Removing Duplicate Records from Data Warehouse by Q-gram and Neural Network

Murtadha M. Hamad [#1], Salih S. Salih [*2]

Department of Computer Science, University of Anbar, Iraq

[1] dr.mortadha61@gmail.com

[2] salih_sami2016@yahoo.com

Abstract-The problem of discovering and removing duplicated records is one of the main problems in the wide area of data cleaning and data quality in the data warehouse. In this paper, researchers try to find a similar data from a set of data records. A similarity grade is assigned to the data records in relation to other data records based on a similarity between tokens of the data records. Data records whose similarity score with respect to each other is greater than a threshold from one or more groups of data records. In this system, a key is created for each record in the database, as shown in suggested algorithms, where this key is input to Q-grams similarity algorithm that calculates the percentage of similarity between each key and another. We have identified the percentage threshold to be 0.68. If the similarity threshold between the key values is exceeded, it enters to the Neural Network algorithm that works with two-phases training data and testing. The suggested approach is tested through several different data warehouse for the evaluating the efficiency. The accuracy acquired from multi DW has been found to be 96.94%.

Keywords-Duplicate Detection, Duplicate Elimination, Similarity score, Q-Gram, Neural Network, Key Generation.

I. Introduction

In the community of data storage and the mission of searching for duplication of records within the data warehouse to a lengthy time continuous problem and has become an area of active research. There take been several Research below takings to talk the difficulties of data duplication caused by duplicate pollution of data. Data warehouse in mostly the existence of unintended duplication of records that was created from millions of data of other database sources is difficult to avoid. [1]. Dirty data provides to bad decisions taken is the result of misuse and the lack of data quality [2].

The difficulty of identifying and eliminating duplicated data is one of the main problems in the wide area of data cleaning and data quality in the data warehouse. Much time, the equal logical actual world entity can have several representations in the data warehouse [3]. Duplicate detection is the task of finding sets of records that mention to the similar entities within a data file.

Duplicate detection is the task of finding collections of records that refer to the same entities within a data file. This mission is not easy when unique persons of the entities are not recorded in the file, and it is especially difficult when the records are topic to errors and missing values [4].

II. Literature Survey

M.Padmanaban et al proposed technique consists of a technique based on the artificial neural network for deduplication technique. A collection of data generated from some similarity measures such (Dice coefficient and Damerau–Levenshtein distance) are used as the input to the feed forward neural network. There are two processes, which illustrate the projected de-duplication system, the training stage, and the testing phase. The planned approach is tested with two different datasets for the evaluating the efficiency the accuracy obtained as 79% [5]. Bilal Khan, et al developed the techniques and methods for a de-duplicator algorithm, they are converted characters to a numeric value that is based on entire data. Then apply data mining technique k-mean clustering is applied on the numeric value that reduces the number of comparisons among the records. To identify and remove the duplicated records, divide and conquer technique is used to match records within a

cluster that further improves the efficiency of the algorithm [6]. M.Padmanaban, et al have developed in their paper the overall steps of the suggested technique are carried out using three different steps, such as 1) feature computation, 2) feature selection, and 3) detection. Initially, the features are computed using Q-gram concept and then, the subset of optimal feature sets is identified by using particle swarm algorithm (PSO) and to classify the records if duplicate or no used naïve Bayes classifier. They implement this system on the data set. and the accuracy obtained in this approach as 89% [7].

III.Problem Description

Discovery and elimination duplicate records Especially that refined constraint specification is a data are ambiguous, there is an important process in data integration and data cleaning process. The presence of more than one record in a data warehouse belonging to the same user has a negative impact on the work, performing operations on the data warehouse is, therefore, necessary to find an efficient technique to find and delete those similar records, and more refined these records even if the database records are not explicitly identical.

IV. Research Objectives

The main objective of the research is to develop detecting and eliminating duplicated data System. The aim of the proposed system is to provide quick and precise efficient system guidance detecting and eliminating duplicated data. Additionally, for training purposes, it helps in reducing the knowledge gap between different individuals in detecting and eliminating duplicated data. The specific objectives of the research are as follows:

- To investigate the related works on detecting and eliminating duplicated data to find the optimal solution.
- To design appropriate representation architecture for the proposed detecting and eliminating duplicated data.
- To design and implement a removal-duplicated system for detecting the duplicated records that found in the data warehouse and remove it by using the intelligent techniques and similarity methods.
- To provide data warehouse without duplicate that leads to minimize the size of DW, reduce the time of searching for the DW and enhancement the decision support system.
- To test and validate the system's performance.

