Using Fuzzy Logic Technique to Eliminate the Duplicates in Large Database

Dr. Mortadha M. Hamad
College of computer
University of Anbar
Ramadi, Iraq
mortadha61@yahoo.com

Alaa Abdulqahar Jihad
Computer Science Department
University of Anbar
Ramadi, Iraq
alaa1985net@gmail.com

***Abstract*— Duplicate records are broad problem in many of the databases. There are wide efforts focusing on elimination of duplicate in data sets, because is it important part of data cleaning. This paper focuses on discovery and removing duplication by using fuzzy logic technique.**

Keywords—Duplicate; data quality; data set; fuzzy logic.

I. INTRODUCTION

Data quality is a key issue in computer-based management systems, it the degree to which data meets the specific needs of specific customers in any department [1][2]. Low quality of data leads to wrong conclusions that ultimately lead to wastage of all kinds of resources and assets [3]. Poor data quality costs businesses vast amounts of money every year.

Data cleaning is the processing of detect and remove errors and inconsistencies from data and improves their quality [4]. The problem of detecting and eliminating duplicated data is one of the major problems the broad area of data cleaning in data warehouse [5]. This paper focuses on records duplicate problem, and use fuzzy logic technique to eliminate it.

II. BACKGROUND

Several researchers in data processing developed data cleaning techniques to optimize data quality, especially for eliminate the duplicates records in data set.

In [6], Rohit Ananthakrishna, Surajit Chaudhuri and Venkatesh Ganti presented Eliminating Fuzzy Duplicates in Data Warehouses. In this paper, they developed an algorithm for eliminating duplicates in dimensional tables in a data warehouse, which are usually associated with hierarchies. They exploited hierarchies to develop a high quality, scalable duplicate elimination algorithm, and evaluated it on real datasets from an operational data warehouse.

In [5,] Kazi Shah Nawaz Ripon, Ashiqur Rahman and G.M. Atiqur Rahaman presented A Domain-Independent Data Cleaning Algorithm for Detecting Similar-Duplicates. This paper presented a domain-independent similar duplicate detection algorithm for data cleaning in data mining. The sorting method proposed in this paper was a novel technique for bringing the similar-records closer and it can sort the records efficiently. The proposed algorithm also explores the idea to make the duplicate detection algorithms faster and more efficient for real-life data. This paper also outlined an algorithm to implement all the proposed methods together.

In [7], Joshua M. Horstman and Roger D. Muller presented Dealing with Duplicates in Your Data. In this paper, they addressed one of the most common data issues: duplicate data. Using simple examples, they described different types of duplicate data and describe strategies for identifying questionable records so they can be evaluated and appropriate action taken. The paper also featured an in-depth discussion of options on PROC SORT (NODUP, NODUPKEY, and DUPOUT) as well as alternative methods such as PROC FREQ and PROC SQL.

In [5], M. Anitha, A.Srinivas, T.P.Shekhar and D.Sagar presented Duplicate Detection Of Records In Queries Using Clustering. In this paper work is to detect exact and inexact duplicates by using duplicate detection and elimination rules.

Some products provide tools for data cleaning. For instance, SQL Server provides a tool for data cleaning called Fuzzy Grouping. The ETL tool performs data cleaning by identifying rows of similar or duplicate data and choosing a canonical row to represent the rows of the data.

III. REASON OF DUPLICATES

Duplicate data is a situation in which multiple copies of data exist scattered around on various systems and in various versions [7].

Duplicated records occur in many databases. The duplicated data are either repeated records (perhaps with some values different), or different identifications of the same real world entity. There are two basic instances where records duplicate:

1. Database receives data from several sources, according to the database system; it gets a repeat of records.
2. Or in local database, data is entered into database more than once.

Whatever, there is a reason that leads to enter data for the same record more than once, and may be entered

incomplete records, and data are completed in the corresponding records.

IV. PARTITION OF ATTRIBUTES

Attributes can be classified into the following:

1. ***Fixed attributes***, such as those attributes like (Customer Name, Birthday, and Gender).
2. ***Variable attributes***, these can be divided into:
   a. ***Largely changing***, such as those attributes like (Region, Marital_Status, and Address) this attributes that be specific in list.
   b. ***Small changing***, such as those attributes like (Total, 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 eliminate the duplicates.

V. FUZZY LOGIC APPLICATION

The fuzzy modeling approach deal with the uncertainty [13]. Fuzzy logic is particularly adapt for environmental problems because of the high level of uncertainty and approximation [9][11].

