I J C T A, 10(8), 2017, pp. 739-744 ISSN: 0974-5572

Processing of Real Time Big Data for Job Classification using Cloud Computing

Ramdas Gawande[1], Sudeep Thepade[2], and Namrata Gawande[3]

ABSTRACT

In the job classification domain[3], accurate classification of jobs to occupation categories is important for matching job seekers with relevant jobs. An example of such a job title classification system is an automatic text job post classification system that utilizes machine learning. Machine learning-based job type classification techniques for text and related entities have been well researched in academia and have also been successfully applied in many industrial settings. In this paper we present a novel approach, a machine learning-based semi-supervised job title classification system[3]. Our method leverages a varied collection of classification and techniques to tackle the challenges of designing a scalable classification system for a large taxonomy of job categories. It encompasses these techniques in a cascade classification architecture. We first present the architecture of our system, which consists of a two-stage Capture with filtration and fine level classification algorithm. The paper concludes by presenting experimental results on real world live data.

Index Terms: Big Data, Hadoop, HDFS, Cloud Computing, data analysis, Machine Learning.

1. INTRODUCTION

The rapid growth of social networks in recent years has developed a new business: the trade of social networks user's data. These social networks data are becoming important for many companies around the world and are often used to determine social networks user's interests for items in order to propose or advertise items to them.

First defined social network sites as web-based services that allow users to construct profile (public or semi-public), to share connections with other users and to view and traverse lists of connections made by others in the system. The personal information posted by users of a social network (which may involve personal description, posts, ratings, but also social links) can be exploited by a recommender system[5].

Second defined recommender systems [6] as software that elicit the interests or preferences of individual consumers for products, either explicitly or implicitly, and make recommendations accordingly. Recommender systems are mainly related to information retrieval, machine learning and data mining.

2. PROPOSED SYSTEM

- To process the live remote feed to prevent unwarranted data loss.
- Analysing the data to make decisions based on real-time processing

1. Real Time Data:
   - Live Feed From Social Network.
   - Data Collected on the Basis of # Hashtag.

[1-3] PCCOE, Pune, Processing of Real Time Big Data for Job Classification using Cloud Computing, *Emails: ramdas.gawande@gmail.com, sudeepthepade@gmail.com, gawande.namrata@gmail.com*

2. Problems to Process Real Time Data:
   - Multiple Languages in the Data feed.
   - Uneven Structure of the Data.
   - High Velocity of Data.
3. Cloud to Process Real Time Data:
   - To store large amount of data in cloud for Further Processing.
   - Cloud helps in Managing this data for Process Scheduling.
   - It helps in-memory cluster computing that increases the processing speed of an application.

2.1. Block Diagram

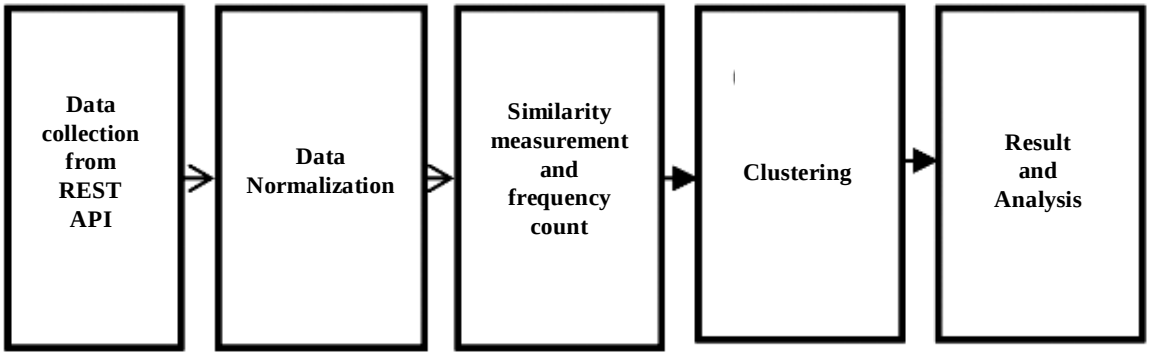

Figure 1: Block Diagram

2.2. Algorithms

Algorithm I. Filtration and Data Load Balancing[1]

Algorithm:

Input: Live Data Feed process data set

Output: filtered data in fixed size block and send each block to processing Mechanism

Steps:

1. Filter related data i.e. Processed data. All other unnecessary data will be discarded.
2. Divide the Data into Appropriate Key Value Pair.
3. Transmit Unprocessed data directly to aggregation step without processing.
4. Assign and transmit each distinct data block of Processed data to various processing steps in Data Processing Unit.