V. Data Integration and Data Cleaning

Data cleansing is the process of modifying or eliminating data in the database that is improper, incomplete, improperly configured, or duplicated [8] [9]. E-Clean method contracts with discovering and eliminating errors, inconsistencies, and duplicate data in a direction to increase data quality. The data quality problems are present in the scientific databases in lone data collections, such as lost information, unacceptable data, and duplicate data. When several data want to be integrated into the alone database, system requirements data cleaning system to increase the importance. This is because the sources often have redundant data in changed performances. In order to provide access to correct and consistent data, combine of different data representations and removal of duplication data become needed. A data cleaning method must satisfy multiple requirements. Firstly, it would discover and eliminate all main errors and inconsistency in the database [9].

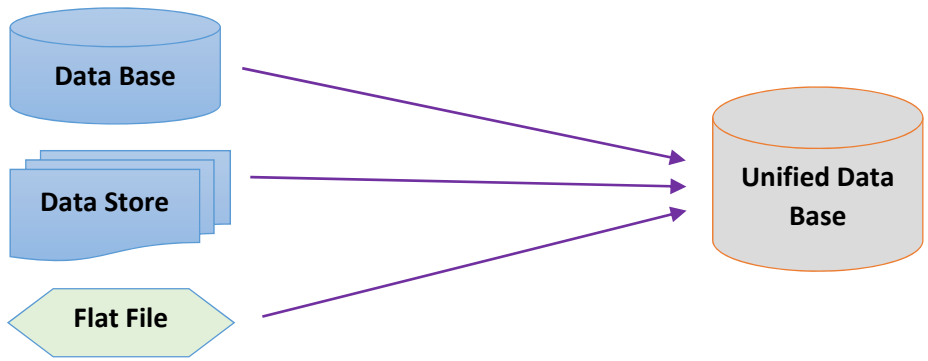


Fig. 1. Data collection and integration [2]

VI. Q-Gram Similarity Algorithm:

This algorithm is a substring of a text, where the length of the substring is q. The idea is to disruption the string into tokens of the q-grams and matching with additional to discover similarities and calculation the number of matches between the strings. Another padding is ready to account for characters at the begin and end to confirm they are not ignored in the comparison. Q-gram processes are not firmly phonetic similar in that they do not work based on a comparison of the phonetic characteristics of words. In its place, q-grams can be supposed to calculate the "distance," or a total of change between two words. Applying the q-gram procedure technique is very favorable as it can competition misspelled or mutated words, even if they are resolute to be "phonetically disparate" [10].

VII. The proposed work

Firstly, the researcher explains the attributes and component of the duplicate Records in the following:

1- ***Fixed attributes***, such as those characteristics like (Customer Name, Blood-Type, and Gender).

2- ***Variable attributes***, these can be divided into:

2.1. ***Largely changing***, such as those characteristics like (Marital-Status, and City) this attributes that be specific on the list.

2.2. ***Small changing***, such as those attributes like (Sales, Unit_Price, Age, Salary, Number_of_Children, Weight, and Length), which are often the attributes that are numerical or quantitative. These fields are helpful in eliminating the duplicates.

Cust-id	Cust- name	B Blood -- type	Gender	Age	Salary	Num_ of _ Children	Weight	Length	City	Marital-Status
1	Sajad Ali khal	O+	M	44	721$	2	120	181	Baghdad	Single
2	Sajad Ali khal	O+	M	48	730$	5	113	179	IRBALE	Married
3	Fatma salam	B+	M	56	400$	0	92	173	Basra	Single
4	Fatma salam	B+	M	60	420$	2	95	172	Karbala	Married

Fig. 2. Sample of Duplicates Records

In this study, the records are passing into a number of steps through the overhead phases. The work of implemented system, as shown in Fig. 3. The records eliminations system consists of five important modules:

