IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-9, Sept. 2018 E-ISSN: 2347-2693

Data Parallelism: A New Approach in Prediction Systems

K.B. Borole[1*], S.D. Rajput[2]

[1]Department of Computer Engineering, SSBT's College of Engineering and Technology, NMU, Jalgaon, India
[2]Department of Computer Engineering, SSBT's College of Engineering and Technology, NMU, Jalgaon, India

[*]*Corresponding Author: kajalborole7@gmail..com, Tel.:+91 9423578270*

Available online at: www.ijcseonline.org

Accepted: 14/Sept/2018, Published: 30/Sept/2018

***Abstract*—** The day-by-day growing data can compromise the performance of the prediction system, because its obvious that the growing data will require more storage and the system will also consume more time for its processing. In prediction system, testing is part where time is consumed. If the entire data is given to the test model, it will run for the entire input size, and becomes time consuming. For this effective reduction strategy for processing time of testing must be introduced. To reduce this processing time introducing parallelism concept can help. The framework used here is based on fork join pool. In this the input size is divided into parts which are small enough to be processed and then the divided parts are given for testing. Thus reducing the time consumed in testing, and making it better than the other system.

***Keywords*—** Fork Join Pool, Open NLP, Sentiment Analysis, Data Parallelism

I. INTRODUCTION

A prediction or prognosis, is a statement about a not known event. It is meaning includes often, but not every time, based upon learning or knowledge. There is no general agreement about the exact difference between the two things, different authors and disciplines describe different ideas. Although guaranteed exact information about the future is in many cases is not possible or not feasible, prediction is useful to help in constructing plans about possible developments in future. Prediction systems are developed for scientific analysis, financial analysis and many more. These systems use a good amount of data collected in form of dataset, database. Prediction System help in predicting results on basis of prior knowledge. The data collected is analysed in many ways according to the analysis on the topic selected. People are becoming increasingly enthusiastic about interacting, sharing, and collaborating through online collaborative media. In recent years, this collective intelligence has spread to many different areas, with particular focus on fields related to everyday life such as commerce, tourism, education, and health, causing the size of the Social Web to expand exponentially. Growing data analysis includes and combines multiple disciplines such as social network analysis, multimedia management, social media analytics, trend discovery, and opinion mining. The increasing data has made the prediction system to have an increased execution time. In future, the growing data has to be managed in such a way that has overcome the problem of increased execution time for the prediction system. Our contribution has constantly focused on such an approach which has a reduced processing time, over growing data. Our prediction system analyse the twitter dataset related to sentiment analysis. Introducing parallelism to this system has divided the input dataset into slots which are small enough to be processed and then are given as a input to the test model, hence reducing the execution time required to process a large data as compared to serial prediction system.

Rest of the paper is organized as follows, Section I contains the introduction of basic Prediction System, Section II contain the related work of various Prediction System, Section III contain the proposed approach used to attain the goal. It also contains architecture and pseudo-code , Section IV contain the describes results and discussion, Section V concludes research work with conclusion and future scope.

II. RELATED WORK

X. Li, Q. Peng, Z. Sun, L. Chai, and Y. Wang [1], said that due to the rapid development of Web, large numbers of documents assigned by readers emotions have been generated through new portals. Comparing to the previous studies which focused on authors perspective, our research focuses on readers emotions invoked by news articles. The research provides meaningful assistance in social media

application such as sentiment retrieval, opinion summarization and election prediction. Here, the readers emotion of news based on the social opinion network are predicted. More specifically, the opinion network based on the semantic distance is constructed. The communities in the news network indicate specific events which are related to the emotions. Therefore, the opinion network serves as the lexicon between events and corresponding emotions. Leveraging the neighbor relationship in network to predict readers emotions is done. As a result, the methods obtain better results. Moreover, we developed a growing strategy to prune the network for practical application. The experiment verifies the rationality of the reduction for application.

Anshuman, S. Rao, and M. Kakkar [2], proposes with the advent to social media the number of reviews for any particular product is in millions, as there exist thousand of websites where that particular product exists. As the numbers of reviews are very high the user ends up spending a lot of time for searching best product based on the experiences shared by review writers. Here it is presented as a sentiment based rating approach for food recipes which sorts food recipes present on various website on the basis of sentiments of review writers. The results are shown with the help of a mobile application: Foodoholic. The output of the application is an ordered list of recipes with user input as their core ingredient.

S. Khatri and A. Srivastava, in [3], considers Sentimental Analysis is one of the most popular technique which is widely been used in every industry. Extraction of sentiments from users comments is used in detecting the user view for a particular company. Sentimental Analysis can help in predicting the mood of people which affects the stock prices and thus can help in prediction of actual prices. Here sentimental analysis is performed on the data extracted from Twitter and Stock Twits. The data is analyzed to compute the mood of users comment. These comments are categorized into four category which are happy, up, down and rejected. The polarity index along with market data is supplied to an artificial neural network to predict the results.