Description: This algorithm takes live data and then filters and divides them into segments and performs load-balancing algorithm.

In step 1, related details filtered out.

In step 2, filtered data are the association of different key value pairs and each pair is different numbers of sample, which results in forming a data block. In Next steps , these blocks are forwarded to processed by Data Processing Unit.

Algorithm II. Processing and Calculation Algorithm

Input: Filtered Data

Output: Normalized data for Job Classification.

Steps:

1. For each event job type or for the job data, Categorical Data related job is extracted.
2. Normalize the extracted data for all the data from live feed.
3. persist the data into data store and forward it.

Description: The processing algorithm[1] calculates results for different parameters against each incoming filtered data and sends them to the next level.

In step 1, the calculation of multiple job classification along with user Furthermore, in the next step, the results are transmitted to the aggregation mechanism.

Algorithm III. Multi job Summarization Algorithm for Multiple user

Input: Normalized job Data of all users.

Output: Final result summary

1. Gather the data from data store in normalized format.
2. Apply Summarization for Individual modal pie from the total data capture.
3. persist the final summary into data store.

Description: here the data is collected and the results from each modal is processed against all and then combines, organizes, and stores these results in NoSQL database.

Description: here the data is collected and the results from each modal is processed against all and then combines, organizes, and stores these results in NoSQL database.

2.3. Flowchart

In proposed system for analyzing real time as well as offline data for real-time applications using term Big Data we have divided real time Big Data processing architecture[6] into three parts, i.e., 1) Data Acquisition Unit 2) Data Processing Unit and 3) Data Analysis and Decision Unit. In these three unit various algorithms or techniques will be implied on data for its analysis.

2.3.1. Data Acquisition Unit

The need for parallel processing of the massive volume of data was required, which could efficiently analyze the Big Data. For that reason, the proposed unit is introduced in the real time Big Data processing framework that gathers the massive volume of data from various available data gathering unit around the world.

2.3.2. Data Processing Unit

In data processing unit[6], has two basic functionalities filtration and load balancer. Filtration mainly involves filtration of data and load balancing of processing power. Filtration makes process of filtering data which is useful to us for analysis and blocks other data. It will surely help to improve performance of system as we are only dealing with the useful data.

2.3.3. Data Analysis and Decision

This unit contains three major functions, such as aggregation and compilation server, results storage server, and decision making server. When results are to be send to the compilation the data is not in aggregated form. So it is necessary to make the given data in aggregated form for proper storage and processing.

This unit contains three major portions, such as aggregation and compilation server, results storage server(s), and decision making server. When the results are ready for compilation, the processing servers in data processing unit send the partial results to the aggregation and compilation server, since the aggregated results are not in organized and compiled form. Therefore, there is a need to aggregate the related results and organized them into a proper form for further processing and to store them.

2.4. Activity Diagram

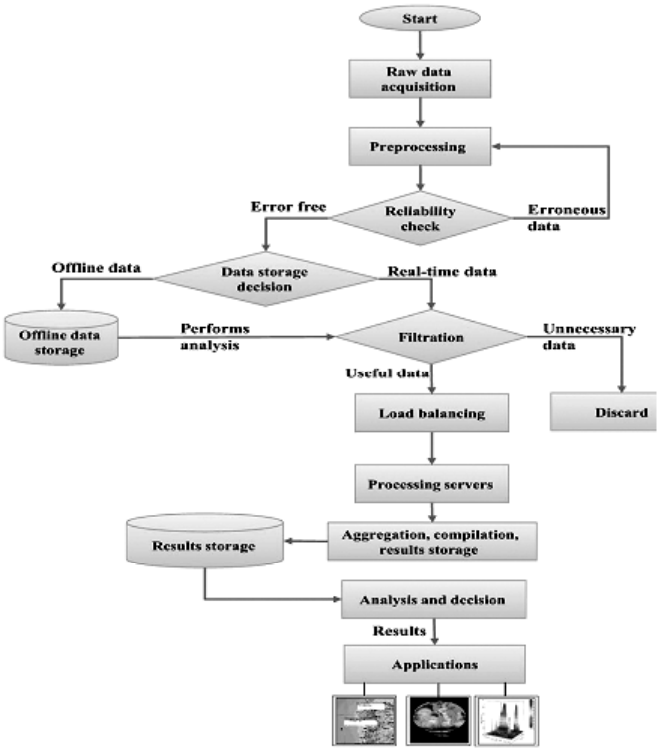

Figure 2: Flowchart of the real time Big Data Processing architecture.

