A Technique for Data Deduplication using Q-Gram Concept with Support Vector Machine

M.Padmanaban
Assistant Professor, Department of CSE
D.R.B.C.C. Hindu College
Dharmamurthy nagar, Pattabiram, Chennai

T.Bhuvaneswari
Assistant.Professor
Department of Computer Science
Government Arts and Science College

ABSTRACT
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, quasi-replicas, 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.In this paper, we have proposed accordingly. In the previous work, duplicate record detection was done using three different similarity measures and neural network. In the previous work, we have generated feature vector based on similarity measures and then, neural network was used to find the duplicate records. In this paper, we have developed Q-gram concept with support vector machine for deduplication process. The similarity function, which we are used Dice coefficient,Damerau–Levenshtein distance,Tversky index for similarity measurement. Finally, support vector machine is used for testing whether data record is duplicate or not. 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 experimental results showed that the proposed deduplication technique has higher accuracy than the existing method. The accuracy obtained for the proposed deduplication 88%.

Keywords
deduplication, support vector machine, training, testing.

1. INTRODUCTION
In digital media the mounting volume of information available has developed into a demanding problem for data administrators. Data repositories such as those used by digital libraries and e-commerce brokers may present records with different structure [1] and digital media is built on datas gathered from these kinds of sources. The competence of an organization to provide constructive services to its users is proportional to the superiority of the data available in its systems. In this situation today, the decision of keeping repositories with "dirty" data goes far beyond technical questions, such as the overall speed or performance of data management systems. The solutions on hand for addressing this problem necessitates more than technical efforts since management and cultural changes are needed [1, 8]. Unneeded copies of information may often contained in the file systems, which may be identical files or sub-file regions, possibly stored on a single host, on a shared storage cluster, or backed-up to secondary storage. Taking advantage of this redundancy, deduplicating storage systems reduce the underlying space needed to contain the file systems (or backup images thereof). Deduplication can work at either the sub-file [5, 9, and 10] or whole-file [11] level. Latest informations reveal that deduplication is considered to be the most-impactful storage technology and it is estimated to be applied to 75% of all backups in the next few years [12].

Categorization of data deduplication strategies can be done according to the basic data units they handle. There are two main data deduplication strategies: 1) File-level deduplication, in which only a single copy of each file is stored. Two or more files are stored as identical if they have the same hash value. A very popular type of service offered in multiple products [6, 13, and 14]; Block-level deduplication, which segments files into blocks and stores only a single copy of each block. The system could either use fixed-sized blocks [15] or variable-sized chunks [16, 17]. There are two basic approaches in terms of the architecture of the deduplication solution. The client is unaware of any deduplication that might occur in the target-based approach deduplication which is handled by the target data-storage device or service. Source based deduplication acts on the data at the client before it is transferred. The client software communicates with the backup server to check for the existence of files or blocks [6] very particularly. To address the above challenges by removing the redundant data chunks before sending them to the remote backup destination, two well-known source de-duplication methods, source local chunk-level de-duplication [7,18,19] and source global chunk-level de-duplication [19,20,21] have been planned in the past.

Due to the out-of-memory fingerprint accesses to massive backed-up data, chunk-level de-duplication has an inherent latency and throughput problem that significantly affects the backup performance is revealed in the latest studies [22-25]. Within the source global chunk-level de-duplication, this overhead of massive disk accesses will strangle the deduplication process and thus increase the backup window. While in source local chunk-level de-duplication, the overhead is alleviated since searching the duplicate chunks is restricted to the same client. This reduced overhead, which increases the backup window due to the increased data transmission cost however, comes at the cost of severely limited compression ratio. As a result it is enviable to attain an optimal tradeoff between de-duplication efficiency and deduplication overhead to continue a shorter backup window than existing solutions. For the purpose of deduplication with effectiveness and accuracy, several other methods are used. The methods are deduplication using genetic algorithm, semantic methods, cloud services etc. and the methods that uses GA is overcome by some difficulties plotted above.

In this paper, a technique for deduplication is plotted based on the support vector machine (SVM). 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 parameter for the documents that are subjected for testing

deduplication. The model parameters calculated are used for the processing with the SVM. The SVM has two phases, one is the training phase and other is the testing phase. In the training phasethe SVM 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,

- The main objective of our paper is Q-gram concept for improving the duplication problem
- A support vector machine 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.
- A set of model parameters are selected from three different similarity measures

The rest of the paper is organized as; the section 2 gives a review of some related works regarding deduplication.Section 3 gives details of the proposed approach with mathematical models. 4th section gives the results and discussion about the proposed approach and with the 5th section we conclude our research work.

2. 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.

