленной работе рассмотрены методы решения некоторых актуальных задач, связанных с определением скрытых пользовательских атрибутов: поиск сообществ пользователей, определение демографических атрибутов пользователей, а также идентификация пользователей в различных социальных сетях.

1 Введение

Онлайновые социальные сети (Facebook, Twitter, YouTube и другие) к настоящему моменту стали неотъемлемой частью Веба и продолжают набирать популярность [1, 2]. За последнее десятилетие социальные сервисы существенно изменились в плане архитектуры, функционала и пользовательского интерфейса. С одной стороны, это обусловлено стремлением сделать их использование более удобным, а с другой - активной коммерциализацией и необходимостью увеличить время, проводимое пользователями на страницах сервисов.

С точки зрения анализа данных, социальная сеть в её современном понимании представляет собой граф с произвольным числом типов вершин и рёбер, весами и атрибутами, допускающий наличие множественных связей между узлами [3]. Возможность создания текстовых и мультимедийных объектов внутри сети делают её уникальным источником данных о личной жизни и интересах реальных пользователей (переписка, дневники, фотоальбомы, видеозаписи, музыкальные композиции и т.д.). Всё это обуславливает повышенный интерес к сбору и анализу социальных данных со стороны компаний (конкурентное преимущество) и исследовательских институтов (новые задачи и точки приложения известных подходов) [4].

Обработка социальных данных требует также разработки соответствующих алгоритмических и инфраструктурных решений, позволяющих учитывать их размерность. К примеру, граф социальной сети Facebook на сегодняшний день содержит более 1 миллиарда пользовательских аккаунтов и более 100 миллиардов связей между ними. Каждый день пользователи добавляют более 200 миллионов фотографий и оставляют более 2 миллиардов комментариев к различным объектам сети. На сегодняшний день большинство существующих алгоритмов, позволяющих эффективно решать актуальные задачи, не способны обрабатывать данные подобной размерности за приемлемое время. В связи с этим, возникает потребность в новых решениях, позволяющих осуществлять распределённую обработку и хранение данных без существенной потери качества результатов.

Помимо большого объёма данных и высокой динамичности социальной сети, нужно принимать во внимание такие факторы, как нестабильность качества пользовательского контента (спам и ложные аккаунты), проблемы с обеспечением приватности личных данных пользователей при хранении и обработке, а также частые обновления пользовательской модели и функционала. В дополнение к перечисленным проблемам, это требует постоянного совершенствования алгоритмов решения различных аналитических и бизнес-задач.

Одной из фундаментальных задач анализа социальных данных является *определение неиз-*

[1]Работа выполнена при поддержке гранта РФФИ №13-07-12134 офи_м “Исследование и разработка методов распределенной обработки больших баз графовых данных”.

Материалы 15-й всероссийской конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции— RCDL-2013, Ярославль, Россия, 14-17 октября 2013 г.

вестных значений атрибутов пользователей. С этой целью специализированные методы анализа сетевых, текстовых и других данных применяются к социальному графу на разных уровнях его организации:

- **на уровне сети** скрытыми атрибутами могут быть группы пользователей (сообщества), в которых состоит пользователь;
- **на уровне пользователя** скрытыми могут быть биографические и демографические атрибуты пользователя, а также его интересы и предпочтения;
- **на межсетевом уровне** скрытым атрибутом пользователя в одной сети может быть идентификатор этого пользователя в другой сети.

В представленной работе рассмотрены некоторые актуальные задачи, связанные с определением атрибутов пользователей, а также разработанные нами методы их решения. Раздел 2 посвящён задаче поиска сообществ пользователей, целью которой является определение набора сообществ, в котором состоит каждый пользователь сети. В разделе 3 рассмотрена задача определения демографических атрибутов пользователей по текстам их сообщений и атрибутам профиля. Раздел 4 содержит описание задачи идентификации пользователей в различных социальных сетях.

2 Поиск сообществ пользователей

Поиск сообществ пользователей является важным инструментом изучения и анализа социальных сетей, позволяющим исследовать мезоскопическую (модульную) организацию сети и использовать полученную информацию для решения различных задач. К примеру, знания о структуре сообществ незаменимы для предсказания связей и атрибутов пользователей, расчёта близости пользователей в социальном графе, оптимизации потоков данных в социальной сети, некоторых аналитических приложений и т.д.

2.1 Задача

С функциональной точки зрения *сообщество* - это группа пользователей, выполняющая общую роль или функцию и обладающая общими свойствами, ценностями и целями. Немаловажным свойством также является тенденция к взаимному влиянию участников сообщества друг на друга.

Поскольку формализация приведённого определения и построение соответствующей модели представляет определённые практические трудности, важно выделить некоторые особенности сообществ пользователей, характерные для социального графа на уровне связей между пользователями.

