An Enhanced Unsupervised Fuzzy Expectation Maximization Clustering for Deduplication of Records in Big data

P. Selvi, D. Shanmuga Priyaa

Abstract: ***The main issue while handling records in data warehouse or cloud storage is the presence of duplicate records which may unnecessarily test the storage capacity and computation complexity. This is an issue while integrating various databases. This paper focuses on discovering records, entirely and partly replicated, before storing them in cloud storage. This work converts whole content of data to numeric values for applying deduplication using radix method. Fuzzy Expectation Maximization (FEM) is used to cluster the numerals, so that the time taken for comparison between records is reduced. To discover and eliminate the duplicate records, this paper used divided-and-conquer-algorithm to match records among intra-clusters, which further enhances the performance of the model. The simulation results have proved that the performance of the proposed model achieves higher detection rate of duplicate records.***

Keywords: Duplication, Data warehouse, Cloud storage, Fuzzy Expectation Maximization, Deduplication.

I. INTRODUCTION

The existence of duplicate records is a chief concern which affects the quality of data in a large database. Detection of duplicates is also referred to as entity resolution or record linkage, involved in discovering records that possibly refer to the same entity. As an illustration, during the process of detecting duplicates, publications are linked to authors' names depending on bibliographic records. Different formats, accents of characters and typing can make cataloguing publications by author names very complex. Repeated occurrences of names of author referring to same author would need to be linked relying on some concepts of resemblance, to attribute publications to their corresponding authors. But there is no wide-ranging pragmatic training that assesses the excellence of the grouping or clustering adapted in these techniques. Thus, this work aims to develop an enriched clustering model for detection of duplicates, within the machine learning paradigms.

In heterogenous sources, the data will be represented in different formats; so, the presence of noise is natural [1] and the essential task of preprocessing is cleansing of data in data warehouse. Cleansing of data includes identifying redundant, incorrect and missing values and correcting them [2]. It also includes completeness, checking format of the data and other related errors in data. Due to the typo error like spelling mistakes, missing integrity constraint, noise entry and duplica ted records [3].

Revised Manuscript Received on October 22, 2019.
***** Correspondence Author
P. Selvi, Research Scholar, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore – 641021.
Dr. D. Shanmuga Priyaa, Professor, Department of CS, CA & IT, , Karpagam Academy of Higher Education, Coimbatore – 641021.

Measure of data quality highly depends on their completeness, consistency, accuracy and redundancy aspects. If the data quality is poor, the strategic decisions done based on these data lead to high misclassification. Thus, this work focuses on record deduplication as a major factor in data quality. It is necessary to alleviate the process of deduplication process for discovering duplicate records and eliminating it.

In this paper, an empowered approach on deduplication using clustering-based record similarity match is performed by converting the field values to numeric ones instead of transforming them to tokens.

II. RELATED WORK

This section talks about some of the existing models of deduplication in big data storage and also maintains the security mechanism to achieve data integrity.

Zheng Yan [4] has developed a model to deduplicate the data which are encrypted and store in cloud based on the proxy re-encryption. This model uses access control and data integrity while performing data deduplication in cloud storage.

Jebamalarand brilly [5] has devised a model for detection and elimination of duplicate data using rule-based duplicate record detection and elimination. This technique is used to improve the effectiveness of the data.

Subramaniyaswamy and Chendthur Pandian [6] present a complete analysis of the various existing models in the field of duplicate record detection; the duplicate record detection is an essential step in the process of integration of data in big data storage.

Lillibridge et al. [7] have formulated a chunk fragmentation-based deduplication model. Instead of comparing the entire field character by character this model fragments the record information and it is encrypted to the reduced size so that the time complexity and comparison complexity are reduced.

Yan et al. [8] propose a thrust-based secure hybrid model which consists of two levels, the first using cryptography keys as revocation and the second relying on validity time of the keys involved in cryptography.

Chuanyi Liu et al. [9] offer a thrust-relationship-based deduplication on cloud storage. The policy-oriented approach is used for discovering duplicate records in data storage. The re-encryption process is applied to protect strongly against intruders, who can't break the secrecy of the data by unlocking the message.