Mois´es G. de Carvalho*et al* [1] have planned a genetic programming approach to trace deduplication that combines a number of different pieces of facts extracted from the data content to discover a deduplication function that is able to recognize whether two entries in a repository are replicas or not. Due to the information, that clean and replica-free repositories not only allow the retrieval of higher-quality information but also lead to more concise data and to potential savings in computational time and resources to process this data. Our approach outperforms an existing state of-the-art method found in the literature which was shown by our experiments. Besides, the recommended functions are computationally less challenging since they use fewer facts. Also, our genetic programming approach is capable of freeing the user from the burden of having to choose and tune this parameter and automatically adapting these functions to a given fixed replica identification boundary.

The structure can indeed have a significant impact on the process of duplicate detection is an argument, proposed by LuisLeitao and PávelCalado [2]. Automatically restructures database objects in order to take full advantage of the relations between its attributes is a method proposed by them. The relative importance of the attributes in the database is reflected by the new structure and the new structure also avoids the need to perform a manual selection. In order to analysis their approach they applied it to an existing duplicate detection system. Using the new learned structure, experiments performed on several datasets show that, they consistently surpass both the results obtained with the original database structure and those obtained by letting a knowledgeable user physically choose the attributes to compare.

Ektefa M *et al*[3] have proposed a threshold-based method which takes into account both string and semantic similarity measures for comparing record pairs. The threshold-based method is experimented on a real world dataset, namely Restaurant and its effectiveness is measured based on several standard evaluation metrics. The proposed similarity method which is based on the combination of string and semantic similarity measures outperforms the individual similarity measures with the F-measure of 99.1% in Restaurant dataset is indicated by the experimental results. In order to detect duplicate records more effectively, semantic similarity should be considered other than string similarity based on experimental results.Elhadi M *et al* [4] have planned method that bring information on experiments performed to investigate the use of a combined part of speech (POS) and an improved longest common subsequence (LCS) in the analysis and calculation of similarity between texts. For the representation of documents, the text's syntactical structures were used. To compare and rank the documents according to the similarity of their representative string, an improved LCS algorithm was applied to such a representation. In detecting duplicate documents within a corpus, and in the filtering of search engine results, the approach was applied and the results obtained were hopeful.

By analyzing the results, it can be seen that, in [1] a genetic programming based deduplication technique is used and it is new technique for the process of deduplication. Sooner than a threshold based method is implemented in [3], a characteristic based technique is described in the [2] for executing the deduplication in databases. Unlike from the other approaches, Elhadi M *et al*[4]implemented a process based on combined part of speech and improved longest common subsequence. With reference to the above researches, in this paper an artificial neural network based deduplication technique is described.

3.NEW TECHNIQUE OF DEDUPLICATION

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, quasi-replicas, 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. Accordingly, in the previous work, duplicate record detection was done using three different similarity measures and neural network. In the previous work, we have generated feature vector based on similarity measures and then, neural network was used to find the duplicate records. In this work, we have developed to improve the existing work by adding the Q-gram concept and the SVM classifier. The overall block diagram is shown in figure 1.

