Анализ социальных сетей: методы и приложения

Антон Коршунов, Иван Белобородов, Назар Бузун, Валерий Аванесов,
Роман Пастухов, Кирилл Чихрадзе, Илья Козлов, Андрей Гомзин,
Иван Андрианов, Андрей Сысоев, Степан Ипатов, Илья Филоненко,
Кристина Чуприна, Денис Турдаков, Сергей Кузнецов
{korshunov, ivbel, nazar, avanesov, pastukhov, chykhradze, kozlov-ilya, gomzin,
ivan.andrianov, sysoev, ipatov, filonenko, chuprina, turdakov, kuzloc}@ispras.ru

Аннотация. В статье описаны основные компоненты разработанного в ИСП РАН стека технологий для анализа пользовательских данных из социальных сетей. Особое внимание уделяется задачам, методам и приложениям анализа сетевых (социальные связи между пользователями) и текстовых (сообщения и профили пользователей) данных: определение демографических атрибутов пользователей, поиск описаний событий в корпусах сообщений, идентификация пользователей различных сетей, поиск сообществ пользователей и измерение информационного влияния между пользователями. Кроме того, рассмотрены подходы к получению исходных данных для анализа: сбор реальных данных путём обращения к веб-интерфейсам социальных сервисов и генерация случайных социальных графов. Для каждого из разработанных инструментов описывается его функциональность, варианты использования, основные шаги используемых алгоритмов и результаты экспериментальных исследований.

Ключевые слова: социальные сети; социальные данные; пользовательские данные; социальный анализ; анализ социальных сетей; анализ содержимого; веб-сервисы; микроблоги; компьютерная лингвистика; теория графов; машинное обучение; распределённые алгоритмы и системы.

1. Введение

Анализ социальных данных стремительно набирает популярность во всём мире [1, 2] благодаря появлению в 1990-х годах онлайновых сервисов социальных сетей (SixDegrees, LiveJournal, Facebook, Twitter, YouTube и другие). С этим связан феномен социализации персональных данных: стали публично доступными факты биографии, переписка, дневники, фото-, видео-, аудиоматериалы, заметки о путешествиях и т.д. Таким образом, социальные сети являются уникальным источником данных о личной жизни и интересах реальных людей. Это открывает беспрецедентные возможности для решения исследовательских и бизнес-задач (многие из которых до этого невозможно было решать эффективно из-за недостатка данных), а также создания

вспомогательных сервисов и приложений для пользователей социальных сетей. Кроме того, этим обуславливается повышенный интерес к сбору и анализу социальных данных со стороны компаний и исследовательских центров.

Аналитическое агентство Gartner в 2012 году опубликовало отчёт под названием "Цикл ажиотажа для развивающихся технологий" [6]. Согласно отчёту, технологии "Социальная аналитика" и "Большие данные" в настоящее время находятся на т.н. "пике завышенных ожиданий". В частности, исследованиями социальных данных активно занимаются университеты Карнеги-Меллон, Стэнфорд, Оксфорд, INRIA, а также компании Facebook, Google, Yahoo!, LinkedIn и многие другие. Компании-владельцы сервисов онлайновых социальных сетей (Facebook, Twitter) активно инвестируют в разработку усовершенствованных инфраструктурных (Cassandra, Presto, FlockDB, Thrift) и алгоритмических (новые алгоритмы поиска и рекомендации пользователей, товаров и услуг) решений для обработки больших массивов пользовательских данных. Возникают и успешно развиваются коммерческие компании, предоставляющие услуги по доступу к хранилищам социальных данных (GNIP), сбору социальных данных по заданным сценариям (80legs), социальной аналитике (DataSift), а также расширению существующих платформ с помощью социальных данных (FlipTop).

Таким образом, специалисты из исследовательских центров и компаний по всему миру используют данные социальных сетей для моделирования социальных, экономических, политических и других процессов от персонального до государственного уровня с целью разработки механизмов воздействия на эти процессы, а также создания инновационных аналитических и бизнес-приложений и сервисов.

Вместе с тем, при работе с социальными данными 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. Всё это требует постоянного совершенствования алгоритмов решения различных аналитических и бизнес-задач.