I. Key Generation.

II. Sort the Records in DW.

III. Blocking DW.

IV. Stage Compression Key Selection.

V. Artificial Neural Network Technique.

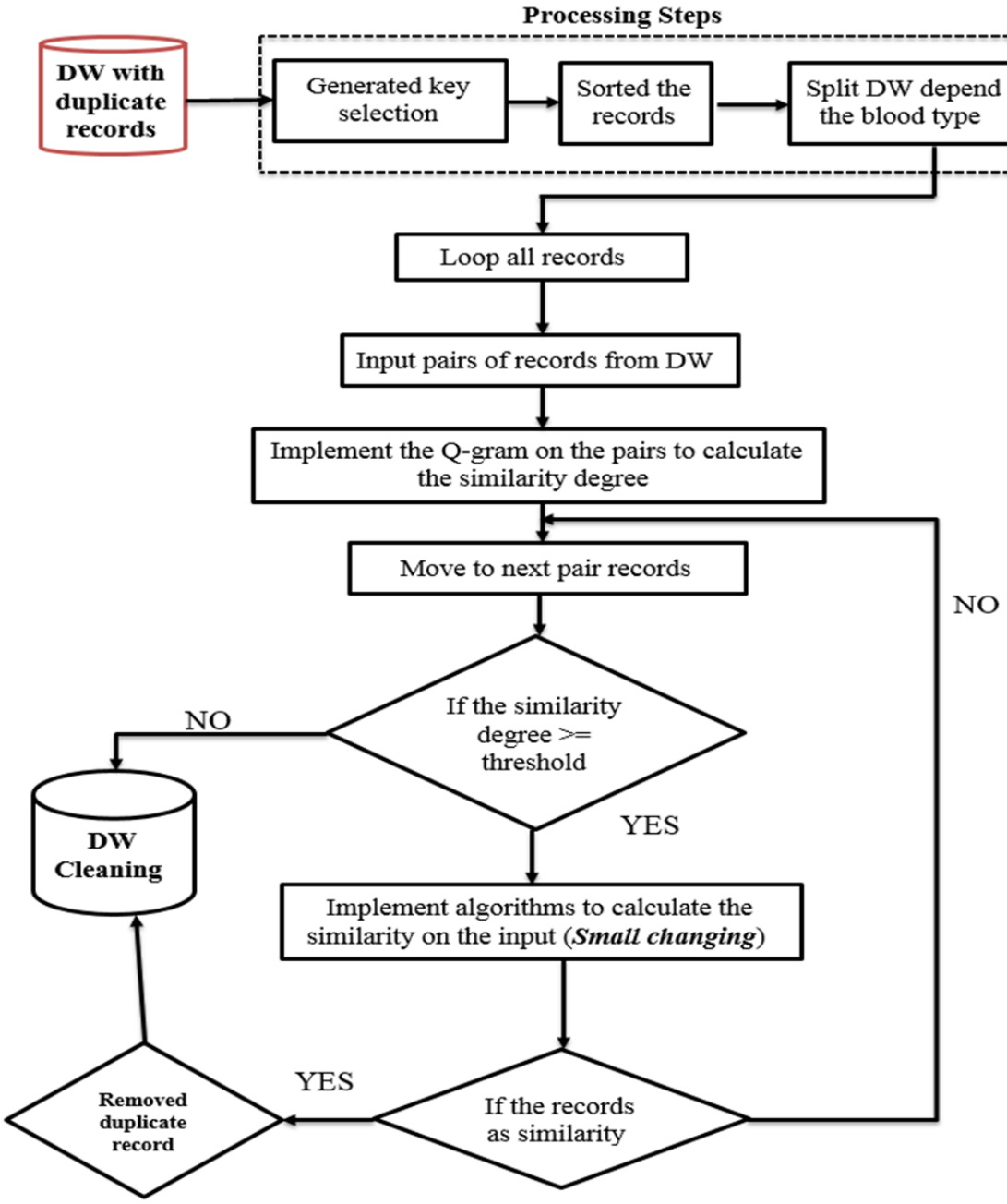
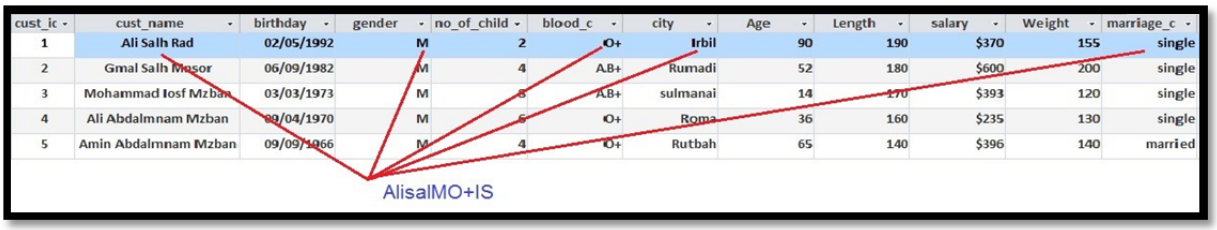


