ПОСТРОЕНИЕ МОДЕЛЕЙ АКТИВНОСТИ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНЫХ СЕТЕЙ

И.А. Рыцарев, А.В. Благов

Самарский государственный аэрокосмический университет им. академика С.П. Королёва
(национально исследовательский университет)

В настоящее время одним из самых перспективных направлений для исследований в различных областях является обработка и анализ данных сверхбольшого объема (Big Data). В данной статье рассматриваются вопросы сбора обработки и анализа данных социальных сетей. Предлагается модель активности пользователей социальных сетей.

Введение

Существует много серий подходов, инструментов и методов обработки структурированных и неструктурированных данных сверхбольшого объема. В эту серию включают средства массово-параллельной обработки неопределённо структурированных данных, прежде всего, решениями категории NoSQL, алгоритмами MapReduce, программными каркасами и библиотеками проекта Hadoop. Для исследования были выбраны именно эти средства сбора и анализа.

Понятие больших данных подразумевает работу с информацией огромного объема и разнообразного состава, весьма часто обновляемой и находящейся в разных источниках в целях увеличения эффективности работы и создания новых.

На данный момент социальные сети находятся на пике популярности, уже сейчас миллионы пользователей используют Facebook и Twitter. Многим компаниям необходимо анализировать данные, полученные из социальных сетей, для оценки отношения пользователей к своим продуктам) [1]. Кроме этого анализ этой области используется в решении вопросов безопасности [2]. Собрав и структурировав текстовые данные из социальной сети, можно проанализировать отношение пользователей к какому-либо выбранному вопросу. Также с помощью анализа можно получить распределение данных по странам, что позволяет оценить популярность выбранной тематики в конкретных локациях.

Сбор данных социальных сетей

Алгоритм работы с данными социальных сетей определяется по следующей схеме:

Сбор данных –>обработка данных –> анализ данных

В настоящее время существует ряд инструментальных средств и решений для сбора и обработки текстовых данных социальных сетей.

Для сбора данных в работе был использовано решение Apache Ambari, данный программный продукт был установлен и сконфигурирован на кластере лаборатории по обработке данных сверхбольшого объема СГАУ. Это позволило осуществлять непрерывный параллельный потоковый сбор данных в течение большого промежутка времени и в больших количествах.

Для проведения эксперимента было необходимо собрать набор данных для обработки. Поэтому была реализована настройка инструмента Flume на сбор абсолютно всех твиттов. Для этого в конфигурационном файле кластерной машины в поле Keywords были указаны наиболее часто употребляемые артикли, частицы, предлоги, цифры и знаки препинания. Далее была произведена настройка хранилища (HDFS) и запуск сбора данных. Срок, в течение которого осуществлялся сбор данных составил ровно семь дней (неделю).

Следующим шагом является этап превращения неструктурированных данных в структурируемые

Обработка неструктурированных текстовых данных социальных сетей

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.

С помощью технологии MapReduce [3] была произведена структуризация путем компоновки и исключения служебных и не представляющих практический интерес даанных. Затем при помощи разработанного программного комплекса была получена конкретная информация, которая была применена для дальнейшего анализа и построения математической модели. Весь этап обработки неструктурированных данных показан на рисунке 1.

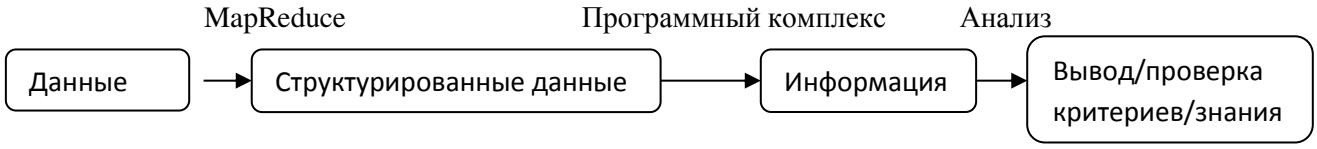

Рисунок 1 – Схема обработки неструктурированных данных

В MapReduce — это фреймворк для вычисления некоторых наборов распределенных задач с использованием большого количества компьютеров (называемых «нодами»), образующих кластер) [3].

Работа MapReduce состоит из двух шагов: Map и Reduce.

На Map-шаге происходит предварительная обработка входных данных. Для этого один из компьютеров (называемый главным узлом — master node) получает входные данные задачи, разделяет их на части и передает другим компьютерам (рабочим узлам — worker node) для предварительной обработки. Название данный шаг получил от одноименнойфункции высшего порядка.

На Reduce-шаге происходит свёртка предварительно обработанных данных. Главный узел получает ответы от рабочих узлов и на их основе формирует результат — решение задачи, которая изначально формулировалась.

