Indian Journal of Science and Technology, *Vol 10(26), DOI: 10.17485/ijst/2017/v10i26/115715, July 2017*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Modified Data Duplication Algorithm to Minimize the Redudancy of Data in Medical Database

M. Rameshkumar[1] and V. Lakshmipraba[2]

[1]Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli – 627012, Tamil Nadu, India; proframeshkumar@gmail.com
[2]Department of Computer Science, Rani Anna Government College for Women, Tirunelveli – 627008, Tamil Nadu, India; vlakshmipraba@rediffmail.com

Abstract

Objective: To secure and reduce the data duplication in medical databases using Modified Sliding and Windowing proposed method. **Method:** Here we proposed Modified Sliding and Windowing technique (MSW). It enables the users to command entry of outsourced information notwithstanding though the proprietorship changes progressively by misusing randomized merged encryption, secures possession aggregate key appropriation. This counteracts information spillage not exclusively to denied user despite the fact that they already possessed that information, additionally to a legit however inquisitive distributed storages servers. **Findings/ Results:** The MSW method gives higher Effectiveness and lesser delay for finding duplication. This method consumes less Memory and provide greater accuracy than the existing duplication algorithms. The comparison is done with NS2 (Network Simulator2) using different dataset. Hence this proposed research can be used to reduce the data duplication in advanced databases with lesser correlations and parameters. **Application/ Improvements:** Data duplication algorithm can be improved in future by adding more parameters with higher accuracy based on severe attacks in future.

Keywords: Data-Mining, Duplication, Navie Bayes Classifier, Privacy Mechanism

1. Introduction

Data-mining assist while the patient's populace expands the restorative database likewise developing each day. The exchanges and examination of patient information is mind boggling without the PC based investigation framework. The PC based investigation framework shows the mechanized therapeutic determination framework. This robotized conclusion framework bolster the therapeutic specialist to settle on great choice in treatments and diseases[1]. Data-mining is the monstrous regions for the specialists to taking care of the enormous measure of patient's informational collections from numerous points of view, for example, comprehend complicated symptomatic tests, deciphering past outcomes, and joining the diverse information together. Customarily Clinics choice is formed by the therapeutic specialist's perceptions and fore learning instead of the information which acquire from the substantial measure of information. This robotized finding framework prompts expands the nature of administration given to the patients and declines the restorative expenditure[2]. Protection is meant by the sensible security of information not the customary security of information e.g. get to control, robbery, hacking and so on. In this place, enemy utilizes authentic strategies to gather touchy data. Different databases are distributed e.g. Census data, Hospital records which allows researchers to effectively study the correlation between various attributes. Assume a clinic has some individual particular patient information which it needs to distribute with

*Author for correspondence

the end goal that, Information remains for all intents and purposes valuable and Identity of an individual can't be resolved. But there are chances that an adversary might infer the secret/sensitive data from the published database[3]. At the point when conventional securities, for example, get to control system is viewed as then a circumstance emerges where in approved client may trade off the protection of others prompting personality revelation. Today numerous associations began gathering and breaking down purchaser information to securely innovate their processing. To avoid identity disclosure it is necessary to satisfy certain privacy requirement. In[4] explores privacy by inconspicuous base and focus on reliable needs for individual authorizations. Improper processes were proposed to utilize the records to fulfill protection with minimum twist and smaller amount of items. In[5] concentrates the issue of fulfilling slow controls and security for user consents. Albeit numerous calculations have been proposed for accomplishing the same yet each has their own.

Healing facilities information volumes are expanding as the administration store and gather enormous measure of information for their own particular use in cloud. As per association technique gathering, by industry examination more associations want to store their information on to database. In this manner obliges association to have more stockpiling and expend more power and vitality for overseeing and dealing with the information, more system assets are used for transmitting the information and additional time is spend on capacities, for example, replication and information reinforcement. A large portion of the data that is put away is copy information, distinctive sources in similar associations for the most part make comparable records or copy documents that as of now exist by which they can work freely. On the off chance that it was conceivable, IT associations would just shield the remarkable information from their reinforcements. Rather than sparing everything more than once, the perfect situation is one where just the new or special substance is spared. Information de-duplication gives this fundamental ability. It offers the capacity to find and expel excess information from inside a dataset. A dataset can traverse a solitary application or traverse a whole association. Repetitive information components can be whole documents or sub-record information sections inside a record. In all cases, the goal of the de-duplication procedure is to store special information components just once, yet have the capacity to reconstitute all substance in its unique frame on request, with 100 percent unwavering quality at circle speeds[6].

