ISSN:0975-9646



# An Approach Based on Artificial Neural Network for Data Deduplication

[1]M.Padmanaban and [2]T.Bhuvaneswari

[1] *Department in Computer Science, D.R.B.C.C. Hindu College,*
*Dharmamurthynagar, Pattabiram, Chennai-6000072*
[2]*Department of Computer Science*
*Government Arts and Science College*

***Abstract*— Data quality problems arise with the constantly increasing quantity of data. The quality of data stored in real-world databases are assured by the vital data cleaning process. Several research fields like knowledge discovery in databases, data warehousing, system integration and e-services often encounter data cleaning problems. The fundamental element of data cleaning is usually termed as deduplication that is the process of identifying the documents signifying the same entity, i.e. redundant data. The proposed technique consists of a method based on artificial neural network for deduplication technique. A set of data generated from some similarity measures are used as the input to the proposed system. There are two processes which characterize the proposed deduplication technique, the training phase and the testing phase. The proposed approach is tested with two different datasets for the evaluating the efficiency. The experimental results showed that the proposed deduplication technique has higher accuracy than the existing method. The accuracy obtained for the proposed deduplication at optimal threshold is 79.8%.**

***Keywords*—Deduplication, artificial neural network, training, testing**

## I. INTRODUCTION

Data warehouses that are archives of data gathered from numerous data sources constitute the foundation of the majority of existing decision support applications and CRM (Customer Relationship Management). Precision of decision support analysis on data warehouses is vital because important business decisions are influenced by such analysis [2]. Independent and practically incompatible standards may be followed by data sources as they are independent. In most cases, query results are a mix of pages containing different entities that share the same name. In an ideal retrieval system, a user would simply input an entity or concept name and receive search results clustered according to the different entities/concepts that share that name. One method to advance such system is to include additional information in the indexed documents [3]. Normally, organizations become aware of sensible precise disparities or inconsistencies while accumulating data from different sources to employ a data warehouse. Such problems belong to the category called data heterogeneity [4]. Erroneous duplication of data occurs when information from diverse data sources that store overlapping information is integrated [5]. But, data received at the data warehouse from external sources have errors like spelling mistakes, inconsistent conventions across data sources, omitted fields etc.,. Arriving data tuples from outside sources need rationale and modification for providing high data quality [2]. An `error-free' procedure in the data warehouse is recommended by data quality. Data cleaning techniques are essential to improve the quality of data [6].

Data mining methods and techniques utilizing software or algorithms called data mining tools, effectively mine and provide an edifying and useful analysis [7]. Predictive and Descriptive are two types of data mining models. The descriptive models like Clustering, Summarization, Association rule, Sequence discovery etc discover the properties of the analyzed data by recognizing the patterns or relationships in data [8]. Clustering classifies a set of objects into a number of subsets called clusters such that objects of the same cluster are exceedingly similar and objects of different clusters are dissimilar to each other [9]. Clustering is recognized as an important process to condense and summarize the information because it can provide a synopsis of the stored data [10]. The predictive model like Classification, Regression, Time series analysis, Prediction etc., using the known values determine the unknown data [8]. Classification technique which comprises many decision-theoretic methods for recognizing data is a wide ranging research field capable of processing a broader category of data than regression [11,12]. The classification algorithm constructs a model by learning from the training set. New objects are classified by using this model [13]. Numerous classification algorithms available are k-nearest neighbor (k-NN) algorithm [14], neural network [15], Naive Bayes, decision tree [16, 17], Bayesian network [18], and support vector machine (SVM) [19]. The exceedingly popular classification algorithm called k-NN exhibits good performance characteristics [20] and it is used in several diverse applications [21] like 3-dimensional object rendering, content-based image retrieval [22], statistics (estimation of entropies and divergences) [23], biology (gene classification) [24-26].

Data cleaning, also called data cleansing or scrubbing, enhance the quality of data by identifying and eradicating errors and inconsistencies from the data [27]. It aims at enhancing the overall data compatibility by concentrating on eradication of changes in data contents and minimizing data repetition. Record duplicates, omitted values, record and field resemblances and duplicate eradications are detected by current data cleaning techniques [28], [6]. Detection of other or several records that signify one distinct real world entity or object is performed by the duplicate record detection process [4, 29]. The difficulty of duplicate detection is really to find whether the same real-

world object is represented by two or more distinct database entries. Record linkage, object identification, record matching etc., are remarkable names for Duplicate detection. It is a greatly researched topic and has high importance in fields such as master data management, data warehousing and ETL (Extraction, Transformation and Loading), customer relationship management, and data integration [4]. The two innate problems that must be addressed by duplicate detection are quick detection of all duplicates in large data sets (efficiency) and proper determination of duplicates and non-duplicates (effectiveness) [30].

Clustering is the categorization of objects into diverse groups, or more exactly, the division of a data set into subsets (clusters), so that the data in each subset (ideally) reveal a few common traits frequently proximity based on certain defined distance measure. The process of splitting database into a set of mutually exclusive subsets (blocks) such that matches do not occur across blocks is commonly termed as blocking. Hence, the efficiency of duplicate detection is increased by blocking which substantially improves the speed of the comparison process [31]. For example, classifying a set of people records based on the zip codes in the address fields, avoids comparing records that have different zip codes [32]. Fingerprint-based and full text-based are the two types of common duplicate detection methods [33]. Textual similarity, typically quantified using a similarity function [34] such as, edit distance or cosine similarity is utilized by most of the current approaches to determine if two representations are duplicates [35].