Обработка социальных данных требует также разработки соответствующих алгоритмических и инфраструктурных решений, позволяющих учитывать их размерность. К примеру, база данных 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.

В статье описаны основные компоненты разработанного в ИСП РАН стека технологий для анализа пользовательских данных из социальных сетей. Раздел 2 посвящён фреймворку для сбора реальных пользовательских данных путём обращения к веб-интерфейсам социальных сервисов. В разделе 3 описан инструмент для генерации случайных социальных графов с заданными структурными свойствами. Разделы 4 и 5 посвящены методам обработки текстовых данных пользователей социальных сетей: определение демографических атрибутов путём лингвистического анализа профилей и текстов сообщений, а также поиск описаний событий в корпусах сообщений. Разделы 6-8 посвящены методам обработки сетевых данных (социальных связей между пользователями). В разделе 6 описан метод идентификации пользователей различных социальных сетей. В разделе 7 описан метод поиска сообществ пользователей. В разделе 8 описан метод измерения информационного влияния и поиска наиболее влиятельных пользователей.

2. Сбор данных из социальных сетей

Веб-интерфейсы социальных сетей являются источниками данных реального времени и предназначены для просмотра и взаимодействия со страницами социальной сети в веб-браузере либо для использования данных пользователей специализированными приложениями. Поскольку сценарии использования интерфейсов социальных сетей не предполагают автоматического сбора данных множества пользователей с целью построения социального графа, то возникает ряд проблем:

1. *приватность данных* - зачастую доступ к данным пользователей разрешён только для зарегистрированных и авторизованных участников сети, что требует поддержки эмуляции пользовательской сессии с помощью специальных учётных записей (*аккаунтов*);
2. *слабая структурированность данных* - во многих случаях программные интерфейсы (API) социальных сетей имеют ограниченный функционал, что требует поддержки получения с помощью пользовательского веб-интерфейса статических копий HTML-страниц, корректной обработки их динамической части (включая исполнение асинхронных запросов к серверу социальной сети), извлечения нужных данных с помощью алгоритма и/или шаблона и построения их структурированного представления, удобного для дальнейшей автоматической обработки;
3. *ограничения доступа и блокировки* - с целью предотвращения несанкционированного автоматического сбора данных и ограничения нагрузки на инфраструктуру сервиса социальной сети владельцы сервисов зачастую вводят явные или скрытые ограничения на допустимое количество запросов от одного пользовательского

аккаунта и/или IP-адреса в единицу времени, что требует учёта количества посылаемых запросов, а также поддержки динамической ротации используемых для сбора данных пользовательских аккаунтов и IP-адресов;
4. *размерность данных* обуславливает необходимость в параллельном методе сбора данных, а также в методах получения репрезентативной выборки пользователей социальной сети (*сэмплирование*).

В связи с постоянной необходимостью получения больших наборов данных из социальных сетей, был разработан фреймворк для сбора данных из различных интернет-сервисов.

Разработанный инструмент поддерживает скачивание данных из социальных сетей Facebook, Twitter, Hunch. Реализовано несколько способов получения репрезентативных выборок пользователей социальных сетей: сэмплирование методом обхода в ширину (breadth-first search, BFS) [1], по Метрополису-Гастингсу (Metropolis-Hastings Random Walk, MHRW) [3] и методом «лесного пожара» (Forest Fire, FF) [2]. Реализован механизм автоматического выбора учетной записи социальной сети для каждого запроса, а также поддержка прокси-соединений. Это обеспечивает устойчивость к блокировкам по IP-адресам и учетным записям. Кроме того, фреймворк поддерживает многопоточное скачивание.

Одной из ключевых особенностей разработанного фреймворка является возможность быстро реализовать новые сценарии скачивания и методы сэмплинга. В частности, на основе фреймворка реализованы алгоритмы сбора данных для задач, описанных в разделах 4-8.