Fuzzy processing in data warehouses can affect many operations, like data selection, filtering, aggregation, and grouping [10]. To avoid the limits of conventional logic in resolution we use fuzzy logic in eliminating the duplicates.

Fuzzy logic resembles human reasoning in its use of approximate information to generate decisions. In relation to matching, the term is used loosely to describe the approach that relies on rules that are imprecise rather than precise and operates on data with boundaries that are not sharply defined [8].

The steps fuzzy logic is:

- Linguistic term selection.
- Membership function selection.
- Generate rules set (if-then rules).
- Apply the rules.
- Defuzzification (getting back to crisp numbers).

This type of fuzzy logic system was first proposed by Mamdani [12]. The main components' are shown in Figure (1):

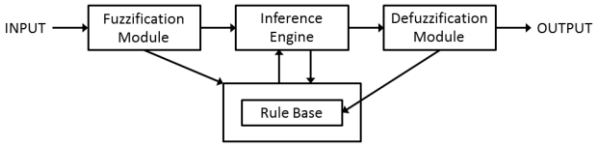


Figure (1) Fuzzy logic system

A. The proposed algorithm:

Input: Data set contains duplicate records
Output: Data set without duplicate records

1. Start.
2. Creating Major Key from major fields only (fixed attributes and from variable attributes that largely changing that be useful only), and place it in a new field.
3. Arrangement of records based on this key.
4. Beginning from the first record to the last record.
   a. If current created key is not equal to previous created key, put this record in the final table directly.
   b. If current created key is equal to previous created key, Apply fuzzy logic:
      i. Calculate the degree of membership for each of the non-fixed fields (variable attributes that small changing) for tow records.
      ii. Apply the proposed rules.
      iii. Find a center for each resulting values.
   c. Check the result of fuzzy logic for these two records, if similar, deletes this record, after taking the data incomplete and complement, but if not similar, put this record in final table.
5. Calculation of ratio the number of duplicates that deleted.
6. Evaluate the system, and display results.
7. End.

B. Creating the Major Key

From major fields only (fixed attributes and from variable attributes that are largely changing that can be of benefit only), for each field:

- Delete any space and characters such as '-'.
- Choice length for each field.
- Copy letters from fields based on length, and replace by '$' if empty.
- Merge all fields to create the Major Key.

C. Linguistic Term Selection (Fuzzy Sets)

Linguistic variables are described for a specific case such as recipes, actions or circumstances. Difference in the time and circumstances execute into difference in some data attributes.

In this paper, we use (Young, Medium, Old) linguistic variables for Age field, (Many, Medium, Few) for Number of Children field, (Many, Medium, little) for Salary and Total field, (Lank, Middling, Fat) for Weight field, (Short, Middling, Tall) for Length field, (Similar, May be Similar, Different) for Similarity function, (Different, Very Similar, Duplicate) for Duplicate function.

D. Membership Function Selection

We have created the following membership functions for Age, Number of children, Salary, Weight and Length attribute in database that used for test. Figures 2, 3, 4, 5 and 6 represent the membership functions.