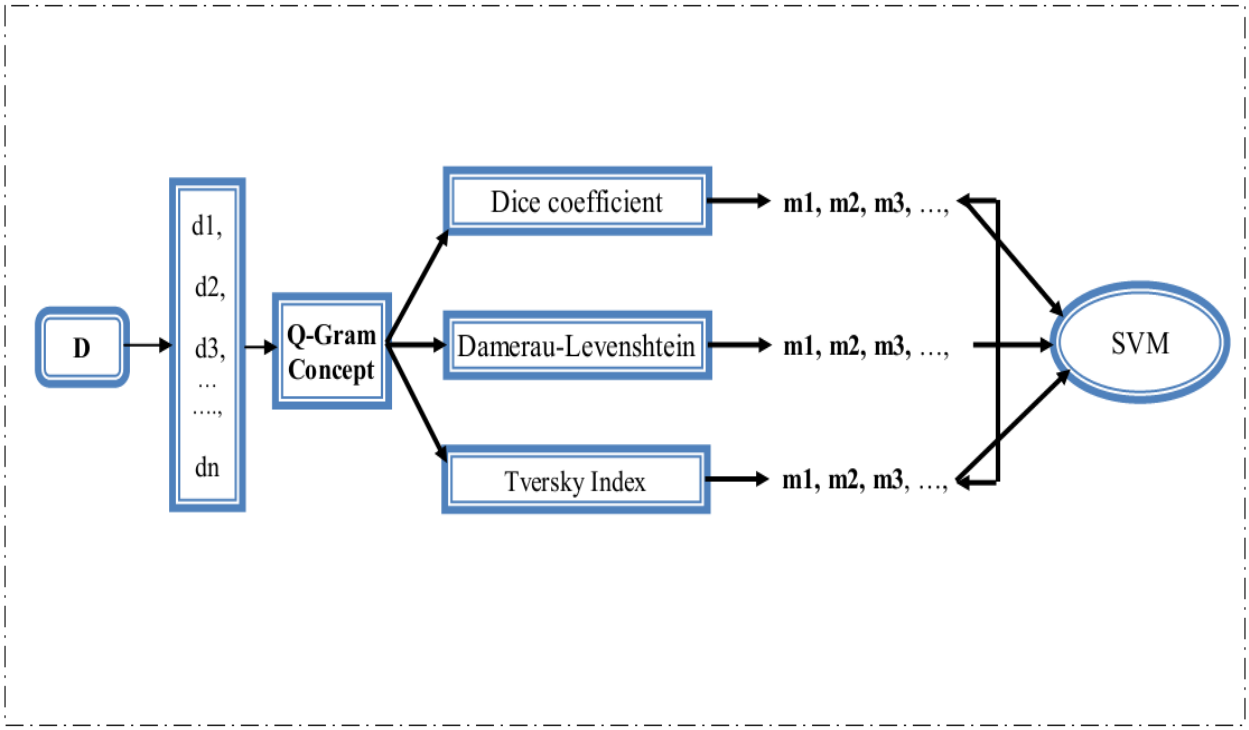

Figure 1. Overall block diagram of our proposed approach

The initial step regarding the deduplication based on the support vector machine is to find the model parameters generated from the similarity functions. The similarity function, which we used are

- Dice coefficient
- Damerau–Levenshtein distance
- Tversky index

The input which is given to the SVM are the value generated from the above plotted similarity distance measures The documents, are processed with similarity measure and each of the measure will produce model parameters which are to be tested for the data redundancy. 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. When compared to the Jaccard index, Dice coefficient is not very different but it has some different properties jaccard index.Like Jaccard the function ranges between zero and one. Unlike Jaccard, the corresponding difference function $d = 1-(2|A \cap B|)/|A|+|B|$ 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,

$$S = \frac{2|A \cap B|}{|A|+|B|}$$

Where,

S- Represents the similarity measure

X and Y - 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, To transform one string into the other it gives the counting needed for the minimum number of operations needed, 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.A set of the model parameters for the processing of neural network are provided by the similarity algorithm of the Damerau-Levenshtein.

3. Tversky Index

The main operation of The Tversky index is to compare a variant to a prototype.As a generalization of Dice's coefficient and Tanimoto coefficient the Tversky index can be seen.For sets A and B of keywords used in information retrieval, the Tversky index is a number between 0 and 1 given by

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

Where, α and β are the parameters of the Tversky index. The similarity measure also provides a set of model parameters.

3.1 Q- Gram Concept

Accordingly, in the previous work, duplicate record detection process was done using three different similarity measures with neural network. In the previous work, we have generated feature vector based on similarity measures and then, neural network was used to find the duplicate records. In this work, we have proposed to improve the existing work by adding the Q-gram concept. In Q gram concept, we are separating the data records into four blocks here we are representing the blocks as set of data's. For example considering a person's address, we are segmenting the data set as person's name, house number; phone number and area name.we are allocating for each block a separate work space. In our Q gram concept, we have implemented the following four ways, which are stated below,

A. 1 gram concept
B. 2 gram concept
C. 3 gram concept
D. 4 gram concept

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

$$F = [f_1, f_2, ..., f_n], \quad f \in F \text{ and } n=1, 2, 3, ...$$

Now the set of documenst are subjected for the processing with the similarity measures

A. 1 gram concept

Consider the dataset contains four data which are name person's name (f1), house number (f2), phone number (f3) and area name (f4)

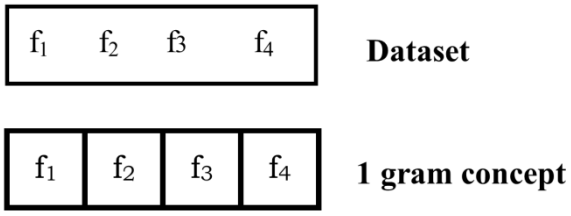

Figure 2: sample diagram of 1 gram concept

From figure 2, the dataset is separated as four blocks and these blocks are taken by individually for deduplication process.1 gram concept sample diagram is shown in figure 2. The similarity measures used in the proposed approach are Dice coefficient (DC), Damerau-Levenshtein (DL) and Tversky Index (TI).This similarity measures are used individually for four separated data record block. Each of the similarity measures produces model parameters individually for the dataset records set F.Three similarity measures is used for the computation of the model parameters .Thus the model parameter calculation should be precise and accurate.