1.1 Sorts of Information De-Duplication

1.1.1 Document Level De-Duplication

It is ordinarily known as single-case stockpiling, record level information de-duplication looks at a document that must be chronicled or reinforcement that has as of now been put away by checking every one of its properties against the list. The record is refreshed and put away just if the document is novel, if not than just a pointer to the current document that is put away references. Just the single case of document is spared in the outcome and applicable duplicates are supplanted by "stub" which focuses to the first record[7].

Duplicate detection, similarly known as entity matching has been a research topic for several decades. The challenge is to effectively and efficiently identify pairs of records that represent the same real world entity8.

1.1.2 Piece Level De-Duplication

Piece level information de-duplication works on the premise of sub-document level. As the name infers, that the document is being broken into portions pieces or lumps that will be inspected for beforehand put away data versus excess. The prevalent way to deal with decide excess information is by allocating identifier to lump of information, by utilizing hash calculation for instance – it creates a one of a kind ID to that identified piece. The identified one of a kind Id will be contrasted and the focal file. On the off chance that the ID is as of now present, now the at that point it speaks to that before just the information is handled and put away before .Thusly a pointer locator is spared to the previous information. In this case the ID not an existing but it is referred as new one. The denoted lump is put away and the referred ID is refreshed in the record of a File[8]. The piece size should be checked and shifts from seller to merchant. Some will have settled square sizes, while some others utilize variable piece sizes moreover few may likewise change the span of settled piece measure for purpose of confounding. Piece sizes of settled size may vary from 8KB to 64KB however the basic contrast with it is the littler the lump, than it will probably have chance to differentiate it as the copy information. In this case less information is put away than it clearly implies more noteworthy decreases in the information on file. The main important issue by utilizing settled size pieces

is that in the event that if the record is adjusted and the de-duplication result utilizes the same already assessed outcome than there will be shot of not distinguishing the same repetitive information fragment, as the pieces in the document would be moved or modified.

1.1.3 Variable Square Level De-Duplication

In[9] attempted to solve the issue of securing protection in miniaturized scale information distributing. Distributing information about people without uncovering delicate data about them is an imperative issue. k-secrecy and I-Assorted qualities has been beforehand utilized instrument for ensuring security however systems are lacking to ensure the protection issues like Homogeneity assault, Skewness Assault, Likeness assault and Foundation Learning Assault so another security measure called "(n, t)- vicinity" is proposed which is more adaptable model it accomplishes more protection and less utility.

In[10] made a novel bunching calculation for vertically parceled information; they test the execution of that calculation in view of investigations and many-sided quality examination. Later they introduced a private rendition of this convention utilizing conventions in view of homo transformed encryption. Our convention is strong against conniving assault.

In[11] utilized diverse way to deal with handle the substance mindful deduplication. In this strategy the information is considered as a protest. Approaching information is changed over into the protest and the same has been contrasted and the as of now put away questions for finding the copy information in successfully. Utilizing of the Byte level examination and the information of the substance of the information, the information document is part into vast information fragments.

In[12] the algorithms called SPC, DPC and FPC. Enable to understand the implementation result of apriori algorithm using Map reduce framework. For every datasets and cluster sizes the parallelization technique is suitable one.

In[13] employed the Dedoop, it is high power and performance based tool for de duplication using Hadoop. It was proved for larger datasets. This methodology implemented by browser specifications that with machine learning to generate match classifiers.

In[14] depicted a framework remarked CBLOCK, to point the de-duplication challenges. The CBLOCK framework is implemented to learn hash functions derived from attribute domains. The hierarchical tree structure constraints ware developed by blocking functions. The method was tested and the utility was proved.

In[15] applies break condition logic and marks the boundary of file. Chunk boundary is computed based on the fingerprint algorithm. File boundary is marked based on break condition. The issue with this approach is the lump measure. The extent of the piece can't be anticipated with this approach, yet it is conceivable to foresee the likelihood of getting a bigger lump or a littler one. This probability is fixed on the basis of the probability of getting a particular fingerprint. A divisor D and the sliding window size define if the probability is bigger or smaller.

2. Research Methodology

The main objective of the proposed work is to secure and reduce the data duplication of hospital management system after the classification of patient's records. In order to do this modified sliding and windowing algorithm is proposed which uses the piggybacking concept. The proposed framework is sub divided into three stages: privacy mechanism, classification and data duplication reduction.