Для оценки производительности фреймворка были проведены эксперименты, в которых скачивались профили пользователей социальных сетей Twitter, Facebook и Hunch. Были достигнуты следующие показатели:

- Facebook: более 500 профилей в час (один поток)
- Twitter: более 3000 профилей в час (один поток)
- Hunch: более 100 профилей в час (один поток)

3. Генерация случайных социальных графов

Несмотря на наличие средств для сбора данных из социальных сетей и большого количества доступных наборов данных, актуальной является задача создания моделей случайных социальных графов и инструментов для генерации случайных графов с заданным набором свойств. Для достоверного тестирования методов анализа социальных данных они должны быть применены к множеству наборов данных с различными свойствами. К примеру, методы поиска сообществ пользователей в социальном графе (раздел 7) могут показывать существенно различные результаты в зависимости от

размера исходного графа, средней степени вершины, коэффициента кластеризации и других структурных свойств. Сбор необходимых для достоверного тестирования реальных данных затруднён не только вследствие временных затрат на скачивание и обработку больших массивов слабоструктурированной информации, но и в силу сложности управления процессом сбора с целью получения набора данных с конкретным набором свойств.

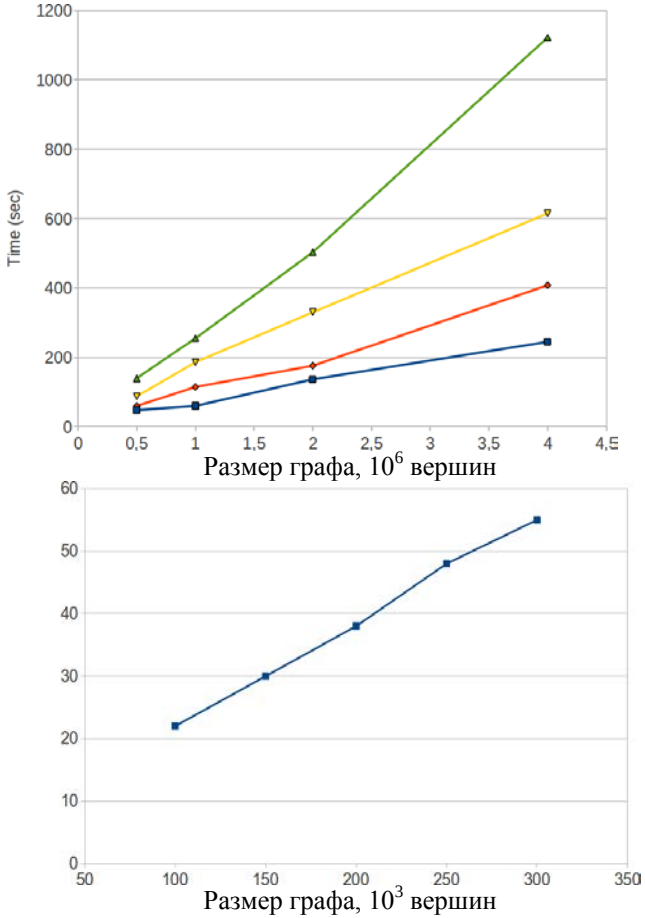

Рисунок 1. Результаты тестирования времени генерации случайных графов с заданной структурой сообществ.
Вверху: на кластерах Amazon EC2 с различным количеством рабочих узлов типа m1.large: зелёная линия – 2 узла, жёлтая линия – 4 узла, красная линия – 8 узлов, синяя линия – 16 узлов.
Внизу: на одном компьютере.

Были разработаны модель и оригинальный метод для генерации случайных графов, обладающих основными свойствами социальных сетей (распределение степеней, диаметр, коэффициент кластеризации и т.д.) и заданной структурой сообществ пользователей. Для каждого пользователя осуществляется генерация атрибутов профиля, социальных связей, сообществ и текстовых сообщений. Предложенный метод имеет распределённую реализацию на основе фреймворка Apache Spark[1], что позволяет создавать случайные графы большой размерности для тестирования производительности и точности методов анализа социальных данных. Результаты тестирования производительности распределённой версии представлены на рис. 1. Генерация графа из 1 миллиарда вершин заняла около 2 часов на кластере Amazon EC2 со 100 рабочими узлами типа *m1.large*.

4. 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.