Using the similarity functions we have modulated the parameters as,

$$M_{DC} = [p_1, p_2, ..., p_n]$$
$$M_{DL} = [p_1, p_2, ..., p_n]$$
$$M_{TI} = [p_1, p_2, ..., p_n]$$

Sorting and combining the three set of model parameters for the processing of the deduplication with SVM is the the next phase of the proposed approach.the weight of these parameters are found by

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

Where,
M_{sort} - the sorted model parameters values and
The set W represents the set with weightage parameters of the neural network.

B. 2 gram concept

Consider the dataset contains four data which are name person's name (f1), house number (f2), phone number (f3) and area name (f4)

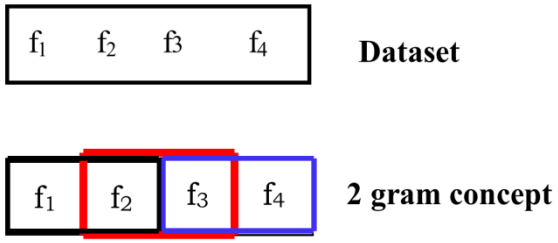

Figure 3: sample diagram of 2 gram concept

The dataset is separated as four blocks and these blocks are taken two by two (shown in figure 3) for deduplication process. Separating the blocks by analyzing the first two blocks and then taking the second and third blocks and then taking the third and fourth blocks is known as 2 gram concept.Here also we have used the similarity function for every two blocks. Three similarity measures are used for the computation of the model parameters and also the model parameter calculation should be precise and accurate.

In this method the parameters are formulated from the similarity functions as,

$$M_{DC} = [q_1, q_2, ..., q_n]$$
$$M_{DL} = [q_1, q_2, ..., q_n]$$
$$M_{TI} = [q_1, q_2, ..., q_n]$$

The weight of these parameters are found by

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

C. 3 gram concept

Consider the dataset contains four data which are name person's name (f1), house number (f2), phone number (f3) and area name (f4)

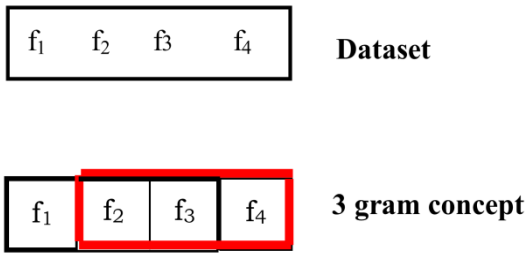

Figure 4: sample diagram of 3 gram concept

The dataset is separated as four blocks and these blocks are taken three by three (shown in figure 4) for deduplication process. In 3 gram concept, we are separating the blocks by analyzing the first three blocks and then taking the second, third and fourth blocks.The similarity measures used in the proposed approach are Dice coefficient (DC), Damerau-Levenshtein (DL) and Tversky Index (TI). Here also we have used the similarity function for every two blocks. The parameters are formulated from the similarity functions as,

$$M_{DC} = [r_1, r_2, ..., r_n]$$
$$M_{DL} = [r_1, r_2, ..., r_n]$$
$$M_{TI} = [r_1, r_2, ..., r_n]$$

The weight of these parameters are found by

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

D. 4 gram concept

Consider the dataset contains four data which are name person's name (f1), house number (f2), phone number (f3) and area name (f4)

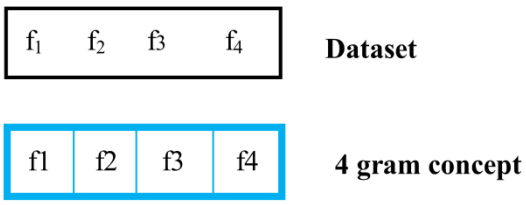

Figure 5: sample diagram of 4 gram concept

The dataset is separated as four blocks and these blocks are taken (shown in figure 5) for deduplication process. In 4 gram concept, we are separating the blocks by analyzing the four blocks. Here also we are using the three similarity measures for analyzing the documents for separating the blocks. Each of the similarity measures produces model parameters individually for the dataset F. Three similarity measures is used for the computation of the model parameters

In this method the parameters are formulated from the similarity functions as,

$$M_{DC} = [s_1, s_2, ..., s_n]$$
$$M_{DL} = [s_1, s_2, ..., s_n]$$
$$M_{TI} = [s_1, s_2, ..., s_n]$$