In this paper, a technique for deduplication is plotted based on the artificial neural network (ANN). The documents are processed initially with some similarity measures namely, dice coefficient, Damerau-Levenshtein distance, and Tversky index. The similarity measures are used to generate the model parameters for the documents that are subjected for testing deduplication. The model parameters calculated are used for processing with the ANN. The ANN has two phases, one is the training phase and other is the testing phase. In the training phase, the ANN is trained to fix some result for the hidden layer according to the input feature and target feature. The training phase is targeted to find the duplicates and non-duplicates from the given inputs. The proposed deduplication technique is evaluated by testing it with two different dataset namely, Restaurant dataset and Cora dataset.

The main contributions of the proposed approach are,

- A set of model parameters are selected from three different similarity measures.
- An artificial neural network is designed in specific to the deduplication.
- Weightage parameter for the neural network is calculated from the training phase.
- In the testing phase, the process of deduplication executed according to the training data.

## II. REVIEW OF RELATED WORK

A handful of researches are available in literature for deduplication. In recent times, deduplication is in distributed manner, has attracted researchers significantly due to the demand of scalability and efficiency. Here, we review the recent works available in the literature for deduplication and the different techniques used for it.

Hong-Jie Dai *et al.* [3] have conducted a survey of advanced Entity Linking techniques. They gave a survey of the EL tasks in the general and the biomedical domain. Results of their EL work were provided for reference, which uncover new EL challenges found in biomedical text mining, along with discussions regarding their possible solutions. K.Deepa *et al.* [35] have proposed an approach to Duplicate Record Detection Using Similarity Metrics and ANFIS. They developed a domain independent approach to detect duplicate records presented in large databases. The approach adopts ANFIS and similarity functions to improve duplicate detection. The main aim of using ANFIS was to reduce the time taken for making decisions in detecting the duplicates. To minimize the number of record comparisons, an appropriate clustering method, known as K-means clustering was used in the duplicate detection phase. Their method was tested on the real-life datasets and the performance was evaluated with the evaluation metrics. They showed that their proposed approach detect duplicates efficiently and accurately. Hanna Kopcke *et al.* [36] have comparatively analyzed 11 proposed frameworks for entity matching. Their study considers both frameworks, which do or do not utilize training data to semi automatically find an entity matching strategy to solve a given match task. Moreover, they considered support for blocking and the combination of different match algorithms. Their study aims at exploring the current state of the art in research prototypes of entity matching frameworks and their evaluations. The proposed criteria should be helpful to identify promising framework approaches and enable categorizing and comparatively assessing additional entity matching frameworks and their evaluations. P. Christen [37] has conducted a survey on Indexing Techniques for Scalable Record Linkage and Deduplication. They presented a survey of twelve variations of six indexing techniques. Their complexity was analysed, and their performance and scalability were evaluated within an experimental framework using both synthetic and real data sets. They aimed at reducing the number of record pairs to be compared in the matching process by removing obvious non-matching pairs, while at the same time maintaining high matching quality. T.A. Faruquie *et al.* [38] have presented a system that was easily configurable to suit the data cleansing needs of an enterprise. They build a system that was scalable, elastic, and configurable. Each enterprise has unique needs, which makes it necessary to customize both the infrastructure and the cleansing algorithms to address these needs. Gianni Costa *et al.* [39] have proposed an incremental technique for discovering duplicates in large databases of textual sequences, i.e., syntactically different tuples, that refer to the same real-world entity. Each newly arrived tuple was assigned to an appropriate cluster via nearest-neighbor classification. This was achieved by means of a suitable hash-based index, which maps any tuple to a set of indexing keys and assigns tuples with high syntactic similarity to the same buckets. An extensive

experimental evaluation on both synthetic and real data has shown the efficacy of their proposed approach.

## III. MOTIVATION BEHIND THE RESEARCH

Several systems that rely on consistent data to offer high quality services, such as digital libraries and e-commerce brokers, may be affected by the existence of duplicates, quasireplicas, or near-duplicate entries in their repositories. Because of that, there have been significant investments from private and government organizations in developing methods for removing replicas from its data repositories. The redundant data exist in the data repository causes problems like excess memory usage, high execution time etc. So in order to overcome the difficulties, techniques like deduplication algorithms are used to find and separate the redundant data in a data repository. Gianni Costa *et al* [1] have proposed an approach to record deduplication using incremental clustering. The incremental clustering based approach to record deduplication that combines several different pieces of evidence extracted from the data content to find a deduplication function that is able to identify whether two entries in a repository are replicas or not. Inspired from the above research, a deduplication method is implemented in this paper with artificial neural network.