2.1 Privacy Preserving Mechanism

Each Hospital needs to transfer persistent records with necessary fields for well required manner. A minimum important protection safeguarding access control system is utilized here too control component. The authorizations and control approach depend on determination predicates on the IQ traits. This method characterize the authorizations along with the precision of data destined for every inquiry. The determination of the imprecision bound guarantees that the approved information has the coveted level of precision. The imprecision bound data is not imparted to the user since knowing the imprecision bound can bring about abusing the protection prerequisite. The security assurance component is required to meet the protection necessity alongside the imprecision destined for every authorization.

2.2 Classifying the Data

Progressed Naive Bayesian conviction classifiers have been utilized as a part of numerous functional applications. They extraordinarily streamline the learning assignment by expecting that qualities are autonomous given the class.

Despite the fact that freedom of properties is an unreasonable suspicion, innocent Bayes classifiers regularly contend well with more advanced models, regardless of the possibility that there is humble relationship between's traits. Credulous Bayes classifiers have noteworthy focal points as far as effortlessness, learning speed, characterization speed, and storage room. They have been utilized, for instance, in content characterization and medicinal analysis. In (non-private) credulous Bayes taking in, the digger is given an arrangement of preparing cases. We expect that every illustration is a characteristic vector of a client together with her class mark. From these cases the excavator takes in a classifier that can be utilized to group new cases. Without loss of simplification, we accept all user' touchy credits should be secured. We have two objectives to accomplish:

- Correctness: the excavator takes in the gullible Bayes classifies precisely.
- Privacy: the excavator adapts nothing about every patient's delicate information with the exception of the learning resultant from the credulous Bayes classifies itself.

  P (Bn given An) = P (An and Bn)/P (A) to figure likelihood of A given B, the calculation tallies the quantity of situations where an and Bn happens together and partitions it by the quantity of situations where A happens alone. Give Y a chance to be an information dual, In Bayesian terms; Y is viewed as "Proof". Give H a chance to be some speculation, with the end goal that the information dual Y has a place with class C. P (H|Y) is back likelihood, of H adapted on X. In contract, P (H) is the earlier likelihood of H

  P (H|Y) =P ((H|Y) P (H))/ (P(Y))

  P (H|X) = (Liklihood*prior)/Evidence

2.3 Reducing the Data Duplication

Sorted Blocks is a speculation of adjusted blocking and windowing calculations for copy discovery. Sorted Blocks initially sorts the records in view of a sorting key. Like for the Sorted Neighborhood the suspicion is that records close in the wake of sorting have a higher likelihood of being copies. Be that as it may, rather than sliding a settled size window over all records, we make disjoint segments and look at all records inside these segments. It is attractive that the sorting keys are one of a kind to get an unambiguous sorting request. To this end, a bigger number of qualities can be incorporated for sorting than for really parceling the information. All things considered, uniqueness is not entirely essential; if there should arise an occurrence of a tie, we utilize the information request of the record.

To guarantee that likewise such copies can be discovered that are shut in the sorting request, however for any reason were allotted to various parcels, an extra segment cover is utilized. This cover is characterized by a physically chosen cover parameter α. It depicts the quantity of records in one parcel to be contrasted and records of the nearby segment. Inside the cover a settled size window with size α+1 is slid over the sorted information and all records inside the window are looked at.

2.3.1 Sorted Blocks Creation

The initial step makes another segment if the most extreme parcel size is come to. This implies, the new parcel is made autonomously of the dividing key. In spite of the fact that records have a similar parcel predicate, they are gathered in various allotments. In any case, because of the cover between the allotments, it is guaranteed that all records are contrasted and its ancestors and successors in the sorting request.

2.3.2 Windowing

This second step utilizes the most extreme parcel estimate as window size to slide a window over the records inside a segment. In the event that the most extreme number of records is reached, for each new record in the parcel, the principal component in the present window is evacuated. This iterative procedure keeps running until the finish of the segment is come to. In this manner, this variation is fundamentally the same as the Sorted Neighborhoods Method which utilizes the piggybacking which is isolated into two channels required for both forward and turned around exchange. Be that as it may, for this situation affirmation are included which squander the data transfer capacity of the turnaround totally

3. Experimental Analysis

In this section we compare the proposed Modified Sliding and Windowing (MSW) method with Incremental adaptive SNM (IA-SNM) and Accumulative Adaptive SNM (AASNM). The proposed algorithms verifies, the data sets are segregated based on its content. The segregated data are classified by

using number of keyword, selected based its feasibility to satisfy the generalized rule. The adopted rule is fine-tuned each iteration, which makes the proposed technique applicable for various spheres.

