International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

User Intent Analysis through Twitter using Big Data Tools

Shantanu Chatterjee[1], Rajat Agarwal[2], Rounak Saraogi[3], Biswaraj Sen[4]
[1, 2, 3, 4]Department of Computer Science & Engineering, Sikkim Manipal University, Sikkim, India-737136

Abstract**—** *In the present era of technological advancements, social media has proven to be a vital tool in gaining an insight about popular opinion. Marketers, product designers and investors have not been far behind in the race for making their product or service excel. This project is aimed at analyzing the intent of the user regarding any topic of his interest by using Twitter as the medium. The output hence obtained can prove very beneficial for the purpose of gaining an insight into popular opinion regarding their product among the competition.*

Keywords**—** *Automation: Big Data: Business Insight: Hadoop: Hive: Flume: Ranking: Twitter.*

I. INTRODUCTION

Big Data represents the large and rapidly growing volume of information that is mostly untapped by existing analytical applications and data warehousing systems. The concept gained momentum in the early 2000s when industry analyst Doug Laney articulated the now-mainstream definition of big data as the '3 Vs' – 'Volume', 'Velocity', 'Variety'. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. . The tools included in Apache Hadoop framework, i.e. HDFS, Apache Flume and Apache Hive, will be used for the retrieval, storage and processing of the data being fetched.

II. PROBLEM DEFINITION

People are often confused before purchasing any new product or service in the market. Similarly, marketing professionals need more and more insight into the likes and dislikes of the people, in order to boost the sales of their products or services. There can be many such instances information related to public opinion is required.

At present, there is no system prevalent which can rate a user-defined set of keywords. There are some systems owned by e-commerce websites which rate products but only based on their own databases, but they fail to analyze Twitter and yield the result based on the data generated through it. Hence, this system fills the void and can prove to be very useful in such scenarios, where the user wants to see which would be the most popular keyword in his intended topic of interest. The difference lies in the input which is being considered, which in this case is real-time unstructured data generated by Twitter.

III. SOLUTION STRATEGY

The system will solve the problem discussed earlier by the help of Hadoop tools which have the ability to work with unstructured, large volumes of data. The aforementioned Hadoop tools which will be used for the analysis are Apache Flume and Apache Hive. Apache Hive is a data warehouse software which facilitates reading, writing, and managing large datasets residing in distributed storage using SQL.
Structure can be projected onto data already in storage using serialiser-deserialiser operation (SerDe). Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability and recovery mechanisms. It uses a simple extensible data model that allows online analytic application.
The proposed solution strategy can be concisely summarized as follows.

- To fetch data through Twitter, a custom source will be created by configuring the Apache Flume tool.
- The data will be continuously stored locally in the Hadoop Distributed File System.
- The data fetched in the previous step will be used for processing thereon using Apache Hive with the help of an SQL-like query.

Because it is a scaled down system, all these activities will be carried out on a single machine with commodity hardware.

IV. OBJECTIVES

This system can prove to be useful in the area of Social Media Analytics. Using the results of the system, better business decisions can be made in the process of marketing a product or service.
The output of the system is in two parts.

- The rating of the 10 keywords based on popularity across Twitter.
- The list of most influential usernames in Twitter pertaining to the intent of the user.

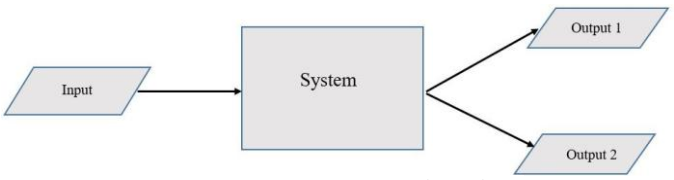

Fig. 1. Block diagram showing flow of I/O.

In the above figure,

- Input – 10 Keywords given by the user.
- Output 1 – Rating of the given keywords.
- Output 2 – List of most influential usernames in Twitter.

