ПРОБЛЕМЫ ПОЛУЧЕНИЯ ДАННЫХ ДЛЯ АНАЛИЗА ИЗ СОЦИАЛЬНЫХ СЕТЕЙ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Семенкевич С. С.

Пилецкий И. И. – кандидат физико-математических наук, доцент

Социальные сети сегодня являются серьезной площадкой для формирования поля обсуждения любого рода вопросов, в том числе – и социальных проблем. Огромные объемы информации и их публичная доступность позволяют проводить разного рода исследования. Однако разнообразие социальных площадок и множество видов информации могут затруднить получение информации для анализа.

Извлечение полезных данных из социальной информации быстро обретает популярность во всем мире в следствии появления веб-сервисов социальных сетей (Facebook, ВКонтакте и др.). С этим связан феномен социализации персональных данных: теперь в общем доступе находятся факты биографии, сообщения, дневники, фотографии, видеоматериалы, аудиоматериалы, истории о путешествиях и т.д. Это дает невероятные возможности для решения исследовательских и бизнес-задач.

Веб-ресурсы социальных сетей считаются аккумуляторами данных в режиме реального времени и используются для просмотра и работы с сервисами социальных сетей с помощью веб-браузера или для взаимодействия с личинной информацией пользователей специальными приложениями. Так как возможности работы с сервисами социальных сетей не предоставляют средств для автоматического получение личных данных большого количества пользователей для того, что бы иметь возможность построить социальный граф, то появляется список возможных проблем. Далее перечислены ключевые проблемы получение данных и возможные способы их решения:

- Приватность данных – часто возможность воспользоваться данными пользователей доступна только для зарегистрированных и авторизованных членов сети, что заставляет поддерживать моделирование пользовательской сессии с использованием специализированных учётных записей (аккаунтов);
- Слабая структурированность данных – в большом количестве случаев открытые программные интерфейсы (API) социальных сетей обладают ограниченным функционалом, поэтому возникает необходимость прибегать к использованию веб-интерфейса для получения копий HTML-страниц, правильной обработки их изменяющейся составляющей (в том числе и выполнение асинхронных запросов к серверам социальной сети), получение необходимых данных с использованием разработанного алгоритма и/или шаблона и приведения их к структурированному виду, для выполнения будущей автоматизированной обработки;
- Ограничения доступа и блокировки – для избежание неразрешённого автоматизированного получения данных и лимитирования нагрузки на ресурсы серверов социальной сети администрация ресурсов часто применяют видимые и иногда негласные лимиты на разрешенное число запросов для одного аккаунта пользователя и/или IP-адреса в течении определенного количества времени, это вызывает необходимость подсчета числа выполненных запросов к серверам, и в том числе разработки механизма динамической ротации выполняемых запросов для получения данных пользовательских аккаунтов и IP-адресов.
- Большое количество информации приводит к необходимости разработки многопоточного приложения для получения информации, в том числе и способов извлечения репрезентативной выборки пользователей из социальной сети (сэмплирование).
- Ограничение в данных – в некоторых социальных сетях (прим. ВКонтакте) реализован механизм по которому, невозможно получить более 1000 вхождений по запросу. Это ограничение сильно затрудняет сбор данных по определенному запросу, и как следствие проведение исследований. Выходом из данной ситуации является написание специализированого ПО для обхода этих ограничений.

Так же при использовании данных из социальных сетей необходимо обратить внимание на некоторые обстоятельства, такие как возможная некорректность и плохое качество генерируемого пользователями контента (спам и ложные аккаунты, дубликаты), возможные трудности с обеспечением сохранности личной информации пользователей при хранении и анализе, в том числе регулярные изменения пользовательской модели и функциональных возможностей. Все эти факторы должны способствовать непрерывному совершенствованию алгоритмов для выполнения многочисленных аналитических и бизнес-задач.

Список использованных источников:

1. Коршунов, А. Анализ социальных сетей: методы и приложения [Электронный ресурс] / А. Коршунов, И. Белобородов, Н. Бузун и др. // Режим доступа: http://cyberleninka.ru/article/n/analiz-sotsialnyh-setey-metody-i-prilozheniya

Библиотека БГУИР