AN ENHANCE INCREMENTAL PTTP ALGORITHM WITH I2MAP REDUCE FOR MINING EVOLVING BIG DATA IN HOSPITAL QUEUING MODEL

S.Vetivel
PG Scholar
Hindustan Institute of Technology,
Chennai.

Dr. R.Krishnaveni
Professor
Hindustan Institute of Technology,
Chennai.

ABSTRACT- Patient overcrowding and waiting delays are one of the major challenge faced by hospitals. For this there is a need of effective queue management system to reduce delays. Unnecessary waiting for long period results in substantial human resource and time wastage and increase the frustration faced by patients. For each patient in the queue, the total treatment time of all the patients before him is the time that he must wait. It would be convenient if the patients could receive the most efficient treatment plan and know the predicted waiting time through a mobile application that updates in real time. In this project we proposed Efficient Patient Treatment Plan (EPTP) prediction algorithm to predict the average waiting time for each treatment task. Based on these predicted data's a new model Efficient Queuing System (EQR) is modeled to find how much time patient has to wait. Real time datasets are collected from various hospitals and combined through hadoop ecosystem. For analyzing this complex data's in this project we are using map reduce framework to select the necessary fields based on key value pairs and the average waiting time is calculated. By using random forest concept for designing EQR and the future treatment time will be predicted in hospitals.

Keywords: *EQR, EPTP, Map Reduce, Random Forest.*

1. INTRODUCTION

1.1 GENERAL

Recent technological advancements have led to deluge of data from distinctive domains (eg Health care and scientific sensors, user generated data, internet and financial companies, and supply chain systems) over the past two decades. The term big data were coined to capture the meaning of this emerging trend. Big data is an evolving term that describes any voluminous amount of structured, semi structured and unstructured data that has the potential to be mined for information. Although big data doesn't refer to any specific quantity, the term is often used when speaking about pet bytes and Exabyte of data, much of which cannot be integrated easily. Because big data takes too much time and costs too much money to load into a traditional relational database for analytics, new approaches to storing and analyzing data have emerged that rely less on data schema and data

quality. Instead, raw data with extended metadata is aggregated in a data lake and machine learning and artificial intelligence(AI) programs use complex algorithms to look for repeatable patterns.

Big data analytics often associated with cloud computing because the analysics of large data sets in real time requires a platform like hadoop to store large data sets across a distributed cluster and map reduce to coordinate, combine and process data from multiple sources, big data is the ocean of information we swim in every day vast zeta bytes of data flowing from our computers, mobile devices, and machine sensors. with the right solutions, organizations can dive into all data and gain valuable insights that were previously unimaginable and to discover how big data technologies and analysis tools can transform your business today.

II. PROPOSED MODEL

In the proposed EPTP model we are using map reduce framework for processing and finding the average of treatment time based on the combination of age and type of treatment for Efficient Patient Treatment Plan (EPTP). For Efficient Queuing System (EQR) is modeled using an enhanced random forest algorithm with classification and regression tree algorithm to predict the future time. The predicted waiting time of each treatment task is obtained by which is the sum of all patients' probable treatment times in the current queue. The waiting time for the patients can be obtained by collecting realistic patient data are analyzed carefully and rigorously based on important parameters, such as patient treatment start time, end time, patient age, and detail treatment content for each different task by using Classification and regression tree. Least waiting time can be obtained from each patient and treatment recommendation with an efficient and convenient treatment plan and recommended.

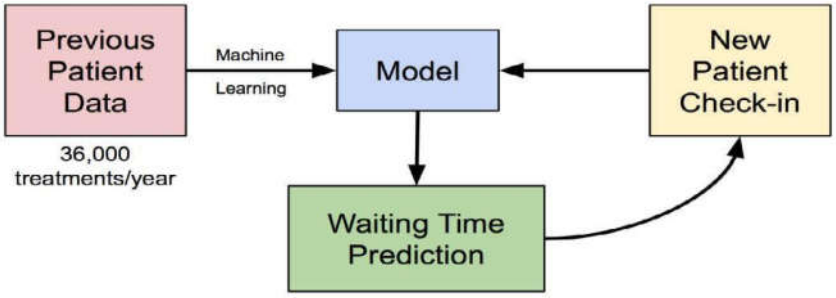


Figure 1: Proposed EPTP model

III. REVIEW OF LITERATURE

A few other recent papers employ statistical learning in wait time prediction. For example, Balakrishna et al. (2014) and Simaiakis and Balakrishnan (2014) estimate aircraft taxi-outimes (the time from gate departure to tak) using reinforcement learning and regression trees, respectively. Senderovich et al. (2014) find that, in a call center, k-means clustering (predicting the wait time for a new arrival based on the average wait for customers that arrived while the queueing system was in a similar state, in the past) is less accurate than simply using the wait time of the last customer to enter service. Providing wait time information can bene_cially inuence customer behavior, according to Armony and Maglaras (2004), Jouini et al. (2011), Allon et al. (2011) and literature surveyed in those papers. For example, in a call center model, providing an accurate wait time prediction and a call back" option shapes arrivals so as to reduce customers' mean and worst-case waits (Armony and Maglaras 2004).

