International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

A Content Level Based Deduplication on Streaming Data using Poisson Process Filter Technique (PPFT)

A. Sahaya Jenitha, V. Sinthu Janita Prakash

Abstract: ***This paper proposed on Poisson process-based algorithm is to carry out content-level deduplication for the streaming data. Since Poisson processes are meant to do the counting of different events happening over a period of time and space, it becomes appropriate to use it for identifying duplications of data as it gets streamed based on time and space, which can allow the deduplication process to be carried out in tandem. Some of the research on deduplication has been focusing on File-level and Block-level deduplication while the focus can be brought to content-level, as data get streamed lively and becomes more dynamic. With this approach, the content-level deduplication will allow the data to be scanned intelligently and at the same time, it can save the deduplication operation time. Also, streaming data has its randomness which is innately there and by having Poisson process based deduplication it will address the random behaviour of the data transfer and can work efficiently in the dynamically connected environment. The proposed method identifies the unique data to store in the Database. Based on the experimental result, the Poisson Process-based algorithm produce 0.912 Area Under Curve (AUC) accuracy on real-world streaming data, which means that if AUC is greater than 0.8 then the performance of algorithm is pretty good. So, the machine intelligence-based deduplication model produced reliable***
and robust deduplication on streaming data compared to existing approaches.

Keywords: ***Deduplication, Poisson Process, Semantic level, Classification Algorithms, Streaming Data***

I. INTRODUCTION

Data deduplication is a process of eliminating the redundant data in a system, typically it is meant for improving effective storage utilization. During the deduplication process, it identifies an extra copy of already existing data in a data set /storage medium, delete the extra copy, leaving only one copy of the data to be stored. Data deduplication is needed, for the following reasons. They are,

1. To reduce the utilization of storage and improve efficiency in Database.
2. For Accuracy and to Reduce cost.
3. To reduce I/O disk operation.

Manuscript published on 30 September 2019
* Correspondence Author
Mrs. A. Sahaya Jenitha, Associate professor in the department of computer science, Cauvery College for Women, Tiruchirappalli, India.
Dr. Janita pursed, Professor and Head in the PG & Research Department of Computer Science, Cauvery College for Women, Tiruchirappalli.

File-level and Block-level Deduplication are already existing mechanisms to handle the duplicates in the storage medium. File-level deduplication is simple to use, implementation and works at the file level by eliminating duplicate files, it takes fewer resources and thus may be deployed over large amounts of the physical storage medium and it's also inexpensive to maintain [1].

Block-level deduplication is much more efficient and more reliable compared with file-level deduplication and it works at fixed-size block or a variable-sized block by eliminating duplicate blocks. It can eliminate chunks of data smaller than its file, these mechanisms are focused on the size of data like file, sub-file, chunk, block, byte and bit and it's not able to address the content of a dataset [2]. So, this kind of algorithms is not able to provide an optimized solution for a specific scenario like content level deduplication of streaming data [3]. In this paper, the Poisson process Algorithm, which is a stochastic process that counts the number of occurrences of a specified event over a period of time is proposed. Here, the Poisson process is introduced into the content-level data deduplication for the streaming data. The major complexity of deduplication on streaming data is its behaviour, which means that streaming data are dynamically changing over a time period (randomness), it's very difficult to handle, because it needs a more intelligent way like Poisson Process (Stochastic) to carry out the deduplication at content level of data which saves crucial data process time and brings forth effective deduplication mechanism. The overall Workflow process and deduplication is discussed. In Section 2, the related work is discussed and Section 3 focuses on Proposed Poisson Process work and Algorithm. In Section 4, Evaluation with datasets are discussed with AUC curve, finally, Section 5 ends up with Discussions and Conclusions for future work.

II. STUDIES ON DEDUPLICATION

Data process is based on Data efficiency which means it should process very fast and avoid the redundancy and improve the throughput of streaming [4]. In the real world, most of the applications have a large amount of data, its cost on storage and process to reduce the capacity of storage and querying time on database [5]. Data duplication is an optimistic technique to reduce the energy or resources conception and improve the storing capacity in the database [6]. Deduplication process will reduce the operations costs, lack of quality in data and performance degradation [7].