Jin Li et al. [10] devised a deduplication model with security mechanism, which uses the privilege of data owner as a proof for authorizing duplication checking process in private cloud storage.

Shweta et al. [11] propose an approach for data deduplication by generating tokens for each file that has to be stored in the private cloud. The token generated for each file will be unique, the file content is hashed and the token is added with the file and stored in the cloud storage. The tokens and the hashed code are compared to detect the presence or absence of duplicate files.

Backialakshmi et al. [12] offer a deduplication technique by providing integrity of data by owners. Using the identity process, the model performs both security paradigm and deduplication process more efficiently.

BhushanChoudhary et al. [13] have a cloud storage deduplication system in order to reduce integrity check and storage size. They transform a predictable file into an unpredictable one. They upgrade the security along with deduplication in a more efficient manner.

James et al. [14] have established a deduplication method by designing Cauchyreed-solomon coding distribution matrix to encode the information of the file content. It is chiefly useful for the big data storage, specifically in cloud storage paradigms.

III. METHODOLOGY OF FUZZY EXPECTATION MAXIMIZATION CLUSTERING FOR RECORD DEDUPLICATION

This work detects and eliminates records that are both completely and partly duplicated. Fuzzy Expectation Maximization method is utilized to decrease the amount of evaluations by establishing clusters by using divide-and conquer-method to match the records in intra-clusters. This work categorizes duplicate records into 3 different groups:

i. fully duplicated records, in which two identical records denote similar real-life dataset
ii. flawed Duplicated Records; due to typo error, duplicate records seem to be different
iii. partially Duplicated Records, in which parts of records are duplicates; but its difference is unique.

To discover these three categories of duplicate records, the proposed method performs data value conversion, clustering and matching.

3.1 Process of Data Conversion

In this stage the model initially converts input data of varied formats into a uniform unique format. There might be abundant configuring problems in data stored in cloud storage collected from various functional systems, specifically while handling the date format which takes various forms like mm-dd-yyyy, dd-mm-yyyy, or yyyy-mm-dd etc. Another example is phone number, where some records may have country and city code and others identical phone number deprived of city code. These kinds of missing values and formatting problems are determined and data are converted to a standard format.

To overcome this problem, the proposed model used radix formula which converts the string, numeric or date format values to numeric format. Thus, it handles the inconsistency in handling different formats of data during the deduplication process. After performing conversion, an additional column is added to store the computed values in the corresponding column with respect to the related row parted by comma.

The table below consists of three different fields, the value of each being changed into numeric value and kept in the attached column.

Table 1: Transformed into Numeric Value and Stored in the Attached Column.

Emp Name	Designation	Salary	Numeric Translation
Anitha	Manager	50000	

The formula for numeric value conversion is as shown in the equation 1:

$$\sum[(rad)^{pos} \times av) \bmod m] \quad (1)$$

where pos refers to position of the character placed in the corresponding filed, whose value is assigned from left to right, starting from 0. Rad denotes radix value which is greater than or equivalent to 36 as it comprises 36 letterings i.e., 10 digits 0-9 + alphabets 26 + special character. Depending on digits, alphabets and special characters, radix value is assigned so that its value is greater than or equal to 36. While using special characters the radix value will be increased as the special character value is also included in alpha value (av). The av is manifest from 0 to 9, aA=10, bB=11,…, zZ=35. m refers to big prime number.

For example: conversion of Sample Record to numeric value using radix method.