Таким образом, актуальны методы частичной идентификации авторов сообщений по значениям их демографических атрибутов. В частности, 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. Помимо интернет-сервисов, такие демографические 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.

[1] http://spark.incubator.apache.org/

Разработанный в ИСП РАН метод определения демографических атрибутов пользователей сети Twitter по текстам их сообщений[2] обладает следующими особенностями [7,8]:

1. широкий набор поддерживаемых атрибутов: пол, возраст, семейное положение, религиозные и политические взгляды;
2. широкий набор поддерживаемых языков: русский, английский, испанский, немецкий, французский, итальянский, португальский, корейский, китайский;
3. полностью автоматический метод сбора и разметки корпусов сообщений пользователей интернета для всех поддерживаемых атрибутов и языков.

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.

На этапе **предварительной обработки текста** к текстам полученного на предыдущем этапе набора данных применяется метод определения языковой принадлежности текста. После этого данные пользователей распределяются в различные наборы данных в зависимости от языка пользователя.

Кроме того, на этом этапе 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.

Таким образом, элементом набора данных для каждого атрибута и языка является набор символьных строк, полученных из текстов сообщений и

[2] https://api.at.ispras.ru/demo/dde

профиля одного пользователя в Twitter, а также значение атрибута у данного пользователя в Facebook.

На этапе **построения признакового описания** из сообщений пользователей извлекаются лингвистические признаки. Из полученных токенов строится набор признаков в виде N-грамм размером от 1 до 3 с учётом порядка токенов. 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* [9]. 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* [10].

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.

Для тестирования использовались наборы данных англоязычных пользователей Twitter, размеченные по полу (мужской/женский), возрасту (моложе 20 лет/от 20 до 40 лет/старше 40 лет), семейному положению (состоит/не состоит в отношениях), политическим (демократ/республиканец) и религиозным (христианин/мусульманин/атеист) взглядам.

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. Результаты оценки качества представлены в табл. 1.

Таблица 1. Результаты тестирования качества результатов метода определения демографических атрибутов пользователей Twitter.

Атрибут	Исходный набор данных		Точность, %
	Количество пользователей	Количество сообщений	
Пол	17937	1147968	83,4
Возраст	10893	697152	74,2
Семейное положение	1901	202175	89,0
Политические взгляды	825	52800	76,4
Религиозные взгляды	2060	131840	85,4

Нужно отметить, что достигнутые результаты в большинстве случаев превосходят другие известные исследования. Например, Rao et al. [11] сообщают о точности 72,33%, Al Zamal et al. [12] - о точности 80,2% для задачи определения пола пользователя в Twitter.

Тестирование метода с использованием сообщений на других языках (русский, английский, испанский, немецкий, французский, итальянский, португальский, корейский, китайский) показало похожие результаты. В общем случае качество результатов во многом зависит от размера обучающего набора данных и его сбалансированности по значениям атрибутов.

5. Поиск описаний событий

Сообщения пользователей социальных сетей составляют существенную долю текстового контента современного Веба. Кроме того, социальные сети зачастую выступают в роли неформальных СМИ, где любой пользователь может опубликовать новостное сообщение о происходящих событиях (информационных поводах).

Вместе с тем, автоматическое извлечение набора сообщений о неизвестном заранее событии является нетривиальной задачей в силу следующих факторов:

- большой объём входных данных (например, пользователи Twitter публикуют несколько тысяч сообщений ежесекундно);
- большое количество нерелевантных/неинформативных сообщений;
- пользователи могут по-разному описывать одно и то же событие;
- различные события могут совпадать по времени;
- сложность разделения события и его подсобытий (например, Олимпийские игры и конкретный футбольный матч в рамках этого

первенства).

Для поиска событий в корпусах сообщений пользователей Twitter была разработана специализированная система, работа которой основывается на последовательном выполнении следующих шагов [13,14]:

- построение сигналов для каждого токена (последовательности символов) с использованием информации о частоте его появления в корпусе в различные моменты времени;
- применение вейвлетного анализа к полученным сигналам;
- удаление незначительных токенов с использованием авто-корреляции сигналов;
- построение матрицы кросс-корреляции сигналов токенов;
- поиск событий как наборов токенов путём кластеризации полученной матрицы;
- поиск сообщений, описывающих каждое событие, с помощью метода мульти-документного реферирования по документам, содержащим токены из каждого набора.

