СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПОТОКОВ ТЕКСТОВЫХ ДАННЫХ В СОЦИАЛЬНЫХ СЕТЯХ

М.И. Хотилин, 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, а также задача определения самых популярных наук во всем мире среди пользователей социальных сетей.

Введение

В настоящее время одним из наиболее активно развивающихся направлений в информационных технологиях являются, так называемые, «большие данные» или Big data. Big data являет собой общее понятие для столь колоссальных объемов данных, что традиционные методы и программные средства для их обработки являются неприемлемыми. Термин «большие данные», в информационных технологиях, подразумевает под собой серию подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объёмов.

За последнее десятилетие социальные сети стали играть огромную роль, будучи с одной стороны предметом социализации людей, а с другой наиболее мощным и доступным политическим, идеологическим и экономическим инструментом [1-2]. 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 – сообщения, в которых так или иначе упоминается о науке и научных исследования, открытиях. Поэтому была реализована настройка инструмента Flume на сбор абсолютно всех твиттов. Для этого в конфигурационном файле кластерной машины в поле Keywords были указаны слова «science», «Science», «SCIENCE» и множествоенное число данных слов. 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.

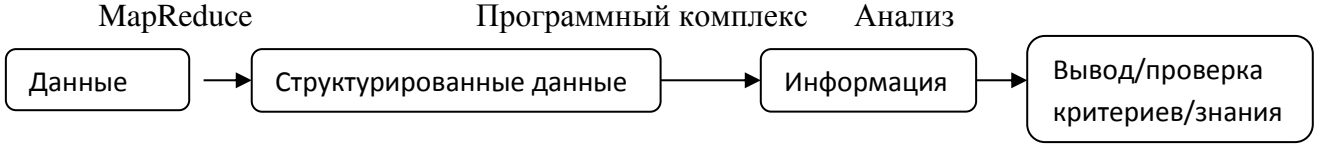

Рисунок 1 – Схема обработки неструктурированных данных

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.

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):

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

Обработка неструктурированных данных заняла приблизительно сутки. После обработки данных было получено 920 Мб интересующей структурированной информации.

Исключив из данного файла все служебные части речи и служебные символы, был получен следующий результат - оставшимися наиболее употребляемыми словами оказались:

- NASA;
- Space;
- Rocket;
- Asteroid;
- Computer;
- Data;
- Information;
- Math;
- Medicine;
- Physics.

Данную совокупность слов можно условно разделить на три кластера: space, computer science и fundamental sciences (math, medicine, physics).

По кластеру space – Sp, K_A количество твиттов в общем множестве текстовых данных S:

$$K_{Sp} = \frac{\sum_i (S_i \in Sp)}{S}. \quad (4)$$

По кластеру computer science – Cs, K_{Cs} количество твиттов в общем множестве текстовых данных S:

$$K_{Cs} = \frac{\sum_i (S_i \in Cs)}{S}. \quad (5)$$

По кластеру fundamental science (math, medicine, physics) – F, K_F количество твиттов в общем множестве текстовых данных S:

$$K_F = \frac{\sum_i (S_i \in F)}{S}. \quad (6)$$

По каждому из этих трех кластеров, при помощи инструмента Flume [4], 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 определение взаимосвязей между различными потоками текстовых данных

Полученная после обработки информация была импортирована в Excel и были выведены распределения употребления твиттов по странам мира, содержащих слова, входящие в категорию «science», в кластер «space», в кластер «computer science» и в кластер «fundamental sciences». В силу больших различий между странами мира было выделено три группы:

- Европа, США, Канада и Австралия;
- Россия и Азия;
- Южная Америка и Африка.

По каждой группе стран по каждому из выделеных кластеров был проведен анализ о соответствии частоты употреблений сообщений о науке и о научных направлениях того или иного кластера. Тем самым был сделан вывод о том, в какой из той или иной группы стран общество, говоря о науке, прежде всего упоминает о научных исследованиях из того или иного кластера.

Для большей строгости эксперимента последовательности распределений твиттов по группам стран по разным кластерам были сравнены с последовательностью распределений твиттов о науке по тем же группам стран, был произведен подсчет значений коэфициента корреляции Пирсона, а также была осуществлена проверка с помощью критерия Колмогорова-Смирнова [5].

Полученные результаты говорят о том, что люди по всему миру, упоминая в социальных стетях о науке, прежде всего имеют ввиду все что связано с космосом, чуть

реже с информационными технологиями и еще реже с математикой, медициной и физикой. Однако для разных групп стран существуют свои особенности.

Заключение

Обработка и анализ данных социальных сетей позволят не только собрать определенную статистику, но и установить ряд зависимостей, определить наиболее популярные научные направления в мире. Полученные на основании подобных результатов выводы могут служить для решения задач социологии, образования, экономики и т.д.

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. Apache Flume [Электронный ресурс] // The Apache Software Foundation URL: https://flume.apache.org/ (дата обращения: 12.05.2015).
5. Критерий согласия Колмогорова [Электронный ресурс] // Академик. Математическая энциклопедия URL: http://dic.academic.ru/dic.nsf/enc_mathematics/2279/КОЛМОГОРОВА (дата обращения: 26.05.2015).