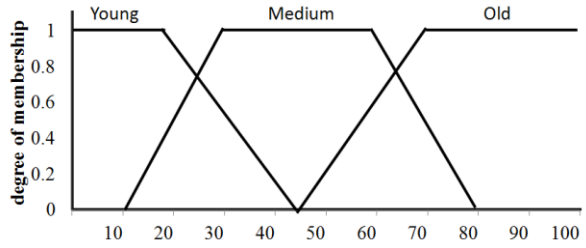


Figure (2) Membership functions for Age

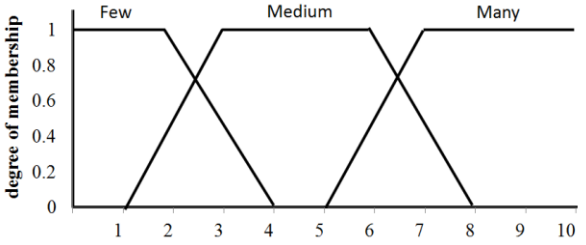


Figure (3) Membership functions for Number of Children

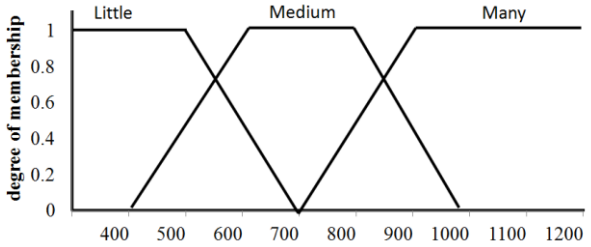


Figure (4) Membership functions for Salary

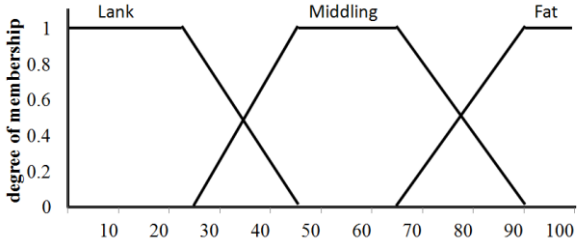


Figure (5) Membership functions for Weight

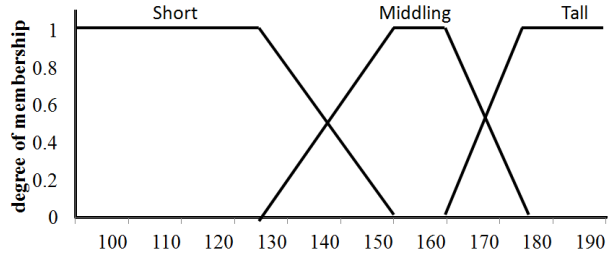


Figure (6) Membership functions for Length

Figures 7 and 8 show Membership functions for Similarity and Duplicate.

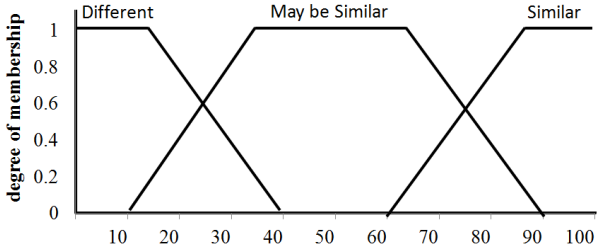


Figure (7) Membership functions for Similarity

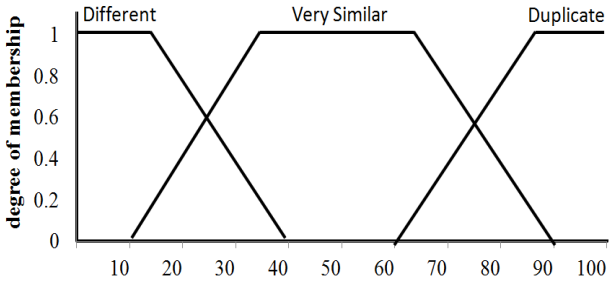


Figure (8) Membership functions for Duplication

E. Fuzzy rule base for the model

After the Fuzzification of input data, processing takes place in rule base of the fuzzy system. In this step, knowledge base for complete the proposed fuzzy logic system is built, where we can build many of rules, such as Figure 9 where Age1 is age field in first record and Age2 is age field in second record and so on. Figure 10 shows the rule that must be build to decide if these two records are duplicate or no.

If **Age1** is **Young** and **Age2** is **Young** Then **Similarity** is **Similar**
If **Age1** is **Medium** and **Age2** is **Medium** Then **Similarity** is **Similar**
If **Age1** is **Old** and **Age2** is **Old** Then **Similarity** is **Similar**
If **Age1** is **Young** and **Age2** is **Medium** Then **Similarity** is **May be Similar**
…
If **Age1** is **Young** and **Age2** is **Old** Then **Similarity** is **Different**
…
If **Salary1** is **Many** and **Salary2** is **Many** Then **Similarity** is **Similar**
…

Figure (9) Knowledge base

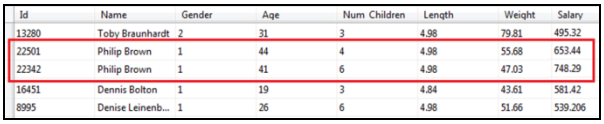
If **Age_Similarity** is **Similar** and **Num_Children_Similarity** is **Similar** and **Salary_Similarity** is **Similar** and **Weight_Similarity** is **Similar** and **Length_Similarity** is **Similar** then **Duplication** is **Duplicate**
If **Age_Similarity** is **May be Similar** and **Num_Children_Similarity** is **Similar** and **Salary_Similarity** is **Similar** and **Weight_Similarity** is **Similar** and **Length_Similarity** is **Similar** then **Duplication** is **Duplicate**
…
If **Age_Similarity** is **May be Similar** and **Num_Children_Similarity** is **May be Similar** and **Salary_Similarity** is **Similar** and **Weight_Similarity** is **Similar** and **Length_Similarity** is **Similar** then **Duplication** is **Very Similar**
…
If **Age_Similarity** is **Different** then **Duplication** is Different
If **Num_Children_Similarity** is **Different** then **Duplication** is Different
…

