Identify and Remove Duplicated Records Using Q-gram and Statistical Techniques from the Data Warehouse

Sura Mahroos[1,*], Rihab Hazim[1], Yaqeen Saad[1], Nadia Mohammed[2]

[1]University of Anbar, College of Computer Sciences and Information Technology, Anbar, Ramadi, 31001, Iraq

[2]University of Anbar, College of Islamic Sciences, Anbar, Ramadi, 31001, Iraq

Emails: surasms917@uoanbar.edu.iq; rehz1991@uoanbar.edu.iq; yaqeen.cs91@uoanbar.edu.iq; nadia.fahad@uoanbar.edu.iq

Abstract

There are several real-world uses for the duplication system or record linkage. In order to help the system make the best judgments, it appears in a broad area of recognizing similar data, joining online papers in the wide web, detecting plagiarism, and allowing several applications to enter it. To improve the financial interest and applicability of logistics project, routing is crucial. The following is the issue with this study: Because duplicate receipts contain the same significant change in data restrictions and limitations, and the data change itself is minor, the duplicate record data is ambiguous to other redacted records that are reassembled with the same customer. The purpose of this study is to use statistical techniques and the Q-gram to discover the best method for the detection and removal of duplicate records. We propose the following goals to help achieve that goal: Reduce the size of the data warehouse (DW) by providing a data warehouse free of duplicates. Decrease the amount of time spent looking for the (DW) and improve the DSS. The approach is divided into two stages: first, identify similarity records based on Q-gram similarity; second, determine whether classification records may be improved by statistical methods. The percentage threshold of 0.68 has been determined. It goes through a statistical process that decides whether this record is duplicated if the key ratio similarity is surpassed. The accuracy of the suggested work is 79%.

Keywords: Duplicate Elimination; Data Cleaning; Similarity Score; Q-Gram Similarity; Statistical Tools

1. Introduction

The task of looking for duplicate records in a data warehouse has been a long-standing issue in the data storage field and is currently the subject of active research. Several studies have been conducted to discuss the challenges of duplication of data brought on by duplicate data pollution. Unintentional duplication of records built from millions of data points from different database sources is a common occurrence in data warehouses. [1]. because of abuse and poor data quality, dirty data leads to poor decisions. [2]. the process of identifying collections of records that refer to similar entities in a data file is known as duplicate detection [3]. One of the primary issues in the broad field of data cleansing and data quality in data warehouses is the challenge of locating and removing duplicate data. There are often several representations of the same logical real-world thing in the data warehouse [4].

The problem and goals of this work are:

Finding and eliminating duplicate records is a crucial step in the data integration and cleaning process, particularly when the modified constraint definition is imprecise. The work is negatively impacted when a data warehouse contains multiple records belonging to the same user. Therefore, it is necessary to perform operations on the DW in order to detect an efficient method for finding and delete alike records, even if the records of database are not explicitly similar.

DOI: https://doi.org/10.54216/JCIM.170101
Received: February 25, 2025 Revised: May 24, 2025 Accepted: July 04, 2025

The development of a system for identifying and removing duplicate data is the primary goal of the study. The goal of the proposed system is to provide a fast, accurate and efficient guide to detect and delete duplicate data. In training terms, it also aids in closing the gap of knowledge between various people in identifying and getting rid of duplicate data. The main contributions of this study are:

- To review into relevant studies on finding the best solution by identifying and removing redundant data.
- To build a suitable representation architecture for the suggested duplicate data detection and removal.
- To create and implement a system to remove redundant records from the DW by employing clever strategies and similarity calculations.
- To offer a data warehouse without duplicates, which will reduce the DW's size, reduce the amount of time spent exploring it, and improve the DSS.
- To verify and test the system functionality