Fig. 3. General System of Duplicate Records Removals

I. Key Generation

Exploiting key to this step is one of the most important steps in this research because it is the foundation, which brings the similarities of each to help us repeating account as well as because of its importance in reducing the time in the research. It also increases the accuracy of the search which is generated from the field (fixed attributes and from variable attributes that are largely changing that can be of benefit only), for all field, as well as from the fields. It took many characters of these fields, and then integrate them to be placed in a new field for eachrecords.



cust_ic	cust_name	birthday	gender	no_of_child	blood_c	city	Age	Length	salary	Weight	marriage_c
1	Ali Salh Rad	02/05/1992	M	2	O+	Irbil	90	190	$370	155	single
2	Gmal Salh Masor	06/09/1982	M	4	AB+	Rumadi	52	180	$600	200	single
3	Mohammad Iosf Mzban	03/03/1973	M	5	AB+	sulmanai	14	170	$393	120	single
4	Ali Abdalmnam Mzban	09/04/1970	M	6	O+	Roma	36	160	$235	130	single
5	Amin Abdalmnam Mzban	09/09/1966	M	4	O+	Rutbah	65	140	$396	140	married

AlisalMO+IS

Fig. 4. Key Generation Work

II. Sort the Records in DW

In this stage, it was arranged (alphabetizing) in the records in a data warehouse in alphabetical order based on the key Which explains in the previous stage, this stage is important in terms of speeding up the search process and the process of comparison resulting in the acquisition system speed in implementation.

III. Blocking DW

After reading the file, database contained a field for each under the blood type for each user, so this study have the process of separation of the contents of the database into four categories depending on the blood type (the fact that blood type fixed and cannot be changed), this process will increase the search speed as well as accuracy in distinguishing So database David to block1= A±, block2=B±, block3=O± and block4= AB±.

IV. Stage Compression Key Selection

After the blocking stage and key generation that generate new fields from records The result will be different keys because it is formed from several fields, including large change for this equality process Bring records so we proposed to use Q-gram methods that divide the strings into multi F patterns depend on the no. of Q such is the string **Salah** and **Q=3 so(#Sa, Sal, ala, lah, ah#)** that compression between the **str1** and **str2** and After comparing the strings will produce a numeric value range between **(0 - 1)** by the equation(1) so the researchers using Q-Gram on this key because it has the capacity to deal with such varieties of change in the Strings, and To get the records that things is duplicate this study are using the q-gram similarity methods on the key generations to calculate the similarity between this key, we apply the threshold 0.68 if the size of q=11, So if the proportion exceeded the threshold limit of similarity between these keys will bring and then will go to another algorithm will work in other fields will be covered to check whether similar or not in the next stage.

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

V. The approach of ANN

In This study, had been applying the **Elman** (feed forward) neural network trained with Back-Propagation (BP). Input vectors for five fields ***(Sales, Unit_Price, Age, Salary, Number_of_Children, Weight, and Length),*** and the corresponding target vectors (duplication) are used to training the network. The common neural network design process has many stages:

1) Random weights given to the correlation between the grid cells.
2) Connect the network in one of the inputs intended for training.
3) Apply the forward deployment process to determine the network outputs.
4) Comparing the actual output with the required outputs and determine the error value.
5) Update the weight and bias that ensures minimized error value using regression deployment.
6) Validate the network (test the network)
7) Extract the result.

In the training phase had been used the number of validation checks that will always adjust the weight to get the higher performance depending on the gradient of the performance. Usually, when the performance is becoming accept the training would terminate.

The gradient will turn out to be little as the training reaches a minimum of the performance. In the event that the extent of the slope is determined with error values, the training will stop. This can be balanced by setting the parameter of the networks. The quantity of approval checks represents the number of successive iterations that the validation performance fails to decrease.

The training procedure contains checkpoint on training phase if you click the Stop Training button in the training window. You may need to do this if the performance function fails to decrease error over many iterations. It is always possible to continue the training by resuming the training command. It will continue to train the network from the completion of the prior run. From the training window, you can access four plots:

performance, training state, error histogram and regression. The rendering plot shows the value of the performance function of each iteration. It plots training, validation and test rendering. The training state plot shows the progress of other training variables, such as the gradient, the number of validation checks, etc. The error histogram plots the distribution of the network errors. The regression plot determines a regression between network outputs and network targets. You can use the histogram and regression plots to validate the network performance.

