Using Q-Gram and Fuzzy Logic Algorithms for Eliminating Data Warehouse Duplications

Dr. Murtadha M. Hamad[1, a] Salih S. Sami[1, b]

[1]College of Computer - Anbar University – Anbar – Iraq
[a] dr.mortadha61@gmail.com, [b] salihsami90@gmail.com

Abstract: ***Context:*** *The duplication system or record linkage has many applications in real life. It seems in a wide area of detecting the similar data join the web documents in wide web, detect the plagiarism and many application enter it, a proper choosing to enhance the data quality that leads to the help system to make the right decisions routing plays a considerable part in order to ameliorate the economic interests and suitability of logistics projects.* ***Problems:*** *In this study, the problem is as follows: Duplicate records data comes with the content of the ambiguity for refined other records that dates back to the same customer, especially since the recipes refined constraint contain the same major change in the data limitations and restrictions as well as contain the same simple change data.* ***Objectives:*** *The aim of this paper is to find an optimal solution for duplicate records detection and elimination by using fuzzy logic (FL) and Q-gram. We suggest achieving that goal with the following.* ***Objectives:*** *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.* ***Approach:*** *The Approach has been presented based on two* ***phases:*** *firstly, find the similarity records by Q-gram similarity; secondly, Classification record whether refined using fuzzy logic. Have identified the percentage threshold of 0.68 the researcher chosen this value based on the results obtained. If the similarity between the key ratio exceeded, it enters to the Fuzzy logic algorithm, which in turn determines if this record duplicated or not. The proposed work has an accuracy of 96%..*

Keywords: *Duplicate Elimination, Similarity score, Q-Gram, Fuzzy logic, Key Generation.*

1. Introduction

Data warehouse in generally the existence of unintended duplicate of records that was generated from millions of data from other database sources are difficult to avoid. In the community of data storage and the task of searching for duplication of records within the data warehouse for a long time continuous problem and has, convert an area of active study. There have remained several Research undertakings to address the problems of data duplication produced by duplicate contamination of data [1]. Data quality problems are trivially, compound and inconsistent. There is no international common standard for reference. So the process of data cleaning is vary from domain to domain but a process used to determine inaccurate, incomplete, or irrational data and then refining the quality done correction of detected errors and omissions. Duplicate discovery plays an importantly part in data clean up and data combination applications.

The problem of identifying and removing duplicated data is one of the main problems in the comprehensive area of data cleaning and data quality in the data warehouse. Many times, the same logical real world entity may have multiple representations in the data warehouse [2] [3]. Duplicate detection is the task of finding sets of records that refer to the same entities within a data file. This task is not trivial when unique identifiers of the entities are not recorded in the file, and it is especially difficult when the records are subject to errors and missing values [4] [5].

Discovery and elimination duplicate records Especially that refined constraint specification is a data are ambiguous, this are 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.

2.Background

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 optimal solution.

To design appropriate representation architecture to the proposed detecting and eliminating duplicated data.

To design and implement removal-duplicated system for detect 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 lead to minimize the size of DW, reduce the time of searching on the DW and enhancement the decision support system.

To test and validate the system's performance.

The problem of identifying and eliminating duplicated data is one of the major problems in the broad area of data cleaning and data quality in the data warehouse. Many times, the same logical real world entity may have multiple representations in the data warehouse [2]. Duplicate detection is the task of finding sets of records that refer to the same entities within a data file. This task is not trivial when unique identifiers of the entities are not recorded in the file, and it is especially difficult when the records are subject to errors and missing values [4].

In this paper, attempt to investigate some of the previous studies and related works on the discovery duplicate data and ways to deleted that are close and connected to our study. Jebamalar et al presented developed a data mining mainly focus on efficient detection and elimination of duplicate data. The main objective of their research work is to detect exact and inexact duplicates by using duplicate detection and elimination rules [5]. Bilal et al developed the techniques and methods for a de-duplicator procedure, which is depended on the numeric translation of entire data. For effectiveness, data mining method k-mean clustering useful on the numeric worth that decreases the number of comparisons between records. To identify and remove the duplicated records, divide and conquer technique is used to match records within a cluster, which further improves the efficiency of the algorithm [6].

