An Approach to Reduce the Data Duplication using Simple Map Reduce Algorithm

Anbarasi. M
Department Of Computer Science And Engineering
M.A.M College Of Engineering
Siruganur, Trichy, India.

Karthika. S K
Department Of Computer Science And Engineering
M.A.M College Of Engineering
Siruganar, Trichy, India.

***Abstract*—Data de duplication is used for reducing the storage space in an organization. In many organizations storage system contains many duplicate copies of data. Information gathered from the user is taken for finding out the duplicate data. Labeled set is one of the critical tasks, which is used to reduce the duplicate data by comparing two sets of data. Two stage sampling selection strategy (T3S) is used for identifying the duplicate data from the large dataset. It consists of three steps. In first step, collect all the candidate pairs for labeling. Block the data according to the alphabetical order. Second step consists of two stages. In its the first step is sample selection strategy. Compare the data of the candidate pairs and produce the small balanced data. The second step consists of fuzzy region identification and classification. Fuzzy region is used to identify the exact match of the candidate data. Classification step classifies the duplicate data from the original data. SIM_MR algorithm is proposed in the paper to improve the accuracy. Checking and conforming the matching pairs of the de duplication in large datasets.**

Keywords— Data de duplication; Two stage sampling selection strategy; Simple Map Reduce.

I.INTRODUCTION

Data mining is examined a dramatic growth in extracting the useful information from large amount of dataset. Data mining is also known as data dredging. The combination data, information and knowledge is known as data mining. Data mining is divided into two types such as text mining and web mining. Web mining is further categorized into three types such as web structure mining, web content mining and web usage mining. Data mining is used in many fields such as marketing, banking and government sectors. Major issues in data mining are noise handling, incomplete data and knowledge fusion. The existing work uses many algorithms such as T3S and fuzzy region identification and classification. Data de duplication is classified into three stages. Blocking stage is used to separate the attributes such as name, sex and age. It is used to match the candidate data and identify the similar data. It is used to reduce the similar data from huge amount of data. The second stage is comparison stage. The reduced data is compared and verified using this stage. The third stage is classification stage. It is used to categorize the duplicate data and original data.The proposed Simple Map reduce algorithm is used to improve the performance by checking the redundant data. The simple map reduce algorithm is based on the edit distance based. The edit distance is used to match the strings of the candidate pairs. This paper is organized as follows: Section 2 presents the literature survey on two stage sampling selection strategy. Section 3 demonstrates our proposed work. The performance evaluation measures are derived in section 4. The experiments carried out and its results are presented in section 5. The observations of the proposed work are concluded in section 6.

II. RELATED WORK

Numerous amount of research works have been carried out in Two Stage Sampling Selection Strategy introducing techniques to remove the duplicate data.A. Arasu et.al.,[1] guarantees the quality of the result than passive learning algorithms.It is capable of producing results over 100 dataset.In [2] A. Arasu et.al., provided a constraints on data mining in order to cleaning domains and improve the quality of the duplicate data. It also provides effective algorithms. Data de duplication is implemented in large scale.R. J. Bayardo et.al.,[3] suggested a problem of finding the vector pairs and the similarity data is above the threshold value. Problems are solved by tuning the parameter. It is applied in many fields such as in query and in filtering process. K. Bellare et.al.,[4] suggests unsuitable metric to match the imbalance. The precision should be larger than the particular threshold. It also maximizes the classifier recall. By using this technique we can minimize the 0-1 loss. A. Beygelzimer et.al.,[5] introduces an technique called testing the weight of the bias and detecting the variance. This techniques provides guarantee compared to other learning algorithms. IWAL reduces complexity. M. Bilenko et.al.,[6] proposed a static active learning algorithm and weakly non duplicate data and improve their efficiency. Research in this paper introduces uniform methodology. It also provides accuracy between the sample and testing data.