### *A. Artificial neural network (ANN)*

An artificial neural network is a system based on the operation of biological neural networks, in other words, is an emulation of biological neural system. ANN is a technique used for specific processes like classification, optimization, etc.A neural network can perform tasks that a linear program cannot.When an element of the neural network fails, it can continue without any problem by their parallel nature. A neural network follows a learning procedure, thus it need not be reprogrammed. An artificial neural network is mainly of two types,

1. **Feed forward neural network**, where the data ow from input to output units is strictly feed forward. The data processing can extend over multiple (layers of) units, but no feedback connections are present, that is, connections extending from outputs of units to inputs of units in the same layer or previous layers.
2. **Recurrent neural networks** that do contain feedback connections. Contrary to feed-forward networks, the dynamical properties of the network are important. In some cases, the activation values of the units undergo a relaxation process such that the neural network will evolve to a stable state in which these activations do not change anymore. In other applications, the changes of the activation values of the output neurons are significant, such that the dynamical behavior constitutes the output of the neural network.

In the current work we are using feed forward neural network for the purpose of deduplication. The neural network we used is a multi-layered neural network. The basic structure of a neural network includes three basic layers a data layer, a hidden layer and an output layer. The structure can be illustrated through the following figure.

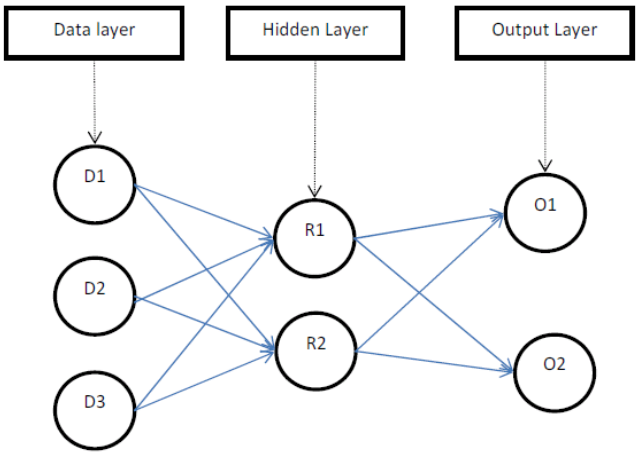


Figure. 1 Basic Structure of multilayered Neural Network

Here the data layer contains data which is given as the input and the hidden layer contain values, which is obtained from the computation of the data inputs and randomly generated weights. The two process executed through the neural network are ***training phase*** and the ***testing phase***. In the training phase, the data input are feed to the nodes in order to find the weights between each node. This will weight value calculated in the training phase will give the output value in the testing phase. 2

## IV. PROPOSED DEDUPLICATION METHOD WITH NEURAL NETWORK

A neural network based method gives high impact to the problem of deduplication with ANN's advance leaning architectures. So once it has executed then nothing can alter the process of ANN. The ANN can be implemented with any application and can be implemented with any problem. The initial step regarding the deduplication based on the artificial neural network is to find the model parameters generated from the similarity functions. The similarity function, which we used are

1. Dice coefficient
2. Damerau–Levenshtein distance
3. Tversky index

The value generated from the above plotted similarity distance measures are used as the input to the ANN. The documents, which are to be tested for the data redundancy, are processed with similarity measure and each of the measure will produce model parameters. These parameters are the basic processing units of the artificial neural network.

### 1. Dice coefficient

Dice coefficient is a similarity measure identical to the Sørensen similarity index, referred to as the Sørensen-Dice coefficient. It is not very different in form from the Jaccard index but has some different properties than the jaccard index. The function ranges between zero and one, like Jaccard. Unlike Jaccard, the corresponding difference function $d = 1 - (2 | X \cap Y |) / | X | + | Y |$ is not a proper distance metric as it does not possess the property of triangle inequality. The similarity function for the dice's coefficient can be given by the following expressions,

$$S = \frac{2 | X \cap Y |}{| X | + | Y |}$$

Where S represents the similarity measure, X and Y are the documents used for the comparison. The resultant od S is a set of model parameters.

**2. Damerau–Levenshtein distance**

In information theory and computer science, the Damerau–Levenshtein distance is a "distance" between two strings, i.e., finite sequence of symbols, given by counting the minimum number of operations needed to transform one string into the other, where an operation is defined as an insertion, deletion, or substitution of a single character, or a transposition of two adjacent characters. The name Damerau–Levenshtein distance is used to refer to the edit distance that allows multiple edit operations including transpositions, although it is not clear whether the term Damerau–Levenshtein distance is sometimes used in some sources as to take into account non-adjacent transpositions or not. The similarity algorithm of the Damerau-Levenshtein distance yields a set of the model parameters for the processing of neural network.

**3. Tversky Index**

The Tversky index is an asymmetric similarity measure that compares a variant to a prototype. The Tversky index can be seen as a generalization of Dice's coefficient and Tanimoto coefficient. For sets X and Y of keywords used in information retrieval, the Tversky index is a number between 0 and 1 given by

$$S(X,Y) = \frac{| X \cap Y |}{| X \cap Y | + \alpha | X - Y | + \beta | Y - X |}$$

Where, $\alpha$ and $\alpha$ are the parameters of the Tversky index. The similarity measure also provide a set of model parameters.