Благодаря проведённым исследованиям эмпирических данных социальных сетей стало возможным составить следующий список *фундаментальных свойств* сообществ пользователей на уровне связей между пользователями в социальном графе [11, 12, 21]:

- вершины в сообществе более тесно связаны друг с другом, чем с вершинами за пределами сообщества;
- количество рёбер в сообществе растёт суперлинейно в зависимости от его размера;
- сообщества могут пересекаться, т.е. один пользователь может относиться к нескольким сообществам, что хорошо согласуется с тем фактом, что человек одновременно может играть несколько социальных ролей в обществе;
- сообщества имеют иерархическую структуру, что можно объяснить тенденцией человеческого общества к формированию иерархии социальных групп;
- размер сообществ распределён по степенному закону;
- количество сообществ, к которым принадлежит вершина, распределено по степенному закону;
- вершины с небольшой степенью чаще входят в небольшое число сообществ, тогда как вершины с большой степенью входят во множество сообществ.

Использование перечисленных свойств позволяет выполнять поиск сообществ пользователей как множеств вершин социального графа. Результатом работы метода поиска сообществ является покрытие — множество сообществ, в котором каждая вершина принадлежит как минимум одному сообществу.

2.2 Метод

В связи с перечисленными особенностями сообществ пользователей многие алгоритмы определения модульной структуры сетей неспособны корректно идентифицировать сообщества в социальном графе. Потенциально применимые алгоритмы можно разделить на классы, основанные на статистической значимости сообществ, случайных блужданиях, локальной оптимизации подграфов, вероятностных моделях и агентских моделях [5, 6]. Из рассмотренных классов только методы, основанные на агентских моделях, позволяют достичь оптимального сочетания качества

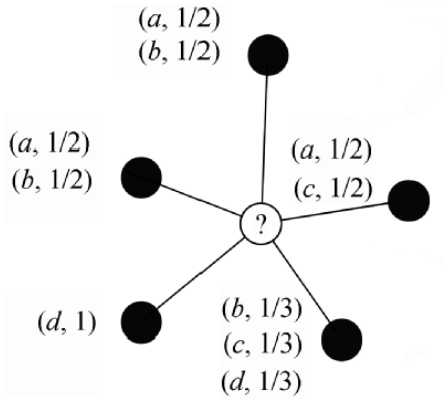

Рис. 1: Алгоритм SLPA: на каждой итерации "говорящие" узлы (окрашены чёрным) посылают "слушающему" узлу (окрашен белым) метки сообществ (a,b,c,d), выбирая их из текущего набора меток для каждого узла. "Слушающий" узел добавляет в свою память самую популярную из полученных меток.

результатов и производительности, необходимого для получения качественных результатов на графах из миллиардов вершин.

Разработанный нами метод представляет собой модификацию алгоритма *SLPA* [10], основанного на агентской модели. Данный алгоритм локально имитирует человеческое общение между парами индивидуумов, а глобально моделирует инфекционный процесс. Основой алгоритма является процесс обмена метками сообществ между вершинами в соответствии с динамическими правилами взаимодействия (рисунок 1):

1. Память каждого узла инициализируется уникальной меткой сообщества;
2. Затем итеративно повторяется последовательность шагов:
   (a) Выбирается "слушающий" узел;
   (b) Каждая из вершин-соседей выбранного узла ("говорящие" узлы) случайным образом выбирает метку с вероятностью, пропорциональной количеству меток данного типа в своей памяти, и посылает выбранную метку "слушающему" узлу;
   (c) "Слушающий" узел выбирает самую популярную из присланных ему меток и добавляет её в свою память.
3. В ходе пост-обработки для каждой вершины выбираются самые популярные метки с помощью заданного порога t;
4. Выбранные метки определяют принадлежность вершин к сообществам.

Вычислительная сложность алгоритма $O(T \cdot |E|)$ для произвольного графа и $O(T \cdot |V|)$ для разреженного социального графа (T - количество итераций).

Кроме того, алгоритм естественным образом формулируется в терминах *Pregel* [10] - вычислительной парадигмы для параллельных вычислений над графовыми данными. Разработанный метод был реализован в рамках *Spark.Bagel* [2] - фреймворка для параллельной обработки данных на кластере из потребительских компьютеров.

Вместе с тем, проведённое нами исследование результирующих покрытий выявило недостаточное качество результатов алгоритма SLPA для случая значительно пересекающихся сообществ. Более детальное исследование получающихся покрытий выявило неспособность алгоритма разделять значительно пересекающиеся сообщества.

Для решения этой проблемы была предложена следующая модификация оригинального алгоритма:

- применить алгоритм поиска *максимальных клик* размером не более 5 вершин к исходному социальному графу;
- всем вершинам, принадлежащим одной клике, назначить одну и ту же метку сообщества;
- вершинам, не принадлежащим найденным кликам, назначить уникальные метки сообществ;
- для каждой вершины найти *локальные сообщества* среди вершин, непосредственно связанных с ней;
- на каждой итерации "говорящий" узел посылает не одну, а несколько случайно выбранных меток каждому из "слушающих" узлов;
- на каждой итерации "слушающий" узел принимает только по одной метке от каждого из своих локальных сообществ (выбирается самая популярная из меток, отправленных вершинами одного сообщества);
- выполнить итерации и пост-обработку по аналогии с оригинальным алгоритмом.

