Методика и приложения для анализа социальных данных

Аннотация. В работе были рассмотрены основные элементы разработанного в ИСП РАН стека технологий, которые применяются для детального анализа данных пользователей социальных сетей. При этом особая роль отводится изучению основных задач и методов, приложений, используемых для анализа текстовой (комментарии и сообщения, а также пользовательские профили) и сетевой (взаимосвязь между юзерами) информации. Главная задача заключается в идентификации пользователей разных Интернет-ресурсов, в выявлении их демографических характеристик (атрибутов), в поиске интересующих описаний в текстовых сообщениях, а также в отборе соответствующих сообществ и установлении информационной связи между пользователями.

В ходе исследования были изучены различные подходы, которые применяются для сбора исходной информации. Реальные сведения удается получить в основном при помощи анализа веб-интерфейсов социальных сервисов и генерирования случайных социальных графов. Каждый предложенный инструмент описывается с точки зрения его функциональности, рассматриваются возможные пути его применения. Кроме того, в статье перечисляются базовые этапы задействованных алгоритмов и отображаются результаты проведенных экспериментов.

Ключевые слова: социальные сети; анализ социальных данных; данные пользователей; анализ социальных сетей; интернет-ресурсы; блоги.

1. Введение

В связи с появлением в 90-х гг. XX века онлайновых сервисов социальных сетей (Facebook, Twitter, LiveJournal, YouTube и прочих) все большую популярность набирает анализ их социальной информации [1, 2]. Данный факт напрямую связан с явлением социализации персональных данных – в общий доступ стали попадать различные детали биографии, личной переписки, фото- и видеоматериалы, сведения о музыкальных предпочтениях, а также отзывы о путешествиях и прочая персональная информация.

В наше время социальные сети содержат подробные сведения о личной жизни и предпочтениях пользователей Интернета, что является уникальным источником для проведения исследований и постановки бизнес-задач (основную массу которых ранее невозможно было осуществить по причине отсутствия исходной информации). Теперь же появилась возможность создания наиболее востребованных приложений и прочих вспомогательных сервисов. В этой связи различные компании и исследовательские центры заинтересованы в получении и анализе пользовательской информации.

В свою очередь в 2012 г. аналитическое агентство Gartner напечатало отчёт "Цикл ажиотажа для развивающихся технологий", согласно которому методики "Большие данные" и "Социальная аналитика" на сегодняшний день находятся на "пике завышенных ожиданий" [6]. Изучением социальных сведений в наше время занимаются такие университеты, как Оксфорд, Стэнфорд, Карнеги-Меллон и INRIA. Данным вопросом занимаются следующие компании: Google, Facebook, LinkedIn, Yahoo! и прочие.

В свою очередь владельцы онлайн-ресурсов (Twitter, Facebook) вкладывают средства в создание улучшенных инфраструктурных (Presto, Cassandra, FlockDB) и алгоритмических (свежие пути поиска и рекомендации пользователей) решений для обработки больших объемов социальных данных. Следовательно, коммерческие организации, которые имеют доступ к хранилищам пользовательской информации (GNIP) и те компании, которые занимаются сбором данных по заданным сценариям (80legs), а также практикуют социальную аналитику (DataSift), на сегодняшний день активно развиваются и увеличивают свои капиталы.

Данные, полученные из социальных сетей, широко используются специалистами исследовательских центров и компаниями с целью моделирования экономических, социальных, политических процессов (характерно для персонального и государственного масштаба) и получения механизмов воздействия на данные явления. Персональная информация пользователей интернет-порталов также применяется для разработки аналитических и инновационных приложений (сервисов) для бизнеса.

При работе с социальными сведениями следует учитывать характер и качество полученного таким образом контента – существует большое количество спама и значительный процент ложных аккаунтов. В некоторых случаях возникают такие трудности, как проблемы с обеспечением приватности личной информации юзеров в процессе ее хранения и обработки, немаловажны также регулярные обновления функционала, изменение пользовательской модели. По этим причинам возникает необходимость в регулярном совершенствовании алгоритмов разрешения аналитических и бизнес-задач.

