Исследование и анализ сообщений пользователей социальных сетей с использованием технологии BigData

И.А. Рыцарев[1,2], А.В. Куприянов[1,2], Д.В. Кирш[1,2]

[1]Институт систем обработки изображений РАН - филиал ФНИЦ «Кристаллография и фотоника» РАН, Молодогвардейская 151, Самара, Россия, 443001
[2]Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086

Аннотация. Настоящая работа посвящена проходившему в городе Самара с 15 июня по 15 июля 2018 года Чемпионату мира по футболу. В рамках работы был организован многопоточный сбор в режиме реального времени, фильтрация и обработка сообщений пользователей социальной сети Twitter в пределах города-организатора и его окрестностей в период с 15 мая по 15 августа 2018 года. Затем было проведено исследование текстов сообщений пользователей на предмет популярности тематик и построения «облака слов». Вторым исследованием стало построение диаграммы динамики количества сообщений на разных языках. В рамках работы были реализованы модули сбора, фильтрации и обработки данных с использованием технологии BigData.

1. Введение

В настоящее время социальные сети переживают бурный рост: каждый день их пользователи генерируют сотни терабайт медиа – изображений и видео. Их анализ имеет огромное значения во многих сферах бизнеса. К примеру, невозможно переоценить влияние интернет-маркетинга на продвижение товаров и услуг на рынке. Однако, для эффективного использования данных механизмов необходимо чётко понимать запросы пользователей. Источником такой информации как раз и могут служить материалы, публикуемые пользователями социальных сетей, а также формируемые в результате их обмена связи между пользователями и целые сообщества. Но в период каких-либо крупных мероприятий контингент интернет-сообществ может сильно изменяться. В данной работе проводится сравнение между потоком сообщений до Чемпионата Мира по футболу, во время и после него.

2. Сбор данных с социальных сетей

Источником данных для исследования была выбрана социальная сеть Twitter. Это было сделано по следующим причинам:

- сеть предоставляет открытый доступ к своим данным (нет ограничения на доступ к данным сервера);
- является второй по популярности социальной сетью (после Facebook, который не предоставляет открытый доступ к своим данным) среди пользователей во всем мире;
- Twitter не является предметной сетью, а значит, отражает общественное мнение более широкого круга пользователей [4].

Сбор данных социальной сети Twitter может осуществляться посредством программных продуктов Apache Ambari и Flume, подробнее данный метод описан в работе[5]. Однако для сбора данных с применением ряда фильтров, зачастую, удобнее разработать свой программный продукт с использованием стандартных библиотек (twitter4j, tweepy и т.п.) [6].

В рамках данного исследования был разработан программный комплекс на языке программирования Python, содержащий модуль авторизации, модуль сбора данных и модуль фильтрации. Данный программный комплекс позволяет собирать данные по геолокации, по ключевым словам, по пользователю. У социальной сети Twitter есть ограничение в виде лимита сообщений который может получить клиент при мониторинге в режиме реального времени. Согласно документации этот порог равен 60 сообщениям в секунду (это примерно 1% от средней скорости твитов). Для исключения перебоев в работе программного комплекса, а так же для минимизации пропусков сообщений была настроена сеть компьютеров расположенных в различных городах и были привлечены облачные сервисы. В каждый экземпляр были внедрено множество уникальных ключей авторизации. Разработанный программный комплекс работает в режиме real-time мониторинга, а также может делать запросы на получение расположенной на серверах информации.

Параметрами фильтрации по геолокации стали координаты города Самара (город принимающий ЧМ по футболу) в виде расширенного геобокса (48.9700523344,52.7652295668,50.7251182524,53.6648329274), который включает в себя не только город Самара, но и город Тольятти (тренировочная база футболистов, город в котором проживали приезжающие туристы), аэропорт Курумоч и ближайшие населенные пункты вблизи г. Самара.

За время работы распределенной сети экземпляров программного комплекса было собрано свыше 1 200 000 сообщений пользователей.

3. Анализ собранных данных с использованием технологии BigData

Слияние собранных данных, их обработка и анализ с использованием традиционных подходов требует огромных вычислительных ресурсов и занимает длительное время. В связи с этим, было принято решение воспользоваться технологией BigData и вычислительным кластером для обработки данных сверхбольшого объема, имеющимся в наличии у Самарского Университета.

В первую очередь был необходимо произвести слияние собранных данных. Для решения этой задачи был реализован модуль слияния данных с использованием технологии MapReduce. В результате работы модуля мы получили более 170000 уникальных сообщений пользователей.

Вторая задача заключается в первичной обработке данных. Потоковые данные, полученные из социальных сетей, содержат в себе множество служебной информации. Для дальнейшего анализа важны лишь те данные, которые представляют интерес, поэтому необходимо отделить служебную информацию от нужной. Для этого был реализован модуль обработки json-response. Данный модуль с помощью технологии MapReduce производит структуризацию путем компоновки и исключения служебных и не представляющих практический интерес данных.

Третьей задачей стал непосредственно анализ собранных данных. Первым исследованием стало построение «облака тэгов» для каждого из трех месяцев отдельно. Результаты исследования вы можете видеть на рисунках 1, 2 и 3 соответственно.

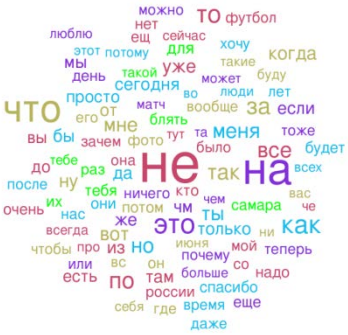

Рисунок 1. «Облако тэгов» за период 15.05-14.06 2018 года.