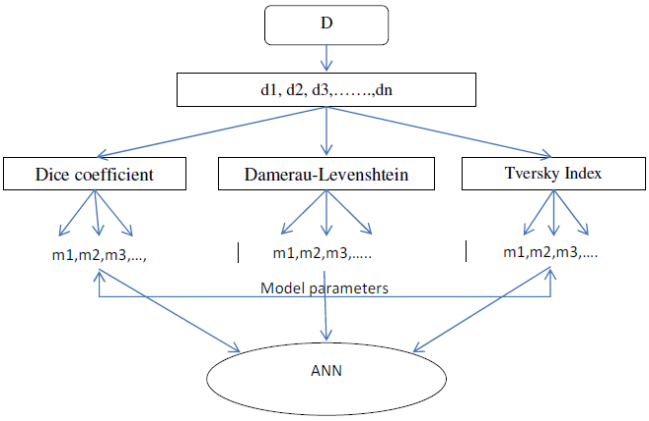


Figure. 2 Process diagram

The above plotted figures 2 shows the processing diagram of the neural network based deduplicate method. As mentioned above, a multilayered technique is used in the proposed neural network based deduplication method.

Consider a document set D which includes a set of duplicate and non-duplicate documents. The set of documents can be represented as,

$$D = [d_1, d_2, ...., d_n], \quad d \in D \text{ and } n=1,2,3,\ldots$$

Now the set of documenst are subjected for the processing with the similarity measures. The similarity measures used in the proposed approach are Dice coefficient (DC), Damerau-Levenshtein (DL) and Tversky Index (TI). The similarity measures process the input documents from the document set *D* and produces the model parameters. Each of the similarity measures produces model parameters individually for the document set D. Three similarity measures is used for the computation of the model parameters because a model parameters are the most significant factor in the proposed neural network method. Thus the model parameter calculation should be precise and accurate. The model parameters produced are listed in the following set of data.

$$M_{DC} = [m_1, m_2, ...., m_n]$$
$$M_{DL} = [m_1, m_2, ...., m_n]$$
$$M_{TI} = [m_1, m_2, ...., m_n]$$

The above listed are the set of model parameter generated based on the similarity measures stated above. The next phase of the proposed approach is to sort and combine the three set of model parameters for the processing of the deduplication with ANN. The input to the ANN is two set, which are, a set of data consist of the sorted model parameters and a set of data consisting of the weightage values. The weightage value is a parameter set for the neural network.

$$M_{Sort} = [m_1, m_2, ...., m_n]$$
$$W = [w_1, w_2, ...., w_n]$$

Where $M_{sort}$ is the sorted model parameters values and the set *W* represents the set with weightage parameters of the neural network.

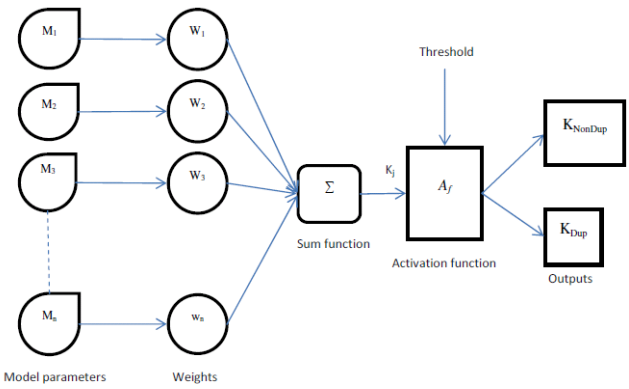


Figure. 3 ANN for Deduplication

The above is the design of the artificial neural network for the deduplication purpose. The $K_j$ is the output obtained by calculating weightage and the model parameters. In this manner we get output for all the model parameters from the set $M_{sort}$. The generalized expression for the calculation of $K_j$ is given by,

$$K_j = \sum_{j=1}^{n} w_j . m_j$$

Where, $K_j$ is the output generated by processing all of the weightage value and the model parameters. The ANN designed for the proposed deduplication technique will generate two output values $K_{NonDup}$ and $K_{Dup}$. The value $K_{NonDup}$ is specific for the non-duplicate documents and

$K_{Dup}$ is specific for duplicate documents.In the figure showing the model of the artificial neural network designed for the proposed deduplication process, we can find a function named as activation functions represented by $A_f$. The activation function acts as a squashing function, such that the output of a neuron in a neural network is between certain values like 0 and 1. In general there are three types of activation functions, but in the proposed approach the following activation function is used. A threshold is set for driving the $K_j$value to some value in therange [0,1].

$$A_f(K_j) = \begin{cases} 1, if \ \ K_j \geq 0 \\ 0, if \ K_j \leq 0 \end{cases}$$

There is the Threshold Function which takes on a value of 0 if the summed input is less than a certain threshold value ($K_j$), and the value 1 if the summed input is greater than or equal to the threshold value. According to the final $K_j$ values we execute the deduplication.