A. Poisson Process to Identify the Patterns

It's kind of anomaly detection to avoid duplicates and classification algorithm like Support Vector Machine (SVM) helps to classify data for detecting the anomaly in the deduplication process [8]. The Poisson process is accountable for finding anomalous event occurrences at a certain period and it can track the activity of each event at a time [9].

Collection of the Poisson process can predict the incidence of the same event on particular time limits and find the hidden states over the data for reorganization whether it is duplicate or not [10].

B. Semantic and Machine Learning approach

The semantic level is kind of grammar to focus on the content of data for addressing whether data is meaningful or not, based on this process, data deduplication will be carried out [11]. Classification algorithm like Support Vector Machine (SVM), k-means, k- Nearest Neighbour (kNN) are used to identify the duplicate records in the digital cognize. It classifies the geographical data whether is it duplicate or not [12]. Binary classification mechanism like Random forest, Decision trees are another way of classifying and identifying the duplicate data based on textual similarity [13]. An enhanced content level de-duplication algorithm identifies the encoded duplicate data with reliability and makes prevention on them in a cloud environment [14]. By introducing progressive sampled indexing, grouped mark and sweep mechanism on single-node to improve the scalability of an algorithm to provide robust deduplication with efficiency and achieve high throughput [15]. A stochastic based framework with EM algorithm to determine the threshold of record linkage model through machine learning algorithms like decision trees, bootstrap aggregating, ada boost, neural nets and support vector machines. Through this paper, it performs the comparison to identify the patterns to reduce the duplicate or make the linkage to improve the throughput of time and memory [16].

III. PROPOSED WORK

Content level deduplication is a proposed efficient method among the already existing approaches in the real world, because it checks the duplicate data at semantic level and it optimize the time and storage complexity of storage medium. This proposed approach, focuses on content level deduplication on streaming data
through Poisson Process Filter based Technique (PPFT). Poisson process has its own intelligence to solve such kind of problems where it happens rarely in the real-world scenario. So, here it solves data deduplication problem where it is based on some characteristic called content or semantic level checking to avoid duplicate data.

Below Fig: 1 will elaborate on deduplication Mechanism using a **Poisson process Filter** approach (PPFT).