Anju et al described and implemented a hybrid method are used for identifying duplicates in hierarchically structured XML data. Most aggressive machine learning procedures are used to derive the provisional probabilities for all new structure arrived. A technique known as binning technique they used to convert the outputs of support vector machine classifiers into accurate posterior probabilities. To improve the rate of duplicate detection network pruning is also employed. Through experimental analysis, it is shown that the proposed work yields a high rate of duplicates thereby achieving an improvement in the value of precision. This technique outdoes other duplicate finding solution in terms of effectiveness [7].

Data integration is the procedure of providing the customer with a united view of data residing at changed sources. The high attention in hesitation in data integration is motivated by many data integration applications in which uncertainty is unavoidable.

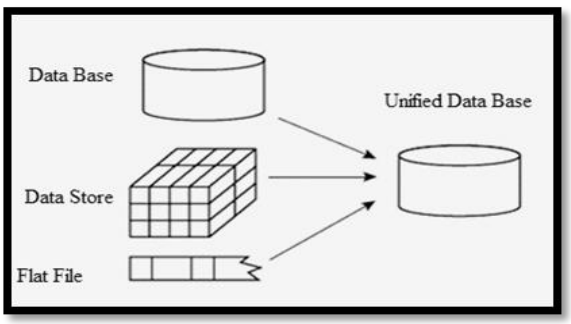


Figure 1. Data Gathering and integration

Data Cleaning is a very important process of the data warehouse [8]. Because there were many mistakes in the data warehouse so, the decision-making process will not be true to that it used many of the algorithms and procedures for data cleansing of homosexuals. The process of identifying and eliminating database defects and duplicates is conversed to as data cleaning [1].

3. Data Duplication

A data warehouse is generated by union large databases acquired from changed sources with heterogeneous representations of info. This raises the topic of data quality, the foremost being discovery, and removal of duplicates, crucial for accurate statistical analyses. Other than using own historical/transactional data, it is not uncommon for large businesses to acquire scores of databases each month, with a total size of hundreds of millions of over a billion records that need to be added to the warehouse. The fundamental problem is that the data supplied by various sources typically include identifiers or string data that are either different among different datasets or simply erroneous due to a variety of reasons including typographical or transcription errors or purposeful fraudulent activity, such as aliases in the case of names. Duplicated Data Occur In Two Ways:

- Repeated records, possibly with some values different.
- Different identifications of the similar real world entity. Repeat records are public and usually easy to detect. The different identification of the alike real- world entities can be a very hard problem to detect [8], [9].

4.Q-gram Function. Filtering and Indexing

Q-gram has many types according to the length of the substrings such as unigram with size of one, bigram with size of two, trigram with size of three, and size four or more is simply called a Q-gram. Q-gram might be defined as shortened substrings of length q then a given string. The substrings might be phonemes, syllables, letters, or words giving to the usage of Q-gram purpose. Letter Q-grams, containing trigrams, bigrams, and/or unigrams, have been used in a variety of ways in text recognition and spelling correction.

The notion of Q-grams for a given string σ, its Q-grams are obtained by "sliding" a window of length q over the characters of σ. Since Q-grams at the beginning and the end of the string can have fewer than q characters from σ we introduce new ∑ characters "#" and "%" not in σ, and conceptually extend the string σ by prefixing or padding it with q – 1 occurrences of "#" and suffixing it with q -1 occurrences of "%". Thus, each Q-gram contains exactly q characters, though some of these may not be from the alphabet ∑.

The intuition behind the use of Q-grams as a foundation for approximate string processing is that when two strings σ1 and σ2 are within a small edit distance of each other, they share a large number of Q-grams.

The Q-grams similarity metric between two strings is created ranging from 0 to 1.0 using a normalized formula.

Filtering algorithms are very sensitive to the error level α:= k=m since this usually affects the amount of text that can be discarded from further consideration. (m =pattern length, k = errors.).