Рисунок 2. «Облако тэгов» за период 15.06-14.07 2018 года.

Рисунок 3. «Облако тэгов» за период 15.07-14.08 2018 года.

Как видно из рисунков 1,2 и 3 наполнение «облаков» кардинально изменилось с началом ЧМ по футболу в Самарской области.

С учетом результатов предыдущего исследования следующим исследованием было решено посмотреть динамику изменения количества сообщений на разных языках. Анализ языка написания сообщения проводился на основе данных представленных социальной сетью Twitter в json-response. Периодом анализа был выбран промежуток в 7 дней. Результаты вы можете видеть на рисунке 4.

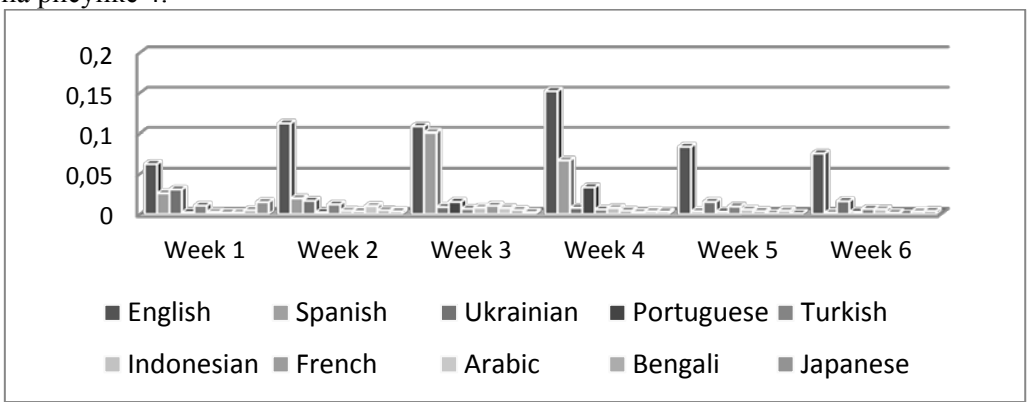

Рисунок 4. Распределение по языкам сообщений за период 11.06-22.07 2018 года.

Как видно из рисунка 4 количество сообщений на отличных от русского языках изменялось в соответствии с матчами, проходящими в городе Самара. Оно начало увеличиваться за неделю до начала турнира, затем уровень количества сообщений сохранялся в течение всего турнира и затем упало до значений близких к нулю в связи с отъездом делегаций.

4. Заключение

В данной работе было проведено исследование активности пользователей социальной сети Twitter Самарской области, а так же активности гостей Чемпионата мира по футболу 2018 приехавших поддержать сборные в город Самара. Исследование показало, что в Самарской

области Twitter не особо популярная социальная сеть и поток гостей из-за рубежа кардинально меняет тематику общения пользователей в рамках Самарской области. Из этого следует, что при анализе данных социальных сетей в период каких-либо крупных мероприятий необходимо учитывать, что тематики сообщений пользователей могут начать кардинально отличаться от данных статистики, которая была собрана до мероприятия. В таком случае необходимо применять методы реактивного анализа потока данных (особенно в даты начала мероприятия).

5. Литература

[1] Rytsarev, I.A. Clustering of social media content with the use of BigData technology / I.A. Rytsarev, A.V. Kupriyanov, D.V. Kirsh, K.S. Liseckiy // Journal of Physics: Conference Series. – 2018. – Vol. 1096(1). – P. 012085.
[2] Blagov, A. Big Data Instruments for Social Media Analysis / A. Blagov, I. Rytcarev, K. Strelkov, M. Khotilin // Proceedings of the 5th International Workshop on Computer Science and Engineering, 2015. – P. 179-184.
[3] Rytsarev, I. Creating the Model of the Activity of Social Network Twitter Users / I. Rytsarev, A. Blagov // Journal of Telecommunication, Electronic and Computer Engineering (JTEC). – 2017. – Vol. 9(1-3). – P. 27-30.
[4] Rytsarev, I.A. Development and research of algorithms for clustering data of super-large volume / I.A. Rytsarev, A.V. Blagov // CEUR Workshop Proceedings. – 2017. – Vol. 1903. – P. 80-83.

Благодарности

Работа выполнена при частичной поддержке Федерального агентства научных организаций (соглашение № 007-ГЗ/Ч3363/26); Министерства образования и науки РФ в рамках реализации мероприятий Программы повышения конкурентоспособности Самарского Университета среди ведущих мировых научно-образовательных центров на 2013–2020 годы; грантов РФФИ № 16-41-630761, № 17-01-00972, № 18-37-00418; в рамках госзадания по теме № 0026-2018-0102 "Оптоинформационные технологии получения и обработки гиперспектральных данных".

Research and analysis of messages of users of social networks using BigData technology

I.A. Rytsarev[1,2], A.V. Kupriyanov[1,2], D.V. Kirsh[1,2]

[1]Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardejskaya street 151, Samara, Russia, 443001
[2]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086

Abstract. This paper is dedicated to the World Cup held in the city of Samara from June 15 to July 15, 2018. As part of the work, a multithreaded collection in real time was organized, filtering and processing messages from users of the social network Twitter within the host city and its surroundings from May 15 to August 15, 2018. Then, a study was conducted of the texts of user messages on the subject of the popularity of topics and the construction of a “word cloud”. The second study was the construction of a diagram of the dynamics of the number of messages in different languages. As part of the work, modules for collecting, filtering and processing data using BigData technology were implemented.