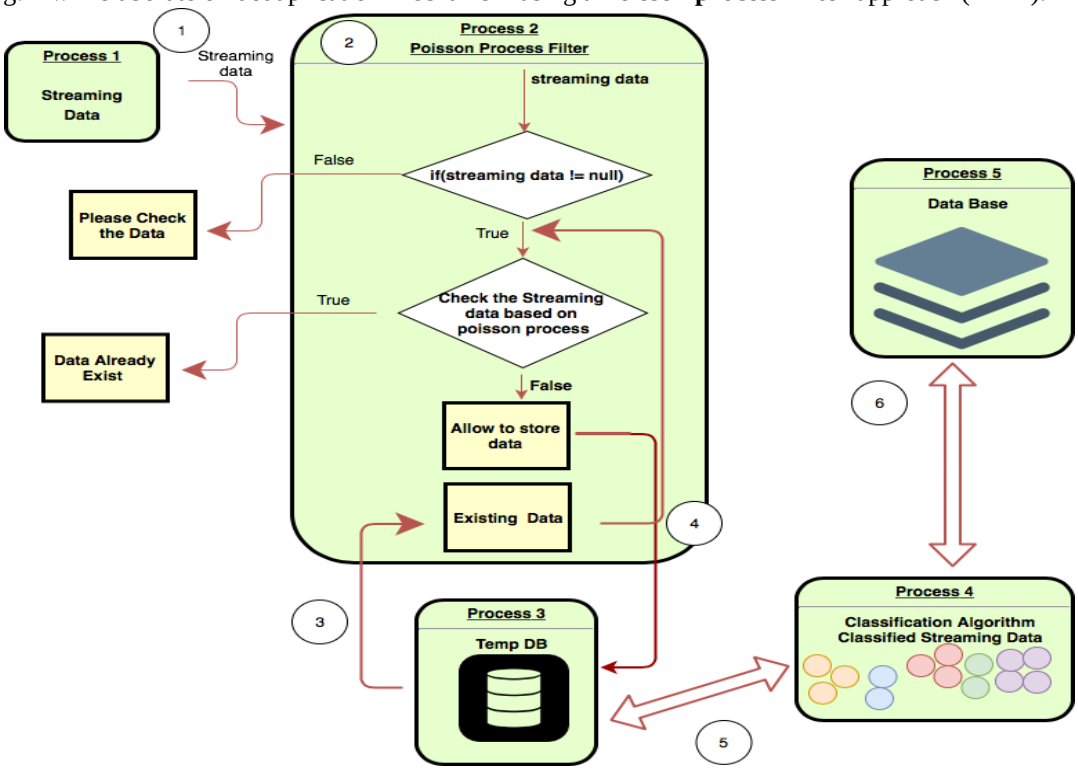

Fig 1: Concept of Deduplication

Process 1: Streaming Data

It continuously generates data from different sources, streaming mechanism helps to produce the streaming data as an input to the system.

Process 2: Poisson Process Filter Technique (PPFT)

PPF checks the number of occurrence same data blocks with the help of temporary Data Base. It acts as a filter to restrict the duplicate file based on the contextual level data deduplication.

It uses the Poisson process to detect the duplicate file with the help of Temp DB and classification algorithm. Streaming data is checked where it is null or not then if is it not null then go to the Poisson process to identify the duplication otherwise it gives an intimation to the system to check the data. In the Poisson process, events are independent of each other. So, the occurrence of one event does not affect the probability another event will occur. The events per time period are constant.

Two events cannot occur at the same time. By these characteristics, Poisson process checks the incoming streaming data with already existing data with help of temp database and classification algorithm model.

If it does not exist, then it stores into temp database otherwise it provides an intimation to the system that "Please check the data" or "Data Already Exist". Here, poison process act as a major role to avoid data duplication. Below formula is for Poisson process,

$$P[N(t) = n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!} \quad \text{for } t \geq 0 \text{ and } n = 0, 1, 2, \ldots$$

n: the exact number of successes
e: Constant equal to approximately 2.71828
λ: mean of successes in the given time interval

Process 3: Temp DB

Temp database stores already existing data from main or cloud database via classification algorithm model and send to PPF to perform data deduplication. And also, it receives new data from PPF and stores temporarily then send to the classification algorithm model to perform classification.

Process 4: Classification Algorithm model

In this process, after checking streaming data received from Temp DB it classifies into number of groups to store in cloud by using any one following classification algorithms like Random Forest, SVM, k-means or kNN [5][9]. It makes data in an organized way to store in the cloud and it is very helpful when it retrieves from the cloud to perform deduplication.

Process 5: Database

Finally, the streaming data will store here and provide data to each phase in the system.

Proposed Algorithm (PPFT)

Input:- String S, S is a streaming data, **string E**, E is an existing data already available in Temp DB

Output:- Deduplicated data is received

1: Declare Variable
2: Initialize variable
3: Read Streaming data and assign into variable S
4: If S != Null then
goto step 5 (Get into Poisson check with existing data).
Else
Check the given streaming data.
End if
5: Check the streaming data based on Poisson process
If E != Null then
If S == E then
goto step 4
Else
temp db ← S
classification model ← temp db
db ← classification model
End if
Else
classification model ← db
temp db ← classification model
E ← temp db
goto step 5
End if
6: End

IV. EVALUATION-EXPERIMENTAL SETUP

In this approach, the evaluation of algorithm is carried out by applying some statistical analysis approach like AUC (Area Under Curve), ROC Curves (Receiver Operating Characteristic) to find the accuracy of algorithm. It uses the video game streaming data set for deduplication and it is stored as a file which can be easily handled by the algorithm. By exploring this dataset, it contains 18625 rows with IGN's (**I**n **G**ame **N**ame) score, so it makes to identify the current trend about the gaming industry. Below Table:1 has the following details like User ID, Name of the steam game, Behaviour name (Purchase/Play) and Hours.