### *A. Training Phase*

The ANN designed for the deduplication purpose goes through a training phase. In the training phase, a set of input is given to fix the weightage value for the ANN based on the deduplication requirements. The training phase is conducted with two input layers, one contains the input values from the model parameters and the other contains the output values stating the duplicates and non-duplicates.The training phase is characterized by two layers input feature and target feature. In accordance with the input feature and output features the neural network will generate the weightage values in specific to the deduplication process. The main processing sequenc of the training processes can be illustrated as,

Input feature → neural network →weight/ threshold adjustment → Error vector → Target feature

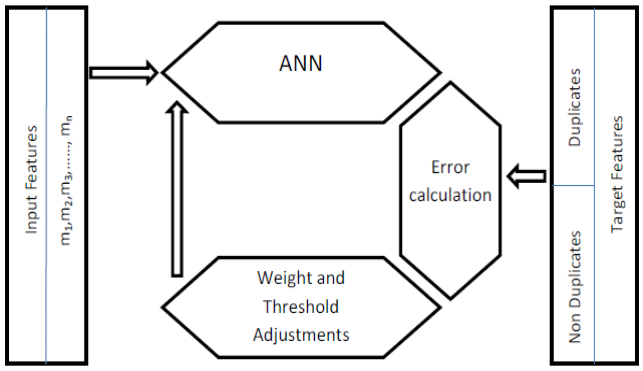


Figure. 4 Training Phase

In the training phase, a supervised learning is implemented i.e.associative learning in which the network is trained by providing it with input and matching output patterns. These input-output pairs can beprovided by an external teacher, or by the system which contains the neural network. In the proposed approach input is the model parameters and the output is the value corresponding to the duplicates and the non-duplicates. So the neural network processes the weight value according to the input features given. This weightage values are then fixed for the neural network designed for the proposed deduplication technique.

### *B . Testing phase*

A sample document is given to the ANN designed for the proposed deduplication technique. The sample document is either duplicate or non-duplicates, but when it is subjected to the trained neural network it will yields a value. If the value is either similar to the values represented for the duplicates and non-duplicates. Thus the neural network produces a more data relevant result for identifying the duplicates from the dataset.

$$K_{test} = \begin{cases} Duplicate, if \ K_{test} \geq K_{Dup} \\ NonDuplicate, if \ K_{test} \geq K_{NonDup} \end{cases}$$

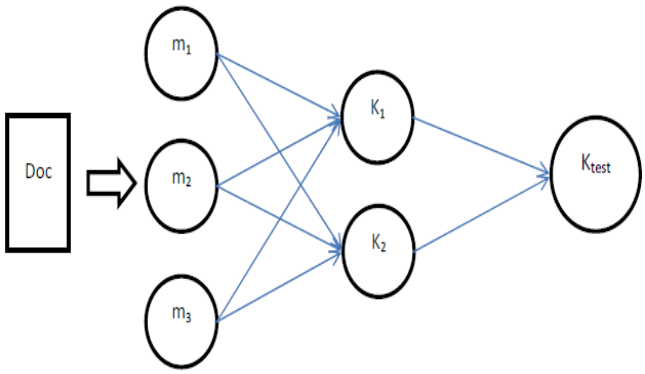


Figure. 5 Testing of a document

## V. RESULT AND ANALYSIS

The performance of the proposed deduplication technique is evaluated in the following section under different evaluation criteria. The algorithms are implemented in MATLAB and executed on a core i5 processor, 2.1MHZ, 4 GB RAM computer.

### *A. Dataset description*

In the experiment we have selected datasets from the Riddle data repository [40] and the datasets used is Restaurant dataset. The datasets, which we are used in our proposed approach, is detailed below.

Dataset1 [Restaurant]: This dataset contains four files of 500 records (400 originals and 100 duplicates), with a maximum of five duplicates based on one original record (using a Poisson distribution of duplicate records), and with a maximum of two modifications in a single attribute and in the full record.

Dataset2 [Cora]: This dataset contains four files of 400 records (300 originals and 100 duplicates), with a maximum of five duplicates based on one original record (using a Poisson distribution of duplicate records), and with a maximum of two modifications in a single attribute and in the full record.

### *B. Evaluation criteria*

In the proposed deduplication technique two criteria are considered for the evaluation purpose, one is accuracy and the other is time for execution. The accuracy defines how precise is the proposed deduplication technique with the above mentioned dataset. Tine for execution is the factor that defines how much time is required for the proposed deduplication technique to record the deduplication.

***1) Accuracy***

The accuracy is the proportion of true results such as true positives and true negatives in the population. It is a parameter of the test. The accuracy value is calculated from the following equation.

$$accuracy = \frac{Number\ of\ true\ positives + Number\ of\ true\ negatives}{number\ of\ true\ positives + number\ of\ false\ negatives + number\ of\ true\ negatives + number\ of\ false\ positives}$$

Here the number of duplicates is considered as the number of true negatives and the numbers of non-duplicates are considered as the true positive. The variance in their value is considered as the accuracy of the proposed deduplication technique.

***2) Time***

Time is the factor that defines the required time for executing the proposed deduplication technique. The time for execution is calculated from the starting of the proposed technique to till the termination of the proposed technique.

*C. Performance Evaluation*

In this section, we plot the performance analysis of the proposed deduplication technique, when the proposed technique is applied to the different datasets namely Restaurant and Cora dataset. The dataset are represented as dataset 1 and dataset 2 for the ease of representation. The evaluation factors used are Time and accuracy. All the experiments are carried out with three threshold values.