2. Related Work

M.Padmanaban et al proposed a method for deduplication that is based on artificial neural networks. The input to the feed forward neural network is a set of data produced by various similarity metrics, such as the Damerau–Levenshtein distance and the dice coefficient. The training and the testing stage are the two procedures that show the suggested de-duplication method. In order to assess the effectiveness of the planned approach, two distinct datasets were used, and the accuracy was (79%) [5]. Bilal Khan, et al. Methods for duplicate deduplication algorithms have been developed. Characters are converted to numeric values based on the complete data. Data mining techniques are then used. The numeric value reduces the number of comparisons between records. is subjected to k-mean clustering. The algorithm's performance is further increased by matching records inside a cluster using the divide and conquer strategy to find and eliminate duplicate records [6]. M.Padmanaban, et al have developed three distinct steps—feature computation, feature selection, and detection—are used to carry out the general steps of the recommended technique. First, the Q-gram idea is used to compute the features. Next, the particle swarm algorithm (PSO) is used to identify the subset of best feature sets. Finally, a classifier of naïve Bayes is used to categorize the records if they are duplicates or not. They use a data collection to construct this system. In addition, 89% accuracy was achieved using this method [7]. The operation of change or eliminate data from a database which is incorrect, lacking, incorrectly duplicated, or configured is known as data cleansing [8], [9]. The E-Clean technique aims to improve the quality of data by identifying and removing mistakes, inconsistencies, and duplicates. Problems with data quality, including duplicate data, lost data, and unacceptable data, can be found in scientific databases with single data collections. A data cleaning system is necessary when many data sets wish to be combined into a single database in order to maximize the significance of the data. This is because the sources frequently provide redundant information in altered performances. It becomes necessary to merge various data depiction and eliminate redundant data in order to supply access to consistent and accurate data. A data cleansing technique needs to meet a number of criteria. It would first find and fix all of the database's major mistakes and inconsistencies [9].

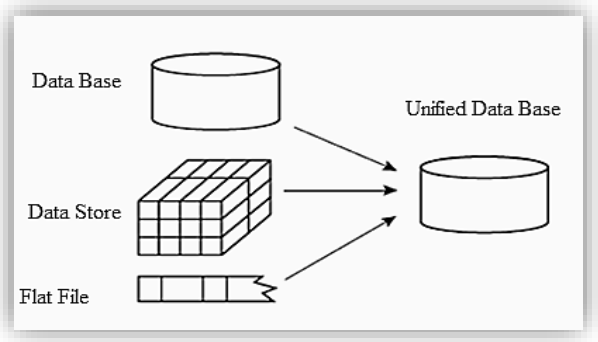

Figure 1. Data collection and integration.

3. Q-Gram Similarity Algorithm

q is the length of the substring that this algorithm is a substring of. The goal is to break the string up into q-gram tokens, match them with extra information to find similarities, and determine how many strings match. To ensure that characters at the start and finish are not overlooked in the comparison, additional padding is prepared to account for them. Since Q-gram processes do not operate by comparing the phonetic properties of words, they are not definitely phonetic similar [10]. Q-grams may be used to determine the "distance," or whole change, between

DOI: https://doi.org/10.54216/JCIM.170101
Received: February 25, 2025 Revised: May 24, 2025 Accepted: July 04, 2025

2 words in its place. It is highly advantageous to use the q-gram method technique since it can compete altered or misspelled words, if they are clearly "phonetically disparate" [11, 12]

4. The Developed Work

This work begins by outlining the characteristics and elements of the redundant records as follows: **1. Stable qualities,** like those that are the same (customer name, gender, and blood type). **2. Variable characteristics**, which can be divided into:

a) Significantly altering, such the identical parts (City and Marital Status), which are specific attributes in the list. **b) Variables that change little,** such sales, age, unit price, salary, weight, number of children, and length; they are in frequent quantifiable or numerical factors. In order to eliminate duplication, these fields are useful.