3.1 Effectiveness

The level of similarity of a framework focused on issue, checked against different informational collection is called as the viability of the framework. The powerful discovery of duplication prompts less demanding de-duplication handle along these lines diminishes the clamor content in the outcome. The proposed work is thought to be on MSW with alternate strategies in its viability measure. The Comparison of effectiveness between existing and proposed MSW algorithm is given in Figure1.

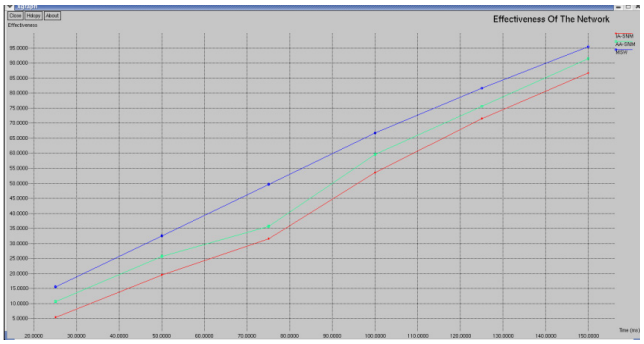

Figure 1. Comparison of effectiveness between existing and proposed algorithm.

3.2 Delay

Any implementation is considered to be efficient, when the time delay is less. The proposed MSW generation is taken experimentally to know the time taken to perform de-duplication. The Comparison of delay between existing and proposed MSW algorithm shown in Figure 2.

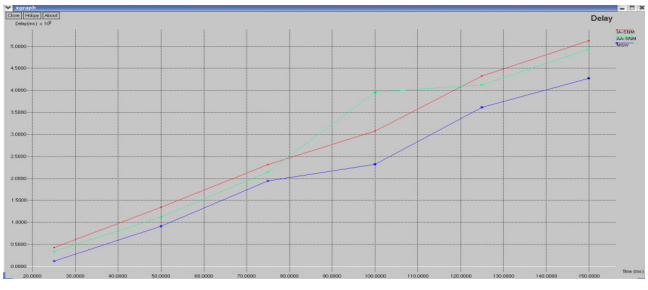

Figure 2. Comparison of delay between existing and proposed algorithm.

3.3 Memory Consumption

Memory consumption is calculated before file upload and after file upload. It is expected from the result that the memory space is increased when we upload the new file in database. But when duplicate file is detected by using hashing algorithm then there is no effect on memory space it is same as before. In this way by using De-duplication memory space is less consumed. The time delay differences between MSW algorithm with other algorithms shown in Figure 3.

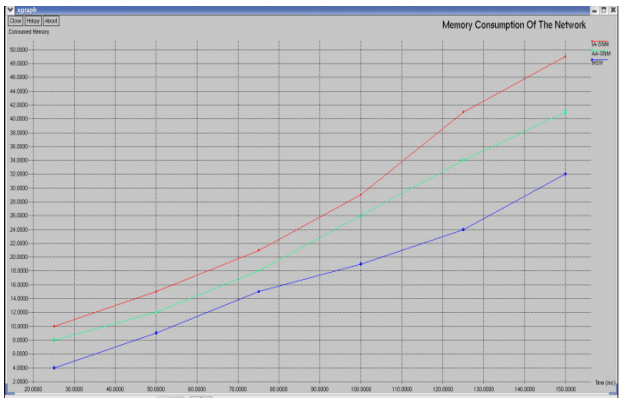

Figure 3. Memory consumption between existing and proposed algorithm.

3.4 Time and Accuracy

Accuracy shows that how accurately our system works to detect the duplicate files. From the below graph we can conclude that duplicator detect the duplicate file in less time and perform accurately. Detection time is a time taken to detect a duplicate file and it is also clear from the graph that it takes very less time to detect a file. The Figure 4 provides the Time and Accuracy between MSW with existing algorithm and means the MSW gives better accuracy. Table 1 illustrates the Analysis of various parameters with proposed MSW approach.