Для обработки социальных сведений требуются соответствующие инфраструктурные и алгоритмические решения, которые позволяли бы работать с данными больших размеров. Например, на сегодняшний день информационная база Facebook вмещает в себя около 1 млрд. аккаунтов и позволяет отследить свыше 100 млрд. социальных связей между пользователями. Ежедневно юзеры социальной сети размещают свыше 200 млн. фотографий и оставляют свыше 2 млрд. комментариев.

Актуальность изучаемой темы заключается в том, что в наше время не существует действенного алгоритма, который позволял бы решить все возникающие задачи – на данный момент нет возможности за приемлемое время обрабатывать сведения соответствующей размерности. Таким образом, возникает необходимость в поиске и разработке новых решений, которые позволяли бы производить распределённую обработку и хранить информацию без потери ее качества.

В данной работе рассматриваются главные составляющие разработанного в ИСП РАН стека методологий проведения анализа социальной информации пользователей социальных сетей. Во втором разделе отражен фреймворк для поиска и накопления реальных персональных сведений посредством анализа веб-интерфейсов сервисов. В статье также описан действенный инструмент, который возможно применять для генерации случайных социальных графов с определенными структурными свойствами.

Были изучены возможные методы обработки текстовой информации юзеров социальных сетей, а именно: выявление демографических признаков посредством лингвистического анализа сообщений и профилей, а также сбор описаний происшествий, отраженных в корпусах текстовых сообщений. В 6-8 разделах приведены методы обработки сетевых данных (социальных связей) пользователей:

- метод идентификации пользователей различных интернет-платформ;
- метод поиска сообществ;
- метод расчета информационного влияния и поиска более влиятельных юзеров.

2. Сбор данных, полученных из социальных сетей

Пользовательские веб-интерфейсы социальных сетей – это в первую очередь источники актуальной информации, которые также используются для ознакомления и взаимодействия с другими страницами определенной социальной сети либо для сбора пользовательских данных специализированными приложениями и сервисами. Так как сценарии применения таких интерфейсов не подразумевают сбор большого количества информации для формирования социального графа в автоматическом порядке, то по этой причине возникают следующие трудности:

1) Недостаточная структурированность информации – зачастую программные интерфейсы (API) социальных сетей характеризуются ограниченным функционалом. В

этой связи возникает необходимость в поддержке получения при помощи веб-интерфейса статических копий HTML-страниц, а также в соответствующей обработке их динамической составляющей (в том числе выполнение асинхронных запросов к серверу). При этом сбор сведений должен осуществляться при помощи определенного алгоритма, а их структурированный вид должен формироваться в удобной форме для автоматической обработки.

2) Приватность информации – обычно получить доступ к страницам могут лишь зарегистрированные и авторизованные участники соответствующей социальной сети. Это невозможно без поддержки эмуляции пользовательской сессии через специальные аккаунты.

3) Блокировка – обычно для предотвращения запрещенного автоматического сбора информации и ограничения нагрузки на инфраструктуру ресурса владельцы площадок устанавливают некие ограничения на определенное количество запросов, производимых собственником учетной записи и/или осуществляемых с одного IP-адреса в установленную единицу времени. По этой причине возникает необходимость в учете количества осуществляемых запросов и в поддержке динамической ротации задействованных при сборе сведений пользовательских аккаунтов.

4) Размерность сведений – обуславливает потребность в параллельном способе генерирования информации, а также в дополнительных способах формирования качественной выборки участников данной социальной сети (процедура сэмплирования).

Так как необходимость в формировании объемных баз данных пользователей социальных сетей возникает регулярно, для генерирования качественной информационной базы был создан фреймворк. Данный инструмент предназначен для скачивания информации из Twitter, Facebook и Hunch. Он предполагает несколько способов формирования репрезентативных выборок участников социальных сетей. К основным относят:

- сэмплирование путем обхода в ширину (BFS) [1];
- метод по Метрополису-Гастингсу (MHRW) [3];
- способ «лесного пожара» (FF) [2].