Empirical evidence from call centers suggests that providing the expected wait time for arriving customers increases the immediate abandonment (balking) rate when the system is congested (Mandelbaum and Zeltyn 2009, Yu et al. 2014), whereas providing simple information (that the wait time is low, medium or high) reduces abandonment rates for all levels of the message (Yu et al. 2014). To beneficially inuence patient behavior in the ED setting, Dong et al. (2015) call for accurate ED wait time prediction, in order to reduce overall waiting by balancing the load among nearby EDs, in a manner similar to the ambulance diversion studied by Deo and Gurvich (2011), Yu et al. (2014) and Xu and Chan (2014).

By analyzing search engine data and the rolling average wait times on websites of 211 U.S. hospitals, Dong et al. conclude that patients increasingly are searching for ED wait times, and are using that information in choosing which local hospital's ED to go to. Unfortunately, however, the rolling average wait times published by hospitals are not informative about the current state of the ED (Armony et al. 2012) so can increase overall waiting by causing patients to choose the wrong hospital (Dong et al. 2015). Another caveat is that a long predicted wait might cause people who should go to the ED to choose not to; thus, publishing the current expected wait time might increase a hospital's profitability but reduce social welfare (Plambeck and Wang 2012).

Batt and Terwiesch (2015) conjecture that ED wait time prediction may reduce the rate at which patients leave without being seen. Empirically, in an ED that does not provide a wait time prediction, Batt and Terwiesch (2015) observe that to the extent that the ED is crowded, patients are more prone to leave without being seen. Patients might wrongly infer that the wait is long by focusing on the queue and not the processing rate, like the retail customers studied by Lu et al. (2013).

The **objectives** of the work are

1. Processing of High dimensional Unstructured patient and business data's produced by hospitals to create a meaningful information for datasets without any loss of information.

2. To calculate the time consumption of patients waiting time based on the tasks under various circumstances.

3. To predict the strict time requirements for the patients processing time during treatment.

IV. MATERIALS AND METHODS

Extensive literature reviews, case studies and discussions with medical experts show that there are number of factors influencing patient treatment time. These factors are identified and taken as attributes for this study. The proposed EPTP algorithm works with modified form of random forest algorithm to predict the treatment waiting time for different treatment tasks based on type of treatment and the type of treatments.

4.1 Random Forest Algorithm

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

The first algorithm for random decision forests was created by Tin Kam Ho using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

An extension of the algorithm was developed by Leo Breiman and Adele Cutler, and "Random Forests" is their trademark. The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho and later independently by Amit and Geman in order to construct a collection of decision trees with controlled variance.

4.2 Data Source:

The data for this study was collected from a local hospital in Chennai, consisting of treatment time for patients with different tasks and different age group data and they are preprocessed to suit this research.

This data consists of 9 attributes such as Patient ID, Age, Gender, Task Name, Department Name, Doctor Name, Date, Starting Time and End Time. These attributes are used to train and develop the system and a part is used to test the significance of the system. These attributes play an important role in diagnosing cancer in all

the cases. This data is stored in a knowledge base which has the ability to expand itself as new data enters the system through front end from which new knowledge is gained and thus the system becomes intelligent.

V. EXPERIMENTAL RESULTS

The experimental results of EPTP algorithm is performed using map reduce algorithm in hadoop framework. The data sets are preprocessed using rapid miner tool and the important features are identified are loaded into the hadoop framework. The input jobs are distributed in the hadoop framework and the average waiting time for the different tasks are identified. These tasks are given into the proposed Efficient Quening System works with Random forest algorithm to create a decision about the future predicted time. Once the average time will be predicted the data's are given into r-programming which predicted the future time consumption for the upcoming patients sucessfuly