В рамках исследования кластере был развернут и настроен инструмент Hortonworks Sandbox. Затем был написан SQL-запрос который отбрасывал всю «системную» информацию, оставляя только «полезные» поля.

Обработка неструктурированных данных заняла 2 дня и в результате на выходе мы получили 60 Гб структурированных «полезных» данных из 400 Гб неструктурированных.

Для извлечения из структурированных данных необходимой информации был разработанн (на языке высокого уровня Java) программный комплекс, который при обработке «выбирал» из структурированных данных необходимые поля: время создания, язык, текст, временную зону.

Общее количество K_j твиттов для каждой L локации (страны) равно:

$$K_L = \sum_i (k_i \in L), \tag{1}$$

где k_i – каждый следующий твитт из обрабатываемого потока

Частота употребления $Count(w)$ каждого уникального слова w определяется из общего множества S текстовых данных:

$$Count(w) = \sum_i (w_i \in S). \tag{2}$$

Настроение каждого твитта $sp(w,d)$ определяется из словаря - d, в котором прописано настроение (отношение):

$$sp(w,d) = \begin{cases} 0, & \text{если } w \text{ имеет негативный окрас,} \\ 1, & \text{если } w \text{ имеет нейтральный окрас,} \\ 2, & \text{если } w \text{ имеет положительный окрас.} \end{cases} \tag{3}$$

В результате доработки с помошью программного комплекса была собрана следующая информация. Самыми популярными словами оказались: «people», «love»,

«time», «life» и «twitter». Полученный результат был сверен и проанализирован на соответствие и с результатами работы [4], проведенной исследователями из Греции в 2014 году. Облако тегов, полученное в результате работы программного комплекса показано на рисунке 2.

Рисунок 2 – Облако самых часто встречающихся слов

По каждому из этих пяти слов, при помощи инструмента Flume, был произведен сбор данных с социальной сети Twitter. Эти данные (так же как и в первый раз) были обработаны при помощи Hortonworks Sandbox. Затем, при помощи разработанного программного комплекса, было определено количество твиттов, в которых употребляются данные слова.

Анализ обработанных данных и построение активностей пользователей

Полученная после обработки информация была импортирована в Excel и было выведено временное X_t распределение употребления твитов (по каждому часу от 0.00 до 23.59), содержащих наиболее популярные слова.

Далее для анализа относительного употребления сообщений той или иной тематики во времени была произведена нормировка каждой полседовательности по следующей формуле:

$$X_{tn} = \frac{X_t}{\sum_{k=0}^{23} X_k} \quad (4)$$

График распределения данной последовательности изображен на рисуноке 3.