Таким образом, на этапе инициализации намечаются центры будущих сообществ с помощью найденных клик. Затем с помощью модифицированных правил взаимодействия вершин между собой поощряется объединение локальных сообществ в глобальные.

[2] http://spark-project.org/

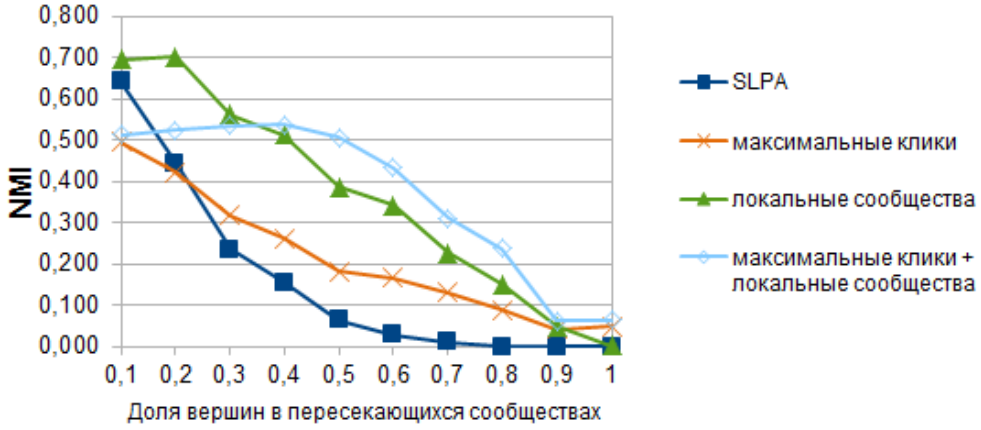

Рис. 2: Результаты оценки качества с помощью неориентированных LFR-графов. Каждый граф состоит из 2000 вершин, часть вершин состоит в пересекающихся сообществах (в данном случае каждая вершина состоит в 6 различных сообществах), остальные вершины состоят в непересекающихся сообществах.

2.3 Результаты

Наиболее распространённым способом оценки качества результатов методов поиска сообществ пользователей является сравнение двух покрытий для некоторого графа: найденного алгоритмом и *референсного*, то есть заранее заданного или известного.

Для оценки качества результатов разработанного метода использовался генератор LFR [11], способный генерировать случайные графы с заданной структурой сообществ.

В качестве количественной меры для сравнения двух покрытий применялась *нормализованная взаимная информация (NMI)* [22], значение которой показывает, в какой степени наличие информации о структуре одного из покрытий уменьшает неопределённость по поводу другого покрытия:

$$N(X:Y) = 1 - \tfrac{1}{2}[H(X|Y)_{norm} + H(Y|X)_{norm}],$$

где $H(X|Y)_{norm}$ - нормализованная условная энтропия X при условии Y (наоборот для $H(Y|X)_{norm}$), а X и Y - случайные величины, ассоциированные со сравниваемыми покрытиями.

С помощью выбранного метода оценки качества было проведено сравнение разработанного метода поиска сообществ пользователей с современными методами поиска пересекающихся сообществ (алгоритмы SLPA [10], MOSES [20], OSLOM [21] и другие). По результатам тестирования значение NMI для предложенного метода в большинстве случаев превосходит аналогичные показатели выбранных для сравнения методов. В остальных случаях лучшее качество показывают методы, обладающие большей вычислительной сложностью и неприменимые к графам большой размерности (миллиарды вершин).

На рисунке 2 приведено сравнение оригинального алгоритма SLPA с предложенными модификациями. Из графика следует, что лучших результатов позволяет добиться сочетание инициализации кликами с модификацией правил взаимодействия вершин с помощью локальных сообществ. Вместе с тем, использование локальных сообществ без клик позволяет исключить значительный объём вычислений, необходимый для поиска максимальных клик. При этом качество результатов ухудшается незначительно и по-прежнему существенно лучше оригинального алгоритма.

Для оценки производительности разработанного метода было проведено тестирование метода в параллельном режиме с помощью сервиса облачных вычислений *Amazon EC2*. По результатам тестирования метод показал линейную масштабируемость от числа вершин в исходном графе, а также от количества параллельно функционирующих вычислительных элементов.

3 Определение демографических атрибутов пользователей

При заполнении своего профиля в социальной сети пользователи зачастую по ошибке или преднамеренно не заполняют некоторые поля либо дают ложную информацию о фактах своей биографии, интересах и предпочтениях. Кроме того, в контентных сетях (Twitter, YouTube) пользовательский профиль часто ограничен набором базовых атрибутов, недостаточным для решения многих задач, предполагающих персонализацию результатов.

3.1 Задача

В системах интернет-маркетинга и рекомендаций особую важность представляет определение *демографических атрибутов* пользователя для таргетированного продвижения товаров и услуг в группах пользователей с одинаковыми значениями атрибутов. К таким атрибутам относятся пол, возраст, семейное положение, уровень образования, профессия, трудоустроенность, религиозные и политические взгляды, место жительства и т.д. Помимо интернет-сервисов, такие социодемографические характеристики находят применение в различных дисциплинах: социология, психология, криминология, экономика, управление персоналом и др.