Механизм подбора аккаунта для каждого запроса осуществляется автоматически, предполагая поддержку прокси-соединений. Такой подход позволяет избежать блокировок по IP-адресам и аккаунтам. При этом возможно и многопоточное скачивание.

Главная особенность фреймворка заключается в возможности оперативно провести новые сценарии скачивания и реализовать методы сэмплинга. Так на базе фреймворка разработаны алгоритмы сбора информации для задач, рассмотренных в параграфах 4-8. Производительность данного метода была оценена на базе профилей участников социальных сетей Твиттер, Фейсбук и Ханч. В результате экспериментов были получены результаты:

- Twitter – обработано свыше 3000 аккаунтов в час (за поток);
- Facebook – свыше 500 аккаунтов в час (за поток);
- Hunch – свыше 100 аккаунтов в час (за поток).

3. Формирование случайных социальных графов

Хотя и существует достаточное количество доступных наборов сведений и разработано множество средств, предназначенных для сбора пользовательских данных, но задача создания моделей случайных социальных графов и инструментов для их генерации с указанным набором свойств остается актуальной. Для того чтобы результаты тестирования методов анализа социальной информации были достоверными, данные методы следует применять к множеству групп сведений с различными характеристиками.

Так, поиск сообществ пользователей в социальном графе может давать совершенно разные результаты, которые отличаются в зависимости от величины исходного графа,

коэффициента кластеризации, средней степени вершины и прочих характеристик. При этом генерирование необходимых для качественного эксперимента реальных сведений затрудняется по причине не только лишь количества времени, потраченного на скачивание и обработку объемной и слабоструктурированной информации, но и в связи с трудностями регулирования выборки.

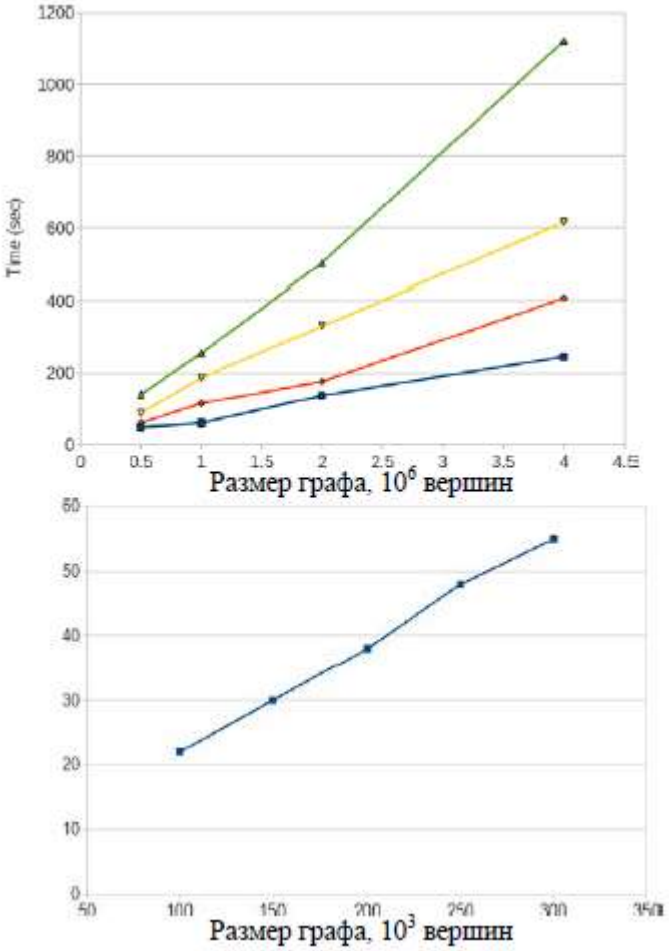

Рисунок 1 – Временные затраты при генерации случайных графов с установленной структурой сообществ[1]

