THE ANALYSYS OF MAIN SOCIAL DATA SOURCES AT RUSSIAN INTERNET

Sukhanov A.[1], Maratkanov A.[2] (Russian Federation)

АНАЛИЗ ОСНОВНЫХ ИСТОЧНИОВ СОЦИАЛЬНЫХ ДАННЫХ В РОССИЙСКОМ СЕГМЕНТЕ СЕТИ ИНТЕРНЕТ

Суханов А. А.[1], Маратканов А. С.[2] (Российская Федерация)

[1]Суханов Александр Александрович / Sukhanov Aleksandr – магистрант;
[2]Маратканов Александр Сергеевич / Maratkanov Aleksandr – магистрант,
кафедра компьютерных систем и сетей, факультет информатики и систем управления,
Московский государственный технический университет им. Н. Э. Баумана, г. Москва

Abstract: *internet users actively share data about themselves. Searching queries, the number of positive and negative responses to certain news are social data, too. In this regard, such an analysis of the data source was produced. Such data can be used for marketing purposes and for research. The sources of social data as an RSS-feeds, search analytics services and social networks are considered in the article. Also, advantages and disadvantages of each way of gathering information are considered in the article.*

Аннотация: *в настоящее время пользователи интернета активно делятся данными о себе и выкладывают личную информацию в сеть. Но кроме этого, к социальным данным можно отнести и такую информацию как поисковые запросы, количество позитивных и негативных откликов на те или иные новости. В связи с этим был проведен анализ основных источников наибольшего числа подобных данных. Подобные данные можно использовать как в маркетинговых целях, так и в исследовательских. В статье рассмотрены такие источники социальных данных, как RSS-ленты, сервисы поисковой аналитики и социальные сети.*

Keywords: *social data, search analytics, data collection, social networks.*
Ключевые слова: *социальные данные, поисковая аналитика, сбор данных, социальные сети.*

Анализ источников данных необходим в связи с тем, что количество информации в сети довольно давно перешло за ту границу, когда вычислительных возможностей уже не хватит для банальной обработки каждого источника информации. Так, согласно данным агентства InternetLiveStats, количество сайтов в сети Интернет приближается к отметке в один миллиард [1].

В связи с этим, одной из главных задач для реализации подобной системы мониторинга является выбор оптимальных источников информации, подходящих под следующие требования:

- большой охват аудитории;
- возможность просмотра обратной связи от аудитории;
- наличие каких-либо объективных метрик для анализа;

В результате был сформирован следующий список ресурсов, которые могли бы служить источниками информации:

- RSS-ленты новостных изданий;
- сервисы поисковой аналитики;
- социальные сети.

1 RSS-ленты

RSS (англ. Rich Site Summary — обогащённая сводка сайта) — семейство XML-форматов, предназначенных для описания лент новостей, анонсов статей, изменений в блогах и т. п. Информация из различных источников, представленная в формате RSS, может быть собрана, обработана и представлена пользователю в удобном для него виде специальными программами-агрегаторами или онлайн-сервисами, такими как: NewsAlloy, FeedBucket и другими [2].

Таким образом, можно прийти к выводу, что любой информационный ресурс, который заинтересован в охвате максимальной аудитории, заинтересован кроме того и в том, чтобы его RSS-лента была максимально полно и подробно заполнена. Кроме того, благодаря тому, что новостные сайты заинтересованы в максимальной доступности своей RSS-ленты, найти кнопку “RSS” программными средствами или же найти программными страницу в каталоге сайта, на которой находится RSS-лента, не составляет труда.

Разделение лент по множеству признаков и машиночитаемость текстов тоже являются одним из основным преимуществ RSS-ленты. Это означает, что единожды настроив свою систему мониторинга на какие-либо RSS-каналы, разработчик этой системы получает не только возможность доступа ко всем данным из этих каналов, но и полностью отсортированные по множеству ключевых показателей данные. То есть, снижается как время разработки всей системы (нет необходимости писать алгоритмы кластеризации новостей), так и нагрузка на уже работающую систему, из-за того, что подобные алгоритмы отличаются достаточно большой ресурсоемкостью.

Однако RSS-лента имеет ряд недостатков, к которым следует отнести то, что, во-первых, на разных информационных ресурсах одни и те же события могут иметь разную трактовку, соответственно, количество позитивных и негативных комментариев к одной и той же новости на разных порталах может

заметно отличаться.

