

INTER-RELATIONSHIP BETWEEN TWITTER DATA FOR RESTAURANT RECOMMENDATION SYSTEM

M. Eliazer¹, Abhyudit Bisht², Rishabh Taneja³

^{2,3}Bachelor of Technology Student

^{1,2,3}SRM Institute of Science and Technology

eliazer.m@ktr.srmuniv.ac.in, bisht.dit@gmail.com,

rishtaneja332@gmail.com

Abstract: With the recent advancements in technology and rapidly growing data, big data systems have emerged a fast-growing sector in almost every field. As the technology for storing, utilizing and analyzing data is advancing, methods to do the same have also advanced to a whole new level. Stream processing is one of the potential applications of big data systems. Stream processing is used to process a continuous flow of data items to gather useful information for the user. Restaurant rating systems are some applications of this method. In this way, this application can be used to provide help to solve end-to-end real-life problems. On this basis, the presented work aims to analyze, the inter-relationship between data related to restaurant information from a different perspective. The proposed work will also help in analyzing patterns in stream processing which can prove to be useful in finding the users requirements, preferences etc. related to restaurants and only taking the most recent data in the account. In this paper, we are analyzing Twitter data by using Hadoop tool along with Hadoop distributed file system (HDFS), MapReduce [4] and hive. The advantage of using this system is that by using these tools we can process the data with no limitation of data and no data loss problems. Also, we can get high throughput with low maintenance cost as the model takes only the required and most recent data. The main motivation for building this model is that it takes only the recent information into consideration and not historical data which is irrelevant in most cases.

Keywords: Big data, Hadoop, Stream processing, Sentiment analysis

1. Introduction

Restaurant data

Data relating to specific restaurant related tweets which either include hashtags or name of restaurants or cuisine type. The data can be further analyzed to find out useful insights like which restaurant is more preferred by the people for a specific location[1] or what type of food people enjoy having. The restaurant

data can have special keywords which may be present in the hashtags or even the tweet sentences. This data is extracted and cleaned for the purpose of finding useful insights from the data from social media.

2. System Configuration

2.1 Hadoop ecosystem

In this model, we will be using a Hadoop environment, which is managed by Apache. It is a very powerful platform that can be used to process large amounts of data in a distributive[5] fashion by forming clusters of computers. All this can be run on simple programming models. The model is designed in such a way that it can scale up to any number of machines and at the same time offering a separate storage and computation demand. The development and this project are used by a global community of users and other contributors. There are specifically four modules to the Hadoop framework by Apache:

Hadoop Common: It has all the utilities and libraries which are needed by other modules in Hadoop. This includes essential Java files and other related scripts that are necessary for proper working of Hadoop.

Hadoop Distributed File System(HDFS): On the community machines it is essential to give the maximum throughput access to application data. In return, we get maximum aggregate bandwidth around all the clusters.

Hadoop YARN: It is equally important to schedule the jobs and managing cluster resources and this is why YARN framework is useful.

Hadoop MapReduce: Since Hadoop works with a large set of data sets a model for parallel processing is required. MapReduce is YARN-based programming model that helps in the simultaneous processing of data.

2.2 Why use Hadoop ecosystem

Hadoop is popular for its ability to store, analyzing and providing access to large datasets. It offers a quick and cost-effective service by utilizing clusters. Along with

other open source products, there are several other which make Hadoop easily accessible and more usable at the same time.

2.3 Map-Reduce model

Hadoop ecosystem also provides us with the power of map-reduce. Map-reduce allow us to perform massive scalability in scenarios where thousands of servers are used in a Hadoop cluster. It is one of the most important algorithms discussed in this model as it is very helpful in situations where we have to deal with large unstructured data. It helps in breaking down the elements into smaller key-value pairs or tuples. This allows us to

2.4 Twitter configuration

One of the main sources of social media data in modern times in twitter and which is why it interests analysts to find out more and more useful insights as possible. The Hadoop environment needs twitter to be configured so it is compatible to work with the proposed model. The Twitter API is used to generate special tokens and keys which are unique for every application. These help in mapping the system configuration with that of twitter configurations. This ultimately guides the in-stream of data to the HDFS storage sink.