Cust-id	Cust-name	Blood-Type	gender	age	salary	Num_of_Childr en	Weight	Length	City	Marital-Status
1	Sadiq Mahdi Mekhlef	O+	M	37	721$	2	120	181	Baghdad	Single
2	Sadiq Mahdi Mekhlef	O+	M	40	730$	5	113	179	Duhok	Married
3	Zahra Mohamed Said	B+	F	28	420$	0	73	174	Basra	Single
4	Zahra Mohamed Said	B+	F	32	431$	1	78	174	Kufa	Married

Figure 2. sample of redundant records.

The records in this system are moving through the overhead phases in a number of steps. The implemented system's work is depicted in Figure (4).There are five key parts that make up the records eliminations system:

I. Generation of Key.

II. Sorting the Records in the DW.

III. Blocking of DW.

IV. Stage Compress Key Selection.

V. Statistical technique.

I. Key Generation

From the primary fields just (fixed element and variables qualities that are subject to significant change and can only be advantageous), for every field:

- The first three letters of each word must be specified in the "cust_name," "sex," and "blood type" fields. Additionally, only the first three letters of each word must be specified in the "city" and "marital status" fields.
- As shown in figure (3), combine the chosen attribute with each key that creates a record.
- Duplicates can be removed, and alphabetical sorting is possible. Every field independently.
- The major key field is produced by merging the output.

DOI: https://doi.org/10.54216/JCIM.170101
Received: February 25, 2025 Revised: May 24, 2025 Accepted: July 04, 2025

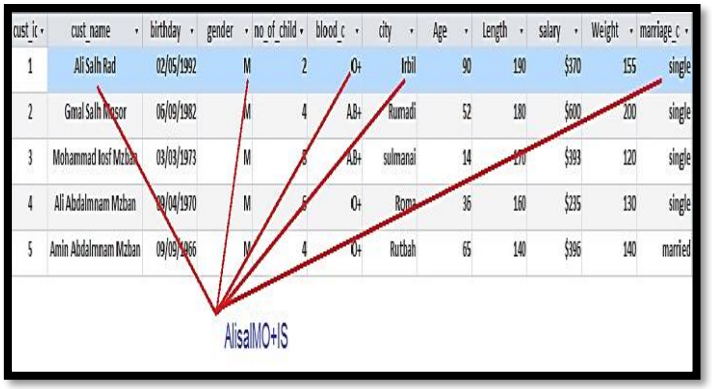

Figure 3. key genration process.