1. Multi-Layer Perceptron (MLP)

Is a feed forward artificial neural network display that maps sets of information onto an arrangement of the output. MLP comprises of numerous layers of nodes in a directed graph, with every layer completely associated with the following one. Aside from the input nodes, every node is a neuron or processing element with a nonlinear activation function. MLP uses an administered learning procedure called back propagation for preparing the network. MLP is an adjustment of the standard of linear perceptron and can recognize information that is not linearly separable. In this study network, MLP includes four layers: input layer, hidden layer, output layer and a context layer. The advantages of using of MLP is mainly is minimize the error function by adjusting weights. The minimization of this error leads to the optimal values of weights. A network can have several layers. Each layer has a weight matrix W, a bias vector b, and an output vector a. The bias used in a neural network to make the chance of node is always 'on'. Its value is set to a random number between 0 and 1 without regard to the data in a given pattern. It is analogous to the intercept in a regression model and serves the same function. If a neural network does not have a bias node in a given layer, it will not be able to produce output in the next layer that differs from 0 (on the linear scale, or the value that corresponds to the transformation of 0 when passed through the activation function) when the feature values are 0 or 'off'. To distinguish between the weight matrices, output vectors, etc., for each one of these layers in the figures (8), the number of the layer is appended as a superscript to the variable of interest. The initial network had a three hidden layer with ten neurons as inputs, tow neurons as outputs, and was trained with six times, using different initial weights for each time.

In this method, the technique as work in two steps (training data and testing data). You can explain in figure (6) that show the best training performance at epoch 701, which characterizes the performance at the 9.73e-06 and the gradient at 0.000366, and with mean squared error 10-7, while the time for the train network was 41 minute and 52 seconds. It is suitable if compared to the percentage accuracy.

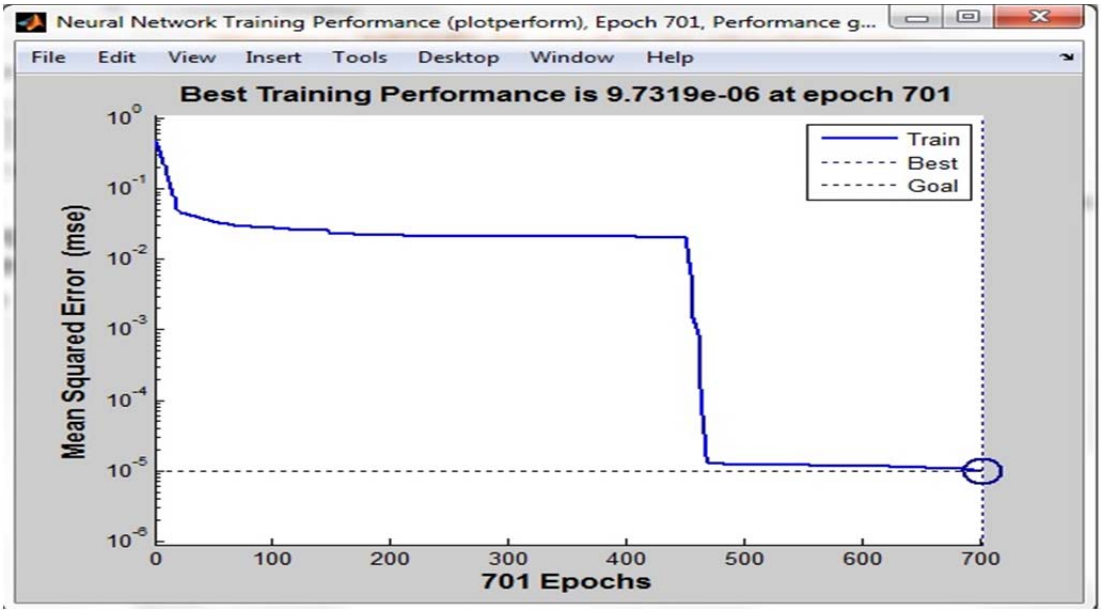


Fig. 6. The Best Training Regression

The algorithm (NN back-propagation) that have to use consists of multiple layers: one layer used for input units, which characterizes 10 neurons that are bearing in mind features. We used three hidden layers, the first hidden layer was consist of 10 neurons, the second hidden layer consist of five neurons and the third hidden layer was two neurons, while the output layer produced 2 neurons that represent duplication state (duplicated

and non-duplicated). The number of records in the data warehouse are (18,000) records for the training phase and Multi DW for the testing phase.