In [7] , a technique called SS join operator is verified by the sample similarity and the data which we choose for matching. The similar joins depends on the alphabetical order or it is not based on the alphabetical order. Verification is case sensitive and it checks whether the specified data is in the first, middle or it may be at the last. P. Christen[8] introduces automatic data pair classification and it is previously discussed in the above two papers. The first process is automatically selects the quality training data with the other matched data pairs. In second process produces effective classification outputs compared to other algorithms. The proposed can be done without the training data and it can

perform with the real time values. P. Christen [9] reports a technique called indexing. Linking the data from the multiple databases provides the same data. By performing the matching operation it eliminates the duplicate data from the original data. This technique separate the dataset into blocks. Matching data are aligned in the same category. Other data are aligned in the different category. D. Cohn et.al.,[11] introduces a technique called learning algorithm which takes the smallest amount compared to what it receives as an input data. It is considered as the better algorithm compared to other sample data. This technique receives the information from the dataset and arises the query, it produces the efficient results. In [12] G. Dal Bianco et.al., suggested a FS-dedup is used for identifying which objects are same in the dataset. This technique is used to reduce the de deuplication process. It is used to identify the similar data using the signature de duplication with less effort. It is the most popular and best technique for identifying the duplicate data with the small dataset.

In [13] a family of genetic programming approach is used to find out the duplicate data. The design pattern which designed for government institution is not suitable for private organizations. Removing the duplicate data needs more effort from large database. The length of the rows and columns is not important for identifying the duplicate data, only the space is reduced. A. Elmagarmid et.al., [14] proposes a technique that data do not have the same key for multiple task. Errors are rectified by many techniques in order to improve the effectiveness and scalability.C. Gokhale et.al., [18] introduces a technique called hands-off-crowd sourcing which is used to limit the workflow of the data. Data matching needs an opening crowd soucing for the masses. Compared to other solutions it produces better results in data matching. This technique involves the complex process in the dataset. In [20] S. Sarawagi and A. Bhamidipaty had developed a technique called a interacting de deuplication data from various datasets. The user presents a coding to find out the classify the matching and non matching pairs. Reteriving the sample number of data is important and produces the accuracy of the data. S. Tejada et.al., [22] produces a technique called Active atlas for identifying the objects. Identifying the object can be done when communicating data from many websites. This identification can be done either by manual process or by transformation of string. The manual process requires more time and it results with many error. User interaction is less and it produces the result of high accuracy.

R. Vernica et.al., [23] proceeds a technique called parallel set similarity joins to reduce the duplicate data. It consists of three types such as self-join R-S joins cases and end to end data. Partition the data with many inputs of the map reduce. It also improves the performance of the datasets in the framework. String similarity join [24] find the similar data from the given candidate data. This approach J. Wang, G. Li, and J. Fe et.al., proposes two similarity joins. Comparing the two data strings and produces the exact data using the fuxxy region and identification and classification. It also improves the performance of the redundant data.J. Wang et.al., [25] introduces the entity matching to find out the similar data. The attributes must be same and it is used for cleaning the error in the data. It also provides the effective algorithms to find out the exact data. It improves the performance and accuracy of the data. C. Xiao, and W. Wang et.al [26] reported that it identifies the more relevant data. The issues are overcome by identifying the duplicate data and improve the accuracy and the performance.

C. Li et.al., [10] proposed a technique called jaccard similarity and cosine similarity. Clean the given data by questing the queries to the database. The required data is retrieved only for small amount of data. It can't able to retrieve the duplicate data from large amount of data. Many algorithms are used to improve the performance. Filtering techniques are used in between the algorithm to improve the performance. Z. Zhang and M. Hadjieleftheriou et.al., [16] presented a B ed - tree and B+ tree uses all types of query similarities based on string distance and normal string distance. String similar search is the basic problem in retrieving the original information. There are many properties which we match the string space and the integer space. Finally it improves the response time and scalability. C. Xiao et.al., [17] introduces a technique called Top –k set similarity join. It is used to retrieve the data from the web page. Data is integrated and analysis of pattern is also recognized. Top k-set is the most important algorithm compared to all other algorithms. By using this technique we can improve the efficiency of memory and valid time. A technique of Trie-join is reported in [21] to identify the data with the small set of data from the database.