Threshold As in the classic q-gram lemma, we define the threshold of a q-gram filter as a function of the length m of the pattern and the distance limit k. That is, the threshold t (m, k) is the smallest number of matching q-grams between a pattern of length m and a substring of the text that is within distance k of the pattern. The quantity of corresponding q-grams is a similarity purpose for strings [10].

5.Proposed Algorithm

The specifications of records can be confidential into the subsequent:

1. Fixed attributes, such as persons characteristics similar (Customer Name, Blood-Type and Gender).

2. Variable attributes, these can be divided into:

2.1 Largely changing, such as persons characteristics like (Marital-Status, and City) this attributes that be specific in list.

2.2 Small changing, such as those characteristics similar (Sales, Unit_Price, Age, Salary, Number_of_Children, Weight, and Length), which are frequently the characteristics that are numerical or measureable. These fields are helpful in remove the duplicates.

In this study the records is transitory into number of parts during the above phases .The structure of implemented system, As shown in the Figure (4).The records eliminations system consists of five important modules:

I. Key Generation.
II. Sorted Neighbored Methods.
III. Blocking.
IV. Stage Compression Key Selection.
V. Fuzzy Logic Technique.

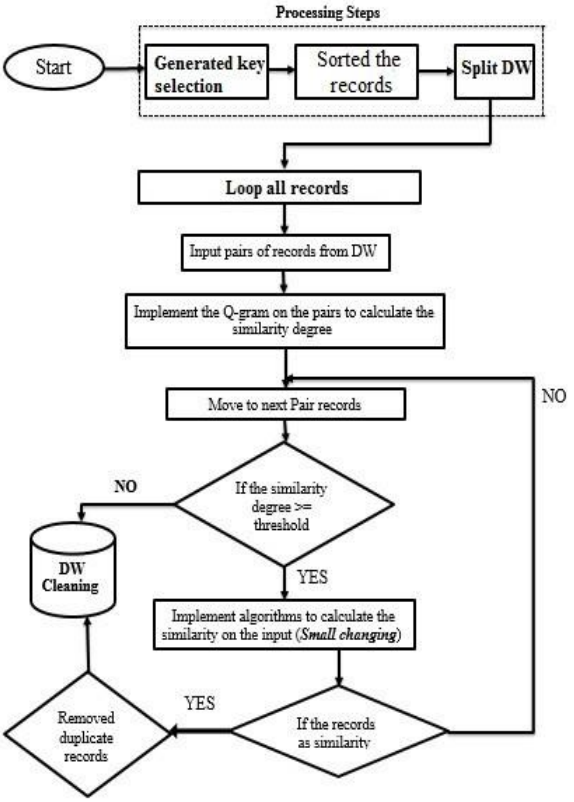


Figure 2. General system of duplicate records removals

Cust-id	Cust- name	ood-Type	gender	age	salary	Num_ of_ Children	Weight	Length	Address	Marital-Status
1	Safi ali samed	O+	M	54	513$	0	116	167	RAMADE	Single
2	Safi ali samed	O+	M	49	518$	2	121	171	IRBALE	Married
3	Sara hane	AB+	F	65	234$	0	87	155	HEET	Single
4	Sara hane	AB+	F	68	228$	1	91	154	Rawaa	Married

Figure 3. Duplicate records

I. Key Generation

From main fields only (fixed elements and from variable attributes that are largely hanging that can be of benefit only), for each field.

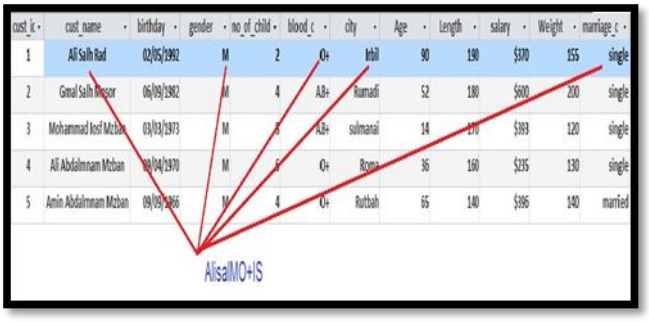


Figure 4. Key generation

- Choose the first three letters of each word from cust name, the word in filed gender, the word in filed blood-type, chose only first chart from filed city and chose only first chart from filed Marital _Status.
- Merge the attribute selected with each that building a key that represent a record see figure (4).
- You can delete duplicates, and sort alphabetically. Each field separately.
- Merge output for produce the main key field.