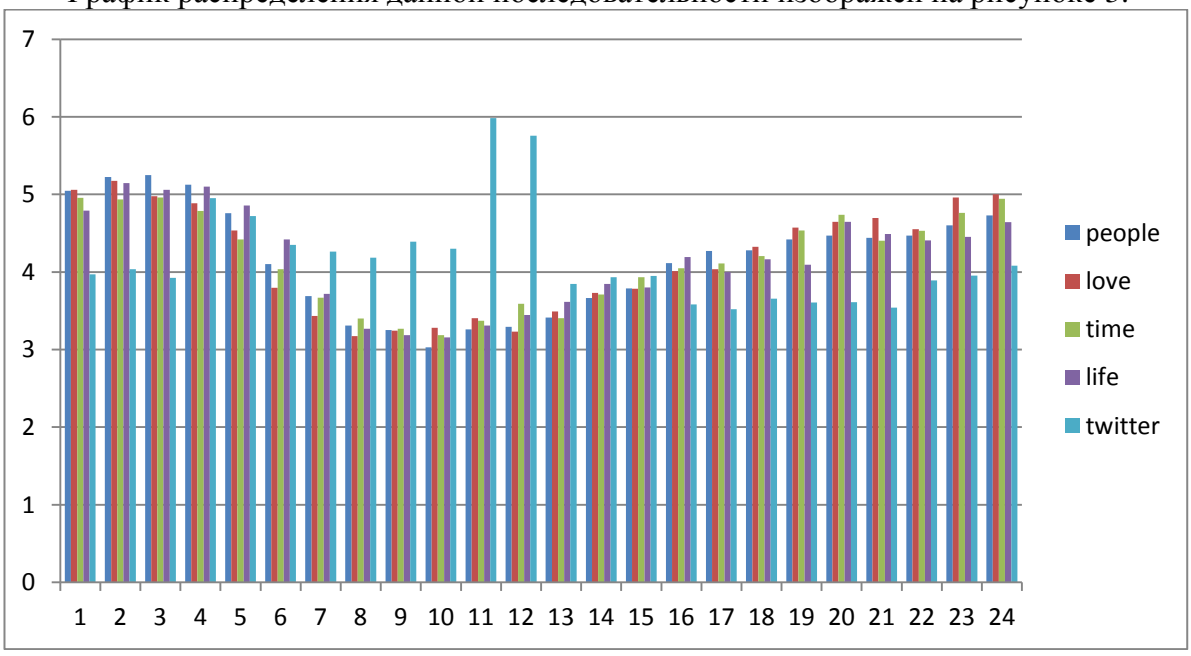

Рисунок 3 – нормированное временное распределение употребления твиттов, содержащих наиболее популярные слова

Построим модель активности пользователей в сети твиттер, основываясь на данные распределения.

Для того чтобы построить модель необходимо определиться с последовательностью, которую можно взять за основу. На рисунке 3 видно, что последовательности “People”, “Love”, “Time” и “Life” примерно одинаковы. Поэтому было взято среднее значение этих четырех последовательностей за основу.

Следующий шаг заключается в аппроксимации проследовательности, для того, чтобы на основе полученной аналитической фукнкции построить математическую модель, описывающую активность пользователей сети twitter.

Аппроксимация была произведена полиномиальной функцией со степенным базисом при использовании матрицы Грамма и метода Гаусса [5].

Таким образом полученная функция, аппроксимирующая временную последовательность (4) имеет следующий вид:

$$y = -0{,}000001x^6 + 0{,}000068x^5 - 0{,}002758x^4 + 0{,}052427x^3 - 0{,}455075x^2 + 1{,}379706x + 3{,}917158. \quad (5)$$

А модель описания активности пользователей социальных сетей:

$$\begin{cases} X(t) = 0, & t \in (-\infty; 0) \\ X(t) = \mathrm{y}, & \mathrm{t} \in [0; 24) \\ X(t) = 0, & \mathrm{t} \in [24; \infty) \end{cases}, \quad (6)$$

где y определяется уровнением 5.

Для проверки адекватности модели были определены значения коэфициента корреляции Пирсона (таблица 1) функции (5) с каждой из временных последовательностей (4). Кроме этого модель была подвергнута проверке с помощью критерия Колмогорова [6]. Полученные результаты приведены в таблице 2.

Таблица 1. Коэффициенты корреляции последовательностей и построенной моделью по критерию Пирсона

ключевое слово последовательности	коэффициент корреляции Пирсон
people	0,979411367
love	0,96626115
time	0,982908869
life	0,937562815
twitter	0,388004818

Таблица 2. Коэффициенты корреляции последовательностей и построенной моделью по критерию согласия Колмогорова

ключевое слово последовательности	λ	Уровень значимости (по таблице Колмогорова)
people	0,406016867	0,996
love	0,846007356	0,4806
time	0,480040795	0,9753
life	1,334817592	0,0582
twitter	7,428028772	0

Гипотеза о том, что модель описывает последовательности сформированные наиболее распространенными словами: people, love, time, life – принимается с высоким уровнем значимости. Последовательность, образованная потоком данных по слову twitter – не подлежит описанию полученной моделью (6) в силу своей специфичности.

Полученные результаты говорят о том, что генерация массовых сообщений во всем мире может иметь определенную временную зависимость,

Заключение

Обработка и анализ данных социальных сетей позволят не только собрать определенную статистику, но и установить ряд зависимостей. Основанные на этих зависимостях модели могут служить для решения задач социологии, экономики и безопасности.

Литература

1. Tan, W., Blake, M. B., Saleh, I., & Dustdar, S. Social-network-sourced big data analytics //IEEE Internet Computing. 2013. №. 5. Pp. 62-69..
2. Васильков А. Как «большие данные» помогают улучшить безопасность [Электронный ресурс] // Компьютерра: сетевой журн. 2014. URL: http://www.computerra.ru/108760/security-n-big-data/ (дата обращения: 24.04.2015).
3. Dean J., Ghemawat S. MapReduce: simplified data processing on large clusters //Communications of the ACM. 2008. Т. 51. №. 1. Pp. 107-113.
4. Semertzidis K., Pitoura E., Tsaparas P. How people describe themselves on Twitter //Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks. ACM, 2013. С. 25-30.
5. Самарский А.А., Гулин А.В. Численные методы М.: Наука, 1989.
6. Критерий согласия Колмогорова [Электронный ресурс] // Академик. Математическая энциклопедия URL: http://dic.academic.ru/dic.nsf/enc_mathematics/2279/КОЛМОГОРОВА (дата обращения: 26.05.2015).