Разработанная система обладает следующими преимуществами:

- не требует данных о пользователях и доступа к внешним базам знаний;
- не требует обучения;
- возможность инкрементальной обработки при поступлении новых сообщений;
- возможность поиска событий в разных временных масштабах – час, день, неделя и т.д.

В качестве примеров найденных событий можно привести следующие наборы токенов из сообщений пользователей:

- «Выборы»: `#electionday, congratulate, decide, elect, Florida, friends, marijuana, nice, people, report, Romney`
- «Спорт»: `award, back, black, final, Friday, game, team, turn, watching, world`

Потенциальной областью применения является поиск и составление краткого реферата реакции пользователей на неизвестные или заранее определённые оффлайн- и онлайн-события. Примерами таких событий могут служить очередной выпуск телевизионного шоу, спортивные события, стихийные бедствия, политические события, запуск нового сервиса для пользователей социальной сети и т.д.

6. Идентификация пользователей различных сетей

Одной из фундаментальных проблем при использовании социальной информации о пользователе является её фрагментированность среди множества различных онлайновых социальных сетей. Каждый год появляется

множество как универсальных, 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[3]), 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 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. Подобная задача возникает также при использовании функции автоматического объединения контактов из различных источников (телефонная книга, социальные сети, мессенджеры), распространённой в современных мобильных устройствах.

Был разработан метод решения задачи идентификации пользователей различных социальных сетей, которая сводится к поиску различных вариантов виртуальных личностей одного и того же пользователя в нескольких социальных сетях[4]. На основе графической вероятностной модели условного случайного поля была разработана оригинальная модель, основанная на похожести виртуальных личностей пользователей по атрибутам их профилей и связям с другими пользователями. Разработанный метод [7,15,16] 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.

[3] http://opensocial.org/

[4] http://uir.at.ispras.ru/uir/demo/

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 (табл. 2). 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 из центральных пользователей. Для сравнения был выбран базовый алгоритм, основанный на расчёте похожести атрибутов профилей пользователей без учёта связей между пользователями.

Таблица 2. Результаты тестирования качества результатов метода идентификации пользователей в социальных сетях Facebook и Twitter.

	Полнота, %	Точность, %	F_1-мера
Разработанный метод	80,0	100,0	89,0
Базовый алгоритм	45,0	94,0	61,0

Таким образом, удалось добиться существенного улучшения точности определения различных виртуальных личностей одного пользователя в эго-сетях пользователей Facebook и Twitter по сравнению с существующими подходами. Был получен патент на изобретение RU 2469389 C1 “Способ интеграции профилей пользователей онлайновых социальных сетей” от 08.11.2011 г.

Предложенный метод может применяться для разработки приложения для мобильных устройств, которое автоматически сопоставляет списки контактов пользователя в различных социальных сетях и предоставляет удобный интерфейс для одновременного чтения всех новостей о том или ином знакомом.

7. Поиск сообществ пользователей

Естественным свойством человеческого общества является тенденция к объединению в различные сообщества. Аналогичная картина наблюдается в социальных сетях, где пользователи объединяются явно (используя средства

сети для создания групп и взаимодействия внутри них) либо неявно (устанавливая связи на основе общей или похожей деятельности, роли, социального круга, интереса или других свойств).

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 [17,18]. 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.

Информация о сообществах (модульной структуре) социальной сети на глобальном уровне находит применение в системах рекомендаций, фильтрации спама и многих других приложениях. Автоматически определённые сообщества ближайших контактов пользователя в социальной сети[5] могут применяться для оптимизации потоков входящей и исходящей информации[6] (отправить сообщение только сообществу "Коллеги", прочитать новости только от сообщества "Близкие друзья").

Был разработан метод поиска неявных сообществ пользователей социальных сетей на основе социальных связей между ними. Предложенный 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, в ходе которого поощряется объединение сообществ ближайших контактов отдельных пользователей в глобальные сообщества. Дополнительным шагом алгоритма является определение сообществ с недостаточной внутренней связанностью и разделение их на более связные подсообщества.