Демографические атрибуты можно условно разделить на *категориальные* (пол, национальность, раса, семейное положение, уровень образования, профессия, трудоустроенность, религиозные и политические взгляды) и *численные* (возраст, уровень доходов). Условность разделения связана с тем, что значения численного атрибута можно отобразить в набор категорий и в дальнейшем рассматривать этот атрибут как категориальный. В частности, значения возраста можно разделить на несколько возрастных категорий, что часто применяется на практике.

Важным вопросом в решении задачи определения скрытых демографических атрибутов является выбор признаков. В представленной работе целевым источником данных была выбрана сеть *Twitter* - контентная сеть с преобладанием текстового содержимого (сообщений пользователей). Таким образом, задача состоит в определении скрытых демографических атрибутов пользователей социальной сети по текстам их сообщений. По смыслу задача эквивалентна классической задаче *социолингвистики*: определению характерных особенностей языка представителей различных социальных групп, позволяющих производить частичную идентификацию человека по принадлежности к этим группам. Такая постановка задачи позволяет использовать предложенный метод для обработки данных многих популярных сетей, поскольку текстовые данные являются наиболее распространённым средством коммуникации.

3.2 Метод

Абсолютное большинство современных методов определения демографических атрибутов пользователей основаны на применении методов машинного обучения с учителем с целью классификации пользователей по лингвистическим и другим признакам в предопределённые классы, соответствующие различным значениям изучаемых атрибутов. Сообщения пользователя рассматриваются как набор символьных строк, из которых извлекаются признаки, а для разметки применяются дополнительные источники данных о пользователе, причём в большинстве случаев разметка производится вручную [16–18].

Разработанный нами метод обладает следующими преимуществами:

- автоматическое построение исходного набора данных;
- извлечение большого количества признаков различных типов как из текстов сообщений, так и из полей профиля пользователя, с учётом особенностей микросинтаксиса *Twitter*;
- использование быстрого и эффективного метода отбора информативных признаков;
- расширяемый набор поддерживаемых атрибутов: все поля *Facebook*-профиля, а также любая информация о предпочтениях и интересах пользователя могут быть использованы в качестве атрибутов [3];
- расширяемый набор поддерживаемых языков благодаря использованию автоматической идентификации языка текста сообщений и применению метода построения исходного набора данных, не зависимого от языка.

Метод состоит из следующих этапов:

- построение исходного набора данных;
- предварительная обработка текста;
- построение признакового описания;
- отбор информативных признаков;
- обучение;
- классификация.

Все этапы, за исключением первого, выполняются отдельно для каждого атрибута.

На этапе **построения исходного набора данных** производится сбор данных пользователей из сети *Twitter*. Для каждого пользователя сначала запрашивается только его профиль в сети *Twitter*. При наличии в нём ссылки на профиль того же пользователя в сети *Facebook* (в которой набор пользовательских атрибутов существенно больше, чем в Twitter) запрашиваются и сохраняются все доступные сообщения пользователя из сети *Twitter*. После чего для текущего пользователя запрашивается и сохраняется его профиль в сети *Facebook*, из которого извлекаются указанные пользователем значения его атрибутов.

[3] https://developers.facebook.com/docs/reference/api/user/

Таким образом, элементом набора данных для каждого атрибута и языка является набор символьных строк, полученных из текстов сообщений и профиля одного пользователя в *Twitter*, а также значение атрибута у данного пользователя в *Facebook*.

На этапе **предварительной обработки текста** к текстам полученного на предыдущем этапе набора данных применяется метод определения языковой принадлежности текста (библиотека *language-detection* [4]). После этого данные пользователей распределяются в различные наборы данных в зависимости от языка пользователя.

Предварительно осуществляется фильтрация сообщений, авторство которых не принадлежит пользователю (*ретвиты*). Поскольку цитирование сообщений других пользователей является весьма популярным способом распространения информации в сети *Twitter*, этот шаг предварительной обработки особенно важен для повышения точности метода.

На этапе **построения признакового описания** из сообщений и полей *Twitter*-профиля пользователей извлекаются лингвистические признаки.

Сначала к исходным текстам применяется токенизация. Для элементов специфического синтаксиса сообщений (*хэштегов*, *@-ссылок*), а также слов из полей профиля создаются токены специальных типов, а для обычных слов из сообщений - токены стандартного текстового типа.

Из полученных токенов сообщений и полей профиля строится набор признаков в виде N-грамм размером от 1 до 7 с учётом порядка токенов. Аналогичный набор признаков строится для всех символов в текстах пользователя. Каждый тип признаков представлен двумя подтипами: с учётом и без учёта регистра символов.

Итоговый вектор признаков для пользователя является бинарным, то есть содержит только информацию о наличии или отсутствии признака в его текстовых данных. Количество экземпляров одного признака игнорируется.