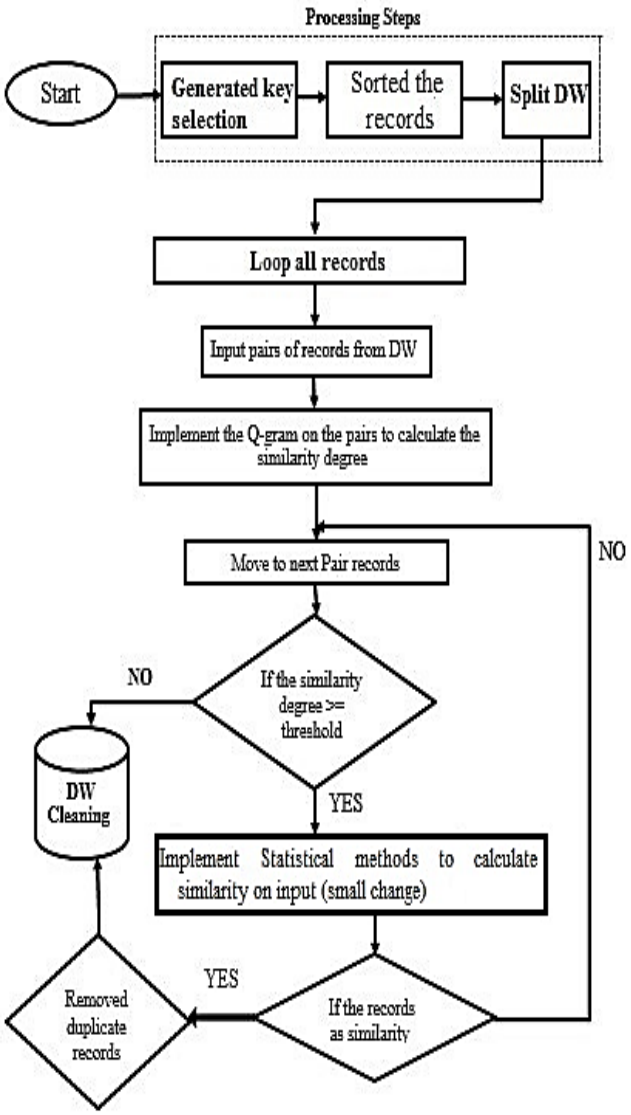

Figure 4. Proposed system diagram of duplicate records eliminations

II. Sort the Records in DW

This step is crucial for expediting the search and comparison processes, which in turn speed up the acquisition system's implementation process. It involved alphabetizing the records in a data warehouse according to the key that was explained in the previous step.

DOI: https://doi.org/10.54216/JCIM.170101
Received: February 25, 2025 Revised: May 24, 2025 Accepted: July 04, 2025

III. Blocking DW

Following the key-generation step, the database included a field for every user's blood type. As a result, the system has a procedure that divides the contents of the data warehouse are separated into four types based on blood type (since blood types are fix and cannot be change). This procedure will improve search speed and accuracy in differentiating data warehouse David to blocks 1 = A±, 2 = B±, 3 = O±, and 4 = AB±.

IV. Stage Compress Key Selection

Following the block phase and the creation of keys that create new files from records Since it is composed of multiple fields, including a significant modification for this equality process, the outcome will be different keys. We suggested using Q-gram techniques, which divide strings to several (F) patterns based on the number of Q. For example, the string (**Salah)** and (**Q=3), so "#Sa, Sal, ala, lah, ah#"** that compare (**str1,** and **str2)** and Equation (1) will generate a numeric value ranging between **(0 ; 1)** after the strings have been compared.In order to determine the similarity between this key, the researchers are utilizing the q-gram simalirty ways on key genration. If the size of q=11, we play the threshold 0.68. If the proportion exceeds this limit of similarity between these keys, then the other algorithm will apply on other fields for checking whether they are same or not in the next phase. It must be that researchers use Q-grama in this key because they have the ability to manage a variety of tales of chord changes.

$$\boldsymbol{Percentage} = \frac{1}{2} * \left(\left(\frac{\boldsymbol{CummonGram}}{\boldsymbol{Gs1}}\right) + \left(\frac{\boldsymbol{CummonGram}}{\boldsymbol{Gs2}}\right)\right) \quad (1)$$

Though:

|Gs1| & |Gs2| is the amount of Q-grams of s1 and s2 respectively.

V. Statistical Approach

Calculate the probability of duplication based on the convergence of values. Unit Price, Sales, Age, Number of Children, Weight, Salary, and Length are input vectors for the five culls. All statistical techniques recommended in the following phase operate on these vectors.

a) Mean

When referring to a single measure of central tendency of a likelihood distribution or random variable that is identified by it, the terms mean and expected value are interchangeable in the fields of probability and statistics.

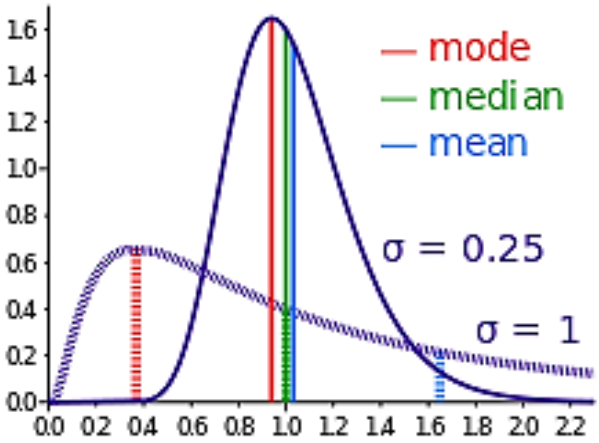

Figure 5. Mean, Median, and Mode

$$\boldsymbol{Mean} = \frac{\boldsymbol{x1} + \boldsymbol{x2} + \cdots + \boldsymbol{x_n}}{\boldsymbol{n}}$$

When comparing two records (two rows), if the difference between the two records for each field is equal to or less than the resulting value (R) and each selected field, the two records are the same. This is done by using the arithmetic mean calculation for a range of fields, the central division of the 100, or any value that specifies produce value (R).