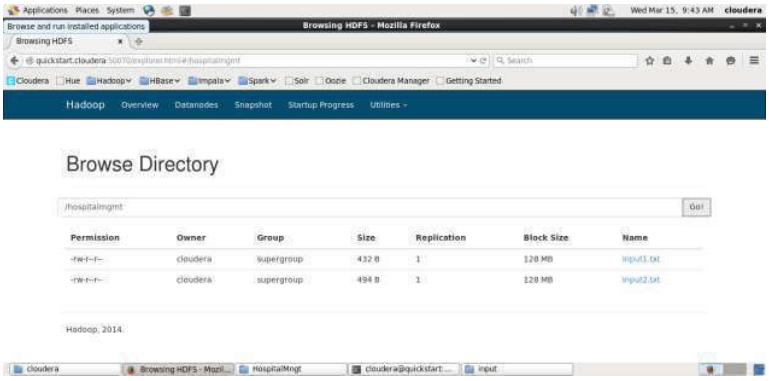
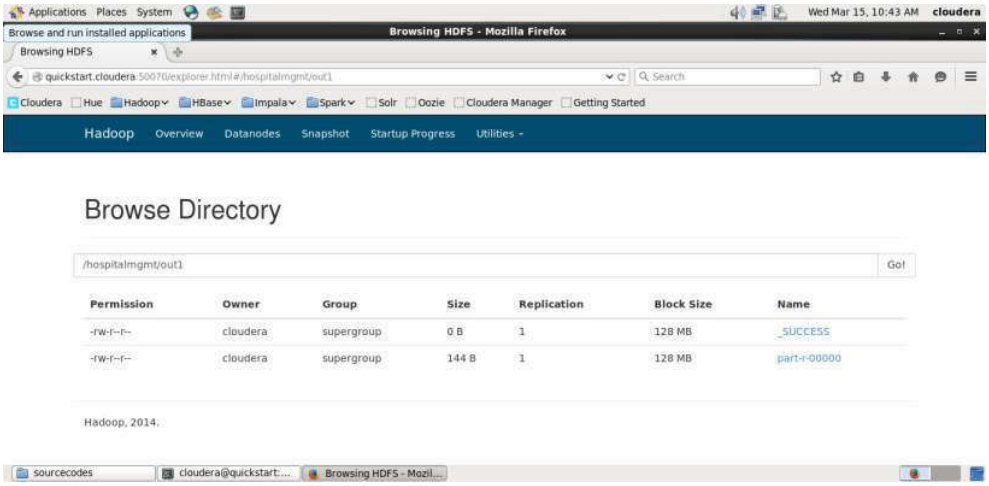
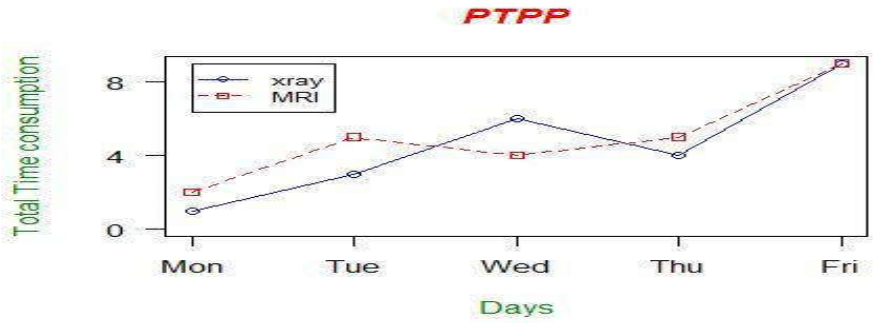


Figure 2: represents the input files are loaded into the hadoop directory

Figureure 3: represents the sucessfull completion of average time calculation in map reduce framework



Figureure 4: represents the output directory created in map reduce framework



Figureure 5: represents the average time for tasks between the days and total time consumption

VI. CONCLUSION AND FUTURE WORK

In this work a novel method using random forest algorithm to predict the patient treatment time. The most effective way to reduce the time delay in hospitals are identified by the tasks completion earlier.This prediction system may provide easy and a cost effective way for identifying the time periods and may play a pivotal role in the earlier prediction and provide effective preventive strategy. This system can also be used as a source of record with detailed patient history in hospitals as well as help patients to make necessary arrangements. In future we are planning to build an android app for giving instant waiting time for patients at a time of appointment itself.

References

[1] JIANGUO CHEN1,2 KENLI LI1,2 ZHUO TANG1,2 KASHIF BILAL3,4, AND KEQIN LI1,2,5,"A Parallel Patient Treatment Time Prediction Algorithm and Its Applications in Hospital Queuing-Recommendation in a Big Data Environment", April 25, 2016 Digital Object Identifier 10.1109/ACCESS.2016.2558199.

[2] S. Tyree, K. Q. Weinberger, K. Agrawal, and J. Paykin, ``Parallel boosted regression trees for Web search ranking," in *Proc. 20th Int. Conf. World Wide Web (WWW)*, 2012, pp. 387_396.

[3] N. Salehi-Moghaddami, H. S. Yazdi, and H. Poostchi, ``Correlation based splitting criterionin multi branch decision tree," *Central Eur. J. Comput. Sci.*, vol. 1, no. 2, pp. 205_220, Jun. 2011.

