Journal of Theoretical and Applied Information Technology
20th October 2015. Vol.80. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

STREAMING TWITTER DATA ANALYSIS USING SPARK FOR EFFECTIVE JOB SEARCH

[1] LEKHA R. NAIR, [2] DR. SUJALA D. SHETTY

[1]Research Scholar, Department of Computer Science, BITS Pilani, Dubai Campus, Dubai, UAE

[2]Assistant Professor, Department of Computer Science, BITS Pilani, Dubai Campus, Dubai, UAE

E-mail: [1]lekharnair@gmail.com, [2]sujala@dubai.bits-pilani.ac.in

ABSTRACT

Near real time Big Data from social network sites like Twitter or Facebook has been an interesting source for analytics by researchers in recent years owing to various factors including its up-to-date-ness, availability and popularity, though there may be a compromise in genuineness or accuracy. Apache Spark, the trendy big data processing engine that offers faster solutions compared to Hadoop, can be effectively utilized in finding patterns of relevance useful for the common man from these sites. Recently many organizations are advertising their job vacancies through tweets, which saves time and cost in recruitment. This paper addresses the issue of real time analyzing and filtering those numerous job advertisements from among the millions of other streaming tweets and classify them into various job categories to facilitate effective job search, utilizing Spark.

Keywords: *Big Data Analytics, Tweet Stream Analysis, Spark Streaming, Social Network Analysis, Streaming Big Data Processing*

1. INTRODUCTION

The Big Data from the enormous sources available now have gained serious attention from researchers in every field with every attempt to maximize the value of knowledge resulting from its processing and analysis. Twitter, a popular social network site, is one such source which contributes huge data which possess values beyond social and commercial interests. Twitter users can express or share their opinions, feelings or information regarding events, products, health or anything in their 140 character restricted short messages termed tweets. Hashtag is the convention of prefixing a word in the tweet with the symbol '#' indicating the keyword or topic of the tweet, which was meant for categorization of tweets based on topics and aids in searching. As per Twitter Inc., 500 million tweets are sent per day. Handling such big streaming data using Spark ecosystem, which is considered as the second generation Big Data processing engine [1], is the topic of discussion of this paper.

Tweets often contain latest information of topics as it is frequently updated. Tweet analysis can reveal useful information and many research papers have been published based on twitter data like finding sentiments related to products, persons or brands [2], real time event detections like earthquake or other disasters [3], twitter spam classification [4] [5], predicting personality from tweets [6] [7], finding network of friends, trending topics [8] etc. In most of the tweet analytic works, analysis has been done on previously collected data and less work been done on streaming twitter data. Apart from the information gathered by these works, tweets contains other valuable information, if extracted on time, is certain to have practical and immediate application in the life of a common man.

Due to the popularity of social network sites, advertisers are now heavily targeting social network users. It is common that many organizations and individuals are now tweeting about job vacancies and hiring details, instead of publishing them in print or online media which saves time, cost and effort in dissemination. These vacancies are intended to be filled on an immediate basis and it can be utilized effectively by job seekers, if they have a near real time access to these tweets.

This paper investigates the problem of real time analysis and filtering of those specific tweets regarding job vacancies from among millions of other ones and to classify them according to the job category, without following any twitter account. A model needs to be proposed for the near real time collection and analysis of job vacancy related tweets that uses machine learning to classify them

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

in accordance with the job type and location for effective job search. Spark Streaming is used to handle the streaming tweets. Since Spark is highly scalable, the model can be deployed on cloud cluster to achieve scalability as per demand.

2. APACHE SPARK CLUSTER FOR BIG DATA ANALYTICS

Big Data processing requirements initiated a paradigm shift from traditional data processing, resulting in the evolvement of Map Reduce based frameworks like Hadoop. Though Hadoop has been extensively used for Big Data processing for years, performance wise a better solution like Apache Spark can be looked upon as a giant step in big data processing. The open source Apache spark ecosystem integrates batch and stream processing and comprise of libraries providing support for machine learning, graph processing and SQL querying.