3. Information retrieval

Information retrieval is the science and innovation of investigating information [6] in the request to find undetermined examples. Some related improvements which are, hypothetical and innovative advancements are added to our insight into the restaurant data.

Information-mining helps in filtering throughout the large databases for intriguing example and connections among various occurrences. The procedure that is involved in this process of information mining follows the following sequence pre-processing, examination and information trading

3.1 Flume for live stream processing

The most commonly used method for establishing a distributed and reliable system is Apache Flume. It proves to effectively collect, aggregate and move log data to a centralized data store from other different sources.

The event can be described as a unit of data flow. The Flume process is responsible for hosting the components which are involved in flowing of data from external source to another.

The events are consumed by Flume which is delivered to it by other external sources in such a format which is recognizable by Flume. Now Flume

acts as agents between the running source and the storage sinks

3.2 Storing live stream data

One of the main problems that we face in analyzing large data stream is the correct storing arrangement and procedures, as data today is massively large and the system should be able to handle it in order to process it.

After setting up the Hadoop environment we use Sqoop which is a component used for transferring of files from the traditional database to Hadoop Distributed File System. This can also be reversed when using for unstructured data.

The semi-structured or unstructured is transformed as a cluster from sources like Twitter.

Document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow. Figure (I) depicts the timeline of the data extraction period. The y-axis represents the total number of tweets extracted and the x-axis represents the time period in days.

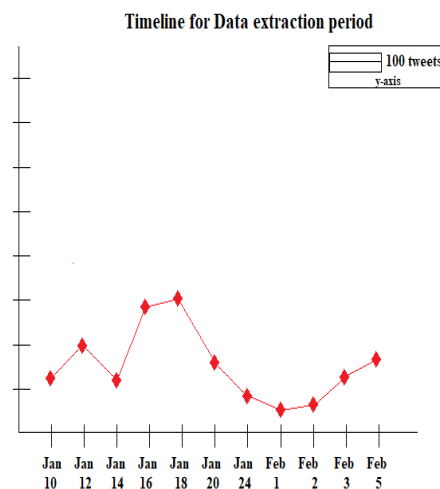


Figure 1. Timeline – Data extraction period

4. Data transformation

Data transformation is an integral part of the Analysis phase. Before this, data is retrieved and different mining methods are used on this collected data. Different information mining methods are Regression, Prediction, Association, Outlier Detection, Clustering and Classification[3]. The expectation finds a connection between free factors and connection

amongst reliant and free factors. There are different calculations of arrangement and expectation. The data collected from different sources are very dirty, that is, it contains a lot of missing information, wrong data, irregular format. This is termed as Unstructured data which is stored as CSV (comma separated values) format where all the entities are separated. For this to be readable by the humans, this unstructured data is converted into structured data using JSON conversion tool.

4.1 Transforming Unstructured Data to Structured Data

A huge aspect concerning Hadoop is that it gives a steady, simple on the pocket and similarly a less complex structure for a get-together, restricting and putting away different information streams that were a few years back not possible.

We make use of Hive for analysis of data and information warehousing device over Hadoop ecosystem system gives multiple functionalities like Information displaying, Information control, Information preparing and Information querying. Information Extraction in Hive implies the formation of tables in Hive and stacking organized and semi-organized information and additionally querying information in light of the necessities.

For cluster handling, we will compose especially characterized contents utilizing a custom guide and lesser contents utilizing a scripting dialect. It gives SQL like condition and support for simple querying. Various steps are followed before the data is finally readable by humans. Steps [2] involved are data cleaning, data pre-processing, and data selection.

4.2 Data cleaning

This is the first step after the data has been collected. This step involves different techniques to remove dirty and noisy data from the stored data. As we are using 'live-streaming data', [12] this step is very important as we will get a lot of unwanted data which needs to be cleaned for better accuracy. This is a preparation for the next step, pre-processing of the data.