Table 2: Conversion of sample record to numeric value using radix method

Emp Name	Designation	Salary	Numeric Translation
Raj	Manager	50000	109, 1103,55

Sample Calculation:

Name = Raj

Let m = 283

$\Sigma\{((36)^2 * 27) \bmod 283) + ((36)^1 * 10) \bmod 283) + ((36)^0 * 19) \bmod 283)\}$

183 + 77+ 19 = 109

Algorithm: Conversion Using Radix method

Input: Fields of each record in a database with different format
Output: Conversion value with uniform format of values with appended numeric value

Begin

- Generate a largest prime number as m
- For record*i*= 1to m (last record in db)
- For attribute *t* = 1to n (last attribute of record)

  For each character in an attribute value

  - If numerals assign values 0 to 9
  - If alphabets assign values from 10 to 35 for Aa to Zz correspondingly
  - If special characters then assign increase the vale starting from 35+
  - Apply the equation to convert the values to uniform format as given in the following equation

$$\sum[(rad)^{pos} \text{ x av}) \bmod m]$$

- Accumulate resultant numeric value into last column of the table by attaching the value parted with comma (,)

End

3.2 Discover record deduplication using Fuzzy Expectation Maximization Clustering

The standard Expectation maximization algorithm is similar to k-means clustering, where the k-means assigns instances to clusters in order to maximize the variance among all variables across the clusters [14, 15]. But EM algorithm performs clustering to maximize the difference in means for continuous variables. It computes probabilities of cluster memberships depending on overall likelihood or probability of the records, given the resultant clusters.

This paper used Fuzzy Expectation Maximization algorithm to cluster the records based on the maximum likelihood among the records to perform deduplication operation. Initially values of records are converted to fuzzy domains and then iterative method of clustering is applied by making use of gaussian mixtures model and estimate likelihood among records and cluster them accordingly. This algorithm mainly uses two different steps, namely expectation (E) and maximization (M).

While observing records for discovering duplicate records, each field is observed and the final output column which sums all the value in the appended field is considered as implicit latent variable z_i used to decide which gaussians it came from. If there are k number of neighborhoods for a specific record which has to be assigned to any one of the cluster in such case, the Z_i is considered as a categorical distribution with the parameters π=[π1,…,πk] and the final output of the kth neighborhood as a Gaussian N(μ_k,σ_k^2), where μ_k as mean value and σ_k^2 as variance.

Once the attribute values are converted to the final output column as values, this proposed work uses fuzzy expectation maximization algorithm, and based on similarity the clustered label is added at the clustered column. Likewise, similar records are stored in single clusters. This results in reduction of comparisons and prominently enhances its performance.

Table 3: Clustering Records using Fuzzy Expectation Maximization

Name	F.name	Job	Salary	Appended Column	Final Output	Clustered
Vinu	Raj	Manager	35250	610, 279, 1103, 346	2338	C1
Kumar	Arun	Accountant	15000	781,613,2535, 99	3929	C2
Uma	Raj	Manager	35250	346, 279, 1103, 346	2074	C1
Anu	Bala	Accountant	18000	398, 534, 1103,265	2300	C1

In table 3, six records are grouped into two. If the table consist of a single group then the number of comparisons will be 15. This is because record 1 is compared with remaining 5 records i.e., records 2, 3,4, 5 and 6, record 2 is compared with 4 different records such as records 3,4,5 and 6, record 3 is compared with 3 different records i.e., records 4,5 and 6, record 4 is compared with 2 different records 5 and 6, finally, record 5 is compared with one different record i.e., record 6.

While there are two different groups, comparisons are done within the same groups reducing the number of comparisons. To achieve this objective, this paper uses fuzzy expectation maximization clustering (FECM) for clustering the records based on their similarity, and it works faster for finding duplicate records within the database.

3.3 Identifying Duplicated Records

Using divide-and-conquer-approach, each record value is divided recursively until it reaches smaller pieces and prolongs the process till a specific size is reached. After that it determines the comparison of single value of a record with single value of other records. If match is found among record's value then the duplication percentage of those two records will be computed.

Table 3 shows clearly that the values of both the records are matched with their corresponding attribute values, and among four attributes, three attribute values are matched, and their duplication is computed as:

¾ * 100 = 75%

It is also revealed that due to typo error, the name filed is mistyped. Such variation is called error and it will be corrected by experts of the

domain. In this case it is known as partly duplicated, and the threshold value checked. If the duplication percentage crosses the specified threshold, the records are considered for examination to analyze whether the difference is original or erroneous. As shown in table 4, if it is erroneous the particular record is corrected, and it becomes fully duplicated with only a single copy of the record being maintained. But if those records hold actual difference, those two records are considered as different and both the copies are maintained in the database.