DOI: https://doi.org/10.54216/JCIM.170101
Received: February 25, 2025 Revised: May 24, 2025 Accepted: July 04, 2025

When comparing two records (two rows), if the difference between the two records for each field is equal to or less than the resulting value (R) and each selected field, the two records are the same. This is done by using the arithmetic mean calculation for a range of fields, the central division of the 100, or any value that specifies produce value (R).

b) Variance and Standard deviation

For the middle arithmetic, the variance is calculated as the sum of the distractions squared by the values. The average deviation values for the middle arithmetic squares serve as a measure of the difference in data dispersion, and the variance quantifies dispersion. The standard deviation is defined by taking the square root of this total, or the sum of the deviations squares.

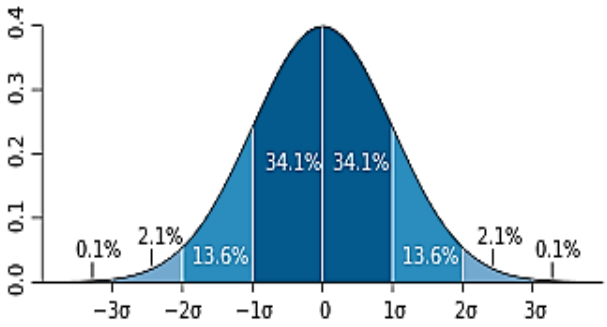

Figure 6. Standard deviation

When comparing two records (two rows), the standard deviation of a set of fields that yield a value calculation (R) is used to determine whether the difference between the two records for each field is equal to or less than the resulting value (R). For each particular field, the two records are duplicated.

a) Correlation

The coefficient of correlation [13]. indicates a linear relationship between two variables and the direction of that relationship as follows. It is a number between -1 and 1.

- +1 Positive correlation.
- -1 Inverse relationship.

A value of zero indicates that the two variables are unrelated.

In addition, be as follows:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left(n\sum x^2 - (\sum x)^2\right)\left(n\sum y^2 - (\sum y)^2\right)}} \quad (3)$$

The correlation between a particular set of fields in two records is determined; if the correlation is positive, we determine whether the two records are duplicates.

5. Results and discussion

This method explains the precision that was inferred from this effort. Consequently, the accuracy result is given.

When assessing the systems and the work's performance, the accuracy of key metrics is taken into consideration. The q-gram layers in the proposed key in this effort, after implementation with a range of threshold values (0.9, 0.8, 0.6 and 0.5), the optimal value to produce results was 0.68.

Throughout this investigation, we utilized q-gram and statistical techniques in tandem to identify and delete duplicates.

DOI: https://doi.org/10.54216/JCIM.170101
Received: February 25, 2025 Revised: May 24, 2025 Accepted: July 04, 2025

Table 1: Number the discover duplicated record

Row count	Statistical		
	Mean	**Standard deviation**	**Correlation**
50000	900	1200	1308
60000	900	1200	1330
70000	900	1300	1640
80000	1000	1404	1801
90000	1100	1612	2100