На этапе **отбора информативных признаков** применяется метод, основанный на расчёте *условной взаимной информации* [19]. Производится итеративный отбор тех признаков, которые содержат наибольшее количество информации о значении атрибута и при этом существенно отличаются от признаков, выбранных на предыдущих итерациях. Таким образом, каждый признак результирующего набора высоко информативен и слабо зависит от остальных признаков.

На этапе **обучения** производится построение модели классификации с использованием *онлайнового пассивно-агрессивного* алгоритма [13].

На этапе **классификации** в качестве входных данных используются тексты сообщений и поля профиля произвольного пользователя. Выполняется алгоритм классификация для заданного языка и атрибута. Результатом является значение атрибута выбранного пользователя.

3.3 Результаты

Для тестирования использовались наборы данных англоязычных пользователей *Twitter*, размеченные по полу (мужской/женский) и возрасту (моложе 20 лет/от 20 до 40 лет/старше 40 лет). Набор данных, размеченный по полу, включает 3 755 пользователей и 180 240 сообщений. Набор данных, размеченный по возрасту, включает 17 050 пользователей и 818 400 сообщений. Все наборы данных сбалансированы по значениям атрибутов.

Для оценки качества результатов используется точность классификации (*accuracy*). Исходный набор данных разделяется на обучающую и тестовую подвыборки. В качестве входных данных используются тексты пользователей сети *Twitter* из тестовой подвыборки исходного набора данных.

Точность классификации по полу составляет 83,3%. Использование для разметки словарей мужских и женских имён английского языка позволяет увеличить исходный набор данных более чем в 4 раза (70734 пользователя) и повысить точность классификации до 89,2%. Rao et al [16] сообщают о точности 72,33%, Al Zamal et al [18] - о точности 80,2% для идентичной задачи.

Точность классификации по возрасту составляет 71,4%.

4 Идентификация пользователей в различных социальных сетях

Одной из фундаментальных проблем проблем при использовании социальной информации о пользователе является её фрагментированность среди множества различных онлайновых социальных сетей. Каждый год появляется множество как общенаправленных, так и нишевых социальных сервисов, и для активных пользователей Интернет типично иметь несколько профилей в различных социальных сетях. Несмотря на то, что существуют попытки по обеспечению единого способа взаимодействия между различными социальными платформами (например, OpenSocial [5]), они не получили широкого применения, а новые социальные сервисы продолжают появляться. Идентификация пользователя в различных социальных сетях позволяет получить более полную картину о социальном поведении данного пользователя в

[4] https://code.google.com/p/language-detection/

[5] http://opensocial.org/

сети Интернет. Обнаружение аккаунтов, принадлежащих одному человеку, в нескольких социальных сетях, позволяет получить более полный социальный граф, что может быть полезно во многих задачах, таких как информационный поиск, интернет-реклама, рекомендательные системы и т.д.

Поскольку поиск аккаунтов пользователя в различных сетях в общем случае требует наличия актуальных данных обо всех пользователях данных сетей, целесообразно ограничить пространство поиска ближайшими соседями какого-либо пользователя, аккаунты которого в исследуемых сетях известны. Таким образом, задача идентификации пользователей в различных социальных сетях в *локальной перспективе* подразумевает сопоставление аккаунтов пользователей в рамках списков контактов некоторого центрального пользователя в различных социальных сетях. Такая задача часто возникает при работе с контактами пользователей в социальных мета-сервисах, которые, в частности, могут служить для объединения новостных потоков в поддерживаемых социальных сервисах или предоставления единой системы обмена сообщениями. Другая область, в которой возникает подобная задача, это функция автоматического объединения контактов из различных источников (телефонная книга, социальные сети, мессенджеры), распространённая в современных мобильных устройствах.

4.1 Задача

Задача идентификации пользователей заключается в поиске как можно большего числа правильно определенных пар аккаунтов (v, u) таких, что $v \in A, u \in B$, принадлежащих одному и тому же пользователю ($< A, B >$ - социальные графы). Сопоставленный аккаунт для аккаунта $v \in A$ обозначается как $\mathrm{pr}(v) \in B$ и называется *проекцией* аккаунта $v \in A$ в B, а множество всех проекций $\{\mathrm{pr}(v)\}_{v \in A}$ аккаунтов из A в B как $\mathrm{PR}(A)$. Если же для аккаунта $v \in A$ не найдено подходящей проекции, то проекция для v называется *нейтральной* и обозначается как $\mathrm{pr}(v) = \mathbf{N}$. Пример двух таких социальных графов $< A, B >$ и сопоставленных пар аккаунтов изображен на рисунке 3.

Поскольку задача идентификации рассматривается в локальной перспективе, то подразумевается, что графы A и B имеют структуру эго-сетей некоторого центрального пользователя. *Эго-сеть* вершины e представляет из себя граф, состоящий из вершины e, ближайших соседей e, а также данных о связях соседей e между собой и с другими вершинами. Такая формулировка отражает реальные ограничения использования социальных сетей, в которых для предоставлении какой-либо информации о социальных связях пользователя требуется его непосредственное разрешение.