The weight of these parameters are found by

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

3.2 Classification Using Support Vector Machine

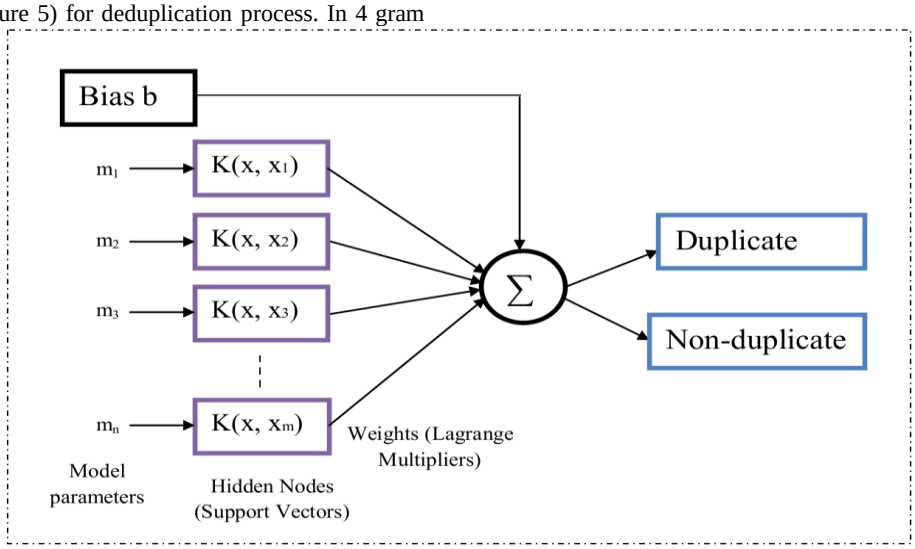

Figure 6. SVM for Deduplication

The above is the design of the support vector machine for the deduplication purpose. The SVM 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 support vector machine designed for the proposed deduplication process.

To train the SVM classifier, we need some data features to identify the duplication and deduplication records in datasets. The data features will then train the classifier and the classifier will find whether the given records are duplication or not. The data features which we have chosen for training the SVM classifier are three similarity measures such as Dice coefficient,Damerau-Levenshtein distance,Tversky index. After computing all the data features, we have to give the values to the classifier. For instance, if we are choosing five duplicate records and five deduplicate records, we need to calculate all the three data features separately for all the duplicate and deduplicate records we had chosen. After calculating all the three data features for every chosen duplicate record and five deduplicate records, we have to give the result to the SVM classifier. Using those results we can train the classifier to identify the duplicate record and non-duplicate record from the given dataset. After the SVM classifier is trained, we can give a new record to find whether it has duplicate or non-duplicate record. Thereafter, the three data features such as Dice coefficient, Damerau–Levenshtein distance, and Tversky index are computed for the new record. The computed values of all the three data features are then give to the SVM classifier.

The SVM classifier is then compare the values of all the three data features with the stored values of duplicate or non-duplicate data. Because during training we have stored all the three data features of the five duplicate records and five non-duplicate records. After comparison, the SVM classifier will identify whether the given MRI image comes under duplicate category or non-duplicate and give the result to us.

Support Vector Machine (SVM)

In most cases, we want to assign an object to one of several categories based on some of its characteristics in our real life situation. For instance, based on the outcome of several data duplication process we want to say whether the record has a duplicate or not. In computer science such situations are explained as classification issue.

The support vector machine (SVM) which was derived from the statistical theory is a powerful supervised classifier and is an accurate learning technique. The SVM was introduced in 1995. It gives successful classification outcomes in different application domains such as medical diagnosis [26, 27]. SVM works under the principle of structural risk reduction from the statistical learning theory. To maximize the margin between the classes and to minimize the true cost [28], its kernel is used to control the empirical risk and categorization capacity. A support vector machine can search an optimal separating hyper plane amid the members and non-members of a given class in a high dimension feature space [29]. There are many general kernel functions such as linear, polynomial of degree and Radial basis function (RBF). Among these kernel functions, a radial basis function proves to be useful because of the fact the vectors are mapped nonlinearly to a very high dimension feature space.

4.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.

4.1 Dataset Description

Datasets from the Riddle data repository was chosen for the experiment [30] and the datasets used is Restaurant dataset. The datasets, which are used in our proposed approach, is detailed below.

Dataset1 [Restaurant]: This dataset consists of 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 limit of two changes in a single attribute in the full record.

Dataset2 [Cora]: This dataset consists of 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 limit of two changes in a single attribute in the full record.

4.2 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.

4.2.1 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.

$$accuracy = \frac{Numberof\ truepositives + Numberof\ truenegatives}{numberof\ truepositives + numberof\ falsenegatives + numberof\ truenegatives + numberof\ falsepositives}$$

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.

4.2.2 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.

4.3 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 evaluation factors used are Time and accuracy. In our approach, we are taken four types of resulting values such as 1 gram concept, 2 gram concept, 3 gram concept and 4 gram concept for two datasets. Figure 7 shows the accuracy graph of restaurant and cora dataset. Figure 8 shows the time graph of restaurant and cora dataset.