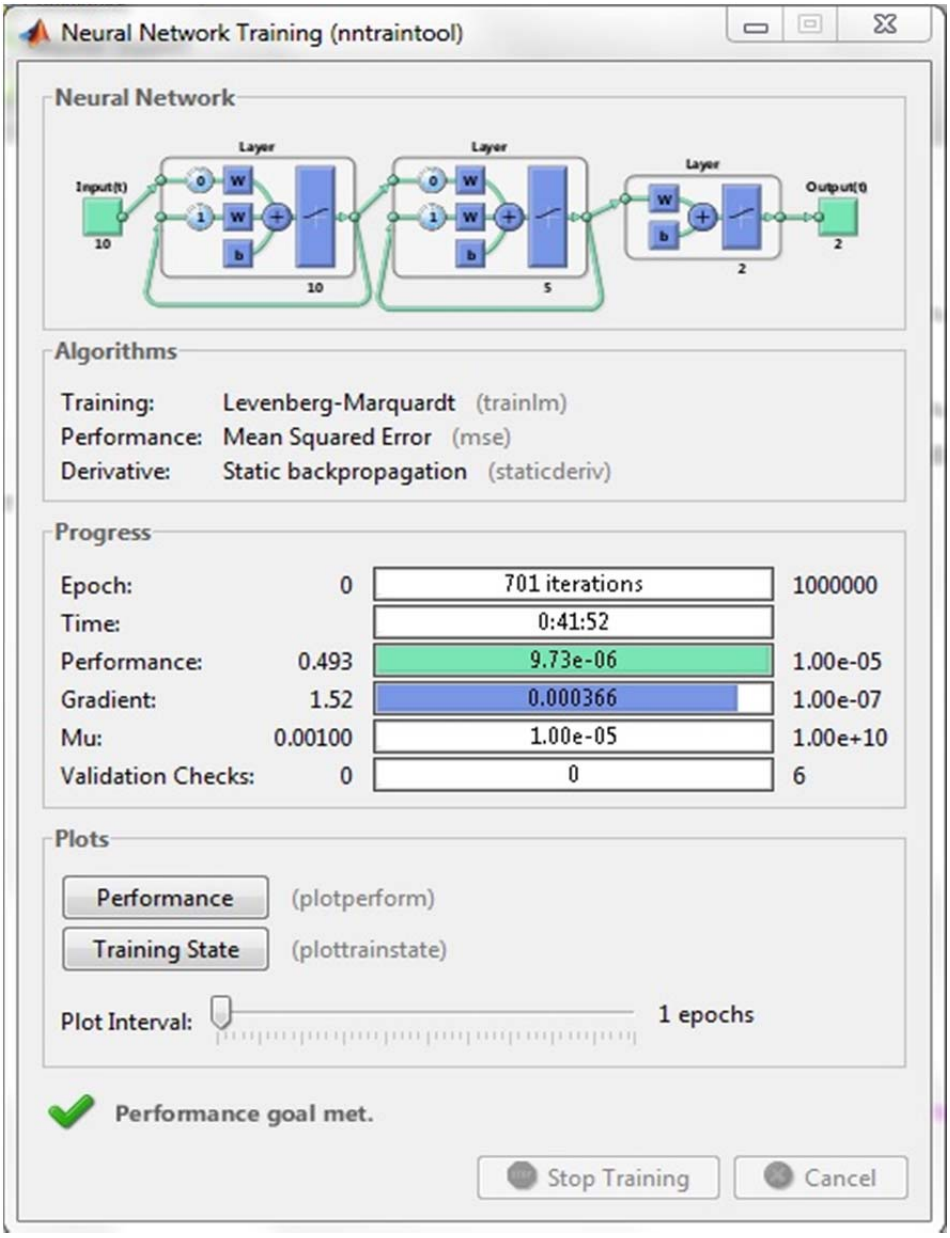


Fig. 7. Result of Neural Network Progress

You can show in the table (1) the results of the classification of the program done the system of a code where every code represents a type of State (duplicate or non-duplicate).

TABLE 1. The Results of the Classification of the **ANN**

Neural network output(target)	Node duplicate	Node non-duplicate
Non-Duplicate	**0**	**1**
Duplicate	**1**	**0**

Levenberg-Marquardt optimization (trainlm) that using to updates weight and bias values.

The networks were trained using the (trainlm) learning algorithm that found as a function in MATLAB.