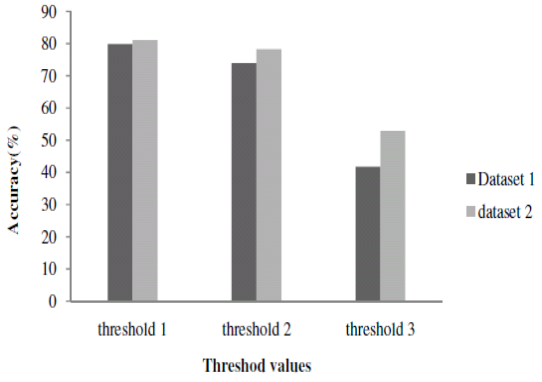


Figure. 6 Analysis based Accuracy

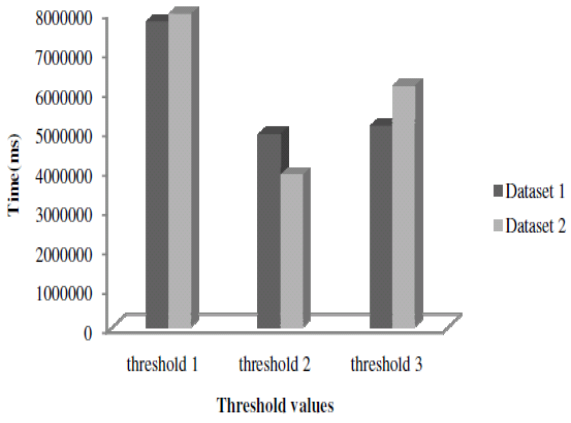


Figure. 7Analysis based on Time

In the above figures 6 and 7 represents the proposed performance evaluation of the proposed deduplication technique. The accuracy is plotted in the figure 6, by analyzing the accuracy values, it can be stated that accuracy is varied in both the dataset according to the different threshold values. The threshold is varied as 1, 1.75 and 2. In the case of time also there is no much different, the time also sensitive according to the threshold value increases. The optimal performance is obtained at threshold value 1.75, i.e. at threshold 2.

*D. Comparative analysis*

The comparative analysis concentrates on the performance analysis of the proposed deduplication with an existing deduplication technique. The existing technique we considered here is an incremental clustering based deduplication method proposed by Gianni Costa *et al.*[1]. In existing method features of the incremental clustering method is used to find the duplicates from a given dataset. The comparison analysis is done by applying the proposed deduplication technique and the existing technique on Restaurant data set on the basis of precision, recall and time, under three threshold values.

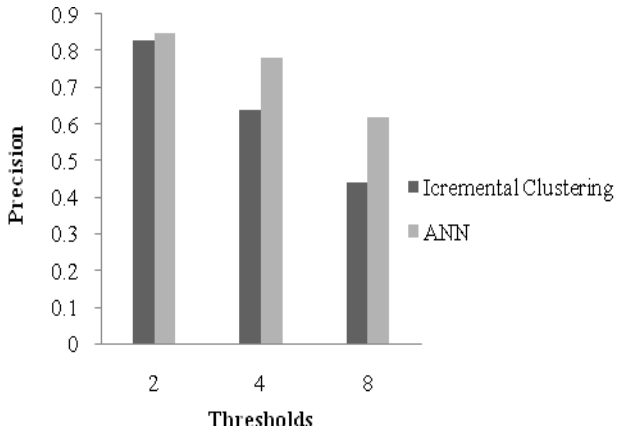


Figure. 8Comparison based on precision

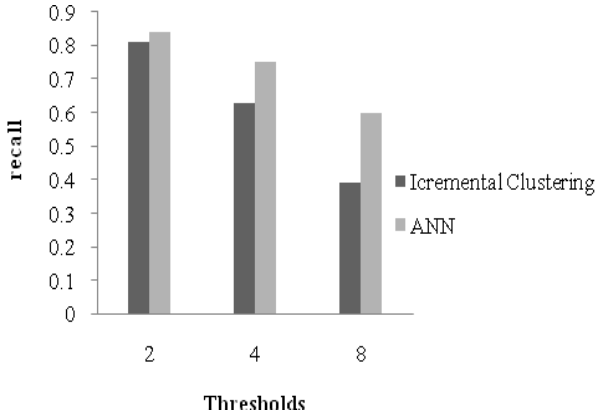


Figure.9.Comparison based recall

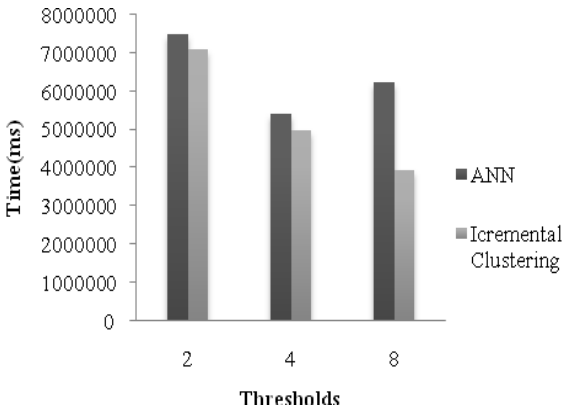


Figure. 10 Comparison based on Time

In the above figures 8 and 9 the comparison analysis of the proposed deduplication technique and the existing incremental clustering based deduplication technique are plotted. In the recall precision based analysis, it can be found that, the proposed deduplication technique achieved considerable increase in accuracy level at different threshold values. The highest value of precision and recall achieved by the proposed technique are 0.85 and 0.84 respectively, while that of the incremental clustering based deduplication technique are 0.83 and 0.81. The figure 10 shows the comparison analysis of the time of execution. In time analysis, the proposed technique utilize more time for the processes of deduplication.