Shantanu Chatterjee, Rajat Agarwal, Rounak Saraogi, and Biswaraj Sen, "User intent analysis through twitter using big data tools," *International Research Journal of Advanced Engineering and Science*, Volume 2, Issue 1, pp. 211-213, 2017.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Ultimately, the objective of the project is to help make better business decisions by narrowing down the scope of options in certain scenarios, and the results aid the user in the same direction.

Rating of the keywords can be relevant in many ways. There may be scenarios where topic/intent specific information may be needed, e.g. in case of a news channel that has to prioritize its broadcast on the basis of current topics of interest. Also, this result can be valuable to the users who may need to have a comparative information among his topics of interest. By looking at the most influential user result and performing a background check manually, marketers can approach the person for promotion of his product.

V. DESIGN

The system is built upon the Hadoop ecosystem, which comprises of the Flume and Hive tools.

Flume is a data ingestion tool mechanism for collecting, aggregating and transporting large amounts of streaming data unstructured twitter JSON (Java Script Object Notation) data from a custom source which accesses the Twitter Streaming API to HDFS(Hadoop Distributed File System). The system uses the custom source to filter the tweets on the set of search keywords supplied by the user to help identify relevant tweets, rather than a pure sample of the entire Twitter firehose.

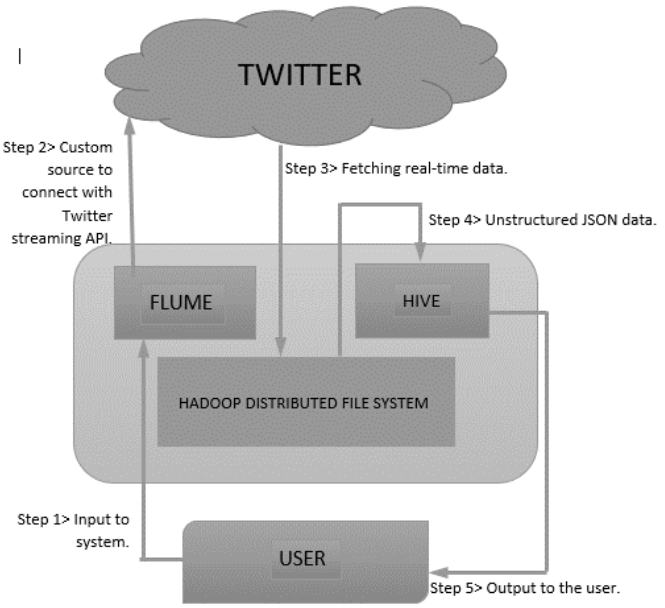

Fig. 2. Control flow of the data pipeline.

Hive provides a query interface that can be used to query data that resides in HDFS. The system incorporates Hive instead of traditional RDBMS (Relational Database Management System) because Hive can process the unstructured twitter JSON data which cannot be processed by traditional RDBMS.A well-defined Hive query will output the ratings of the user supplied keywords in descending order of most number of tweets.

VI. IMPLEMENTATION

To achieve the solution to the problem stated, the following steps are to be undertaken, which are discussed in brief –

1. Creation of a Twitter application using an existing account. A set of keys will be generated exclusive to the Twitter™ application created, which will be used later on.
2. Configuring the 'flume.conf' file, which is basically a configuration of the Apache Flume tool which links it to Twitter™ through the application created earlier.
3. Specifying the 10 keywords which have to be rated. For this, a simple C program will be written which will take the input from the user and write the same to the 'flume.conf' file.
4. Running the 'flume -ng' command, which initiates the crawling of real time data from Twitter™ through the custom source created by configuring the Flume tool. Data which is being fetched is stored on a continuous basis in the HDFS, and this will continue to happen until the execution of the 'flume –ng' command is stopped.
5. Performing 'SerDe' operation on the data that has been locally stored in the machine. As a result, the data which was previously unstructured will now be organized in a partial schema and hence can be considered semi-structured.
6. Running the query on Apache Hive tool to analyze the data and obtain the output, which will be the rating of the 10 keywords in descending order.