Разработанный метод обладает следующими особенностями:

- применимость к ориентированным и неориентированным графам;
- учёт весов на рёбрах;
- поиск как пересекающихся, так и непересекающихся сообществ;
- поиск как локальных (среди ближайших контактов пользователя), так и глобальных сообществ;
- низкая вычислительная сложность: $O(|E|)$, где $|E|$ - количество рёбер в графе;
- возможность распределённой реализации в рамках вычислительной модели *Pregel* [19].

[5] http://lcd.at.ispras.ru/demo/
[6] http://lcd.at.ispras.ru/feed/

Для оценки качества результатов разработанного метода использовался разработанный генератор случайных графов (раздел 3), способный генерировать случайные графы с заданной структурой сообществ.

Наиболее распространённым способом оценки качества результатов методов поиска сообществ пользователей является сравнение для некоторого графа двух наборов сообществ: найденного алгоритмом и *референсного*, то есть заранее заданного или известного. В качестве количественной меры для сравнения двух покрытий применялась *нормализованная взаимная информация (NMI)* [20]. Результаты представлены на рис. 2.

Для оценки производительности разработанного метода было проведено тестирование распределённой реализации на основе фреймворка Apache Spark 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 узлов.

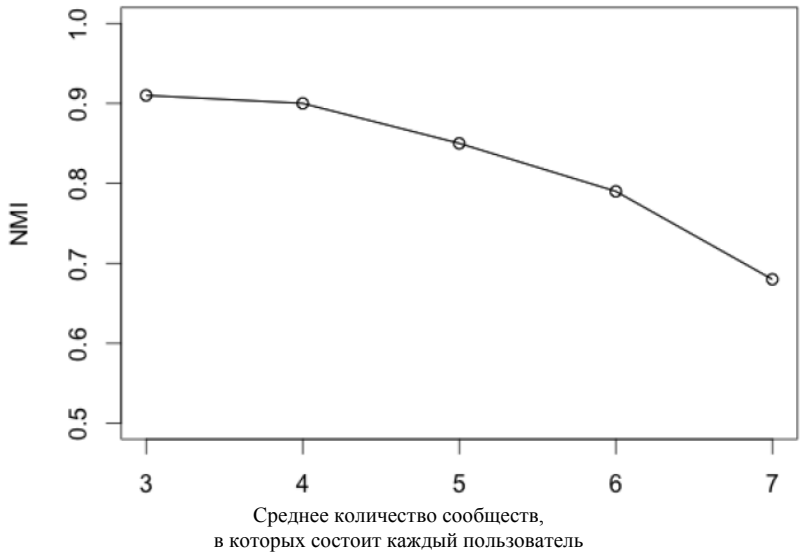

Рисунок 2. Результаты тестирования качества метода поиска глобальных сообществ пользователей.

В результате удалось добиться беспрецедентного сочетания низкой вычислительной сложности, масштабируемости и точности результатов, что позволяет применять предложенный метод для поиска сообществ пользователей к графам социальных сетей с популяцией свыше 1 миллиарда пользователей.

8. Измерение информационного влияния

Был разработан метод[7] измерения информационного влияния между пользователями в социальных сетях с ориентированными связями и преобладанием текстового содержимого (на примере Twitter). Основой метода является модель, учитывающая такие индикаторы информационного влияния, как близость интересов пользователей, количество оригинальных сообщений и цитирований, опубликованных пользователем под влиянием других пользователей, близость пользователей в социальном графе, а также факт нахождения пользователей в одних и тех же сообществах. Кроме того, разработанный метод обладает низкой вычислительной сложностью и имеет распределённую реализацию на основе фреймворка Apache Spark, позволяющую обрабатывать графы социальных сетей с популяцией свыше 1 миллиарда пользователей.

Предложенный метод может применяться в системах социальной рекомендации, а также для поиска тематических экспертов и знаменитостей, обладающих значительным информационным влиянием на конкретного пользователя либо в масштабе всей сети.