Motivated by the fact that the contribution in the data de duplication is to retrieve the duplicate data from the original data as specified this section, have the limitation of focusing only on certain issues and each has its own drawbacks. A simple map reduce technique is used to retrieve the duplicate data from large amount of dataset is proposed in the paper. Also this algorithm is the most of the two stage sampling selection strategy. The detailed description of the proposed method is presented in the next section.

III. DATA DEDUPLICATION TECHNIQUES

a. Basic terminologies

1) *De Duplication:* De Duplication is defined as the compression of data for eliminating the duplicate copies of data. Storage space can be increased and reduce the no bytes that must be sent.
2) *Fuzzy region:* Fuzzy region is defined as the application boundaries that considerably reduce the context. Fuzzy region is to find out the exact match of the data.
3) *FS-DE DUP Framework:* It is used to reduce the potential size of the training set. By performing the classification process can obtain the useful information.
4) *Active Learning:*Active learning is defined as improving the accuracy. It is used to segregate the matching and non-matching pairs.
5) *Inverted Index:*Inverted index is consists of two types. The first type consists of small set of records and it identifies the duplicate data. The second type

consists of large amount of data and using this data it can't able to retrieve the original data.

6) *SIM_MR:*Map reduce is defined as calculating the edit distance from the large datasets of strings. One of the best alignment of string is the edit distance.

b. Algorithm

Algorithm 1:*SSAR: Rule-based Active Selective Sampling*
Input: *Unlabeled set T and ∑ min*
Output: *The training set D*
Begin
Step 1: Unlabeled pair filters the irrelevant elements from the database.
Step 2: The database is projected and useful rules are extracted.
Step 3: Unlabeled pair with similar features are compared with current training set.
Step 4: Unlabeled pair with dissimilar produce projection and compared the training set.
Step 5: Re projection for the dissimilar features.
Step 6: Similarity true matching pairs are found.
Step 7: If the true matching pair is not found the non-matching pair with high similarity is return.
End

Algorithm 2: *Active fuzzy region selection*

Input: Level L1,L2

Begin

Step 1: Initialize the value of minimum false pair, minimum true pair.

Step 2: The condition must be started from the value zero to

ten.

Step 3: The initialized value is incremented.

Step 4: The cpi denotes the false value and initialize the value as minimum true pair.

Step 5: If the minimum true pair is low then print the lowest true pair.

Step 6: The Lpi denotes the true value and initialize the minimum true pair.

Step 7: If the minimum false pair is high then print the highest false pair.

End

IV. DATA DEDUPLICATION USING THE TWO STAGE SAMPLING SELECTION STRATEGY

Redundancy removal of a two-stage sampling strategy aimed at reducing the user labeling effort in large scale de duplication tasks. This process considerably reduces the number of candidate pairs since pairs below the fuzzy region can be pruned out. As indicated in the diagram a de duplication data is divided into three stages. The stage is blocking stage. The second stage is comparison stage. Finally the third stage is classification stage. In blocking stage the user details are collected from the dataset. The dataset consists of all the details of the customer who is having account in the organization. The sub branches collect all the details of the customer and submit it to the main branch. Redundancy removal of a two-stage sampling strategy shown in fig 1.