## VI. CONCLUSION

The deduplication has been one of the most emerging techniques for data redundancy and duplication. The duplication creates lots of problems in the information retrieval system. In this paper, a deduplication technique based on artificial neural network is implemented. The technique uses a set of similarity values for calculating the deduplication among the dataset. The proposed deduplication technique uses a training phase to calculate the weight parameter to the neural network and then a testing phase is implemented to find the redundant or duplicate data. The proposed deduplication technique is used with two dataset to evaluate it performance in the deduplication purpose. The performance analysis showed that the proposed technique performswell in finding the deduplication. The performance evaluation is based on two different factors, one is accuracy and other is time, the highest accuracy obtained for the proposed deduplication technique is 79.8%. A comparison study has done to the proposed deduplication technique with the existing genetic program based deduplication technique. The accuracy achieved for proposed deduplication technique is 79.8% while the existing technique achieved only 76.8% a fixed threshold value.Future up-gradation to the proposed deduplication technique can be done with the help of advanced learning algorithms.

## REFERENCE

[1] Gianni Costa, Giuseppe Manco, Riccardo Ortale, " An incremental clustering scheme for data de-duplication", Transactions on Data mining knowledge discover, Vol 20, pp: 152- 187, 2010.

[2] Surajit Chaudhuri, Kris Ganjam, Venkatesh Ganti and Rajeev Motwani, \Robust and Efficient Fuzzy Match for Online Data Cleaning", In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 313 - 324, New York, USA, 2003.

[3] Hong-Jie Dai, Chi-Yang Wu, Richard Tzong-Han Tsai and Wen-Lian Hsu, "From Entity Recognition to Entity Linking: A Survey of Advanced Entity Linking Techniques", The 26th Annual Conference of the Japanese Society for Artificial Intelligence, 2012.

[4] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis and Vassilios S. Verykios, \Duplicate Record Detection: A Survey", IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No. 1, pp. 1 - 16, January 2007.

[5] Surajit Chaudhuri, Anish Das Sarma, Venkatesh Ganti and Raghav Kaushik, \Leveraging Aggregate Constraints for Deduplication", In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 437 - 448, New York, USA, 2007.

[6]J. Jebamalar Tamil selvi and Dr. V. Saravanan, \A Unified Framework and Sequential Data Cleaning Approach for a Data Warehouse", International Journal of Computer Science and Network Security, Vol. 8, No. 5, May 2008.

[7] Lin Chang and Xue Bai," Data Mining: A Clustering Application", In proceedings of PACIS 2010.

[8] S. P. Deshpande and V. M. Thakare, "Data Mining System And Applications: A Review," International Journal of Distributed and Parallel systems, Vol. 1, No. 1, pp. 32-44, 2010.

[9] Aynur Dayanik, Craig G. Nevill-Manning, "Clustering in Relational Biological Data", ICML-2004 Workshop on Statistical Relational Learning and Connections to Other Fields, pp: 42-47, 2004.

[10] Pham, D.T. and Afify, A.A. "Clustering techniques and their applications in engineering", Proceedings- Institution of Mechanical Engineers Part C Journal of Mechanical Engineering Science, Vol: 221; No: 11, pp: 1445-1460, 2007.

[11] Thair Nu Phyu," Survey of Classification Techniques in Data Mining," In Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong Kong , Vol. 1, pp 18-20, 2009.

[12] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand and Dan Steinberg," Top 10 Algorithms in Data Mining," Knowledge and Information Systems, Vol. 14, No. 1,pp. 1-37,2007.

[13] Ganesh Kumar. M and Arun Ram. K," Controlling Free Riders in Peer to Peer Networks by Intelligent Mining", International Journal of Computer and Electrical Engineering, Vol. 1, No. 3, pp. 288-292, 2009.

[14] Chuanyao Yang , Yuqin Li , Chenghong Zhang and Yunfa Hu," A Fast KNN Algorithm Based on Simulated Annealing", In Proceedings of the International Conference on Data Mining, Las Vegas, Nevada, USA, pp . 25-28, 2007.

[15] J. J. HOPFIELD,"Neural networks and physical systems with emergent collective computational abilities", In Proceedings NatL Acad. Sci, USA, Vol. 79, pp. 2554-2558, 1982.

[16] W. Buntine," Learning classification trees", In D. J.Hand, editor, Artificial Intelligence frontiers in statistics, Chapman & Hall, London, pp 182–201,1993.

[17] J. R. Quinlan. C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[18] Neapolitan, R.E., Learning Bayesian Networks, Prentice Hall, Upper Saddle River, NJ, 2004.

[19] C. J. C. Burges," A tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery, Vol. 2, No 2,pp. 121–167, 1998.