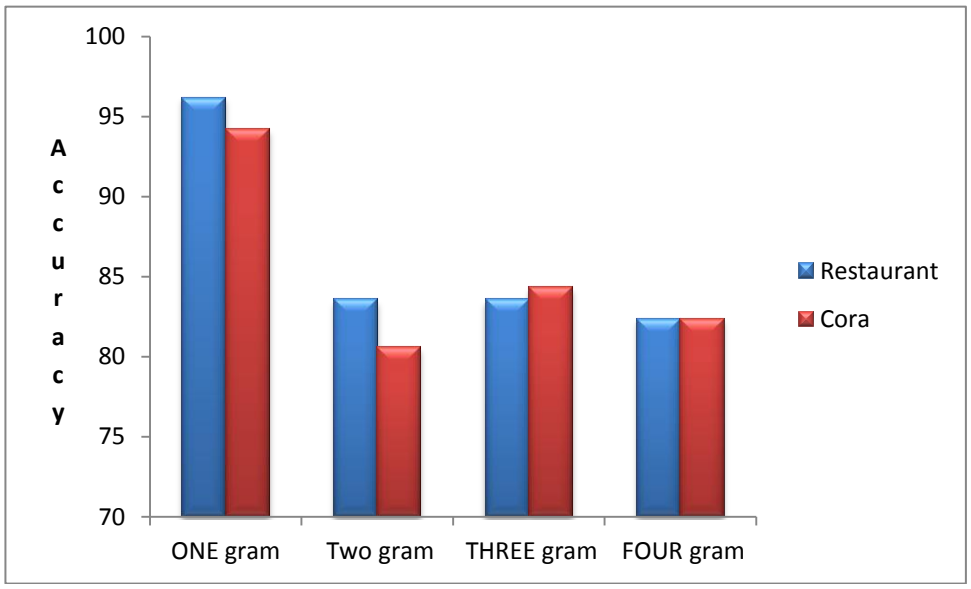

Figure 7:Accuracy graph of Restaurant and cora datasets

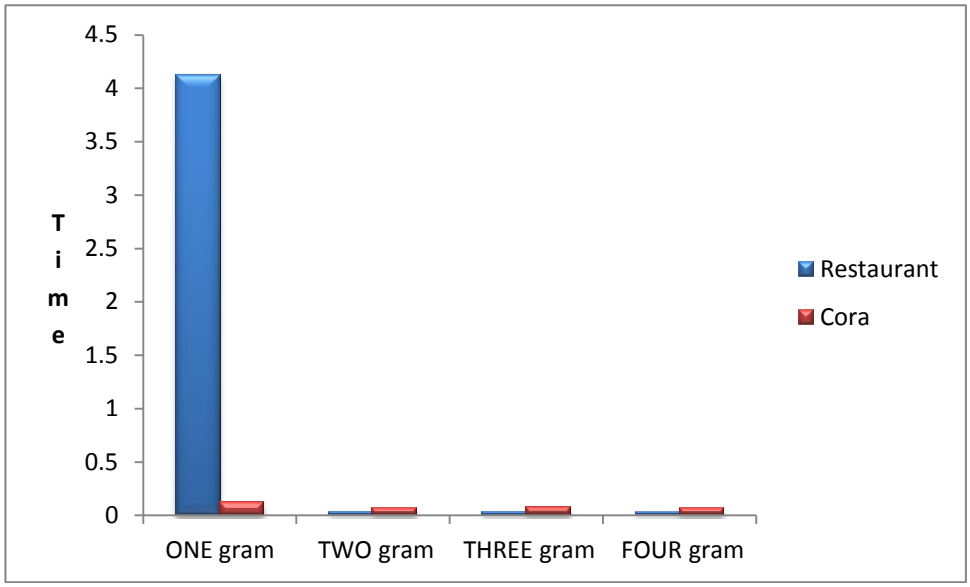

Figure 8:Time graph of Restaurant and cora datasets

4.4 Comparative analysis

The comparative analysis concentrates on the performance analysis of the proposed deduplication with a neural network snd fuzzy technique. The performance analysis has been made by plotting the graphs of evaluation metrics such as accuracy and time. The comparison analysis is done by applying the proposed deduplication technique and the existing technique on cora data set on the basis of accuracy and time. By analyzing the figure 9, our proposed approach is better accuracy performance (88% for ONE gram concept and TWO gram concept) compared to the neural network and fuzzy techniques. By analyzing the figure 10,our proposed approach is better time performance compared to the neural network and fuzzy techniques. Our proposed approach is having good performance compared to other techniques.