Table 4 Partly Duplicated Records

Name	F.name	Job	Salary	AppendedColumn
Kumar	Arun	Accountant	15000	781,613,2535,99
Umar	Arun	Accountant	15000	561,613,2535,99

Table 4 shows two records with 75% of duplication. There is a unique variance among these records; only their name field differs and if it is confirmed that both the records belong to different individuals their data will be maintained separately; else they are considered as duplicate, and only a single copy of the record will be maintained.

Table 5 Fully Duplicated Records

Name	F.name	Job	Salary	Appended Column
Vinu	Raj	Manager	35250	610,279, 1103, 346
Vinu	Raj	Manager	35250	610, 279, 1103, 346

As in table 5, if both records under investigation are fully duplicated, the duplicated records will be discarded and the original copy will be maintained in the database. For example, if the comparison value among two records is 100%, it is confirmed that these two records are fully duplicated.

Algorithm for Deduplication using Fuzzy Expectation Maximization

Input: Restaurant Dataset with Duplicate records (RDS)
Output: Dataset with Unique Records (FRDS)
Begin
For each record in RDS
For each field in record
Convert the characters in field to numeric values using radix conversion

$$\sum[(rad)^{pos} \text{ x av}) \bmod m]$$

Append each converted field value to append column
End for
End for
For each record in RDS
Sum the values in append field and store it in final output field
End for
// Clustering process
For each record [final output] in RDS
//convert to fuzzy value
Assign membership function within the interval value [0, 1]

$$\mu A(x) = \begin{cases} 0 & (x < a) or (x < d) \\ \dfrac{x-a}{b-a} & a \le x \le b \\ 1 & b \le x \le c \\ \dfrac{x-a}{b-a} & c \le x \le d \end{cases}$$

Generate an initial model Ml = (Cl_1, …, Cl_k) which randomly selects the centroids
repeat
// reassign records clustering
Calculate Pb(r|Ci), Pb(r) and Pb(Ci |r)
for every object r from RDS and each cluster = (Gaussian) Cl_i
// (recompute the models
Calculate a new-fangled model $Ml' = \{Cl_1', \ldots, Cl_k'\}$ by recalculating W_i , μ_{Cl} and σ_{Cl} for each Cluster Cl_i
Substitute Ml by Ml'
until |Ed(Ml') – Ed(Ml)| <∈
return Ml'
Store FRDS = Ml'
End

IV. EXPERIMENTAL RESULT

This proposed Fuzzy Expectation Maximization Clustering (FEMC) based Deduplication model is simulated using matlab. It uses two different datasets, namely Cora dataset [18] which consists of 1295 records and the restaurant dataset [19] with 864 records for performing the deduplication using unsupervised learning paradigm by clustering similar records and discovering effectively records that are duplicated fully and partially . This proposed model is compared

with other two existing model K-means [16] and DBSCAN [17] algorithm.

Assessment Metrics

To evaluate the performance of FEMC-based deduplication model three different metrics are used in this work. The proposed FEMC deduplication method is compared with K-Means and DBSCAN clustering models.

Precision

It is a measure of correctly clustered records, with actual number of records that truly belong to their corresponding classes.

***Precision* = Number of Correctly Clustered Records /Total number of actual records in each cluster.**

Recall

It is the ratio of the correctly clustered records to the number of records in a specific cluster.