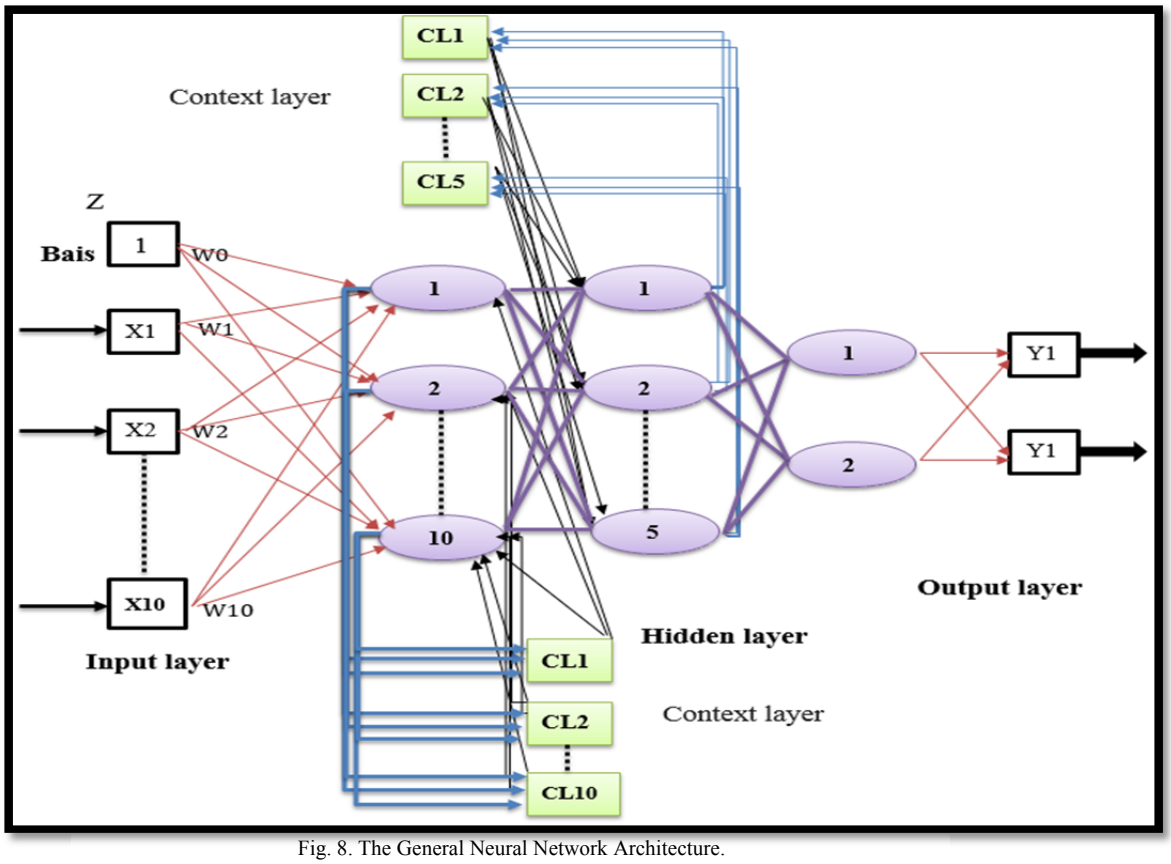


Fig. 8. The General Neural Network Architecture.

The outputs of these networks classified the record into two states. Duplicate, non-duplicate.

VIII. Results and discussion

In this approach, we have explained the accuracy that which were drawn from this work. Therefore, we will show the accuracy result.

The accuracy of the important metrics for evaluating the performance of the system and the work is considered

i. In this work the layers q-gram on the proposed key and after implementation using several values of the threshold (0.9, 0.8, 0.6, and 0.5) after work was the best value in giving output is 0.68.

ii. After this we have implemented Q-Gram and ANN on the existing data have been obtained on the accuracy as illustrated in the chart (10) in discovering and removing duplicate records process.

TABLE 2. Results of Executing Suggested (Q-Gram & ANN) Algorithm

No. of Data Warehouse records before reduction	No. of DW records after reduction	Step1: Discovery by Q-gram	Step2: Number of deleted records by (ANN)	Rate of reduction
100000	90021	11894	9979	10 %
200000	160217	58031	39783	20 %
250000	203470	61204	46530	20 %
300000	201140	124092	98860	33.33 %
500000	219313	387625	280687	60 %