Таким образом, разработана модель и уникальный метод для формирования случайных графов, которые отражают главные свойства социальных сетей (распределение степеней, коэффициент кластеризации, диаметр и прочее) и обладают установленной структурой сообществ участников социальных сетей. Относительно отдельно взятого пользователя проводится генерация характеристик профиля, его связей, выбранных сообществ и размещенных сообщений. Данный способ характеризуется также распределённой реализацией, проводимой на базе фреймворка Apache Spark[2].

Так для тестирования результативности и точности методов анализа социальной информации удается разрабатывать случайные объемные графы. Из рисунка 1 можно предположить, что на генерацию графа из 1 млрд. вершин уйдет около 2-х часов (с применением 100 узлов m1.large).

[1] На верхней части рисунка: на кластерах Amazon EC2 с разным количеством рабочих компьютеров типа m1.large (зелёная линия содержит 2 устройства, жёлтая – 4, красная – 8, синяя – 16). На нижней: в рамках одного устройства.

[2] Официальный сайт Apache Spark – http://spark.apache.org/.

4. Выявление демографических признаков участников социальных сетей

Заполняя профиль социальной сети, некоторые участники умышленно или по ошибке указывают недостоверную информацию либо вообще не заполняют отдельные поля. В этой связи получить достоверные сведения касательно их биографии, предпочтений и интересов практически невозможно. Проблема усугубляется также тем, что в таких сетях, как Twitter или YouTube, профиль пользователей содержит ограниченный набор атрибутов. При этом полученных сведений зачастую недостаточно для разрешения задач, основывающихся на персонализации результатов.

Следовательно, актуальными становятся методы частичной идентификации авторов сообщений по характеру их демографических признаков. К примеру, при интернет-маркетинге особую значимость обретает установление демографических качеств участника социальной сети для таргетированного продвижения товаров (услуг) в группах интернет-пользователей, характеризующихся одинаковыми признаками. Подобного рода демографические атрибуты применяются также в таких дисциплинах: экономика, социология, криминология, психология, управление персоналом.

Демографические признаки условно подразделяются на следующие:

- численные: возраст, размер доходов;
- категориальные: пол, семейное положение, национальность, образование и профессия, религиозные и политические взгляды.

Числовую информацию можно представить в виде отдельной категории. Так, пользователей можно распределить по нескольким возрастным группам. Созданный в ИСП РАН метод выявления демографических признаков профилей социальной сети Twitter по сообщениям пользователей характеризуется следующим образом [7,8]:

- обширным набором поддерживаемых атрибутов: возраст, пол, семейное положение, политические и религиозные мировоззрения;
- широким спектром языков: ресурс поддерживает английский, русский, испанский, французский, немецкий, китайский и прочие;
- доступностью автоматического сбора и разметки корпусов пользовательских сообщений для всех возможных атрибутов и языков.

Метод выявления демографических признаков пользователей предполагает следующие стадии:

1) Формирование исходной базы сведений – осуществляется сбор данных профилей Twitter. Для каждого участника запрашивается только лишь его профиль. Если в сети Facebook содержится ссылки на его страницу (более обширный набор атрибутов), то в дальнейшем извлекаются и сохраняются все доступные сообщения юзера в сети Twitter. Далее запрашивается и сохраняется профиль этого же пользователя в Facebook (генерируются указанные им значения атрибутов). Следовательно, составляющим набора сведений для каждого признака выступает набор символьных строк, которые собраны из текстов профиля одного пользователя в Twitter и получены из сведений этого же пользователя в Facebook.

2) Подготовительная обработка текстовой информации – к полученному на первом этапе набору сведений применяется метод определения языковой принадлежности текста. После чего информация распределяется в разные группы по признаку использованного пользователем языка. При этом предполагается также фильтрация сообщений, авторство которых не принадлежит данному участнику социальной сети (так называемые ретвиты). Таким образом, данный этап играет важную роль при повышении точности метода.