Apache spark [9], originated from Berkeley, now licensed under Apache foundation offers much faster performance and a variety of features in comparison with the most sought out Hadoop Big Data Processing System. Though Hadoop is a matured batch processing system with many projects being completed and much expertise being available, it has its limitations. Hadoop is written in java and mainly rely on two functions, the Map and the Reduce, all operations are to be represented in terms of these two functions which makes the programming a little complicated. Spark program can be written using Java, Python or Scala and it offers more functions other than just the map and reduce and above all it provides an interactive mode, the spark shell, which makes programming much simpler for Spark compared to Hadoop. Hadoop persists data back to the hard disk after a map or reduce operation, while spark performs in-memory data processing and hence repetitive operations on same data will be done much faster. Hence memory requirement of Spark is higher compared to Hadoop but if the data fits in the memory, spark works faster or else it has to move data back and forth the disk which deteriorates spark's performance. Being a batch processing system, Hadoop users have to depend on other platforms like Storm [10] for real time data processing, Mahout for machine learning or Giraph for graph processing. But Spark ecosystem includes Spark streaming, MLLib, GraphX and Spark SQL for real time data processing, machine learning, graph processing and SQL querying respectively, which gives competitive advantage for Spark over Hadoop. Table 1 summarizes the comparison of spark against Hadoop.

Table 1: Comparison of Spark against Hadoop

Spark	Hadoop
Second generation Big Data processing engine, with extended features.	First generation Big Data processing engine, matured, with much expertise available.
Availability of functions other than the Map and the Reduce, the option to write program in java, python or Scala and provision of interactive mode - the spark-shell makes programming easy.	Rely on just the Map and the Reduce functions, which makes programming difficult
Up to 100 times faster to Hadoop, especially in iterative operations, as intermediate data/result is persisted in memory	Slower as intermediate data/result is stored in hard disk
Spark being a batch processing engine also includes spark streaming for streaming data processing, MLLib for machine learning, GraphX for graph processing and spark SQL for querying thus providing an all-in-one solution.	Mainly a batch processing engine where users can depend on other compatible platforms for performing stream processing, machine learning or database querying.
Compatible with Hadoop Distributed File System(HDFS)	
Memory requirement is higher. Degradation in performance if data not fit in the memory	Lesser memory requirement

Spark application will be having a driver program that runs the main function and performs parallel operations on various nodes in a spark cluster. By introducing the concept of Resilient Distributed Dataset (RDD) [11], the collection of immutable objects partitioned across the nodes of a cluster for performing parallel operations which can be persisted in memory for repetitive/iterative use, Spark outperforms Hadoop with100 times faster performance by saving time of read/write from disk, especially in running machine learning applications where iterative operations on data is common. RDDs are formed by transformation from other RDDs or file and RDDs retain the information by which it is formed

Journal of Theoretical and Applied Information Technology
20th October 2015. Vol.80. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Since Big Data analytics involves application of machine learning/data mining techniques on Big Data, Spark offers MLLib, the machine learning library that includes popular machine learning algorithms for classification, clustering and association. Integration of MLLib in spark ecosystem is another advantage that spark is having while Hadoop struggles with Mahout, the machine learning platform.

Spark streaming facilitates stream data processing, though spark is basically a batch processing engine. Incoming data stream is grouped into batches of interval less than a second and processed by the batch processing spark engine integrating the powerful features to near real time processing. In this paper we discuss the usage of Spark engine with MLlib, Streaming and Spark SQL for processing, classification and retrieval of career information.

3. THE APPLICATION MODEL

Figure 1 shows the general model of the application which can be run either on premise or can be deployed on cloud. The application connects to Twitter and collect tweet streams which are job advertisements and make them available to the job seeker on a near real time basis.

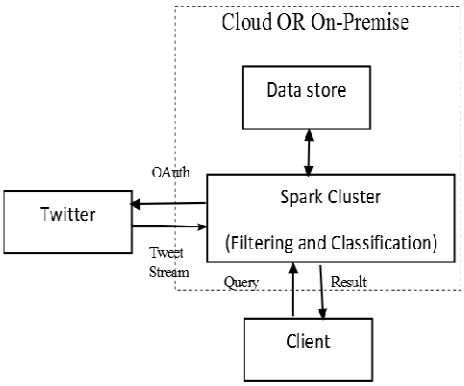

Figure 1: The Application Model

3.1 Twitter Authentication

Since the dataset is from Twitter tweet stream, OAuth authentication with a twitter account needs to be configured. For this a Twitter based application is created, and the consumer key, consumer secret key, access token and access token secret key generated is used for authentication.