II. Sorted Neighbored Methods

In this stage, it was arranged (alphabetizing) in the records in a data warehouse in alphabetical order based on the

key Which explain in the previous stage, this stage are 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

Since the database contained a field for each under the blood type for each user, so we 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 Segmentation 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 filds 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 divade the strings in to multiy F paterrins depend on the no. of Q such is the string **Saleh** and **Q=3 so(#Sa, Sal, ale, leh, eh#)** that compresion between the **str1** and **str2** and After comparing the strings will produce a numeric value range between **(0 - 1)** by the equation(1) so we using Q-Gram on this key because it has the capacity to deal with such varieties of change in the Strings, and to become the records that things is duplicate. In this study using the q-gram simalirty methods on the key genrations to calculate the simalirty betwean this key, we aplay 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 on other fields will be covered to check whether similar or not in the next stage.

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

Whereas:

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

Consider the key generation (section **(i)**) as an input is the proposed algorithm (q-gram).

V. Fuzzy Logic Technique

FL is generated in 1965. By Lotfi A. Zadeh [11], [12]. With regard to uncertainty the best in dealing with it is a technique fuzzy logic. The technique fuzzy logic has the power to solve the problem that nature high adaptation of uncertainty and approximation. Fuzzy logic resembles anthropoid rational in its use of imprecise information to create decisions [12]. In this approch we used fuzzy inference technique is the so-called Mamdani method. The Mamdani-style fuzzy inference process is performed in four stages: fuzzification of the input variables, rule evaluation, aggregation of the rule outputs, and finally defuzzification.

❖ **Fuzzification**

The main stage is to take input and clear , in this approach , we use (Young , Medium, Old), linguistic variables to the field of age, (low , medium, many) for a number of children's field , (Little , Medium, many) for Salary and total field , (skinny , middling , fat) to the field of weight, (Short, middling , tall) for the length of the field.

❖ **Membership Function**

This phase is to take the fuzzified inputs, and we have produced the next membership functions for Age, Number of children, Salary, Weight and Length attribute in database that display in the figures 5,6,7,8,9.