Кроме того, многие сайты не отдают весь свой контент в RSS. Причина этого проста - новостные агентства ведут постоянную борьбу за аудиторию и из-за этого вынуждены всеми силами бороться с такими явлениями как воровство контента. Поэтому в RSS публикуют, как правило, только короткие новости, которые редко получают большое количество комментариев и просмотров.

2. Сервисы поисковой аналитики

Статистика запросов фактически представляет собой механизм, позволяющий проводить исследования, которые невозможно провести никаким другим способом. Подобная статистика является наиболее достоверным источником современного языка, в отличие от анализа поисковых результатов, которые являются приблизительными, в силу того, что информация в интернете быстро устаревает. Кроме того, запросы к поисковой системе считаются одним из наиболее репрезентативных источников живого языка.

Другими словами статистика запросов – это количество поисковых запросов пользователей по «ключевым словам» за определенный промежуток времени.

2.1. Yandex

Яндекс предоставляет доступ к своей статистике всем желающим в рамках системы по продаже рекламы Яндекс.Директ. Кроме стандартной информации о количестве запросов в месяц, а также словосочетаниях и близких темах, поисковик предоставляет возможность отсеивать результаты по регионам, городам в хронологической последовательности.

Учитывая тот факт, что Яндекс является самой популярной в Рунете поисковой системой, подобная статистика является наиболее репрезентативной при оценке положения дел в Рунете.

Следует отметить, что в статистике поисковых запросов Яндекса приводятся не только производные от введенных вами слов (в левой колонке как раз будут показаны эти самые расширенные варианты запросов с добавлением других слов), но еще дополнительно в правой колонке будут показаны ассоциативные запросы, которые набирали те же самые пользователи в Яндексе вместе с введенными вами словами за одну и ту же сессию поиска [3].

2.2. Rambler

Система статистики имеется и у Рамблера. Она менее репрезентативна в силу меньшей популярности поисковой системы, чем статистика Яндекса, но её преимуществом является более подробная информация. К примеру, сервис выдает информацию о количестве запросов не только с заглавной страницы, но также и со всех остальных. Кроме того, статистика Рамблера позволяет использовать несложный язык запросов для уточнения или, наоборот, расширения результата.

Данный механизм отличается от статистики запросов в Яндексе тем, что в ней не объединяются результаты для разных словоформ. Т.е. можно без дополнительных операторов получить статистику частотности запроса именно по словам в нужном падеже и требуемом числе.

2.3. Google

Крупнейшая в мире поисковая система *Google* также предоставляет открытый доступ к своей статистике запросов. В отличие от двух предыдущих, количественная статистика доступна в формате *csv*. Визуально статистика представляется лишь относительно – в виде графика. Отчёты выделяются особой подробностью: например, кроме обычной статистики запросов пользователей, можно посмотреть степень конкуренции рекламодателей за конкретный поисковый запрос, просмотреть историю трафика для выбранных ключевых слов; предоставляется подсказка возможно полезных минус-слов.

В особом виде статистику отображают графики *Google Trends*. Сервис позволяет вводить до 5 разных запросов, изучать и сравнивать изменение интереса к ним в мире в виде графика за прошедшие 2-3 года.

3. Социальные сети

Феномен социализации персональных данных 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.

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) [4].

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.

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 распределенную обработку и хранение данных без существенной потери качества результатов.

Литература

1. Онлайн статистика агентства LiveStats [Электронный ресурс]. Режим доступа: http://www.internetlivestats.com/ (дата обращения: 08.12.2016).
2. Веб-сайт Википедия. RSS. [Электронный ресурс]. Режим доступа: https://ru.wikipedia.org/wiki/RSS/ (дата обращения: 07.12.2016).
3. Веб-сайт SEO-Pult. Яндекс.Вордстат. [Электронный ресурс]. Режим доступа: https://seopult.ru/library/Яндекс.Вордстат/ (дата обращения: 08.12.2016).
4. Веб-сайт TopTenReviews. DataSift Review. [Electronic resource] URL: http://www.toptenreviews.com/services/internet/best-social-media-monitoring/datasift-review/ (date of access: 06.12.2016).