9. Заключение

Рассмотрены основные компоненты разработанного в ИСП РАН стека технологий для анализа пользовательских данных из социальных сетей. Описаны задачи, методы и приложения анализа сетевых и текстовых данных: определение демографических атрибутов пользователей, поиск описаний событий в корпусах сообщений, идентификация пользователей различных сетей, поиск сообществ пользователей и измерение информационного влияния между пользователями. Кроме того, рассмотрены подходы к получению исходных данных для анализа: сбор реальных данных путём обращения к веб-интерфейсам социальных сервисов и генерация случайных социальных графов.

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 [4,5]. 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 более эффективно решать задачи, связанные с обработкой персональной информации [21].

[7] http://imdemo.at.ispras.ru/demo

Список литературы

[1] Najork M., Wiener J. L. Breadth-first crawling yields high-quality pages // Proceedings of the 10th international conference on World Wide Web. – ACM, 2001. – С. 114-118.

[2] Leskovec J., Faloutsos C. Sampling from large graphs //Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2006. – С. 631-636.

[3] Gjoka M. et al. Practical recommendations on crawling online social networks //Selected Areas in Communications, IEEE Journal on. – 2011. – Т. 29. – №. 9. – С. 1872-1892.

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

[5] 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

[6] Key Trends to Watch in Gartner 2012 Emerging Technologies Hype Cycle. http://www.forbes.com/sites/gartnergroup/2012/09/18/key-trends-to-watch-in-gartner-2012-emerging-technologies-hype-cycle-2/

[7] Антон Коршунов. Задачи и методы определения атрибутов пользователей социальных сетей // Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL’2013

[8] Антон Коршунов, Иван Белобородов, Андрей Гомзин, Кристина Чуприна, Никита Астраханцев, Ярослав Недумов, Денис Турдаков. Определение демографических атрибутов пользователей микроблогов // Труды Института системного программирования РАН, том 25, 2013 г. DOI: 10.15514/ISPRAS-2013-25-10.

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

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

[11] 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

[12] 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

[13] Jianshu Weng, Bu-Sung Lee: Event Detection in Twitter // ICWSM 2011

[14] Zhu, Xiaojin and Goldberg, Andrew and Gael, Jurgen Van and Andrzejewski, David. Improving Diversity in Ranking using Absorbing Random Walks // HLT-NAACL, 97--104, 2007

[15] 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)

[16] 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.

[17] 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), Brussel, Belgium
[18] 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.
[19] Grzegorz Malewicz, Matthew Austern, Aart Bik, James Dehnert, Ilan Horn, Naty Leiser, Grzegorz Czajkowski. Pregel: a system for largescale graph processing // Proceedings of the 2010 ACM SIGMOD International Conference on Management of data
[20] Andrea Lancichinetti, Santo Fortunato, Janos Kertesz. Detecting the overlapping and hierarchical community structure in complex networks // New J. Phys. 11 033015, 2009
[21] Social Network Data Analytics. Editors: Charu C. Aggarwal. Springer, 2011

Social network analysis: methods and applications

Anton Korshunov, Ivan Beloborodov, Nazar Buzun, Valeriy Avanesov, Roman Pastukhov,
Kyrylo Chykhradze, Ilya Kozlov, Andrey Gomzin, Ivan Andrianov, Andrey Sysoev,
Stepan Ipatov, Ilya Filonenko, Christina Chuprina,
Denis Turdakov, Sergey Kuznetsov
{korshunov, ivbel, nazar, avanesov, pastukhov, chykhradze, kozlov-ilya, gomzin, ivan.andrianov, sysoev, ipatov, filonenko, chuprina, turdakov, kuzloc}@ispras.ru
ISP RAS, Moscow, Russia