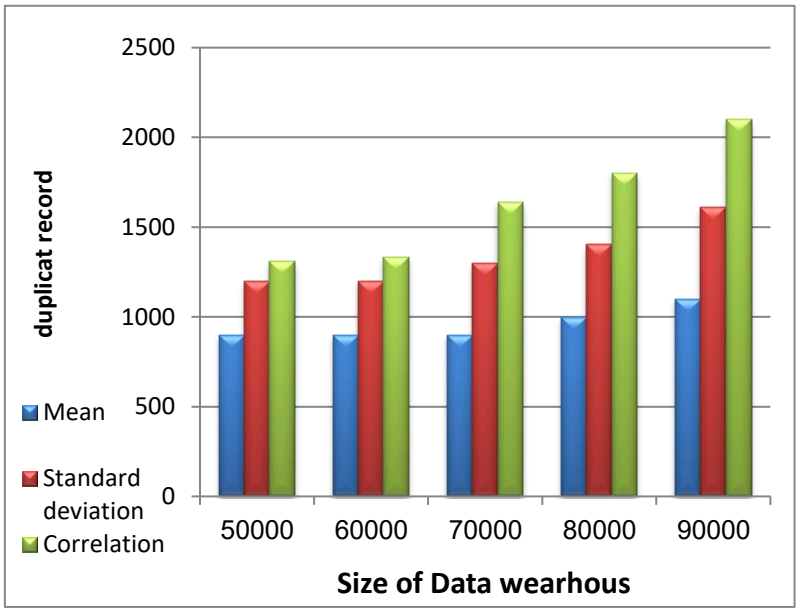

Figure 7. Data cleaning approach

Table 2: Compaction of data cleaning approach

Parameters	Statistical		
	Mean	**Standard deviation**	**Correlation**
Accuracy	High	Medium	Medium
Processing Time	Medium	Medium	High
Model complexity	Low	Low	High
Flexibility	Yes	Yes	No It needs to overlapping processors

DOI: https://doi.org/10.54216/JCIM.170101
Received: February 25, 2025 Revised: May 24, 2025 Accepted: July 04, 2025

6. Conclusion

In this study work, an efficient duplicate discovery and removal approach is developed to obtain good results of duplicate detection and removal by dipping false positives. This study's performance shows that compared to the present procedure, it significantly reduced time and improved duplicate results. This method of identifying and eliminating duplicate records uses statistical techniques and the q-gram. Eliminating duplicates takes time and is influenced by file size and record count. The time it takes to find and handle duplicates rises with data size. Although statistics vary according on the type of statistical function utilized, correlation is preferable. Depending on the size and kind of data that needs to be cleaned, one can select how to eliminate duplicates. Able to handle duplicates using the dimension hierarchies, the figure (8) shows the simple implementation of the system.