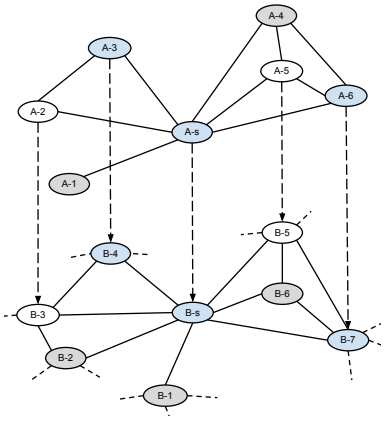

Рис. 3: Результат идентификации пользователей. Пунктирные стрелки обозначают проекции между аккаунтами. Для вершин, закрашенных синим, проекции были известны заранее, проекции для незакрашенных вершин были установлены алгоритмом, для вершин, закрашенных серым, проекции не были найдены

4.2 Метод

Большинство современных методов идентификации пользователей в различных социальных сетях ограничивается лишь анализом атрибутов профилей пользователей, поскольку они зачастую содержат информацию, помогающую идентифицировать пользователя. Общая схема работы таких методов выглядит следующим образом [14]:

1. приведение данных из полей профилей из двух социальных сетей к некоторому общему виду (например, вектору, элементами которого являются атрибуты профилей);

2. попарное применение различных способов сравнения к атрибутам профилей из анализируемых сетей;

3. подсчет результирующего показателя *похожести* между профилями и отсечение всех парных результатов, для которых этот показатель ниже некоторого порогового значения.

После этого все оставшиеся пары считаются сопоставленными между собой и принадлежащими одному пользователю.

Очевидно, что информация, содержащаяся в профилях, достаточно ненадежна, так как данные, указанные пользователем в разных социальных сетях, могут существенно отличаться, быть скрытыми из-за настроек приватности или не поддерживаться в актуальном состоянии.

Одним из способов улучшения результатов описанного подхода является привлечение дополнительных источников данных, в частности информации о социальных связях между пользователями.

Разработанный нами метод [7,8] использует социальные связи обеих рассматриваемых социальных сетей путем сравнения оригинальных списков контактов, естественным образом комбинируя их с информацией атрибутов профилей, благодаря чему лишен многих недостатков существующих методов идентификации пользователей.

Метод основывается на двух основных принципах:

1. задачи выбора проекций для связанных вершин в графе A взаимосвязаны, иначе говоря, выбор проекции для некоторой вершины зависит от значений проекций связанных с ней вершин;

2. если две вершины в графе A связаны, их проекции должны иметь наиболее высокое значение графовой близости.

В качестве функции графовой близости $0 \leq \mathsf{network\text{-}similarity}(\mathrm{pr}(v), \mathrm{pr}(u)) \leq 1$ используется модифицированный коэффициент Дайса:

$$\mathsf{network\text{-}similarity}(v,u) = \frac{2 \cdot w(L_v \cap L_u)}{w(L_v) + w(L_u)}, v, u \in B,$$

где L_v и L_u - множества вершин, связанных с v и u соответственно, а $w(L) = |L|$ - вес этих множеств.

Также предполагается, что один из графов $< A, B >$ является ненаправленным. В дальнейшем без ограничений общности таким графом считается A.

На основе графа A строится модель *условных случайных полей* [15], в которой множество наблюдаемых переменных представлено вершинами графа A: $\mathbf{X} = V(A) = \{\mathbf{x}_v, v \in A\}$, с каждой из которых ассоциирована одна скрытая переменная $\mathbf{Y} = \{\mathbf{y}_v,\ v \in A\}$, определяющая проекцию данной вершины $\mathbf{y}_v = \mathrm{pr}(v) \in B$. Скрытые переменные могут принимать в качестве значения одну из вершин графа B. Связи же наследуются из графа A: $E = E(A)$. Данная модель порождает следующее вероятностное распределение:

$$p(\mathbf{Y}|\mathbf{X}) = \exp(-E(\mathbf{Y}|\mathbf{X})),$$
$$E(\mathbf{Y}|\mathbf{X}) = \sum_{v \in V} \Phi(\mathbf{y}_v|\mathbf{x}_v) + \sum_{(v,u) \in E} \Psi(\mathbf{y}_v, \mathbf{y}_u|\mathbf{x}_v, \mathbf{x}_u),$$

где E - *функционал энергии*, моделируемый функцией *унарной энергии* Φ и функцией *бинарной энергии* Ψ. Обе энергетические функции вещественны и неотрицательны.

Унарная энергия характеризует похожесть вершины в A и его проекции в B на основании полей профилей в этих двух социальных сетях:

$$\Phi(\mathbf{y}_v|\mathbf{x}_v) = \alpha(v) \cdot (1 - \mathsf{profile\text{-}similarity}(v, \mathrm{pr}(v)))$$

Бинарная энергия отвечает за близость между проекциями вершин v и u в графе B:

$$\Psi(\mathbf{y}_v, \mathbf{y}_u|\mathbf{x}_v, \mathbf{x}_u) = 1 - \mathsf{network\text{-}similarity}(\mathrm{pr}(v), \mathrm{pr}(u))$$

Здесь $0 \leq \mathsf{profile\text{-}similarity} \leq 1$, и $\alpha(v) = \log(degree(v)) \geq 0$ - коэффициент баланса между унарной и бинарной энергией.