Abstract. Social data analysis is rapidly gaining popularity worldwide due to the emergence of online social networking services in 1990-s. That also relates to the phenomenon of personal data socialization: biography facts, correspondence, diaries, photos, videos, audios, travel notes, etc became available to the public. Thereby, social networks are a unique source of data about real people's personal lives and interests. This offers an unprecedented opportunity to address research and business objectives, as well as to create auxiliary services for social network users. The paper describes the basic components of ISPRAS technology stack for social network data analysis. Particular attention is given to tasks, methods, and applications of network (social connections between users) and textual (user messages and profiles) data analysis: demographic attribute detection, event detection in messages corpora, user identity resolution, community detection, and influence measurement. Distributed implementations of certain methods using Apache Spark are also described. Collecting social data is associated with a number of well-known issues, including privacy, lack of structure, access restrictions, data size, etc. Therefore, means for input data acquisition are also considered: collecting real data through web-interfaces of social services (Facebook, Twitter, Hunch) and generating random social graphs including profile attributes, social ties, community memberships, and textual messages for each user. For each of the developed tools we describe its functionality, use cases, basic steps of the underlying algorithms, and experimental results.

Keywords: social networks; social data; user data; social analysis; social network analysis; content analysis; web-services; microblogs; computational linguistics; graph theory; machine learning; distributed algorithms and systems.

References

[1]. Najork M., Wiener J. L. Breadth-first crawling yields high-quality pages. Proceedings of the 10th international conference on World Wide Web. – ACM, 2001. – С. 114-118.

[2]. Leskovec J., Faloutsos C. Sampling from large graphs. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2006. – С. 631-636.

[3]. Gjoka M. et al. Practical recommendations on crawling online social networks. Selected Areas in Communications, IEEE Journal on. – 2011. – Т. 29. – №. 9. – С. 1872-1892.

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

[5]. 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

[6]. Key Trends to Watch in Gartner 2012 Emerging Technologies Hype Cycle. http://www.forbes.com/sites/gartnergroup/2012/09/18/key-trends-to-watch-in-gartner-2012-emerging-technologies-hype-cycle-2/

[7]. Anton Korshunov. Zadachi i metody opredeleniya atributov pol'zovatelej sotsial'nykh setej [Problems and methods for attribute detection of social network users]. Trudy 15-j Vserossijskoj nauchnoj konferentsii «EHlektronnye biblioteki: perspektivnye metody i tekhnologii, ehlektronnye kollektsii» [The Proceedings of the National Russian Research Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections»] - RCDL'2013. (in Russian)

[8]. Anton Korshunov, Ivan Beloborodov, Andrey Gomzin, Christina Chuprina, Nikita Astrakhantsev, Yaroslav Nedumod, Denis Turdakov Opredelenie demograficheskikh atributov pol'zovatelej mikroblogov [Detection of demographic attributes of microblog users]. Trudy ISP RAN [Proceedings of ISP RAS], vol. 25, 2013. DOI: 10.15514/ISPRAS-2013-25-10. (in Russian)

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

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

[11]. 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

[12]. 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

[13]. Jianshu Weng, Bu-Sung Lee: Event Detection in Twitter. ICWSM 2011

[14]. Zhu, Xiaojin and Goldberg, Andrew and Gael, Jurgen Van and Andrzejewski, David. Improving Diversity in Ranking using Absorbing Random Walks. HLT-NAACL, 97--104, 2007

[15]. 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)

[16]. Sergej Bartunov, Anton Korshunov. Identifikatsiya pol'zovatelej sotsial'nykh setej v Internet na osnove sotsial'nykh svyazej [Joint Link-Attribute User Identity Resolution In Online Social Networks]. Doklady Vserossijskoj nauchnoj konferentsii «Analiz izobrazhenij, setej i tekstov» [Analysis of Images, Social Networks, and Texsts conference] AIST'2012. Ekaterinburg, March, 16-18 2012 (in Russian)

[17]. 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), Brussel, Belgium

[18]. Nazar Buzun, Anton Korshunov. Vyyavlenie peresekayushhikhsya soobshhestv v sotsial'nykh setyakh [Identifying overlapping communities in social networks]. Doklady Vserossijskoj nauchnoj konferentsii «Analiz izobrazhenij, setej i tekstov» [Analysis of Images, Social Networks, and Texsts conference] AIST'2012. Ekaterinburg, March, 16-18 2012 (in Russian)

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

[20]. Andrea Lancichinetti, Santo Fortunato, Janos Kertesz. Detecting the overlapping and hierarchical community structure in complex networks. New J. Phys. 11 033015, 2009

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