User ID	Name of the Steam Game	Behavior Name	Hours
151603712	The Elder Scrolls V Skyrim	purchase	1
151603712	The Elder Scrolls V Skyrim	play	273
151603712	Fallout 4	purchase	1
151603712	Fallout 4	play	87
151603712	Spore	purchase	1
151603712	Spore	play	14.9
151603712	Fallout New Vegas	purchase	1
151603712	Fallout New Vegas	play	12.1
151603712	Left 4 Dead 2	purchase	1
151603712	Left 4 Dead 2	play	8.9
151603712	HuniePop	purchase	1
151603712	HuniePop	play	8.5
151603712	Path of Exile	purchase	1
151603712	Path of Exile	play	8.1
151603712	Poly Bridge	purchase	1
151603712	Poly Bridge	play	7.5
151603712	Left 4 Dead	purchase	1
151603712	Left 4 Dead	play	3.3
151603712	Team Fortress 2	purchase	1
151603712	Team Fortress 2	play	2.8
151603712	Tomb Raider	purchase	1
151603712	Tomb Raider	play	2.5

Table 1: Streaming Data

Below Fig: 2 shows the accuracy or goodness of a proposed algorithm in train data set,

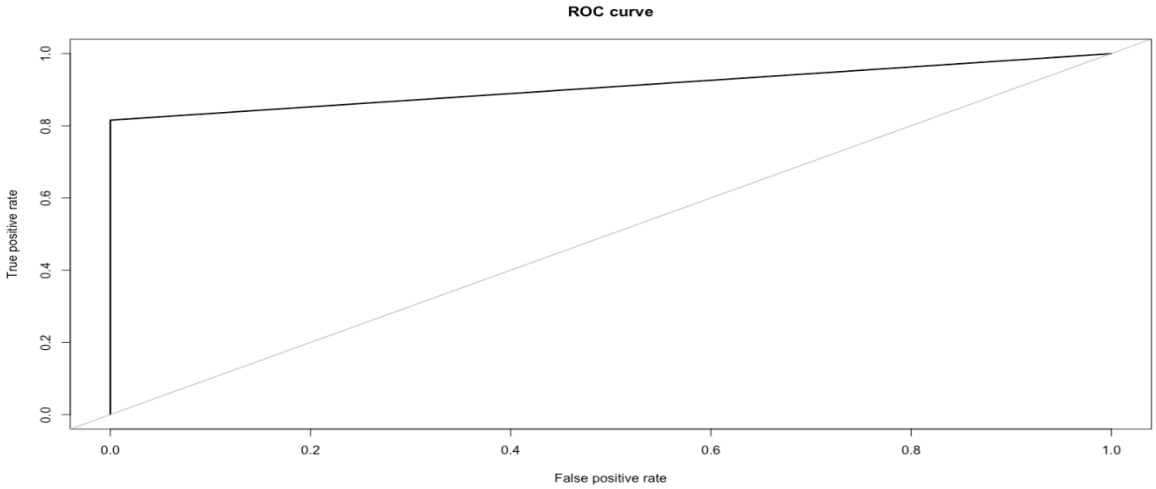

Fig 2: ROC curve on train data (AUC: 0.908)

Below Fig 3: shows the accuracy or goodness of the proposed algorithm on test data set,

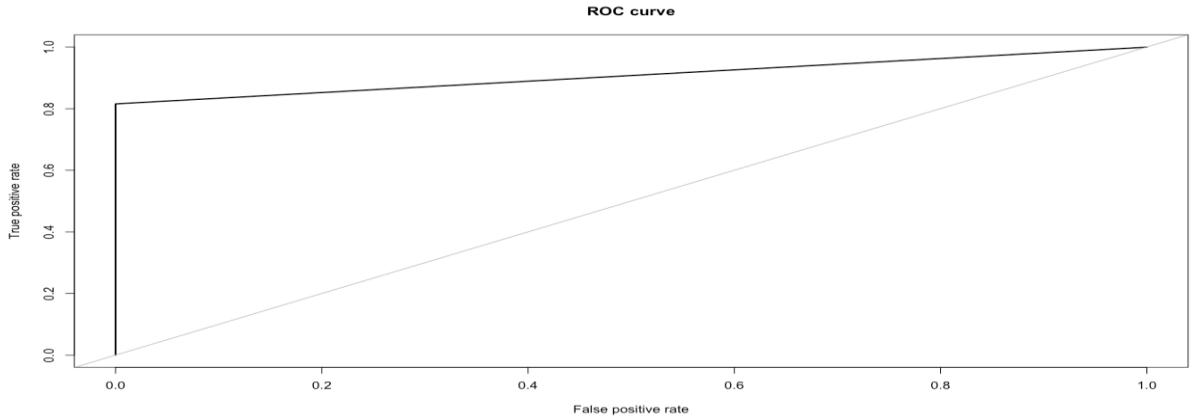

Fig 3: ROC curve on test data (AUC: 0.912)