Figure (10) Knowledge base for decide if the records are duplicate or no

VI. CONCLUSIONS

To reach correct decisions, the data used must be correct. The process of removing matched records is complex and difficult. The proposed algorithm focuses on the fields that are of few changes such as age, number of children salary and includes fixed fields such as name, gender, etc. The concept of fuzzy logic has been applied to these

fields and we got good results. The algorithm does not only depend on the major fields, but also takes the secondary fields into consideration. Figure (11) shows some of discovered duplicates.



Id	Name	Gender	Age	Num Children	Length	Weight	Salary
13280	Toby Braunhardt	2	31	3	4.98	79.81	495.32
22501	Philip Brown	1	44	4	4.98	55.68	653.44
22342	Philip Brown	1	41	6	4.98	47.03	748.29
16451	Dennis Bolton	1	19	3	4.84	43.61	581.42
8995	Denise Leinenb...	1	26	6	4.98	51.66	539.206

Figure (11) Discovered duplicates.

The proposed algorithm using fuzzy logic gives clear flexibility in dealing with data, away from limits that go to desuetude many of duplicates. Recommendation is to apply the fuzzy logic to discover the similarity in textual data.

REFERENCES

[1] Vasarhelyi, M., and M. Greenstein, Underlying principles of the electronization of business: a research agenda, International Journal of Accounting Information Systems, 2003.

[2] Dr. Linda F. Ettinger, Improving the Data Warehouse with Selected Data Quality Techniques: Metadata Management, Data Cleansing and Information Stewardship, University of Oregon, December 2005.

[3] R. Arora, P. Pahwa, S. Bansal, Alliance Rules of Data Warehouse Cleansing, IEEE , International Conference on Signal Processing Systems, Singapore, May 2009, Page(s): 743 – 747.

[4] Kazi Shah, Ashiqur Rahman and G.M. Atiqur Rahaman, A Domain-Independent Data Cleaning Algorithm for Detecting Similar-Duplicates, Khulna University, Bangladesh, 2010.

[5] M.Anitha, A.Srinivas, T.P.Shekhar and D.Sagar, Duplicate Detection Of Records In Queries Using Clustering, Karimnagar, India, International Journal of Research in Computer Science eISSN 2249-8265 Volume 2 Issue 2, 2012.

[6] Rohit Ananthakrishna, Surajit Chaudhuri and Venkatesh Ganti, Eliminating Fuzzy Duplicates in Data Warehouses, Hong Kong, China, 2002.

[7] Joshua M. Horstman, Roger D. Muller, Dealing with Duplicates in Your Data, MWSUG 2011.

[8] Jean-Pierre Dijcks, Matching and Merging data – Black Art or Exact Science, Oracle Corporation, January 2008.

[9] So S.S., Cha S.D., Kwon Y.R. Empirical evaluation of a fuzzy logic-based software quality prediction model, Fuzzy Sets and Systems, 127 (2), pp. 199-208, 2002.

[10] Dariusz Mrozek, Fuzzy Data Warehouse and Fuzzy OLAP Project Home Page, Institute of informatics, Poland, Accessed in 26/01/2015.

[11] Zabeo A., Semenzin E., Torresan S., Gottardo S., Pizzol L.1, Rizzi J., Giove S., Critto A. and Marcomini A., Fuzzy logic based IEDSSs for environmental risk assessment and management, International Environmental Modelling and Software Society (iEMSs), 2010, Canada.

[12] Kunwar Babar Ali, Anjana Gosain, Predicting The Quality of Object-Oriented Multidimensional (OOMD) Model of Data Warehouse Using Fuzzy Logic Technique, International Journal of Engineering Science & Advanced Technology, 2012.

[13] Chi-Yuan Yeh, Wen-Hau Roger Jeng, and Shie-Jue Lee, Data-Based System Modeling Using a Type-2 Fuzzy Neural Network with a Hybrid Learning Algorithm, IEEE TRANSACTIONS ON NEURAL NETWORKS, Taiwan, 2011.