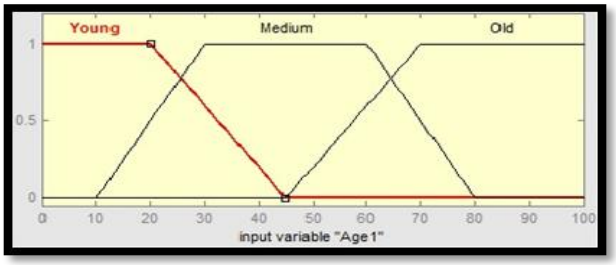


Figure 5. Membership functions for Age

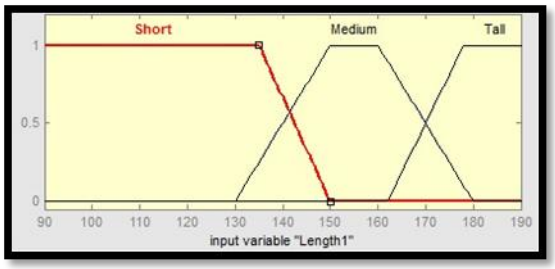


Figure 6. Membership functions for Length

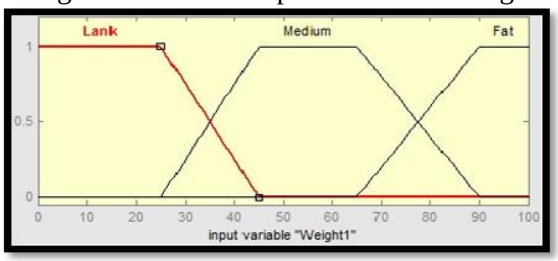


Figure 7. Membership functions for

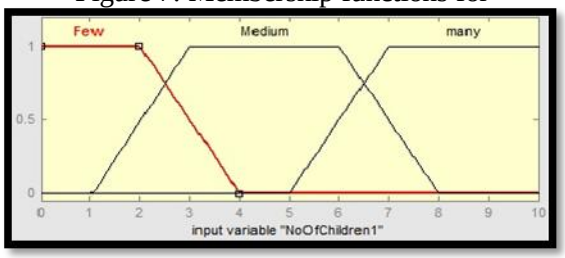
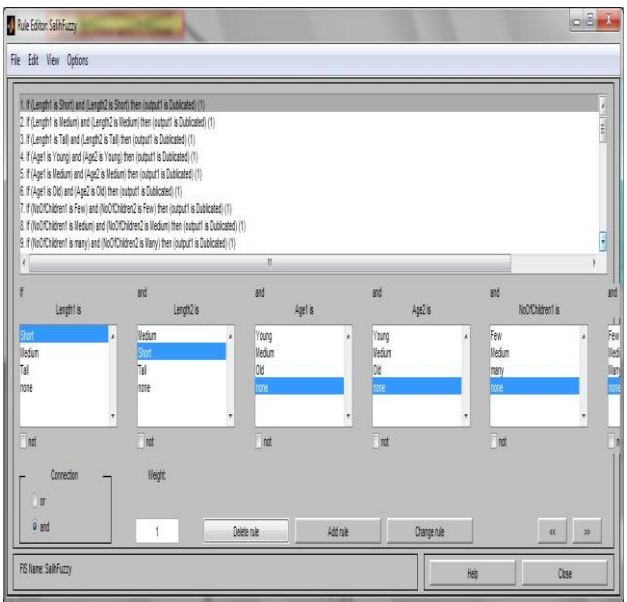


Figure 8. Membership functions for Num. of Child



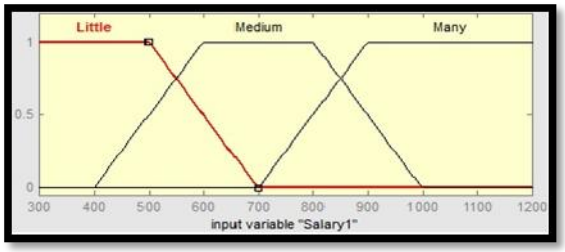


Figure 9. Membership functions for Salary

Figure 10. Knowledge base

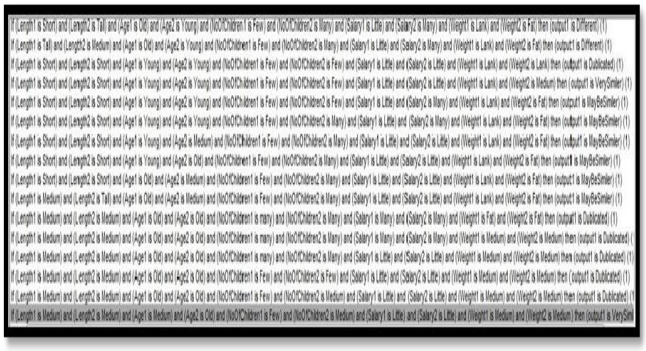


Figure 11. knowledge base for decide if the records are duplicate or no

❖ **Aggregation of the rule outputs**

Aggregation is the procedure of unification of the yields of all rules. In this study, produce the membership on the input (fuzzification) in the preceding stage, so in this step we need knowledge base is used to build many rules to continuo this system (fuzzy logic) we can see some of the rules that built about this purpose in figure (10), and figure (13) that work to classify if the two records duplicate or no.

❖ **Defuzzification**

At this stage, we have used the Centroid deffuzification method where this process is the most commonly used process

$$Z = \frac{\sum_{i=1}^{n} ci\mu Ai\,(x)\mu Bi\,(y)}{\sum_{i=1}^{n} \mu Ai\,(x)\mu Bi\,(y)} \qquad (2)$$

Where ci is the center of Ci, are universal approximates, i.e. so, in this step we take the center of all the values by the Equation (2).