[4] G. Chrysos, P. Dagritzikos, I. Papaefstathiou, and A. Dollas, ``HC-CART: A parallel system implementation of data mining classi_cation and regression tree (CART) algorithm on a multi-FPGA system," *ACM Trans. Archit. Code Optim.*, vol. 9, no. 4, pp. 47:1_47:25, Jan. 2013.

[5] N. T. Van Uyen and T. C. Chung, ``A new framework for distributed boosting algorithm," in *Proc. Future Generat. Commun. Netw. (FGCN)*, Dec. 2007, pp. 420_423.

[6] Y. Ben-Haim and E. Tom-Tov, ``A streaming parallel decision tree algorithm," *J. Mach. Learn. Res.*, vol. 11, no. 1, pp. 849_872, Oct. 2010.

[7] L. Breiman, ``Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5_32, Oct. 2001.

[8] G. Yu, N. A. Goussies, J. Yuan, and Z. Liu, ``Fast action detection via discriminative random forest voting and top-K subvolume search," *IEEE Trans. Multimedia*, vol. 13, no. 3, pp. 507_517, Jun. 2011.

[9] C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes, ``Robust and accurate shape model matching using random forest regression-voting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1862_1874, Sep. 2015.

[10] K. Singh, S. C. Guntuku, A. Thakur, and C. Hota, ``Big data analytics framework for peer-to-peer botnet detection using random forests," *Inf. Sci.*, vol. 278, pp. 488_497, Sep. 2014.

[11] S. Bernard, S. Adam, and L. Heutte, ``Dynamic random forests," *Pattern Recognit. Lett.*, vol. 33, no. 12, pp. 1580_1586, Sep. 2012.

[12] H. B. Li, W. Wang, H. W. Ding, and J. Dong, ``Trees weighting random forest method for classifying high-dimensional noisy data," in *Proc. IEEE 7th Int. Conf. e-Business Eng. (ICEBE)*, Nov. 2010, pp. 160_163.

[13] G. Biau, ``Analysis of a random forests model," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 1063_1095, Apr. 2012.

[14] S. Meng, W. Dou, X. Zhang, and J. Chen, ``KASR: A keyword-aware service recommendation method on MapReduce for big data applications,"*IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 12, pp. 3221_3231, Dec. 2014.

[15] Y.-Y. Chen, A.-J. Cheng, and W. H. Hsu, ``Travel recommendation by mining people attributes and travel group types from community-contributed photos," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1283_1295, Oct. 2013.

[16] X. Yang, Y. Guo, and Y. Liu, ``Bayesian-inference-based recommendation in online social networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 4, pp. 642_651, Apr. 2013.

[17] G. Adomavicius and Y. Kwon, ``New recommendation techniques for multicriteria rating systems," *IEEE Intell. Syst.*, vol. 22, no. 3, pp. 48_55, May/Jun. 2007.

[18] G. Adomavicius and A. Tuzhilin, ``Toward the next generation ofrecommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734_749, Jun. 2005.

[19] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, ``Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97_107, Jan. 2014.
[20] Apache. (Jan. 2015). *Hadoop*. [Online]. Available: http://hadoop. apache.org

[21] Apache. (Jan. 2015). *Spark*. [Online]. Available: http://spark-project.org

[22] J. Dean and S. Ghemawat, ``MapReduce: Simpli_ed data processingon large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107_113, Jan. 2008.

[23] M. Zaharia *et al.*, ``Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," in *Proc. USENIX NSDI*, 2012, pp. 1_14. 1782 VOLUME 4, 2016.

24] Apache. (Jan. 2015). *Mahout*. [Online]. Available: http://mahout. apache.org.

[25] Y. Xu, K. Li, L. He, L. Zhang, and K. Li, ``A hybrid chemical reaction optimization scheme for task scheduling on heterogeneous computing systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 12, pp. 3208_3222, Dec. 2015.

[26] K. Li, X. Tang, B. Veeravalli, and K. Li, ``Scheduling precedence constrained stochastic tasks on heterogeneous cluster systems," *IEEE Trans. Comput.*, vol. 64, no. 1, pp. 191_204, Jan. 2015.

[27] D. Dahiphale *et al.*, ``An advanced MapReduce: Cloud MapReduce, enhancements and applications," *IEEE Trans. Netw. Service Manage.*, vol. 11, no. 1, pp. 101_115, Mar. 2014.

[28] M. Zaharia *et al.*, ``Fast and interactive analytics over hadoop data with spark," in *Proc. USENIX NSDI*, 2012, pp. 45_51.