3) Создание описания атрибутов – извлекаются лингвистические атрибуты. Из собранных токенов формируется набор характеристик в виде N-грамм величиной от 1 до 3 с учётом их порядка. Отельный вид атрибутов представлен 2-мя подвидами: без учёта и с учетом регистра символов. При этом итоговый вектор характеристик для профиля содержит лишь сведения о наличии или отсутствии атрибута в его авторских текстах.

4) Выбор наиболее информативных атрибутов – используется метод, который базируется на вычислении условной взаимной информации [9]. Осуществляется итеративный отбор атрибутов, содержащих максимальное количество информации о значении признака. При этом они значительно отличаются от атрибутов, отобранных на предыдущих стадиях. Так обеспечивается высокая информативность и независимость от других признака результирующего набора.

5) Обучение – осуществляется построение модели классификации с применением онлайнового пассивно-агрессивного алгоритма [10].

6) Систематизация (классификация) – в качестве входных сведений применяются тексты сообщений и заполненные поля профиля произвольного юзера. Проводится алгоритм систематизации для установленного признака и языка. В качестве результата выступает значение признака выбранного профиля.

Все этапы, кроме начального, реализуются отдельно для каждого признака. Для эксперимента применялись наборы сведений англоязычных пользователей социальной сети Twitter, которые также были разделены по полу, возрасту (до 20 /старше 20 лет, до 40/от 40 лет), семейному положению, политическим (республиканец/демократ) и религиозным (христианин/атеист/мусульманин) мировоззрениям. Для оценивания качества результатов применяется точность классификации (accuracy). Используемый набор сведений группируется в зависимости от обучающей и тестовой выборки. Оценка качества эксперимента отображена в таблице 1.

Таблица 1 – Качество результатов тестирования метода определения демографических признаков профилей социальной сети Twitter

Признак	Исходные данные		Точность, %
	Число юзеров	Число сообщений	
Половая принадлежность	17937	1147968	83,4
Возраст	10893	697152	74,2
Семейный статус	1901	202175	89,0
Политические предпочтения	825	52800	76,4
Религиозные предпочтения	2060	131840	85,5

Следует заметить, что полученные результаты в большинстве случаев превосходят другие имеющиеся на данный момент исследования. К примеру, Rao et al. говорят о достоверности в 72,33% [11], в Al Zamal et al. – 80,2% [12] при выявлении пола участников сети Twitter.

Эксперименты, проведенные для тестирования способа с применением сообщений на других языках, показали аналогичные результаты. Можно сделать вывод, что качество результатов зачастую зависит от величины обучающего набора сведений и его сбалансированности по характеристикам признаков.

5. Сбор описаний событий

Текстовый контент современного Веба главным образом состоит из сообщений участников социальных сетей. При этом социальные сети обычно выступают в качестве неформальных средств массовой информации. Каждый пользователь может разместить новость о происходящих вокруг событиях. Автоматическое формирование набора сообщений о неизвестном изначально происшествии выступает в качестве нетривиальной задачи по следующим причинам:

- объемность входных данных – участники сети Twitter ежесекундно размещают несколько тысяч постов;
- большое количество малоинформативных сообщений;
- разнообразие точек зрения на одно и то же событие;

- несколько происшествий могут происходить в одно и то же время;

- трудности при разделении происшествия и его составляющих (к примеру, Олимпийские игры и отдельный матч по футболу).