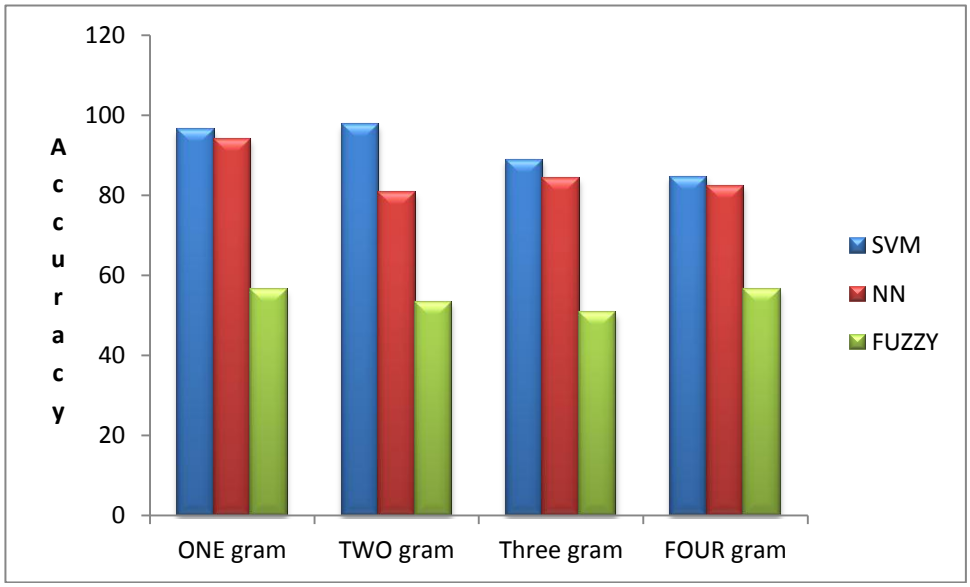

Figure 9:Accuracy graph of comparatrive analysis

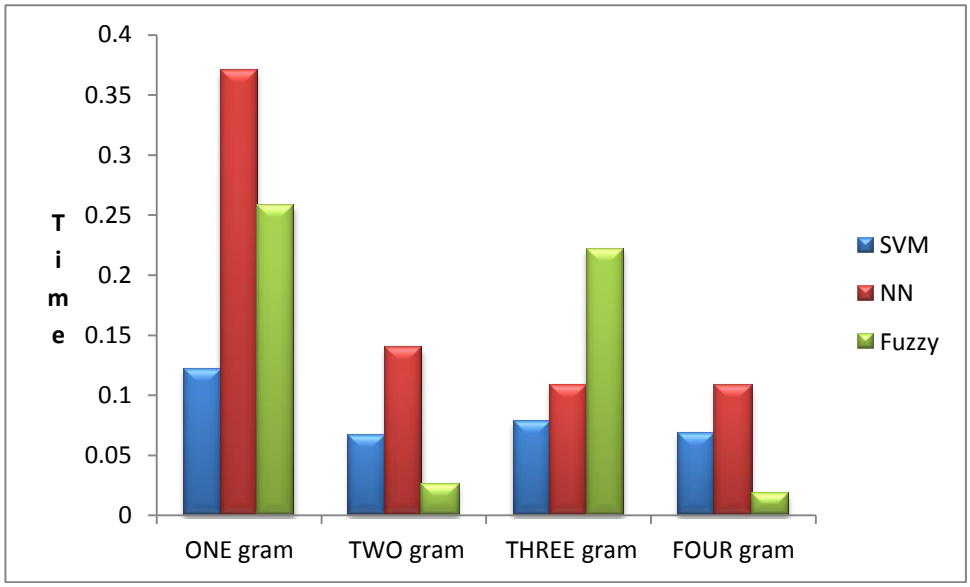

Figure 10:Time graph of comparatrive analysis

5. CONCLUSION

In this paper, we have developed Q-gram concept with support vector machine for deduplication process. The similarity function, which we are used Dice coefficient,Damerau–Levenshtein distance,Tversky index for similarity measurement. Finally, support vector machine is used for whether data record is duplicate or not. 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 experimental results showed that the proposed deduplication technique has higher accuracy than the existing method. The accuracy obtained for the proposed deduplication 88%.

6. REFERENCE

[1] Moises G. de Carvalho, Alberto H. F. Laender, Marcos Andre Goncalves, Altigran S. da Silva, “A Genetic Programming Approachto Record Deduplication”, IEEE Transaction on Knowledge and Data Engineering, 2011.

[2] LuísLeitão and PávelCalado, "Duplicate detection through structure optimization", ACM international conference on Information and knowledge management, pp: 443-452, 2011.

[3] Ektefa, M, Sidi. F,Ibrahim. H, Jabar. M.A., Memar. S, Ramli. A, "A threshold-based similarity measure for duplicatedetection ", Ieee conference on Open systems, pp: 37-41, 2011.

[4] Elhadi. M, Al-Tobi. A, " Duplicate Detection in Documents and WebPages Using Improved Longest Common Subsequence andDocuments Syntactical Structures",International Conference on Computer Sciences and Convergence Information Technology,pp: 679-684, 2009.

[5] Dutch T. Meyer and William J. Bolosky, " A Study of Practical Deduplication", Computer and Information Science,pp: 1-13,2011.