VI. Results and Discussion

In this paper have been implemented actual measures to evaluate the performance of the system

work to get optimal criteria for 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.

a)Execution time

The run of proposed system shown the importantance to detect efficiency of the proposed system, the following figures illustrates the Execution time how long taken to detect and delete duplicates time also shown the size of database and calculate the average time of all implementation in sec.

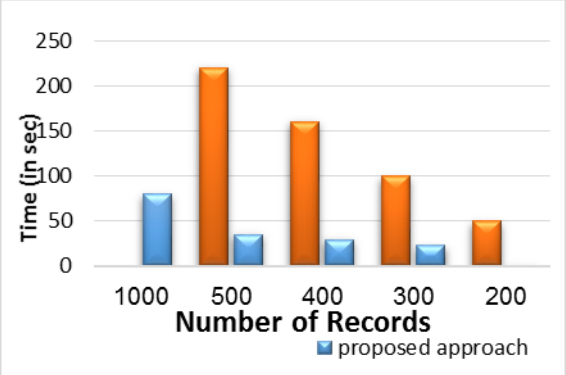


Figure 12. Performance in terms of computation time with other approach.

(b) Accuracy

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 (0.4, 0.5, 0.6 and 0.7) after work was the best value in giving output is 0.68, display in chart (13).

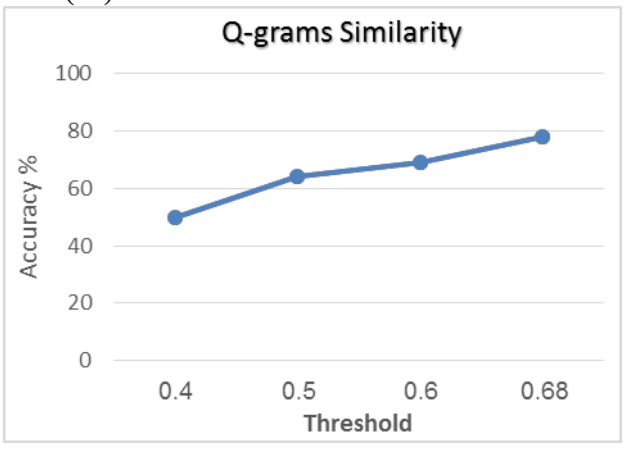


Figure 13. Accuracy on detect duplicate records by using Q-grams Similarity

ii. Also we have implemented fuzzy logic on the same numeric values fields and after the change threshold values was the best value (0.6) as planned described in (14).

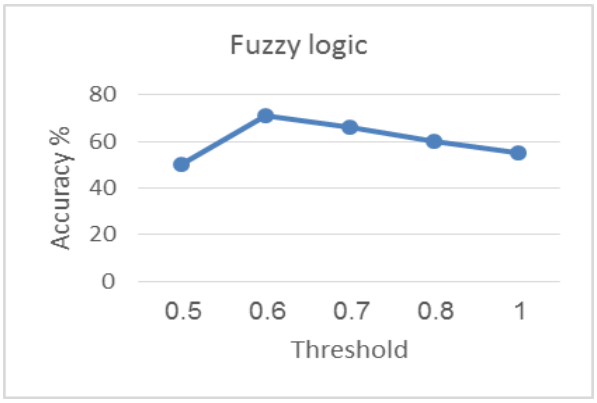


Figure 14. Accuracy on detect duplicate records by using Fuzzy logic

iii. After this we have implemented Q-Gram and fuzzy logic on the existing data have been obtained on the accuracy as illustrated in the chart (15) in detecting and deleting duplicate records process.

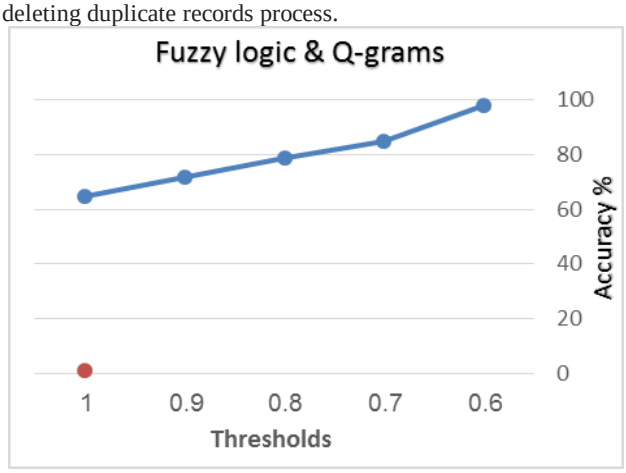


Figure 15. Accuracy on detect duplicate records by using **Fuzzy logic & Q-grams**

By employed after using the Q-Gram well as when using fuzzy logic alone does not give the desired results, but after the merger in the use of Q-gram and fuzzy logic, it gave good results.