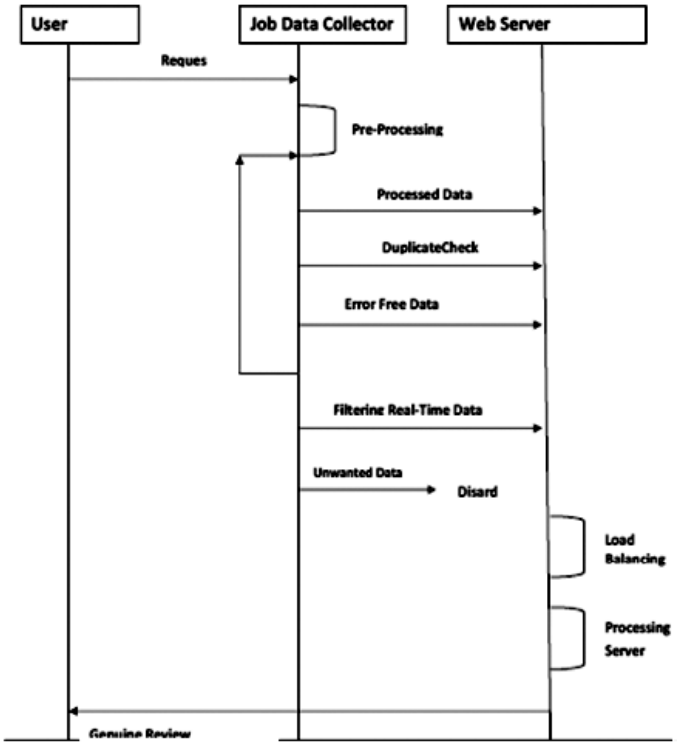

Figure 3: Activity Diagram

3. RESULTS

The experimental results of this project will process the real time social networking job postings and will do preprocessing, feature extraction and classification into various job categories .The job seekers will be able to get the job postings as per their expertise and area of interest.

Feature extraction will be done based upon below parameters.

- Name
- Location
- Job Type
- Job Location
- Time and Date
- ID

Figure 4: Snapshot of big data from Social networking site

Preprocessing, Feature extraction and classification will be performed on above data.

4. CONCLUSION

This work proposes behavioural approach to detect job Postings from live data feed. We derive an aggregated behaviour mining methods for jobs classification according to the name that they demonstrate pattern behaviours, so as to evaluate our proposed methods that conducts user evaluation on a live dataset containing reviews of different types of jobs. We found that here proposed methods generally outperform the baseline method based votes.

As part of future work, we can incorporate job feed detection into the various other useful job aggregators and vice versa. Exploring ways to learn behaviour patterns related to that mining so as to improve the accuracy of the current regression model is also an interesting research direction.

REFERENCES

[1] Rathore, Muhammad Mazhar Ullah, Anand Paul, Awais Ahmad, Bo-Wei Chen, Bormin Huang, and Wen Ji. “Real-Time Big Data Analytical Architecture for Remote Sensing Application”, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2015.

[2] Diaby, Mamadou, Emmanuel Viennet, and Tristan Launay. “Toward the next generation of recruitment tools : an online social network-based job recommender system”, Proceedings of the 2013 IEEE/ACM International onference on Advances in Social Networks Analysis and Mining - ASONAM 13, 2013.

[3] Javed, Faizan, Qinlong Luo, Matt McNair, Ferosh Jacob, Meng Zhao, and Tae Seung Kang. “Carotene: A Job Title Classification System for the Online Recruitment Domain”, 2015 IEEE First International Conference on Big Data Computing Service and Applications, 2015.

[4] Lim, Ee-Peng, Viet-An Nguyen, Nitin Jindal, Bing Liu,and Hady Wirawan Lauw. “Detecting product review spammers using rating behaviors”, Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM 10 CIKM 10, 2010.

[5] Namrata Gawande, Ramdas Gawande, “Processing of Real Time Big data for Machine Learning”, May 2016 International Journal of Advanced Research in Computer and Communication Engineering

[6] PuneetGarg, Rinkle Rani, SumitMiglani, “Mining Professional’s Data from LinkedIn”, 2015 Fifth International Conference on Advances in Computing and Communications.

[7] Ahmed AbdeenHamed, Xindong Wu, James R Fingar., “A Twitter-based Smoking Cessation Recruitment System”,2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

[8] MamadouDiaby , Emmanuel Viennet, “Taxonomy-based Job Recommender Systems On Facebook and LinkedIn Profiles”, 2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS).

[9] Emmanuel Malherbe, MamadouDiaby , Mario Cataldi , Emmanuel Viennet , Marie- AudeAufaure, “Field Selection for Job Categorization and Recommendation to Social Network Users”, 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014).