III. METHODOLOGY

In this section, the proposed approach for the prediction system is described. The proposed approach focus on reducing the computational time required to analyse the dataset. For this we have considered two models, which are Data parallelism (DP) model and other is serial model which will focus on sentiment analysis using dataset. The assumptions made for the models, the architecture for the data parallelism and serial model are shown and explained.

A. Architecture

Following Figure 1 shows the basic architecture of the proposed data parallelism model for prediction system. Its architecture consists of the dataset, Training Phase, Fork Join Pool Framework, division logic used, Testing Phase. The dataset is downloaded from "data for everyone" website, having 40,000 tweets along with their emotion from the tweet. Training Phase involves a training model in which the training tweets are trained using training iterations and cutoff. For data parallelism, Fork Join Pool Framework is used. It is applied to the testing tweets just after the training phase. These trained tweets are used later in testing phase. For parallelism, Fork Join Pool framework act like a media which provides a environment in which all the testing tweets are loaded and is then splited into parts of smaller size. These divided parts are then given to the test model simultaneously. In Test Model, tweets are classified using Document Categorizer provided by Open NLP. Lastly, Accuracy, Precision, and Recall is calculated.

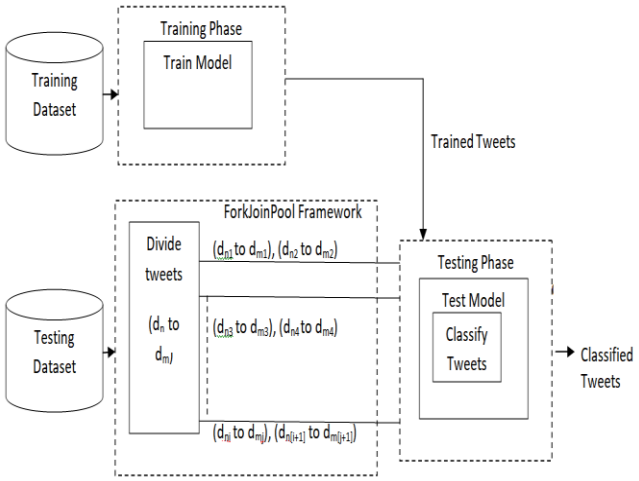

Figure 1. Architecture of proposed Data Parallelism Model.

Figure 2 shows the basic architecture of the serial model for prediction system. Its architecture consists of the dataset, Training Phase, Fork Join Pool Framework, division logic used, Testing Phase. The dataset is downloaded from "data for everyone" website, having 40,000 tweets along with their emotion from the tweet. Training Phase involves a training model in which the training tweets are trained using training iterations and cutoff. These trained tweets are used next in testing phase. Testing tweets are used in Test Model. In Test Model, tweets are classified using Document Categorizer provided by Open NLP. Lastly, Accuracy, Precision, and Recall is calculated. The running timings are stored in the database.

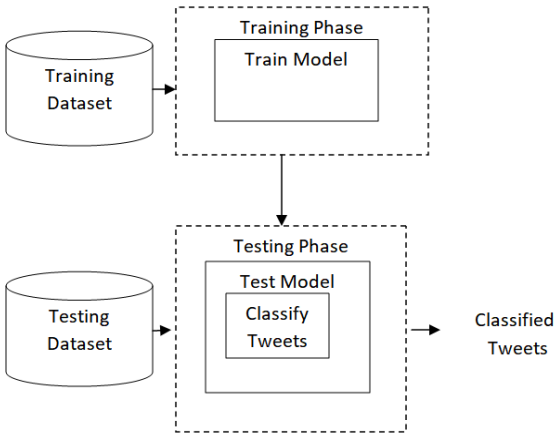

Figure 2. Architecture of Serial Model.

B. Pseudo-Code

Step 1: Assume minimum size to be 1000

Step 2: Let the input tweets be from dntodm

Step 3: If input tweets are greater than minimum size, divide it into slots using,

Calculate; mid = (dn+dm) / 2 ;

Step 4: invokeAll[(dni ; dmj ,v, model), new(dni+1; dmj+1,v, model)];

Step 5: Once the input is divided into smaller slots, testmodel() is called.

Step 8: End.

IV. Results and Discussion

This section includes experimental results. Table 1 shows the experimental results for data parallelism model and serial model. It presents speedup analysis between serial and data parallelism model, in which we have tested the models with increasing no of tweets and studied and recorded its execution time. It is observed that, the average speedup obtained with the growing input is about 9.25 percent.

Table 1. Experimental Results

Sr No	No. of Tweets	Execution Time for Serial (ms)	Execution Time for Data Parallelism (ms)	Speedup (%)
1	5000	7688	6068	21
2	10000	15613	13485	13
3	15000	23970	20518	14
4	20000	29243	26617	8
5	25000	42261	40580	3
6	30000	51727	50598	4
7	35000	69138	66848	3
8	40000	90715	81555	10
9	-		Average Speedup	9.25

Figure 3 shows the results obtained from both the data parallelism and serial model prediction system. In which X-Axis has No. of Tweets and Y-Axis has Execution time. The time is taken from the database, in which the testing time is being saved. It is calculated as presented in performance metrics.