Recall* = Number of Correctly Clustered Records /*Number of records in a specific cluster

F-Measure: It highly depends on the precision and recall value's harmonic mean to detect duplicate records

F-Measure = 2 * (precision * recall) /(precision + recall)

Table 6: Performance Analysis of Three Different Clustering Models in Deduplication Process

Clustering Approaches	Correctly Clustered Instances (%)	Incorrectly Clustered Instances (%)
K-means	75.7	24.3
DBSCAN	82.6	17.4
FEMC	97.98	2.02

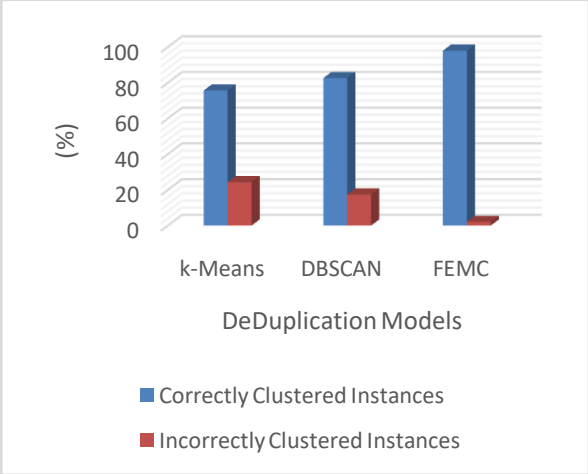

Figure 1: Performance Analysis of three different clustering models in deduplication process

Table 6 and figure 1 reveal that the proposed Fuzzy Expectation maximization algorithm with the help of gaussian mixture method can cluster a greater number of correct similar records with a high percentage of accuracy; it works based on the likelihood and probability bases instead of randomness as done by the k-means and dbscan algorithm. The highest correctly clustered percentage is 97.98, and the minimum misclustering percentage is 2.02%.

Table 7: Performance comparison of Deduplication models based on evaluation measure

Deduplication models	Precision	Recall	F-measure
K-means	75.6	78.2	76.88
DBSCAN	80.4	84.7	82.49
FEMC	95.8	98.6	97.18

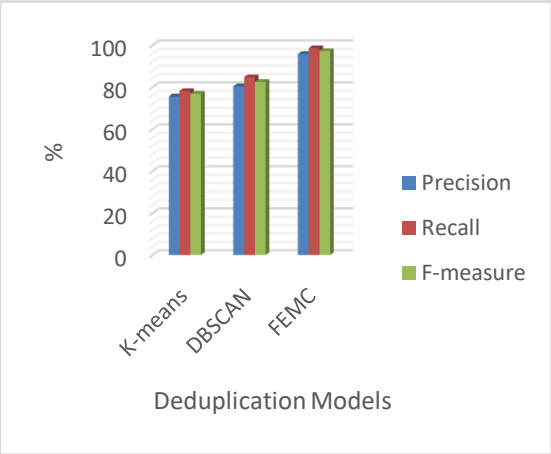

Figure 2: Performance comparison of Deduplication models based on evaluation measure

Table 7 and figure 2 reveal that the similarity of records with the presence of characters and numbers are handled well in the propose FEMC model by using radix numeric conversion method. In this the representation of the field values are denoted in fuzzy values and the similarity among the records is determined using Expectation maximization clustering model, which groups the more similar records within a cluster using the gaussian mixture model or distribution. Where K-means and Dbscan clustering models clusters the instances or records based on the centroids and the distance measure they produce poor results while comparing FEMC deduplication model.

V. CONCLUSION

This work aims to develop an enhanced deduplication model to increase the efficiency of the storage capacity and decrease the computation complexity while working with files or records either in data warehouse or in cloud storage. Before storing all the records there is a need for discovering whether there is an existing copy of record in the data storage, to avoid redundancy and duplication. To overcome this issue, the proposed work develops an unsupervised clustering model fuzzy expectation maximization which clusters the similar records to their corresponding cluster.

In standard deduplication model, while a record has to be stored, the process of verifying its existence is essential to avoid duplicates. So, the entire data storage is searched and each record is compared against it. But this proposed method searches only within the clusters which it belongs to thereby reducing the time for comparison and computation complexity. Additionally, to handle both alphabets, numeric and special characters presented in each field are converted into numeric value using radix conversion method, so that the comparison of records also becomes very effective. The proposed model FECM performance is compared with k-means and DBSCAN, and the results prove that FECM achieves higher accuracy, while comparing other two models.