Для облегчения поиска происшествий в корпусах сообщений участников сети Twitter была создана специализированная концепция, принцип работы которой базируется на осуществлении следующей последовательности этапов [13, 14]:

- формирование сигналов для каждой последовательности символов (токенов) на основе информации о частоте их появления;

- вейвлетный анализ полученных знаков;

- устранение несущественных токенов при помощи автокорреляции сигналов;

- построение матрицы кросс-корреляции полученных знаков;

- поиск происшествий путём кластеризации построенной матрицы;

- поиск текста, который отражает соответствующее событие при помощи мульти-документного реферирования по документам, отображающим последовательности символов из каждой группы. Такая система обладает множеством преимуществ, основными из которых являются:

• отсутствие данных участников сети и доступа к внешним базам;

• отсутствие необходимости в обучении;

• возможность поиска происшествий в любых временных рамках (час, сутки, неделя и пр.);

• возможность инкрементальной обработки при получении свежих сообщений.

Таким образом, потенциальной сферой использования данного метода является поиск и составление краткой картины реакции участников социальных сетей на любые события: выпуск определенного телевизионного шоу, результат спортивных матчей, политические события, внедрение нового ресурса для юзеров социальной сети и пр.

6. Идентификация участников различных социальных сетей

Основополагающей проблемой, с которой приходится сталкиваться при использовании социальной информации об интернет-пользователях, является фрагментированность данных. Ежегодно число универсальных и нишевых онлайн-сервисов стремительно увеличивается. Сегодня иметь несколько профилей в разных социальных сетях обычное явление. Осуществлялось множество попыток обеспечить взаимодействие различных платформ (к примеру, OpenSocial[3]), однако, на практике это не эффективно. Число социальных сервисов постоянно пополняется новыми.

Распознавание интернет-пользователя и поиск его аккаунтов в нескольких социальных сетях дает возможность сформировать наиболее точное представление о предпочтениях и поведении этого человека в сети. Социальный граф данного пользователя получается более полным. Это дает свои плоды при решении таких задач, как информационный поиск, разработка эффективной интернет-рекламы и др.

Главная задача при идентификации пользователя в разных социальных сетях заключается в сопоставлении профилей тех центральных пользователей и их списков контактов, учетные записи которых в исследуемых сетях известны. Данная ситуация зачастую случается при работе с контактами юзеров таких мета-сервисов, которые могут быть полезны для объединения информационных потоков в поддерживаемых сервисах. Подобная цель преследуется и при автоматическом обмене контактов между несколькими источниками (телефонная книга, мессенджеры или различные социальные сети). Данная функция достаточно распространена в современных устройствах.

Идентификация участников нескольких социальных сетей главным образом сводится к поиску возможных вариантов виртуальных профилей одного и того же

[3] Официальный сайт Фонда OpenSocial – http://opensocial.org/

пользователя. Графическая вероятностная модель условного случайного поля послужила основой для разработки уникального инструмента, базирующегося на схожести виртуальных личностей участников социальных сетей по социальным характеристикам их учетных записей и взаимосвязям с другими профилями сервиса [7,15,16]. Данный способ анализирует социальные связи двух исследуемых социальных сетей посредством сопоставления уникальных списков контактов, объединяя их с данными социальных атрибутов страниц. Так удается избежать основных недостатков, которые встречаются в разработанных ранее методах распознавания пользователей.

Предложенный метод был проверен на сведениях, полученных из Facebook и Twitter. Так, 16 центральных юзеров, которые имели учетные записи в этих сетях, открыли доступ к своим профилям и указали несколько пар аккаунтов, которые принадлежали одному человеку. Были также загружены профили друзей (и их друзей) всех участников эксперимента. В социальную сеть Twitter профиль удавалось загрузить лишь при наличии между участниками взаимосвязей следования для установления семантики отношений, характерных для сети Facebook. Общее количество аккаунтов в Twitter и Facebook – 398 и 977 соответственно. Что касается числа взаимосвязей, то для Твиттера их количество приравнивалось к 108, а для Фейсбука – к 641. При этом число соотнесенных пар участников – 102.

Чтобы оценить точность полученных результатов применялись показатели точности, полноты и F_1-меры. Более подробная информация отражена в таблице 2. Имеющаяся база сведений была поделена на две выборки: обучающую и тестовую. Показатели качества рассчитывались при помощи кросс-валидации с делением имеющихся данных на три непересекающиеся части. Сравнение осуществлялось с помощью базового алгоритма, который предполагает выявление похожести признаков учетных записей пользователей без определения взаимосвязей между ними.