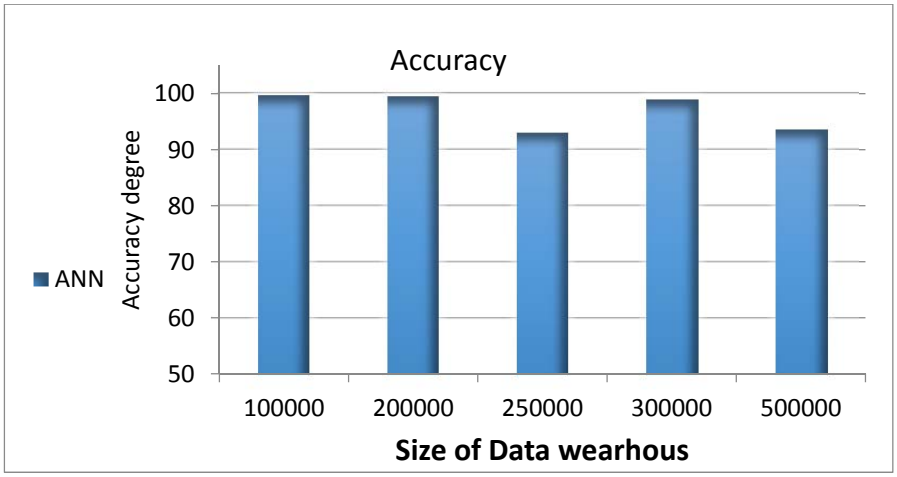


Fig. 9. The Accuracy Ratio for De-duplication using Neural Network

The good result that acquired after merging between the Q-Gram and Artificial Neural Network.

CONCLUSIONS

We have presented an approach that includes many steps to duplicate records discovery and removal in the several data warehouse size and the accuracy showing in the figure (9). In this approach, we used Q-Gram and ANN to eliminations the duplicate records and we used a high threshold to remove the incorrect pairs of records detected as duplicates. Along with, blocking technical depend on the blood type method was utilized to increase the accuracy taken to improve duplicate detection and key generation work to reduce the time of compressions Furthermore, the q-gram Bring suspected records duplicate. In addition, the neural network takes a decision if the records as duplicate or no. This work is efficient in terms (q-gram- neural network) of accuracy, as well as a higher speed in comparison with the work of his predecessor discrimination**.**

REFERENCES

[1] P. Ponniah, Data Warehousing Fundamentals for It Professionals. Second Edition, Copyright #, 2010.
[2] R. Agusthiyar and K. Narashiman," A Simplified Framework for Data Cleaning and Information Retrieval in Multiple Data Source Problems", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 3, Issue 8, August 2014.
[3] M.Anitha, A.Srinivas, T.P.Shekhar and D.Sagar, Duplicate Detection of Records in Queries Using Clustering. International Journal of Research in Computer Science eISSN 2249-8265 Volume 2 Issue 2 pp. [29-32], 2012.
[4] C. Mellon, Detecting Duplicates In A Homicide Registry Using A Bayesian Partitioning Approach. Institute of Mathematical Statistics, Vol. 8, No. 4, pp. [2404–2434], 2014.
[5] M.Padmanaban and T.Bhuvaneswari, An Approach Based on Artificial Neural Network for Data Deduplication", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (4), 2012.
[6] B. Khan, A. Rauf, H. Javed, S. Khusro, and H. Javed, Removing Fully and Partially Duplicated Records through K-Means Clustering. IACSIT International Journal of Engineering and Technology, Vol. 4, No. 6, 2012.
[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, pp. [0975 – 8887], 2013.
[8] N. Choudhary, A Study of Problems and Approaches of Data Cleansing/Cleaning. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 2, February 2014.
[9] H. H. Mohamed, T. L. Kheng, C. Collin, and O. Siong Lee, E-Clean: A Data Cleaning Framework for Patient Data. First International Conference on Informatics and Computational Intelligence, IEEE, 2011.
[10] E. Ukkonen, Approximate string matching with q-grams and maximal matches. Theoretical Computer Science, 92 (1), pp. [191-211], 1992.

Authors

[1]**Murtadha M. Hamad** received his MSc. degree in computer science from University of Baghdad, Iraq in 1991, received his Ph.D. degree in computer science from University of Technology, Iraq in 2004, and received the Assist Prof. title in 2005. Currently, he is a dean of College of Computer, University of Anbar. His research interested includes Data Warehouse, Software Engineering, and Distributed Database.

[2]**Salih S. Salih** has received his BSc. degree in computer science, Iraq in (2011-2012), he has been MSc. student (2014- tell now) in Computer Science Department, College of Computer, Anbar University. Fields of interest: Applying The Concept of a duplications elimination in the data warehouse. He has taught many subjects such as Database, Data Warehouse, Big Data and Data Mining.