JES
Journal of Engineering Sciences
Vol 11, Issue 3, MARCH/2020
ISSN NO:0377-9254

AN EMPHERICAL APPORACH FOR DETERMINING CLUSTERING BASED DUPLICATION DETECTION AND ELIMINATION PROCESS

[1]Sri Rama Lakshmi Reddy, Research scholar,
JJT University,Jhunjhunu ,Rajasthan.
[2]Dr.K Rajendra Prasad, Professor
CSE Department, Institute of Aeronautical Engineering, Hyderabad

ABSTRACT:

We propose a clustering technique for entropy based text dis-similarity calculation of de-duplication system. The process of detecting and removing database defects and duplicates is referred to as data cleaning. The fundamental issue of duplicate detection is that inexact duplicates in a database may refer to the same real world object due to errors and missing data. The proposed framework uses six steps to improve the process of duplicate detection and elimination. The new method offers more accuracy dis-similarity measure for each cluster data without manual intervention at the time of duplicate deduction. This research work will be efficient for reducing the number of false positives without missing out on detecting duplicates. In this study we propose a Multi-Level Group Detection (MLGD) algorithm which produces a most accurate group with most closely related object using Alternative Decision Tree (ADT) technique.

Keywords: Clustering Algorithm, Alternative Decision Tree Algorithm, Duplicate Detection, Efficient Method, Cluster Data, Clustering Formation

INTRODUCTION:

A data warehouse is basically a database and having unintentional duplication of records created from the millions of data from other sources can hardly be avoided. In the data warehousing community, the task of finding duplicated records within data warehouse has long been a persistent problem and has become an area of active research. There have been many research undertakings to address the problems of data duplication caused by duplicate contamination of data. Here we need transformation logic for converting source data into target database for standardize the record format as well as record value for detect a duplicate record. Therefore, data should be transformed and cleansed before loading into a target database. There are two issues to be considered for duplicate detection: Accuracy and Speed. The measure of accuracy in duplicate detection depends on the number of false negatives (duplicates that were not classified as such) and false positives (non-duplicates which were classified as duplicates). The algorithm's speed is mainly affected by the number of records compared, and how costly these comparisons are. Generally CPUs are not able to do duplicate detection on large databases within any reasonable time, so normally the number of record comparison needs to be cut down. In this research work, a framework is developed to handle any duplicate data in a data warehouse**.**

LITERATURE REVIEW:

Lavanya Pamulaparty et al (2014) Detecting near Duplicates is very difficult in large collection of data like "internet". The presence of these web pages plays an important role in the performance degradation while integrating data from heterogeneous sources. These pages

either increase the index storage space or increase the serving costs. He concerns detecting, and optionally removing duplicate and near duplicate documents which are used to perform clustering of documents .We demonstrated our approach in web news articles domain. The experimental results show that our algorithm outperforms in terms of similarity measures. The near duplicate and duplicate document identification has resulted reduced memory in repositories.

Bilal Khan et al (2012) Records duplication is one of the prominent problems in data warehouse. This problem arises when various databases are integrated. This research focuses on the identification of fully as well as partially duplicated records. In this paper we propose a de-duplicator algorithm which is based on numeric conversion of entire data. For efficiency, data mining technique k-mean clustering is applied on the numeric value that reduces the number of comparisons among records.

Lohman et al (2010) proposed an architecture for quickly detecting the recurrences of crashes. They also proposed a call stack matching metric. Similar to our work, they also considered the distance to the top frame and alignment offset. However, our method differs in the way the similarity metrics are formulated and in the way the parameters are tuned.

Manikandan, G et al (2011) presents the analysis of leader- follower, k-means and k-medians clustering algorithms in outlier detection based on some statistical models and spatial proximity. Clustering, which is so much used in pattern recognition, reduces the searching load. Outliers, the one which is different from norm, should be detected and handled properly. Otherwise, it will affect the original data in clustering in a great manner. Dataset for simulation has been generated using "weka" software.

OBJECTIVES:

1. To improve data quality and increase speed of the data cleaning process.
2. To Study the Duplicate Record Detection Using Different Techniques.
3. To Study the Performance Analysis of Clustering Algorithms.

METHODOLOGY:

In this study we propose a two new algorithm, First one is a clustering algorithm, which will overcome the existing clustering dis-advantage partition and hierarchical that may be either partition or hierarchical. Second one is de-duplication algorithm, which will produce the dis-similarity, percentage of the pair of string in each cluster. Here we introduced an efficient clustering mechanism as Multi-Level Group Detection using AD Tree for splitting a data into cluster, with most closely related object. Then we are applying the de-duplication mechanism in each clustered data, though this proposal method we can reduce the total time consumption for clustering formation and data comparison for deduplication than existing traditional clustering mechanism and de-duplication mechanism.