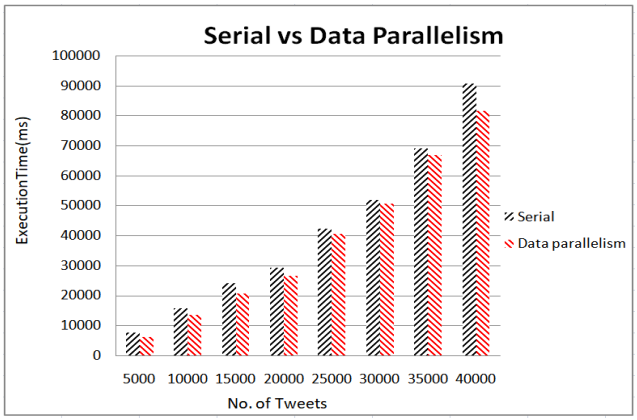

Figure 3. Graphical representation of Results obtained for proposed Data Parallelism Model and Serial Model

V. Conclusion and Future Scope

The proposed work which is based on parallelism has given results which are better than that of the serial model. Here, the model is used for reducing the testing time consumed during the sentiment analysis of twitter data. Also, the model is tested with the increasing data size and corresponding results were recorded and further analysed. After analysis, satisfactory results were observed which has reduced the testing time. Hence, attaining our objective. In Future, the model can be used in developing other prediction system which are based on scientific analysis like medical diagnosis, election polls. Also, this development can be extended in prediction system which are based on natural language processing where tweets in regional languages can also be considered.

Acknowledgment

This is a great pleasure and immense satisfaction to express my deepest sense of gratitude and thanks to everyone who has directly or indirectly helped me in completing my project work successfully. I present my sincere thanks to the Principal for moral support and providing excellent infrastructure in carrying out the Project work. I am very thankful to acknowledgement to the Head of Department, Computer Engineering. I express my gratitude towards Project guide who is Assistant Professor, Computer Engineering and who guided and encouraged me in completing the project work in scheduled time. No words are sufficient to express my gratitude to my family for their unwavering encouragement. I also thank all friends for being a constant source of my support.

REFERENCES

[1] X. Li, Q. Peng, Z. Sun, L. Chai, and Y. Wang, "Predicting social emotions from readers perspective", IEEE Transactions on Affective Computing, no. 1, pp. 1-1, 2017.

[2] Anshuman, S. Rao, and M. Kakkar, "A rating approach based on sentiment analysis," Proceeding of 2017 7th International Conference on Cloud Computing, Data Science and Engineering Confluence, pp. 557-562, 2017

[3] S. Khatri and A. Srivastava, "Using sentimental analysis in prediction of stock market Investment , " proceeding of 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 566-569,2016.

[4] Y. Shen, W. R. Yazhi Gao, and Z. Xiong, "Convolutional neural network based sentiment analysis using adaboost combination," Proceeding of 2016 International Joint Conference on Neural Networks (IJCNN), pp. 1333-1338, 2016.

[5] R.Hong, M. Chuan He, Yong Ge, and X. Wu, "User vitality ranking and prediction in social networking services: a dynamic network perspective," IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 6, pp. 1343-1356, June 2017.

[6] I. Smith, "A parallel artificial neural network implementation," Proceedings of The National Conference On Undergraduate Research (NCUR, pp. 1-4, April 2006.

[7] HS.Kisan, HA.Kisan, and AP.Suresh, "Collective intelligence and sentimental analysis of twitter data by using standfordnlp libraries with software as a service saas," Proceeding of 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-4, 2016.

[8] R.Krchnavy, M.Krchnavy, and M. Simko, "Sentiment analysis of social network posts in slovak language," 2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), pp. 20-25, July 2017.

[9] D. Cenni, G. P. Paolo Nesi, and I. Zaza, "Twitter vigilance: a multi-user platform for cross-domain twitter data analytics, nlp and sentiment analysis," , Proceeding of 2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (Smart World / SCALCOM / UIC / ATC / CBD Com / IOP / SCI), pp. 1-8, Aug 2017.

[10] F. Nausheen and S. H. Begum, "Sentiment analysis to predict election results using python," Proceeding of 2018 2nd International Conference on Inventive Systems and Control (ICISC), pp. 1259-1262, Jan 2018.

Authors Profile

Mr Satpalsing Rajput , pursing PHD in Computer Science from NMU. Completed ME in Computer Science from NMU in the year 2013. Completed BE in Computer Engineering in the year 2008.Having 10 years of experience as Industrial as well as Academic

Ms Kajal Borole,pursing ME in Computer Science from NMU. Completed BE in Computer Engineering from NMU in the year 2016 .