[6] Danny Harnik, Benny Pinkas, Alexandra Shulman-Peleg, " Side channels in cloud services, the case of deduplication in cloud storage",vol. 8, no. 6, pp: 40-47, 2010.

[7] Yujuan Tan, Hong Jiang, Dan Feng, Lei Tian, Zhichao Yan, Guohui Zhou, " SAM: A Semantic-AwareMulti-Tiered Source De-duplication Framework for Cloud Backup",International Conference on Parallel Processing (ICPP), pp: 614-623, 2010.

[8] N. Koudas, S. Sarawagi, and D. Srivastava, “Record linkage: similarity measures and algorithms,” in Proceedings of the2006 ACM SIGMOD International Conference on Management of Data, pp. 802–803, 2006.

[9] C. Dubnicki, L. Gryz, L. Heldt, M. Kaczmarczyk, W. Kilian, P. Strzelczak, J. Szczepkowski, C. Ungureanu, and M. Welnicki.Hydrastor: a scalable secondary storage. In Proc. 7th USENIX Conference on File and Storage Technologies, 2009.

[10] C. Ungureanu, B. Atkin, A. Aranya, S. Gokhale, S. Rago, G. Cakowski, C. Dubnicki, and A. Bohra. Hydrafs: A high-throughputfile system for the Hydrastor content-addressable storage system. In Proc. 8th USENIX Conference on File and StorageTechnologies, 2010.

[11] W. Bolosky, S. Corbin, D. Goebel and J. Douceur.Single instance storage in Windows 2000.In Proc. 4th USENIX WindowsSystems Symposium, 2000.

[12] S. Dorward and S. Quinlan.Venti: A new approach to archival data storage. In Proc. 1st USENIX Conference on File andStorage Technologies, 2002.

[13] H. S. Gunawi, N. Agrawal, A. C. Arpaci-Dusseau, R. H. Arpaci- Dusseau, and J. Schindler, “Deconstructing commodity storageclusters,” in ISCA ’05: Proceedings of the 32nd annual international symposium on Computer Architecture. Washington, DC, USA:IEEE Computer Society, pp. 60–71, 2005.

[14] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer, “Reclaiming space from duplicate files in a serverlessdistributed file system,”International Conference on Distributed Computing Systems, vol. 0, p. 617, 2002.

[15] S. Quinlan and S. Dorward, "Venti: a new approach to archival storage," in First USENIX conference on File and StorageTechnologies, Monterey,CA, 2002.

[16] A. Muthitacharoen, B. Chen, and D. Mazieres, "A low-bandwidth network file system," in Symposium on Operating SystemsPrinciples, 2001, pp. 174–187.

[17] M. Vrable, S. Savage, and G. M. Voelker, " Cumulus: Filesystem Backup to the Cloud," in FAST'09, Feb. 2009.

[18] Syncsort Backup Express and NetApp, "http://www.syncsort. com."

[19] EMC Avamar, "http://www.emc.com."

[20] NetBackupPureDisk, "http://www.symantec.com."

[21] CommvaultSimpana, "http://www.commvault.com."

[22] B. Zhu, K. Li, and H. Patterson, "Avoiding the disk bottleneck in the Data Domain deduplication file system," in FAST'08, Feb. 2008.

[23] S. Rhea, R. Cox, and A. Pesterev, "Fast, inexpensive contentaddressed storage in Foundation," in USENIX'08, Jun. 2008.

[24] M. Lillibridge, K. Eshghi, D. Bhagwat, V. Deolalikar, G. Trezise, and P. Campbell, "Sparse Indexing: Large scale, inlinededuplication using sampling and locality," in FAST'09, Feb. 2009.

[25] D. Bhagwat, K. Eshghi, D. D. Long, and M. Lillibridge, "Extreme Binning: Scalable, Parallel Deduplication for ChunkbasedFile Backup," HP Laboratories, Tech. Rep. HPL-2009- 10R2, Sep. 2009.

[26] Guyon I., Weston J., Barnhill S., Vapnik V.," Gene Selection for Cancer Classification using Support Vector Machines," Machine Learning, vol.46, no.1-3, pp. 389-422, 2002.

[27] Zhang J., Liu Y., "Cervical Cancer Detection Using SVM Based Feature Screening," Proc of the 7th Medical Image Computing and Computer-Assisted Intervention,vol.2, pp.873-880,2004.

[28] Zhang K., CAO H.X., Yan H., "Application of support vector machines on network abnormal intrusion detection," Application Research of Computers, vol.5, pp.98-100, 2006.

[29] Kim D., Park J., "Network-based intrusion detection with support vector machines," Lecture Notes in Computer Science, vol. 2662, p. 747-756, 2003.

[30] http://www.cs.utexas.edu/users/ml/riddle/data.html