4.3 Data pre-processing

Arranging: Rank information by significance. Organizing information into important requests so we can examine it more adequately.

Gathering and division: Compress information and section distinctive gatherings. Separating the information into gatherings or fragments that speak to various classifications or classes, which can be additionally used to effectively distinguish the information we are working with.

4.4 Data Selection

After the information cleaning stage, it's the ideal opportunity for choice of vital and important information for the following stage that is information change. Here fitting information [11] is picked and prepared. The chosen information is chosen in such a way, to the point that it has an entire arrangement of data required for best outcomes. The information with missing data is disposed of.

5. Data analysis

Spotlights on demonstrating and learning revelation for prescient as opposed to absolutely expressive [10] purposes, while business knowledge covers information investigation that depends vigorously on conglomeration, concentrating on business data.

5.1 Hive for data selection

Hive is an information distribution center foundation apparatus to process organized information in Hadoop. It dwells over Hadoop to abridge Huge Information and makes questioning and examining simple. Apache Hive (Hive Query Language) with Hadoop Appropriated record Framework is utilized for Investigation of information. Hive gives a SQL-like interface [8] to process information put away in HDP. Due to its SQL-like interface, Hive is progressively turning into the innovation of decision for utilizing Hadoop. To set up HIVE in Hadoop: Fabricate or Download the JSON SerDe.

5.2 Setting up environment and software tools

Investigation administrations with exceed expectations can be utilized as a front end for inquiry process utilizing dialects like SQL. The PowerPivot and numerous different apparatuses can be utilized to better distinguish the answers to the current issue.

5.3 Finding useful insights

Finding the inter-relationship between the acquired data and new insights is the main motive of any analysis program.

This phase is very much dependent on the user and not on the system, as it is the user who defines in what way he wishes to find patterns in the data depending on his own personal requirements. In Figure (II) we can clearly see the four major phases that together combined help in finding useful insights to the user. The user's choice of data or algorithm may differ but these all these phases are to be followed in the same order to get the best results.

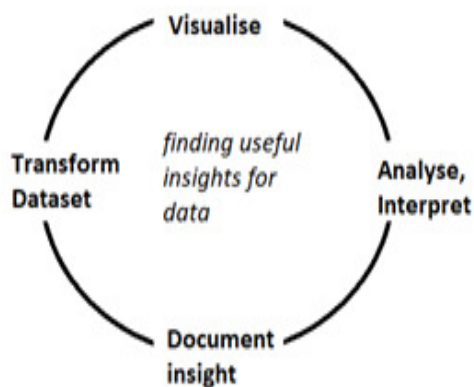


Figure 2. Process for finding insights

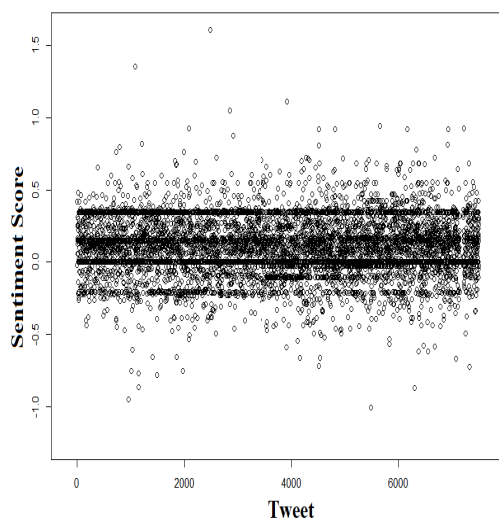


Figure 3. Scatter plot for Sentiment score for every Tweet

6. Results and Graphs

In this paper, we have presented a new approach called Learning Analytics and Predictive analytics to identify the likes of customers living in an area with respect to the food they enjoy. This helps us in identifying the trends such as most popular restaurant and most liked food items. This research develops a new methodology using Hadoop to determine the inter-relationship between people and the restaurant data. The model can be very useful in cases like surveying for new restaurants.

We can clearly see in Figure (IV) each distinct attribute like Total number of tweets, Total number of followers represented on the x-axis with the measure of amount of data on the y-axis. This bar plot visualization helps us in understanding the data in a better way.