REFERENCES

1. P. Poonniah, Data Warehousing Fundamentals- A comprehensiveguide for IT professionals, 1st ed., 81-265-0919- 8, Glorious Printers:New Delhi, India, 2006.
2. R. Arora, P. Pahwa, and S. Bansal, “Alliance Rules for DataWarehouse Cleansing,”in Proceeding of International Conference ofSignal Processing Systems, IEEE, 2009.
3. M. Rehman and V. Esichaikul, “Duplicated Record Detection forDatabase Cleansing,” in Proceeding of Second InternationalConference on Machine Vision, 2009.
4. Zheng Yan, Wenxiu Ding, Xixun Yu, Haiqi Zhu, Robert H. Deng,Deduplication on Encrypted Big Data in Cloud, IEEE Transactions on Big Data, VOL. 2, NO. 2, Pp 138-150,2016
5. J. JebamalarTamilselvi, C. BrillyGifta, Handling Duplicate Data in Data Warehouse for Data Mining, International Journal of Computer Applications (0975 – 8887) Volume 15– No.4, February 2011
6. V.Subramaniyaswamy and S. Chendthur Pandian, A Complete Survey of Duplicate Record Detection using Data Mining Techniques, Information Technology Journal 11(8), 941-945, 2012.
7. M. Lillibridge, K. Eshghi, and D. Bhagwat, ―Improving Restore Speed for Backup Systems That Use Inline ChunkBasedDeduplication,‖ Proc. 11th Usenix Conf. File and Storage Technologies, 2013, pp. 183–198.
8. Z. Yan and M.J. Wang, ―Protect Pervasive Social Networking Based on Two Dimensional Trust Levels,‖ IEEE Systems J., Sept. 2014, pp. 1–12; doi: 10.1109/JSYST.2014.2347259.
9. Chuanyi Liu, Xiaojian Liu and Lei Wan “Policy-based deduplication in secure cloud storage,” in Proc. Trustworthy Compute. Serv., 2013, pp. 250–262, doi: 10.1007/978-3-642- 35795-4_32.
10. Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou “A Hybrid Cloud Approach for Secure Authorized De-duplication” IEEE Transactions on Parallel and Distributed Systems: PP Year 2014.
11. Shweta D. Pochhi, Prof. Pradnya V. Kasture, “Encrypted Data Storage with De-duplication Approach on Twin Cloud” International Journal of Innovative Research in Computer and Communication Engineering.
12. Backialakshmi, NManikandan. “M Secured Authorized De-Duplication in Distributed System”, IJIRST International Journal for Innovative Research in Science and Technology— Volume 1 — Issue 9 — February 2015.
13. BhushanChoudhary, AmitDravid, “A Study on Secure Deduplication Techniques In Cloud Computing” International Journal of Advanced Research in Computer Engineering and Technology (IJARCET) Volume 3, Issue 12, April 2014
14. Dempster, A.P.; Laird, N.M.; Rubin, D.B. "Maximum Likelihood from Incomplete Data via the EM Algorithm". Journal of the Royal Statistical Society, Series B. 39 (1): 1–38, (1977).
15. Diffey, S. M; Smith, A. B; Welsh, A. H; Cullis, B. R, "A new REML (parameter expanded) EM algorithm for linear mixed models". Australian & New Zealand Journal of Statistics. 59 (4), (2017).
16. Campello, Ricardo J. G. B.; Moulavi, Davoud; Zimek, Arthur; Sander, Jörg, "Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection". ACM Transactions on Knowledge Discovery from Data. 10 (1): 1–51, (2015).
17. Ling, R. F, "On the theory and construction of k-clusters". The Computer Journal. 15 (4): 326–332. 1972
18. https://relational.fit.cvut.cz/dataset/CORA
19. https://www.cs.utexas.edu/users/ml/riddle/data.html

AUTHORS PROFILE

P. Selvi, Research Scholar, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore – 641021.

Dr. D. Shanmuga Priyaa, Professor, Department of CS, CA & IT, , Karpagam Academy of Higher Education, Coimbatore – 641021.