3.2 Real Time Data Collection and Processing Using Spark Streaming

After authentication, the streaming application receives the tweet stream and group them into batches with a suitable selection of batch interval. Job vacancy related advertisements are filtered at real time from the streaming tweets based on hashtags. Using windowing function, all the relevant advertisements collected over a chosen interval of time is written to a text file. This intermediate result with listed job vacancies itself can serve as a source of information to the job seekers, though it is not categorized into various types.

3.3 Job Categorization Using Machine Learning

Filtering of job advertisements based on allied hashtags gives good results, but many unethical practitioners use popular tags like '#job' or '#career' in their tweets with totally unrelated content for gaining attention. To remove such tweets and to categorize the relevant advertisements in various classes like IT, Construction, Driving, Healthcare etc. as given in Table 1, for ease of search, machine learning based classification is to be performed. For sake of anonymization, some characters in the tweet are replaced with 'x' in the given table.

Table 2: Tweet Classification Based on Job Type

Tweet	Job Category
#IT #Job alert: VMware Quality System Engineer | VMware | #PaloAlto http://txxxx #Jobs	IT
Blown an interview? Maybe not. Here's how to recover: http://t.co/xxx #career	invalid
#Job #Germantown STxxx: Systems Administrator (Sxx): xx Project Overview: ... http://txxx	IT
JOB OPENING: Project Financial Controls Specialist - IRC at Mxxx (Minneapolis, MN) http://txxxxj #job	Finance
Axx Txx TRAINING #Transportation #Job: DRIVERS (#OKLAHOMA, OK) http://t.coxx	Driving

Several works on topic based tweet classification has been done before, like using Bag

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of Words model or Term Frequency/Inverse Document Frequency model etc. [12] [13]. The article [14] gives an overview on various methods for short text classification.

Most of the job vacancy tweets will be structured ones with minimum typos and colloquial language content, which simplifies the process of classification. Supervised learning can be used to predict the category of the job being advertised. A set of 3000 filtered tweets collected as in previous section is to be manually annotated by two annotators with appropriate class labels to be used as the training set.

3.4 Searching With Spark SQL

Tweets classified under various job categories will be stored in the database, which can be queried to find vacancies belonging to a particular Job category. Spark SQL provides querying functionality. Job seeker can query about a category and the advertisements regarding that category can be given as the result, which is useful if the client is using mobile phones for job search

4. DEPLOYMENT IN CLOUD ENVIRONMENT

This spark application can be run on standalone spark cluster with a master and a few worker nodes. It can also be deployed on a cloud like Amazon EC2. Spark-ec2 script allows launching, managing and shutting down of multiple named spark clusters in Amazon EC2 cloud. Each cluster's machines are placed in corresponding EC2 security groups. Amazon EC2 security keys needs to be created which is meant for secure shell (ssh) connection. EC2 cluster can be launched by running the command from spark-ec2 directory of local machine, ./spark-ec2 -k <keypair> -i <key-file> -s <num-slaves> launch <cluster-name>. To ssh into the cluster the command, ./spark-ec2 -k <keypair> -i <key-file> login <cluster-name> is used.

5. RESULTS

Spark application was run on a cluster with one master and 2 slaves with the hardware configuration of Intel core i5 processor and having Ubuntu 14.04.02 environment. Streaming tweets are collected in batch interval of 2 seconds after an initial filtering using hashtags. All job advertisement tweets arrived in a time interval of 30 minutes are coalesced and saved as a text file. This intermediate result file as such can be used for job search, though it is not categorized. Tweets which are job advertisements can be further classified into various job categories and/or various locations and stored in a database for effective querying.

From the experiment, it is observed that on an average, over 1000 valid job vacancies are tweeted per hour. It is also observed that job vacancies in the IT category is tweeted the most, contributing to 23 % of total job vacancy tweets, followed by Health care and Finance categories. Figure 2 shows the percentage of job categories whose vacancies are tweeted the most. These job vacancies advertised though tweets are more likely to be filled on a first come first serve basis, which necessitates immediate dissemination of this information to the job seeker.