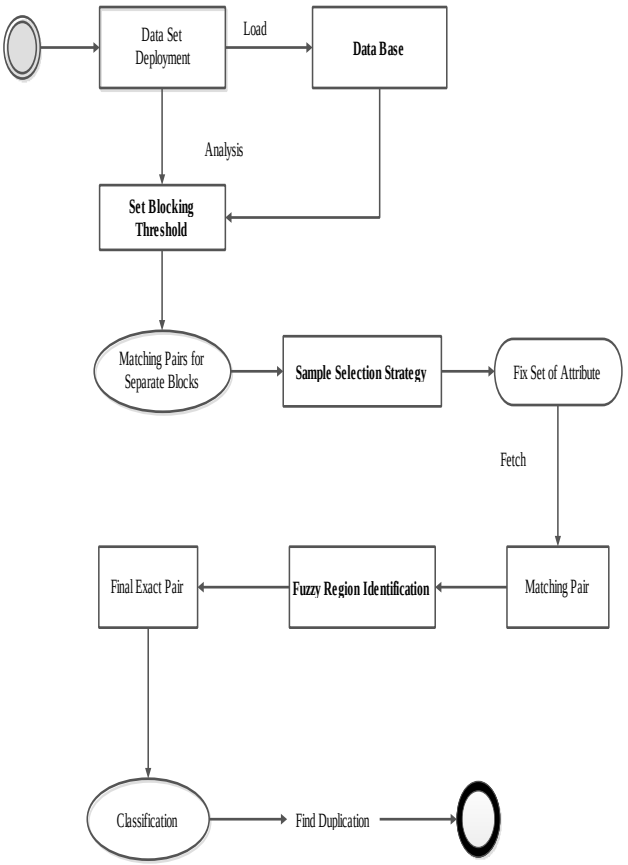

Fig. 1. Redundancy Removal

The blocking stage is used to separate the customer data according to the alphabetical order. In comparison stage, a two stage sampling selection strategy technique is used. It is further classified into two stages. The first stage is sampling selection strategy. The reduced data from blocking stage is moved to the sampling selection strategy. The ten customers are then reduced to the six customers. The redundant data is then sent to the second stage. The second stage is fuzzy region identification and classification. The redundant data is then moved to the fuzzy region identification and classification. Fuzzy region is used to find out the exact match of the candidate data. The six customers are then reduced to the four customers. The third stage is the classification stage. It is used to classify the candidate data whether it is the original data or the duplicate data. In classification stage, categorize and separate the data from the duplicate data from the original data. The output is produced.

The proposed work is used to verify the performance analysis. To check, whether found out data is correct or not. SIM_MR algorithm is used to check and improve the performance. SIM_MR is based on the Edit distance based algorithm. This algorithm is used to find the similar data based upon the strings. The candidate data must be based on the matching strings. Finally it verifies the original data and produces the result. Fig 2. represent the system architecture.

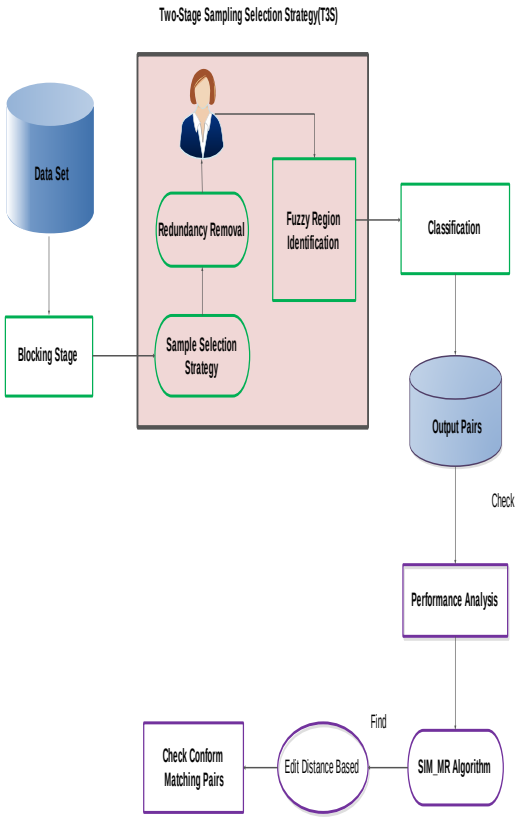

Fig. 2. System Architecture

V. SYSTEM DESCRIPTION

This section presents the procedure for finding out the duplicate data. The various components used here are: (A) Dataset Deployment, (B)Blocking Stage, (C)Sample Selection Strategy, (D) Detecting the Fuzzy region and Classification, (E) Redundancy Removal, (F) Performance Evaluation.