В качестве функции близости profile-similarity между полями профилей используется вероятность, с которой бинарный классификатор (*C4.5* с *MultiBoosting*) считает их принадлежащими одному пользователю на основании сравнения между строковыми значениями атрибутов профилей. Используемые для сравнения поля профилей *Facebook* и *Twitter* приведены в таблице 1.

Таким образом, для графов $< A, B >$ существует оптимальная конфигурация проекций:

$$\mathbf{Y}^* = \underset{Y}{\mathrm{argmin}} E(\mathbf{Y}|\mathbf{X}),$$

которая минимизирует функционал энергии, максимизируя сумму функций близости и правдоподобие модели.

Разумно допустить, что не для всех вершин необходимо выбирать проекции. В разработанном методы проекция вершины v считаются *заранее известной*, если $\mathsf{profile\text{-}similarity}(v, \mathrm{pr}(v)) \geq \Delta$. Кроме того, проекции некоторых вершин могут быть указаны явно. Использование известных проекций позволяет уменьшить объём вычислений и повысить качество результатов за счёт добавления априорной информации о модели.

Поскольку выбрать разумные фиксированные значения функций близости для *нейтральных* проекций не представляется возможным, для фильтрации неправильно выбранных проекций используется схема обучения бинарного классификатора (*C4.5* с *MultiBoosting*). Используя информацию о контексте каждой вершины в A, классификатор решает, правильно ли для неё выбрана проекция. Для этого используются следующие признаки:

1. $\mathsf{profile\text{-}similarity}(v, \mathrm{pr}(v))$;

2. средняя графовая близость к проекциям смежных вершин;

3. доля заранее известных проекций среди смежных вершин;

4. взаимная согласованность смежных вершин с заранее известными проекциями:

$$\frac{1}{n} \cdot \sum_v \frac{1}{n-1} \sum_{u \neq v} \mathsf{network\text{-}similarity}(\mathrm{pr}(v), \mathrm{pr}(u)|v, u)$$

Таблица 1: Поля профилей Facebook и Twitter, участвующие в сравнении

Поле в Facebook	**Поле в Twitter**
Имя (name)	Имя (name)
	Псевдоним (screen_name)
Веб-сайт (website)	URL

Таблица 2: Результаты экспериментов

метод	полнота	точность	F_1
взвешенная сумма	0.45	0.94	0.61
profile-similarity	0.51	**1.0**	0.69
предложенный метод	**0.80**	**1.0**	**0.89**

4.3 Результаты

Разработанный метод был протестирован на данных из социальных сетей *Facebook* и *Twitter*. 16 *центральных* пользователей, имеющих профиль в обеих сетях, предоставили доступ к своим эго-сетям, а также указали пары аккаунтов, принадлежащих одному и тому же пользователю. Для всех участников эксперимента были загружены профили их друзей (вместе со связями между ними), а также друзей их друзей. В *Twitter* профиль загружался только при наличии между пользователями взаимных связей *следования* для поддержания семантики связей *дружбы*, характерных для *Facebook*. Суммарное число профилей в *Twitter* и *Facebook* 398 и 977, а число связей 108 и 641 соответственно. Общее число сопоставленных пар пользователей - 102.

Для оценки качества результатов используется точность, полнота и F_1-мера. Исходный набор данных разделяется на обучающую и тестовую выборки. Для расчёта показателей качества применяется кросс-валидация с разбиением исходных данных на 3 непересекающихся блока. В качестве входных данных используется пара эго-сетей в *Facebook* и *Twitter* какого-либо центрального пользователя.

Для сравнения были выбраны два базовых алгоритма, основанных на расчёте похожести профилей пользователей. Первый алгоритм использует взвешенную сумму значений функций строковых близостей между полями профилей, коэффициенты для которых подбирались при помощи линейной регрессии из предположения, что между правильно сопоставленными профилями сумма близостей должна быть равна 1. Второй алгоритм использует функцию profile-similarity. Результатом работы базового алгоритма считается *максимальное паросочетание* между графом A и B с некоторым порогом близости профилей, ниже которого проекция не включалась в результаты и считалась нейтральной.

Результаты тестирования приведены в таблице 2.

5 Заключение

В работе были рассмотрены основные особенности социальных сетей как источников данных, а также некоторые задачи и методы анализа разнородных пользовательских данных из социальных сетей, связанные с определением неизвестных значений пользовательских атрибутов.

Одной из доминирующих тенденций развития социальных сетей как социокультурного феномена является более глубокое понимание особенностей социального поведения человека и, как следствие, создание новых средств для самовыражения, а также обмена информацией и опытом. Разумно ожидать дальнейшего расширения пользовательской модели и функционала социальных сетей, что приведёт к появлению новых типов данных в виде объектов и связей социального графа и, как следствие, возможности решать новые задачи, связанные с обработкой персональной информации.