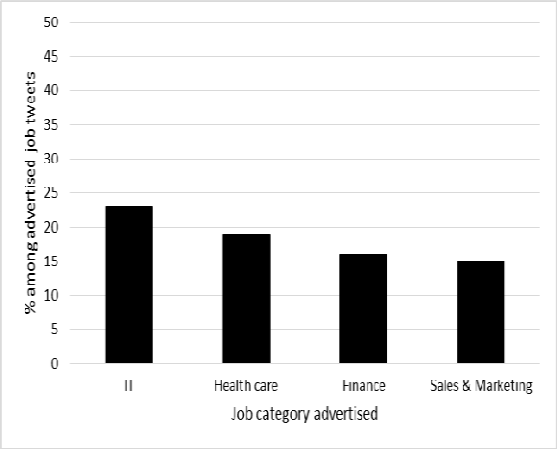

Figure 2: Popular Job Categories Tweeted, Expressed as a % of Total Job Vacancy Tweets.

6. DISCUSSIONS

The contribution from this work involves usage of open source and scalable software, the Spark ecosystem, to process the twitter streaming data to extract timely information of Job vacancies without the need for the seeker to follow of any twitter account. Though this paper deals with a specific application, the model can be used as a general one for finding interesting patterns from other streaming data sources. As data size increases, more nodes can be added to the cluster for effective processing. Chances of error in this setup include miss categorization of jobs or omission of valid tweets, which will be minimal.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

7. CONCLUSION

In this Big Data era, social network sites like Twitter are profoundly used for information dissemination of various breeds including job vacancy announcements. This research work succeeded in developing and implementing a scalable model for real time analysis and filtering of job vacancy related tweets from among millions of other streaming tweets and classify them into different job categories that can lead to improving the effectiveness of job search. Access to these tweets without being a follower of the advertiser's twitter account opens up thousands of job opportunities around the globe for a job seeker, which would have been impossible otherwise. In this work, Spark Streaming was utilized for handling the streaming tweets. Spark being open source and highly scalable, it can easily cater the needs of ever growing data size as well.

8. FUTURE WORK

The model explained in this paper can be generalized to serve variety of streaming applications. It can be extended to provide as a web service which sends near real time job vacancy details to registered users at fixed intervals of time. Also, the classified data from the present work can be further mined to find trends in job opportunities, for different categories and different locations, at different periods of time in a year, as required.

REFERENCES:

[1] F. Gebara, H. Hofstee and K. Nowka, "Second-Generation Big Data Systems," IEEE Computer, vol. 48, no. 1, pp. 36-41, 2015.

[2] S. Liu et al., "TASC: Topic-Adaptive Sentiment Classification on Dynamic Tweets," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 6, pp. 1696 - 1709, 2015.

[3] T. Sakaki, O. Makoto and M. Yutaka, "Tweet analysis for real-time event detection and earthquake reporting system development," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 4, pp. 919-931, 2013.

[4] K. Thomas et al., "Suspended Accounts in Retrospect: An Analysis of Twitter Spam," in ACM SIGCOMM conference on Internet measurement conference, 2011.

[5] J. Song, L. Sangho and K. Jong, "Spam filtering in twitter using sender-receiver relationship," in Recent Advances in Intrusion Detection, Springer-Verlag Berlin Heidelberg, 2011, pp. 301-317.

[6] D. J. Hughes et al., "A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage," Computers in Human Behavior , vol. 28, no. 2, pp. 561-569, 2012.

[7] C. Sumner et al., "Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets," in 11th International Conference on Machine Learning and Applications (ICMLA), 2012.

[8] G. Dehong et al., "Sequential Summarization: A Full View of Twitter Trending," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 2, pp. 293 - 302, 2014.

[9] [Online]. Available: https://spark.apache.org/docs/latest/. [Accessed 5 May 2015].

[10] W. Yang, X. Liu, L. Zhang and L. T. Yang, "Big Data Real-time Processing Based on Storm," in 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, 2013.

[11] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker and I. Stoica, "Spark: Cluster Computing with Working Sets," in USENIX conference on Hot topics in cloud computing, 2010.

[12] K. Lee et al., "Twitter Trending Topic Classification," in IEEE 11th International Conference on Data Mining Workshops (ICDMW), 2011.

[13] B. Sriram et al., "Short text classification in twitter to improve information filtering," in Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010.

[14] S. Ge et al., "Short Text Classification: A Survey," Journal of Multimedia, vol. 9, no. 5, pp. 635-643, 2014.