a. DATASET DEPLOYMENT

The dataset deployment is used to extract the data from the large dataset of an organization. The dataset is divided into many attributes such as First name, Last name, sex date of birth, email-id, contact number, street name, city, address, blood group and date of joining. The customers used to fill these attributes. Some customers may not fill some of the details. Collect all the data from the four branches and submit it to the main branch. The total customer count is 106. The data is given for extraction.

b. BLOCKING STAGE

The blocking stage is used to set the threshold. Separate the alphabetical letters into the four blocks. The first block is separated from A –F . The second block is separated from the G-L. The third block is separated from the M-R . The fourth block is separated from the S-Z. The collected data from the dataset deployment is separated according to the blocks. The reduced candidate data is then moved to the sample selection strategy.

c. SAMPLE SELECTION STRATEGY

The sample selection strategy is used to two attributes. Set two attributes such as First name and the sex. These two attributes are most commonly filled by the customers. So these two attributes are taken for testing the data. The blocked data is taken and is verified using the blocks. The first block is analyzed and it results the count of the customer. Similarly the second block is analyzed and produces some similar data. The third block is analyzed and produces some similar data. Finally the fourth block separated and produces the similar data. After analyzing each block the customer count is reduced and the customer count is 39.

d. DETECTING THE FUZZY REGION AND CLASSIFICATION

The fuzzy region and classification is divided into four attributes such as Street name, city, address, and blood group. These are the attributes which are most commonly filled by the customers. By analyzing using these attributes the customer data will be reduced to six customers. Among the six customers, three customers are the duplicate data and three customers are the original data. The next stage is redundancy removal.

e. REDUNDANCY REMOVAL

The redundancy removal produces the final result. This stage consists of the total customer, redundancy removal and total. After identifying the duplicate data, the total customer count is 106. The redundancy data count is 3. Finally the total count is 103. In the above details the total customer count is the original data which we receive during the dataset deployment. Duplicate data is identified as three members from the 106 members. The original customer details are stored in the server of the main branch.

f. PERFORMANCE EVALUATION

In this module, evaluate the performance of the system using time and accuracy metrics. By using the simple map reduce algorithm improving the performance using the edit distance based algorithm. The edit distance is one of most important technique to improve the performance.

VI. CONCLUSION

In this paper a two stage sampling selection strategy and simple map reduce concept is proposed. The technique is built on blocking, comparison and classification stage. By analyzing the data in these three stages we can identify the duplicate data from the original data. T3S is used to reduce the effort and improve the effectiveness. Maximum false pair and minimum true pair are used to reduce the boundary

value and produces the ideal values. Future work, output. verifying the performance analysis and checking for the previous output. Implementing an algorithm which is tailored to the map reduce framework architecture platforms better than the simple parallel implementation.