Благодарности

Автор благодарит научного руководителя д.т.н. Кузнецова С.Д., а также коллег из отдела информационных систем ИСП РАН за помощь в разработке и реализации представленных в работе методов: Аванесова В.С., Андрианова И.А., Бартунова С.О., Белобородова И.Б., Бузуна Н.О., Гомзина А.Г., Ипатова С.А., Турдакова Д.Ю., Филоненко И.И.

Список литературы

[1] D. M. Boyd, N.B. Ellison. Social network sites: Definition, history, and scholarship // Journal of Computer-Mediated Communication, 2007, 13(1), article 11

[2] George Pallis, Demetrios Zeinalipour-Yazti, Marios D. Dikaiakos. Online Social Networks: Status and Trends // New Directions in Web Data Management 1, Studies in Computational Intelligence Volume 331, 2011, pp 213-234

[3] Facebook Open Graph. `https://developers.facebook.com/docs/opengraph/`

[4] Social Network Data Analytics. Editors: Charu C. Aggarwal // Springer, 2011

[5] Nazar Buzun, Anton Korshunov. Innovative Methods and Measures in Overlapping Community Detection // Proceedings of the International Workshop on Experimental Economics and Machine Learning (EEML 2012). — Leuven, Belgium, 6 May 2012

[6] Назар Бузун, Антон Коршунов. Выявление пересекающихся сообществ в социальных сетях // Доклады Всероссийской научной конференции «Анализ изображений, сетей и текстов» (АИСТ'2012). — Екатеринбург, 16-18 марта 2012 г.

[7] Sergey Bartunov, Anton Korshunov, Seung-Taek Park, Wonho Ryu, Hyungdong Lee. Joint Link-Attribute User Identity Resolution in Online Social Networks // Proceedings of The Sixth SIGKDD Workshop on Social Network Mining and Analysis (SNA-KDD'12). — Beijing, China, 12 August 2012

[8] Сергей Бартунов, Антон Коршунов. Идентификация пользователей социальных сетей в Интернет на основе социальных связей // Доклады Всероссийской научной конференции «Анализ изображений, сетей и текстов» (АИСТ'2012). — Екатеринбург, 16-18 марта 2012 г.

[9] J. Xie, B. Szymanski. Towards Linear Time Overlapping Community Detection in Social Networks // PAKDD 2012

[10] Grzegorz Malewicz, Matthew Austern, Aart Bik, James Dehnert, Ilan Horn, Naty Leiser, Grzegorz Czajkowski. Pregel: a system for large-scale graph processing // Proceedings of the 2010 ACM SIGMOD International Conference on Management of data

[11] Andrea Lancichinetti, Santo Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities // Physical Review E80, 016118 (2009)

[12] Jaewon Yang, Jure Leskovec. Defining and Evaluating Network Communities based on Ground-truth // Proceedings of 2012 IEEE International Conference on Data Mining (ICDM), 2012

[13] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, Yoram Singer. Online Passive-Aggressive Algorithms // JMLR, 7(Mar):551–585, 2006

[14] I. Veldman. Matching profiles from social network sites: similarity calculations with social network support // Master's thesis, University of Twente, 2009

[15] J. Lafferty, A. McCallum, F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data // Proc. 18th International Conf. on Machine Learning, 2001. Morgan Kaufmann. pp. 282–289

[16] Delip Rao, David Yarowsky, Abhishek Shreevats, Manaswi Gupta. Classifying Latent User Attributes in Twitter // Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents, 2010

[17] John D. Burger, John Henderson, George Kim, Guido Zarrella. Discriminating Gender on Twitter // Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011

[18] Faiyaz Al Zamal, Wendy Liu, Derek Ruths. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors // Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, 2012

[19] François Fleuret. Fast Binary Feature Selection with Conditional Mutual Information // JMLR, 5:1531–1555, 2004

[20] Aaron McDaid, Neil Hurley. Detecting Highly Overlapping Communities with Model-Based Overlapping Seed Expansion // Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM '10)

[21] Andrea Lancichinetti, Filippo Radicchi, José J. Ramasco, Santo Fortunato. Finding statistically significant communities in networks // PLoS ONE 6, e18961 (2011)

[22] Andrea Lancichinetti, Santo Fortunato1, János Kertész. Detecting the overlapping and hierarchical community structure in complex networks // New J. Phys. 11 033015, 2009

Problems and methods for attribute detection of social network users

Anton Korshunov
Institute for system programming of RAS

The increasing popularity of online social network services — the main sources of personal data of Internet users — brings unprecedented opportunities for solving research and business problems, and also for building auxiliary services and applications for social network users. Detection of latent user attributes constitutes a fundamental problem of social data analysis. In the paper, methods for solving several actual problems related to latent user attribute detection are considered: user community detection, detection of demographic user attributes, and user identity resolution in different social networks.