So, Performance (throughput) of each existing algorithm like Random forest and k-means is low, when compared with the proposed approach PPFT. The accuracy and the Errors of the proposed algorithm is compared with the existing classification algorithms and its variations is shown in fig :4 below and comparison shown in Table:2

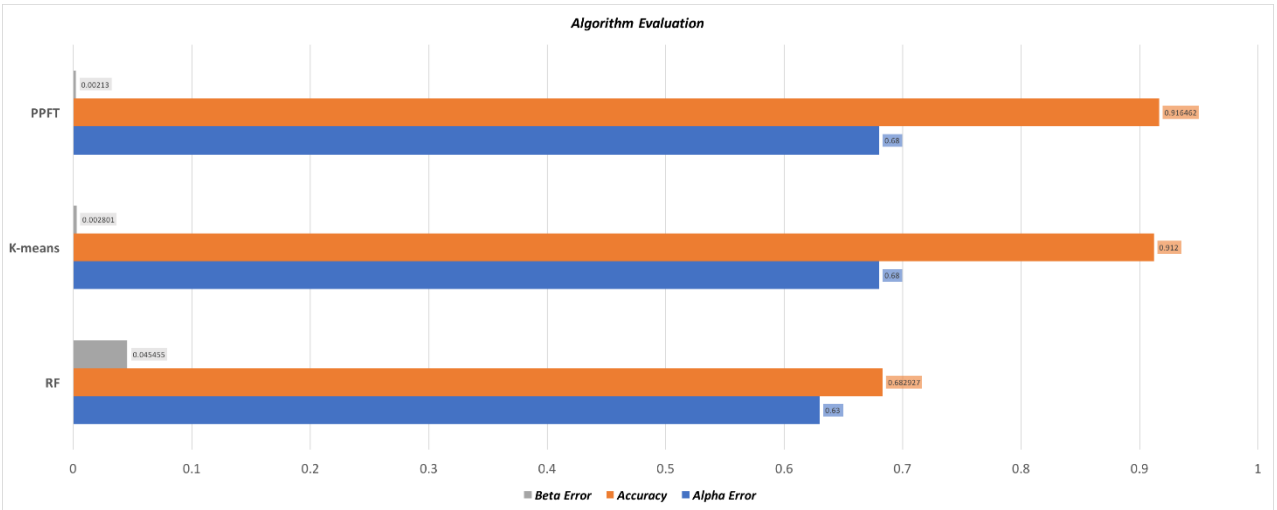

Fig 4: Algorithm Evaluation

S.no	Algorithm		Alpha Error	Beta Error	Accuracy
1.	**Random forest-based model**		0.631579	0.045455	**0.682927**
2.	**K-means based algorithm model**		0.68	0.002801	**0.9124005**
3.	**PPFT (Stochastic based approach)**	**Train dataset**	0.631579	0.000000	0.707317
		Test data set	0.680000	0.000000	**0.916462**

Table: 2 Comparison between Existing and Proposed algorithms

V. RESULTS AND CONCLUSION

Data deduplication is a very important technique in the storage mechanism. It reduces the Storage cost, Access time and Produces Accuracy. File-level and block-level deduplication are already existing mechanism but in this proposed scheme done on content level using Poisson Process Filter Technique produces much Accuracy. So here streaming data are random events and system does not know which one comes first .so that, it need Poisson process to find the data whether it is already existing or not with the help of classification algorithm. Classification algorithm used to classify the streaming data into number of groups to perform Content level deduplication which helps to store in the database. According to this paper, Deduplication on streaming data at the semantic level is performed. The streaming data can be of any type, but it should be checked at time instance, since all data are classified and stored in the database. Existing Methods like File-level and Block-level deduplication, did not look at content or semantic level of available data and it does not suit for random event like streaming data. In this paper, PPFT did data deduplication on streaming data in future it can be further optimized by using Machine Intelligence.