REFERENCES

[1] A. Arasu, M.Gotz, and R.Kaushik, "On active learning of record matching packages," in Proc. ACM SIGMOD Int. Conf. Manage. Data, pp. 783–794, 2010.
[2] A. Arasu, C.R_e, and D.Suciu, "Large-scale deduplication with constraints using dedupalog," in Proc. IEEE Int. Conf. Data Eng., pp. 952–963, 2009.
[3] R. J. Bayardo, Y. Ma, and R. Srikant, "Scaling up all pairs similarity search," in Proc. 16th Int. Conf. World Wide Web, pp. 131–140,2007.
[4] K. Bellare, S. Iyengar, A. G. Parameswaran, and V. Rastogi, "Active sampling for entity matching," in Proc. 18th ACM SIGKDD Int.Conf. Knowl. Discovery Data Mining, pp. 1131–1139, 2012.
[5] A.Beygelzimer, S.Dasgupta, and J.Langford, "Importance weighted active learning," in Proc. 26th Annu. Int. Conf. Mach.Learn., pp. 49–56, 2009.
[6] M. Bilenko and R. J. Mooney, "On evaluation and training-set construction for duplicate detection," in Proc. Workshop KDD,pp. 7–12, 2003.
[7] S. Chaudhuri, V. Ganti, and R. Kaushik, "A primitive operator for similarity joins in data cleaning," in Proc. 22nd Int. Conf. Data Eng.,p. 5, Apr. 2006.
[8] P. Christen, "Automatic record linkage using seeded nearest neighbour and support vector machine classification," in Proc.14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining,pp. 151–159, , 2008.
[9] P. Christen, "A survey of indexing techniques for scalable record linkage and deduplication," IEEE Trans. Knowl. Data Eng., vol. 24,no. 9, pp. 1537–1555, Sep. 2012.
[10] C. Li, J. Lu, and Y. Lu. Efficient merging and filtering algorithms for approximate string searches. In *ICDE*, 2008.
[11] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," Mach. Learn., vol. 15, no. 2, pp. 201–221, 1994.
[12] G. Dal Bianco, R. Galante, C. A. Heuser, and M. A. Gonalves,"Tuning large scale deduplication with reduced effort," in Proc.25th Int. Conf. Scientific Statist. Database Manage., pp. 1–12, 2013..
[13] M. G. de Carvalho, A. H. Laender, M. A. Goncalves, and A. S. da Silva, "A genetic programming approach to record deduplication," IEEE Trans. Knowl. Data Eng., vol. 24, no. 3, pp. 399–412, Mar. 2012.
[14] A. Elmagarmid, P. Ipeirotis, and V. Verykios, "Duplicate record detection: A survey," IEEE Trans. Knowl. Data Eng., vol. 19, no. 1,pp. 1–16, Jan. 2007.
[15] Z. Zhang, M. Hadjieleftheriou, B. C. Ooi, and D. Srivastava. Bed-tree: an all-purpose index structure for string similarity search based on edit distance. In SIGMODConference, pages 915–926, 2010.
[16] C. Xiao, W. Wang, X. Lin, and H. Shang. Top-k set similarity joins. In ICDE, pages 916–927, 2009.
[17] C. Gokhale, S. Das, A. Doan, J. F. Naughton, N. Rampalli, J. Shavlik,and X. Zhu, "Corleone: Hands-off crowdsourcing for entity matching," in Proc. ACM SIGMOD Int. Conf. Manage. Data, pp. 601–612, 2014.
[18] H. K€opcke and E. Rahm, "Training selection for tuning entity matching," in Proc. Int. Workshop Quality Databases Manage. Uncertain Data, pp. 3–12, , 2008.
[19] S. Sarawagi and A. Bhamidipaty, "Interactive deduplication using active learning," in Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 269–278, 2002.
[20] J. Wang, G. Li, and J. Feng. Trie-join: Efficient trie-based string similarity joins with edit-distance constraint *PVLDB*, 3(1):1219–1230,2010.
[21] S. Tejada, C. A. Knoblock, and S. Minton, "Learning domain-independent string transformation weights for high accuracy object identification," in Proc. 8th ACM SIGKDD Int. Conf. Knowl. DiscoveryData Mining, pp. 350–359, 2002.
[22] R. Vernica, M. J. Carey, and C. Li, "Efficient parallel set-similarity joins using mapreduce," in Proc. ACM SIGMOD Int. Conf. Manage. Data, pp. 495–506, 2010.
[23] J. Wang, G. Li, and J. Fe, "Fast-join: An efficient method for fuzzy token matching based string similarity join," in Proc. IEEE 27th Int. Conf. Data Eng., pp. 458–469, 2011.
[24] J. Wang, G. Li, J. X. Yu, and J. Feng, "Entity matching: How similar is similar," Proc. VLDB Endow., vol. 4, no. 10, pp. 622–633, Jul.2011.
[25] C. Xiao, W. Wang, X. Lin, J. X. Yu, and G. Wang, "Efficient similarity joins for near-duplicate detection," ACM Trans. Database Syst.,vol. 36, no. 3, pp. 15:1–15:41, 2011.