Таблица 2 – Качество тестирования на основе метода распознавания пользователей в Facebook и Twitter

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

Проанализировав результаты тестирования, можно сделать вывод, что в результате эксперимента удалось получить улучшенные показатели точности идентификации разных виртуальных профилей одного человека в социальных сетях Facebook и Twitter в сравнении с имеющимися подходами (данный метод запатентован 08.11.2011 г. и предполагает изобретение RU 2469389 C1 "Способ интеграции профилей пользователей онлайновых социальных сетей"). Предложенный подход будет также полезен при разработке приложения для мобильных устройств, принцип которого заключается в автоматическом сопоставлении списков контактов виртуальных профилей в нескольких социальных сетях. При этом облегчается одновременное чтение новостей имеющихся друзей.

7. Отбор сообществ пользователей социальных сетей

В социальных сетях наблюдается аналогичная естественным явлениям картина – социум объединяется в сообщества разного характера посредством сетевых средств. Результатом такого взаимодействия являются группы и отношения внутри них. Взаимодействие может обретать и неявный характер – формируются связи на основе схожей или общей деятельности, совместных интересов и предпочтений.

Через поиск сообществ пользователей можно не только изучить и проанализировать социальные сети, но и исследовать модульный принцип действия сети, применив в дальнейшем полученную информацию для достижения разных целей [17, 18]. Так,

познания о характере сообществ полезны для предугадывания связей и признаков пользователей, расчёта взаимосвязи профилей в социальном графе, улучшения потока сведений в социальной сети и т.п.

В глобальном плане информация касательно модульной структуры социальной сети незаменима для сферы рекомендаций и фильтрации спама, а также в работе прочих ресурсов. На основе социальной взаимосвязи пользователей сети был получен алгоритм поиска неявных сообществ участников социальных сетей. Данный метод заключается в имитации виртуального общения между участниками и практически воссоздает инфекционный процесс.

Основа алгоритма заключается в обмене метками сообществ между вершинами согласно динамическим правилам взаимодействия. При этом поощряется создание глобальных сообществ из групп ближайших контактов определенных юзеров. Предложенный метод характеризуется такими чертами:

- может применяться как к ориентированным, так и к неориентированным графам;
- происходит отбор пересекающихся и непересекающихся групп пользователей;
- осуществляется поиск как локальных (из ближайших контактов участника социальной сети), так и глобальных групп;
- низкая расчётная сложность: $O(|E|)$, где $|E|$ – число ребер в графе;
- допускается распределённая реализация в пределах модели Pregel [19].

Чтобы оценить эффективность исследуемого метода был применен генератор случайных графов, который позволяет создавать их с установленной заранее структурой сообществ сети (параграф 3). Зачастую для оценки точности результатов поисковых методов прибегают к сравнению для конкретного графа двух наборов сообществ: отобранного предложенным алгоритмом и референсного (предварительно известного). Количественная мера определялась при помощи нормализованной взаимной информация (NMI) [20]. Результаты отображены на рисунке 2.