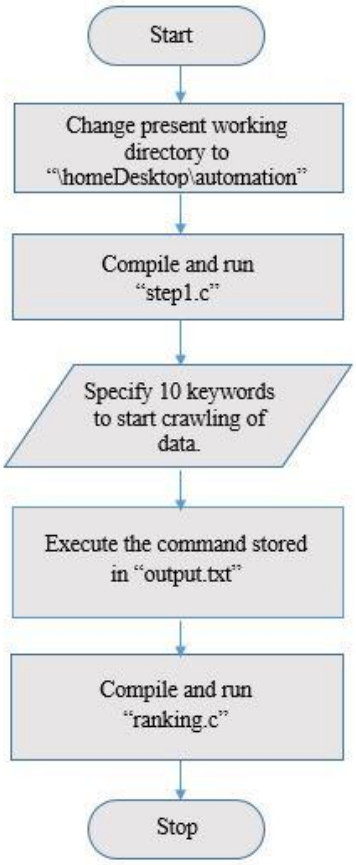

Fig. 3. Steps in the execution of the system.

Shantanu Chatterjee, Rajat Agarwal, Rounak Saraogi, and Biswaraj Sen, "User intent analysis through twitter using big data tools," *International Research Journal of Advanced Engineering and Science*, Volume 2, Issue 1, pp. 211-213, 2017.

VII. Conclusion and Future Scope

The output of the system is successful in understanding the intent of user to a certain extent. By referring to the output, i.e. the rating of the keywords and also the most influential user, a better strategy can be made for promoting a product or service.

Using these results, the scope of decision making will be narrowed down in certain scenarios, as the user will know which other product comes nearest in the competition. After manually analyzing the sentiment of the tweet(s), the most influential person can be approached for promotion of the product.

VIII. Future Scope

The system can be made more autonomous to make it a one step process. It will free the user from the hassle of modifying the configuration file as opposed to the requirement of him to do so now. It will provide a higher level of abstraction which will be better for the end user.

This system can be combined with natural language processing systems to yield better results, like performing 'sentiment analysis'. Sentiment analysis, also called opinion mining, refers to the use of natural language processing to identify and extract subjective information in source materials. The output of such a system aims to determine the attitude of the speaker, whether positive, negative or neutral. It will eliminate the need of a manual analysis that may be required to understand the opinion of the user in a better way.

References

[1] T. White, *Hadoop: The Definitive Guide*, 4th Edition. O'Reilly Media, 2015.
[2] Hadoop tips: Analyse Tweets using Flume, Hadoop and Hive (16/08/2016) Retrieved from http://www.thecloudavenue.com/2013/03/analyse-tweets-using-flume-hadoop-and.html
[3] Apache Flume Tutorial (20/08/2016). Retrieved from http://www.tutorials point.com/apache_flume
[4] Sentiment Analysis with Hadoop (01/09/2016). Retrieved from http://hortonworks.com/hadoop-tutorial/how-to-refine-and-visualize-sentiment-data
[5] The Evolution of Searcher Intent Markers: a New Way to Look at SEO (06/10/2016). Retrieved from https://www.searchenginejournal.com/the-evolution-of-searcher-intent-markers-a-new-way-to-look-at-seo/61541/
[6] Hadoop: What It Is and How It Works (22/10/2016). Retrieved from http://readwrite.com/2013/05/23/hadoop-what-it-is-and-how-it-works/

Shantanu Chatterjee, Rajat Agarwal, Rounak Saraogi, and Biswaraj Sen, "User intent analysis through twitter using big data tools," *International Research Journal of Advanced Engineering and Science*, Volume 2, Issue 1, pp. 211-213, 2017.