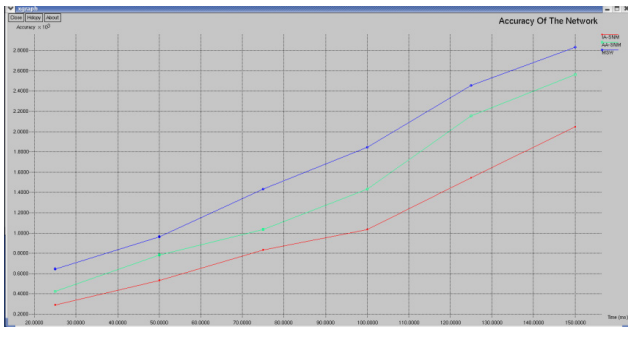

Figure 4. Accuracy calculation between existing and proposed algorithm.

Table 1. Analysis of various parameters with proposed approach

PARAMETER	IA-SNM	AA-SNM	MSW
Effectiveness	86.642%	91.456%	95.3841%
Delay	202.632ms	187.563ms	119.954ms
Memory consumption	49%	41%	32%
Time and accuracy	2046	2564	2831
Computation overhead	5125	4934	4275

4. Conclusion

Productive copy recognition is an essential assignment particularly in substantial datasets. In this paper, we have contrasted two strategies and changed blocking and windowing, for lessening the quantity of correlations. Also, we have presented piggybacking which is a speculation of blocking and windowing. Explores different avenues regarding a few genuine datasets demonstrate that Sorted Blocks beats the two different methodologies. A challenge for Sorted Blocks is finding the correct arrangement settings, as it has a larger number of parameters than the other two methodologies.

5. References

1. Sanakal R, Jayakumari T. Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine. International Journal of Computer Trends and Technology. 2014 May; 11(2):98–9. Crossref
2. Wang K, Chen R. Privacy-Preserving Data Publishing: A Survey of Recent Developments. ACM Computing Survey. 2010; 42(4):1–53. Crossref
3. Rizvi S, Mendelzon A. Extending Query Rewriting Techniques for Fine-Grained Access Control. Proceeding ACM SIGMOD International Conference Management of Data, 2004. p. 551–62. Crossref
4. Chaudhuri S, Sudarshan S. Fine Grained Authorization through Predicated Grants. Proceeding IEEE International Conference Data Engineering, 2007. p. 1174–83. Crossref
5. Elmagarmid KA, Ipeirotis PG. Duplicate Record Detection: A Survey. IEEE Trans. ON Knowledge and Data Engineering. 2007 Jan; 19(1):1–16. Crossref
6. Maddodi S, Girija AV. Data Deduplication Techniques and Analysis. 3rd International Conference on Emerging Trends in Engineering & Technology, 2010. p. 581–5. Crossref
7. Rohit A, Surajit C. Eliminating fuzzy duplicates in data warehouses. Proceedings of the 28th International Conference on Very Large Databases (VLDB), 2002. p. 586–97.
8. Draisbach U, Naumann F. DuDe: The duplicate detection toolkit Proceedings. International Workshop on Quality in Databases (QDB). 2010. p. 1–7.
9. Prakash M, Singaravel G. A new model for privacy preserving sensitive Data Mining. IEEE 3rd International Conference on Computing Communication & Networking Technologies (ICCCNT), 2012. p. 1–8. Crossref
10. Elmisery AM, Huaiguo Fu. Privacy preserving distributed learning clustering of healthcare data using cryptography protocols. 34th Annual Computer Software and Applications Conference Workshops (COMPSACW), 2008. p. 140–5.
11. Bloom BH. Space/time tradeoffs in hash coding with allowable errors. Communication ACM. 1970 Jul; 13(7):422–6. Crossref
12. Lin MY, Hsueh SC. Apriori-based frequent item set mining algorithms on Map Reduce. Proceedings of the 6th international conference on ubiquitous information management and communication, 2012 Feb. p. 1–12
13. Kolb L, Thor A, Rahm E. Dedoop: efficient deduplication with Hadoop. Proceedings of the VLDB Endowment. 2012; 5:1878–81. Crossref
14. Dassarma A, Jain A, Machanavajjhala A, and Bohannon P. An automatic blocking mechanism for large-scale de-duplication tasks. Proceedings of the 21st ACM international conference on Information and knowledge management, 2012. p. 1055–64.
15. Muthitacharoen A, Chen B, Mazieres D. A low-bandwidth network file system. ACM SIGOPS Operating System. 2001 Dec; 35(5):174–87. Crossref