[20] Lei Wang, Latifur Khan and Bhavani Thuraisingham," An Effective Evidence Theory based K-nearest Neighbor (KNN) classification," In proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, NSW ,Vol.1, pp . 797-801, 2008.

[21] Qi Yu, Antti Sorjamaa, Yoan Miche, Eric Severin and Amaury Lendasse," Optimal Pruned K-Nearest Neighbors: OP-KNN – Application to Financial Modeling", In proceedings of 8th International Conference on Hybrid Intelligent Systems, Barcelona, Spain, No. 1, 2008.

[22] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition", In International Conference on Computer Vision and Pattern Recognition, New York (NY), USA, 2006.

[23] M. N. Goria, N. N. Leonenko, V. V. Mergel, and P. L. Novi Inverardi, “A new class of random vector entropy estimators and its applications in testing statistical hypotheses”, J. Nonparametric. Stat., vol. 17, pp. 277–297, 2005.
[24] F. Pan, B. Wang, X. Hu, and W. Perrizo, “Comprehensive vertical sample-based knn/lsvm classification for gene expression analysis”, J. Biomed. Inform, vol. 37, pp. 240–248, 2004.
[25] Vincent Garcia, Eric Debreuve, Frank Nielsen, and Michel Barlaud, “k-nearest neighbor search: fast GPU-based implementations and application to high-dimensional feature matching”, In Proceedings of the IEEE International Conference on Image Processing (ICIP), Hong Kong, China, pp. 3757-3760, September 2010.
[26] Mehdi Moradian and Ahmad Baraani, ” KNNBA: K-Nearest-Neighbor-Based-Association Algorithm,” Journal of Theoretical and Applied Information Technology, Vol.6, No. 1, pp. 123-129, 2009.
[27] Erhard Rahm and Hong Hai Do, \Data Cleaning: Problems and Current Approaches", IEEE Data Engineering Bulletin, Vol. 23, No. 4, December 2000.
[28] Arthur D. Chapman, \Principles and Methods of Data Cleaning | Primary Species and Species Occurrence Data", Version 1.0, Report for the Global Biodiversity Information Facility, Copenhagen, 2005.
[29] Mikhail Bilenko and Raymond J. Mooney, \Adaptive duplicate detection using learnable string similarity measures", in proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, 2003.
[30] Uwe Draisbach and Felix Naumann, \A Comparison and Generalization of Blocking and Windowing Algorithms for Duplicate Detection", In Proceedings of the 7th International Workshop on Quality in Databases at VLDB, Lyon, France, 2009.
[31] J. Jebamalar Tamilselvi and V. Saravanan, \Token-Based Method of Blocking Records for Large Data Warehouse", Advances in Information Mining, Vol. 2, No. 2, pp. 5 - 10, 2010.
[32] Steven Euijong Whang, David Menestrina, Georgia Koutrika, Martin Theobald, and Hector Garcia-Molina, \Entity Resolution with Iterative Blocking", In Proceedings of the 35th SIGMOD international conference on Management of data, Providence, pp. 219 - 232, Rhode Island, USA, 2009.
[33] Hui Yang and Jamie Callan, \Near-Duplicate Detection by Instance-Level Constrained Clustering", In Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 421 - 428, Seattle, WA, USA, 2006.
[34] Israr Ahmed, Abdul Aziz, \Dynamic Approach for Data Scrubbing Process", International Journal on Computer Science and Engineering (IJCSE), Vol. 2, No. 2, pp. 416 -423, 2010.
[35] K.Deepa and Dr.R.Rangarajan, "An Approach to Duplicate Record Detection Using Similarity Metrics and Anfis", Journal of Computational Information Systems, vol. 8, no. 6, pp. 2231-2243, 2012.
[36] Hanna Kopcke and Erhard Rahm, "Frameworks for entity matching: A comparison", Data & Knowledge Engineering, vol. 69, no. 2, pp. 197-210, 2010.
[37] P. Christen, "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, no. 99, pp. 1, 2011.
[38] T.A. Faruquie, K.H. Prasad, L.V. Subramaniam, M. Mohania, G. Venkatachaliah, S. Kulkarni, and P. Basu, "Data cleansing as a transient service", IEEE 26th International Conference on Data Engineering, pp. 1025-1036, 2010.
[39] Gianni Costa, Giuseppe Manco and Riccardo Ortale, "An incremental clustering scheme for data de-duplication", DATA MINING AND KNOWLEDGE DISCOVERY, vol. 20, no. 1, pp. 152-187, 2010.
[40]Site: Riddle dataset, http://www.cs.utexas.edu/users/ml/riddle/data.html

**M.Padmanaban** is a Research Scholar in Bharathiar University, Coimbatore, and working as an Assistant Professor in Computer Science Department in D.R.B.C.C. Hindu College Dharmamurthy Nagar, Pattabiram, Chennai-6000072.

**T.Bhuvaneswari’s** area of research is Data Mining. She has completed PhD in June 2007 from SCSVMV University, Enathur. Guiding Scholars in Bharathiar University, MS University, MGR University, SCSVMV University. She has published papers in national and international Journals, National and international conferences. Having 12yrs of Teaching experience in Deemed University and currently working as Assistant Professor in Government Arts and Science College in Department of Computer Science.