REFERENCES

1. A.K. Elmagarmid, P.G. Ipeirotis, and V.S. Verykios, "Duplicate Record Detection: A Survey", IEEE Trans. Knowledge and Data Eng., Vol: 19, Iss: 1, Pg:1-16, Jan. 2007.
2. N. Koudas, S. Sarawagi, and D. Srivastava, "Record linkage: similarity measures and algorithms," ACM SIGMOD International Conference on Management of Data, Pg: 802–803, 2006.
3. Peter Christen, "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication", IEEE Transactions On Knowledge And Data Engineering, Vol:24, Iss:9, September 2012.
4. Cheng, S. T., Chen, J. T., & Chen, Y. C. "Fast Deduplication Data Transmission Scheme on a Big Data Real-Time Platform". (2017).
5. Hou, Z., Chen, X., & Wang, Y. "Content-level deduplication on mobile internet datasets". In AIP Conference Proceedings (Vol: 1836, No: 1, Pg: 020086). AIP Publishing. (2017, June).
6. He, Q., Li, Z., & Zhang, X. "Data deduplication techniques". In 2010 International Conference on Future Information Technology and Management Engineering (Vol: 1, Pg: 430-433). IEEE. (2010, October).
7. Pande, N. A., & Ghuse, N. D. "Record Deduplication Approaches and Algorithm for Removing Duplicate Data".
8. Luca, S., Kars makers, P., & Vanrumste, B. "Anomaly detection using the Poisson process limit for extremes". In 2014 IEEE International Conference on Data Mining (Pg: 370-379). IEEE. (2014, December).
9. Ihler, A., Hutchins, J., & Smyth, P. "Adaptive event detection with time-varying Poisson processes". In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (Pg: 207-216). ACM. (2006, August).
10. Huang, H, "Recurrent Poisson Process Unit for Automatic Speech Recognition" (Doctoral dissertation, Hong Kong University of Science and Technology). . (2018).
11. Tan, Y., Jiang, H., Feng, D., Tian, L., Yan, Z., & Zhou, G. (2010, September). "SAM: A semantic-aware multi-tiered source de-duplication framework for cloud backup". 39th International Conference on Parallel Processing (Pg: 614-623). IEEE. 2010
12. Martins, B. "A supervised machine learning approach for duplicate detection over gazetteer records". In International Conference on Geospatial Semantics (Pg: 34-51). Springer, Berlin, Heidelberg. (2011, May).
13. Lazar, A., Ritchey, S., & Sharif, B. "Improving the accuracy of duplicate bug report detection using textual similarity measures". In Proceedings of the 11th Working Conference on Mining Software Repositories (Pg: 308-311). ACM. (2014, May).
14. Periasamy, J. K., & Latha, B. "An enhanced secure content de-duplication identification and prevention (ESCDIP) algorithm in cloud environment". Neural Computing and Applications, 1-10. (2019).
15. Guo, F., & Efstathopoulos, P. "Building a High-performance Deduplication System". In USENIX annual technical conference. (2011, June).
16. Sariyar, M., & Borg, A. "The Record Linkage package: Detecting errors in data".
17. The R Journal 2(2), Pg: 61-67. (2010).

AUTHORS PROFILE

Mrs. A. Sahaya Jenitha is presently working as Associate professor in the department of computer science, Cauvery College for Women, Tiruchirappalli, India. She has received her Bachelor degree in computer science in the year 1998 and Master's degree in 2001 and Master of philosophy in 2007. Her research interests include Neural Networks, Big Data, Machine Learning etc. She is life member of ISSE. She has 19 years of teaching experience.

Dr. Janita pursed Master of Computer Applications and Ph.D from Bharathidasan University. She is currently working as a Professor and Head in the PG & Research Department of Computer Science, Cauvery College for Women, Tiruchirappalli. She is a member of IEEE, Computer Society of India and a life member of ISSE. She has published more than 35 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE, Elsevier, ACM, Springer and it's also available online. Her main research work focuses on Communication Networks, Network Security, Big Data Analytics, Data Mining and Mobile Communications. She has 23 years of teaching experience and 14 years of Research Experience.