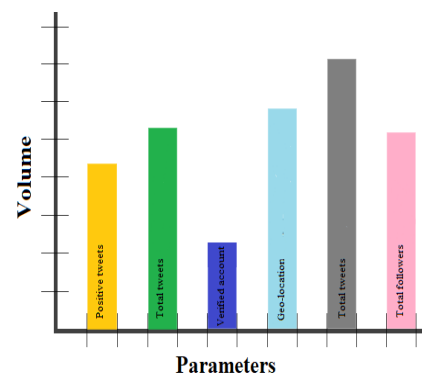


Figure 4. Representation of individual attribute

7. Future scope and Conclusion

The model uses spark that allows us to run in-memory processing on the clusters. This proves to be more efficient and more flexible as compared to Hadoop's MapReduce paradigm. This allows us to make repeated access to data that is being extracted from the live stream of data.

The model can be further enhanced by refining the data retrieval process to only extract the required data which would save lot of memory, time and computation power. The presented model is unique in its own way because it takes into account only recent data that is specified by the user. Also, the data storage is more optimized as we deal with a less amount of data and the results obtained are much more relevant to recent trends as the model completely ignores using historical data in data analysis.

References

- [1] R. Krüger, D. Thom, and T. Ertl, "Visual analysis of movement behaviour using web data for context enrichment," in Proc. TelePacific Vis. Symp., 2014, pp. 193–200.
- [2] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," Transp. Res. Part C: Emerging Technol., vol. 19, no. 4, pp. 557–568, 2011.
- [3] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban Computing: Concepts, methodologies, and applications," ACM Trans. Intell. Syst. Technol., vol. 5, no. 3, pp. 1–55, 2014, Art. no. 38.
- [4] I. Ceapa, C. Smith, and L. Capra, "Avoiding the crowds: Understanding tube station congestion patterns from trip data," in Proc. ACM SIGKDD Int. Workshop Urban Comput., 2012, pp. 134–141.

- [5] W. Zeng, C. Fu, S. M. Arisona, A. Erath, and H. Qu, "Visualizing mobility of the public transportation system," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1833–1842, Dec. 2014.
- [6] G. L. Andrienko and N. V. Andrienko, "Spatio-temporal aggregation for visual analysis of movements," in *Proc. IEEE Symp. Vis. Anal. Sci. Technol.*, 2008, pp. 51–58.
- [7] A. Slingsby, J. Wood, and J. Dykes, "Treemap cartography for showing spatial and temporal traffic patterns," *J. Maps*, vol. 6, no. 1, pp. 135–146, 2010.
- [8] Z. Wang, T. Ye, M. Lu, X. Yuan, H. Qu, J. Yuan, and Q. Wu, "Visual exploration of sparse traffic trajectory data," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1813–1822, Dec. 2014.
- [9] C. Tominski, P. Schulze-Wollgast, and H. Schumann, "3D information-visualization for time-dependent data on maps," in *Proc. 9th Int. Conf. Inf. Vis.*, 2005, pp. 175–181.
- [10] S. Thakur and A. J. Hanson, "A 3D visualization of multiple time series on maps," in *Proc. 14th Int. Conf. Inf. Vis.*, 2010, pp. 336–343.
- [11] H. Doraiswamy, N. Ferreira, T. Damoulas, J. Freire, and C. T. Silva, "Using topological analysis to support event-guided exploration in urban data," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 2634–2643, Dec. 2014.
- [12] T. Padmapriya and V. Saminadan, "Improving Performance of Downlink LTE-Advanced Networks Using Advanced Networks Using Advanced feedback Mechanisms and SINR Model", *International Conference on Emerging Technology (ICET)*, vol. 7, no. 1, pp. 93, March 2014.
- [13] S. V. Manikathan and V. Rama, "Optimal Performance of Key Predistribution Protocol In Wireless Sensor Networks" *International Innovative Research Journal of Engineering and Technology*, ISSN NO: 2456-1983, Vol-2, Issue –Special –March 2017.