6. CONCLUSION

We implemented the Duplicate records system to detection and elimination in the large database. In this approach, we used Q-Gram and Fuzzy logic to eliminations the duplicate records. Fuzzy logic algorithm is applied to solve the problem because it is

efficient in solving high computational complexity problems, especially when the solution space is large such as duplication data. It is clear that the algorithm has succeeded in achieving the paper objectives through optimizing the classification to detect duplicate that the fuzzy logic Algorithm is efficient in solving this type of problem it is capable in solving many other problems in real life. Along with, blocking technical depend on the blood type method was utilized to reduce the time taken on each comparison to improve duplicate detection. Furthermore, the q-gram Bring suspected records duplicate. In addition, the fuzzy logic take decision if the records as duplicate or no. This work is efficient in terms (q-gram- fuzzy logic) of accuracy, as well as a higher speed in comparison with the work of his predecessor discrimination.

REFERENCES

[1] Paulraj Ponniah and Sons **"Data Warehousing Fundamentals For It Professionals",** Second Edition, 2010.

[2] Shital Gaikwad and Nagaraju Bogiri, "**A Survey Analysis on Duplicate Detection in Hierarchical Data**", International Conference on Pervasive Computing (ICPC), IEEE, 2015.

[3] Rohit Ananthakrishna, Surajit Chaudhuri and Venkatesh Ganti, "**Eliminating Fuzzy Duplicates in Data Warehouses",** Proceedings of the 28th VLDB Conference, Hong Kong, China, 2002**.**

[4] Mauricio Sadinle," **Detecting Duplicates In A Homicide Registry Using A Bayesian Partitioning Approach ",** The Annals of Applied Statistics, Vol. 8, No. 4, pp. [2404–2434], 2014**.**

[5] Tamilselvi, and Brilly, "**Handling Duplicate Data in Data Warehouse for Data Mining",** International Journal of Computer Applications (0975 – 8887) Vol. 15, No.4, 2011.

[6] Khan, Rauf, Javed, Khusro, and Javed, " **Removing Fully and Partially Duplicated Records through K-Means Clustering",** IACSIT International Journal of Engineering and Technology, Vol. 4, No. 6, 2012.

[7] Abraham, and Kanmani," **A Novel Approach for the Effective Detection of Duplicates in XML Data",** International Journal of Computational Engineering Research, Vol. 04, Issue, 3, 2014.

[8] MATTEO MAGNANI and DANILO MONTESI, "**A Survey on Uncertainty Management in Data Integration**", ACM Journal of Data and Information Quality, Vol. 2, No. 1, Article 5, Pub. Date: July 2010.

[9] R.Kavitha Kumar And, Dr. Rm.Chadrasekaran ,"**Attribute Correction-Data Cleaning Using Association Rule And Clustering Methods**", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.2, 2011.

[10] J. Kärkkäinen,"**Computing the Threshold for q-Gram Filters**", M. Penttonen and E. Meineche Schmidt (Eds.), pp. [348–357], Springer-Verlag Berlin Heidelberg, 2002.

[11] Vandna Kamboj, Amrit Kaur **"Comparison of Constant SUGENO Type and MAMDANI-Type Fuzzy Inference System for Load Sensor",** International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Vol. 3, Issue-2, May 2013**.**

[12] Mohammed, M.A., "**Design and Implementing an Efficient Expert Assistance System for Car Evaluation via Fuzzy Logic Controller**". International Journal of Computer Science and Software Engineering (IJCSSE), 4(3), pp. [60-68], 2015.