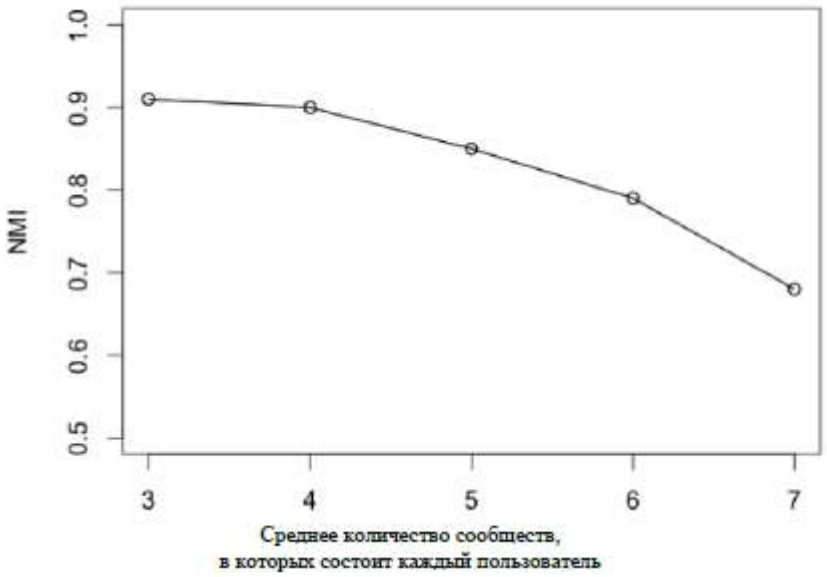

Рисунок 2 – Результаты оценки качества способа поиска глобальных сообществ участников социальной сети

Производительность предложенного метода оценивалась при помощи тестирования распределённой реализации на базе фреймворка Apache Spark на основе сервиса облачных вычислений Amazon EC2. Удалось достичь небывалого сочетания масштабируемости, низкой вычислительной сложности и достоверности. Таким образом, разработанный метод будет полезен для поиска сообществ пользователей к графам социальных сетей с численностью более 1 млрд. участников.

8. Определение степени информационного воздействия

На примере Twitter был протестирован алгоритм измерения информационного влияния между участниками социальных сетях в условиях преобладания текстового содержимого. Главная черта данного метода – модель, которая позволяет учитывать следующие сигналы информационного влияния:

- схожесть предпочтений и интересов юзеров;
- нахождение пользователей в одних и тех же сообществах;
- число оригинальных сообщений, которые участник социальной сети разместил под воздействием других пользователей.

Преимущество данного метода заключается в том, что он легко рассчитывается и характеризуется распределённой реализацией на базе фреймворка Apache Spark. Разработанный метод может использоваться при функционировании систем социальной рекомендации, а также при отборе экспертов или знаменитостей, которые имеют значительное влияние на пользователей социальной сети.

9. Заключение

В статье были изучены основные составляющие созданного в ИСП РАН стека технологий для оценки и анализа информации пользователей различных социальных сетей. Были также перечислены задачи, методы и приложения, необходимые для анализа текстовой информации сети. К основным следует отнести следующие:

- выявление демографических признаков участников социальной сети;
- поиск описаний происшествий в текстовых сообщениях;
- генерация данных о профилях одного человека в разных сетях;
- оценка информационного воздействия среди пользователей.

Среди подходов, применяемых для получения исходных сведений, были рассмотрены следующие:

- поиск достоверных сведений посредством обращения к веб-интерфейсам социальных сервисов;
- генерирование случайных социальных графов.

Основополагающей тенденцией развития социальных сетей в качестве социокультурного явления можно назвать углубленное понимание специфик социального поведения социума и, как следствие, образование новейших инструментов для самовыражения и обмена информацией [4,5]. В дальнейшем можно предположить постепенное расширение пользовательской модели и возможностей социальных сетей, что может привести к образованию новых видов сведений, представленных в качестве объектов и связей социального графа. В этой связи решение задач, которые главным образом связаны с обработкой социальных данных, будет осуществляться на более высоком уровне [21].

Список использованных источников

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. Коршунов А. 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. Коршунов А., Белобородов И., Гомзин А., Чуприна К., Астраханцев Н., Недумов Я., Турдаков Д. 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. Bartunov S., Korshunov A., 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 (SNAKDD' 12).
16. Бартунов С., Коршунов А. 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. Buzun N., Korshunov A. Innovative Methods and Measures in Overlapping Community Detection // Proceedings of the International Workshop on Experimental Economics and Machine Learning (EEML' 2012), Brussel, Belgium.
18. Бузун Н., Коршунов А. 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.