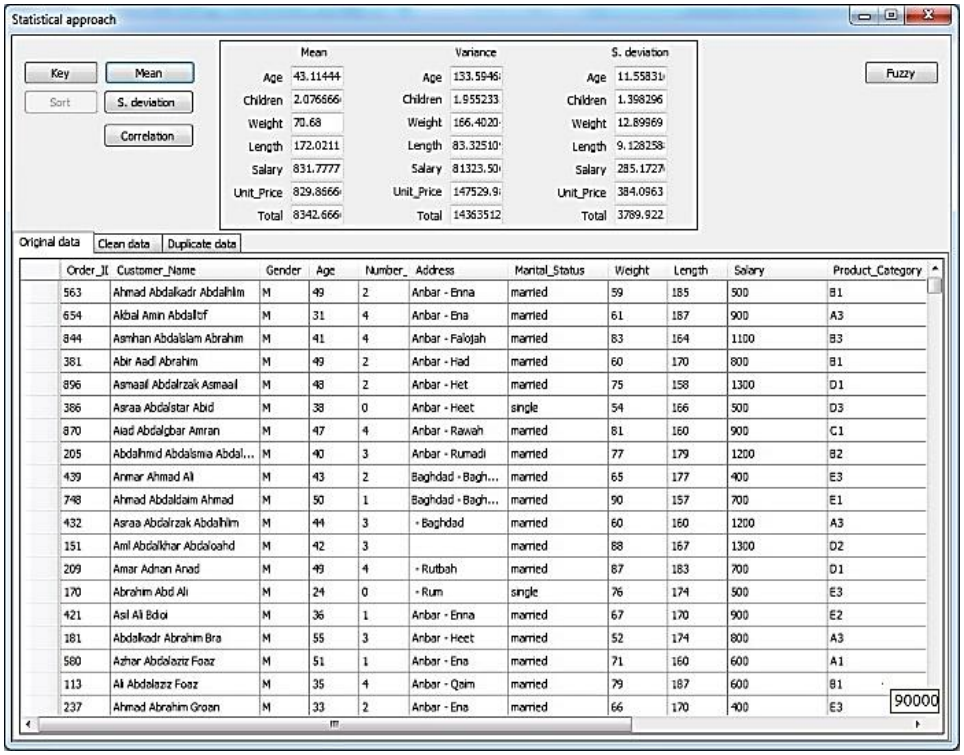

Figure 8. Implementation of the proposed work.

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] P. Bhatia, *Data Mining and Data Warehousing: Principles and Practical Techniques*. Cambridge University Press, 2019.

[2] Darch Abed Dawar, "Enhancing Wireless Security and Privacy: A 2-Way Identity Authentication Method for 5G Networks," *International Journal of Mathematics, Statistics, and Computer Science*, vol. 2, pp. 183–198, 2024. doi: 10.59543/ijmscs.v2i.9073.

[3] Kennedy et al., "Epidemiology of homicide in community-dwelling older adults: a systematic review and meta-analysis," *Trauma, Violence, & Abuse*, vol. 24, no. 2, pp. 390-406, 2023.

[4] M. Anitha et al., "Duplicate Detection of Records in Queries Using Clustering," *International Journal of Research in Computer Science*, eISSN: 2249-8265.

[5] M. Padmanaban and T. Bhuvaneswari, "A technique for data deduplication using Q-gram concept with support vector machine," *International Journal of Computer Applications*, vol. 61, no. 12, 2013.

[6] B. Khan et al., "Identification and removal of duplicated records," *World Applied Sciences Journal*, vol. 13, no. 5, pp. 1178-1184, 2011.

DOI: https://doi.org/10.54216/JCIM.170101
Received: February 25, 2025 Revised: May 24, 2025 Accepted: July 04, 2025

[7] M. Padmanaban and R. Radha, "PSO Algorithm to Select Subsets of Q-Gram Features for Record Duplicate Detection," *International Journal of Computer Applications*, vol. 82, no. 12, 2013.

[8] O. Azeroual, A. Nikiforova, and K. Sha, "Overlooked Aspects of Data Governance: Workflow Framework For Enterprise Data Deduplication," in *2023 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS)*. IEEE, 2023.

[9] O. Alotaibi, S. Tomy, and E. Pardede, "A Framework for Cleaning Streaming Data in Healthcare: A Context and User-Supported Approach," *Computers*, vol. 13, no. 7, p. 175, 2024.

[10] L. Yang et al., "Authenticating q-gram-based similarity search results for outsourced string databases," *Mathematics*, vol. 11, no. 9, p. 2128, 2023.

[11] M. Sagheer, S. S. Salih, and S. M. Searan, "Design and Implementation of Secure Stream Cipher Algorithm," *International Journal of Computing and Digital Systems*, vol. 7, no. 03, pp. 127-134, 2018.

[12] S. Yakhni et al., "Using fuzzy reasoning to improve redundancy elimination for data deduplication in connected environments," *Soft Computing*, vol. 27, no. 17, pp. 12387-12418, 2023.

[13] C. H. Lindquist et al., "When pre-release optimism meets post-release reality: Understanding reentry success through a longitudinal framework assessing pre-and post-release perceptions," *Crime & Delinquency*, vol. 71, no. 1, pp. 144-174, 2025.

DOI: https://doi.org/10.54216/JCIM.170101
Received: February 25, 2025 Revised: May 24, 2025 Accepted: July 04, 2025