Clustering/Blocking of records: Sorting the large datasets and data duplicate elimination process with this large database faces the scalability problems. The clustering techniques are used to cluster or group the dataset into small groups based on the distance values or some threshold values to reduce the time for the elimination process. The blocking methods are used for reducing huge number of comparisons. This step is used for grouping the records that are most likely to be duplicated based on the similarity of block-token-key.

Formation of Tokens: This step makes use of the selected attribute field values to form a token. The tokens can be created for a single attribute field value or for combined attributes. For example, contact name attribute is selected to create a token for further cleaning process. The contact name attribute are split as first name, middle name and last name. Here first name and last name are combined as contact name to form a token. Tokens are formed using numeric values, alphanumeric values and alphabetic values by selecting some combination of characters.

Detection and elimination of duplicate data: In step (v), the rule based duplicate detection and elimination approach is used for detecting and eliminating the records. During the elimination process, only one copy of duplicated records are retained and eliminated other duplicate records. The elimination process is very important to produce a cleaned data. The above steps are used to identify the duplicate records.

Attribute Selection with parameters: Wrong selection of attribute affects the performance and accuracy of data cleaning process. Hence key fields are selected in such a way that fields should contain sufficient information to identify the duplication of the record. The name attribute has the highest threshold value compared to other attributes and seven attributes are selected for the next process of data cleaning.

Duplicate Detection Tools:

In the past decade, various data cleaning tools were sold out in market and they were available as public software packages mainly for duplicate record detection.

Febrl: The Febrl (Freely Extensible Biomedical Record Linkage) is an open-source data cleaning tool kit.

Tailor: Tailor is a flexible record matching toolbox. The main feature of this toolbox is that it enables users to apply different duplicate detection methods on the data sets. This tool is termed to be flexible because multiple models are supported**.**

Whirl: Whirl is an open source duplicate record detection system used for academic and research purposes. Similar strings within two lists identified using a token-based similarity metric

Duplicate Detection

Multi-Level Group Detection using ADTree

MLGD forms a tree for the clustering process. In the tree structure, the height of each level of nodes represents the dis-similar degree between each cluster. MLGD incorporate the futures of ADTree features and overcome the existing hierarchical clustering problem and reduce the time consumption for duplicate detection and number of record comparisons. ADTree divide the data based on short name; if cluster is already available with the short name then insert a record into a same cluster else create a new cluster with the new name of short name then insert into a new cluster. Here we did not use any split algorithm for splitting data into a cluster; instead we are using ADTree technique for splitting a whole data into cluster. A condition predicates the attribute comparison value, here we are checking clustering index value contains the short name value or not. In each cluster sub-set short name pointing to the

whole record. If cluster is already available then starts the de-duplication process else create a new cluster and then exit from the process.

RESULTS:

Table 1. Total number of Iteration between with grouping and without grouping

Record volume	No of iteration (Croce)	Without grouping	With grouping
5 K	100	2.5	0.16
10 K	300	10.0	0.37
35 K	500	122.5	4.56
50 K	700	250.0	9.30
70 K	900	490.0	18.23
80 K	1100	640.0	23.80
100 K	1300	1000.0	37.19

None of the existing algorithm has this kind of fuctionality to find a duplicate value wihin that specified time line.

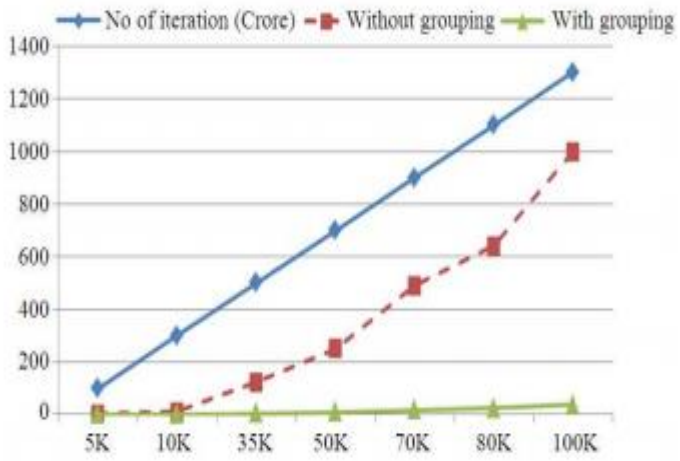

Figure: Number of Iteration between with grouping and without grouping

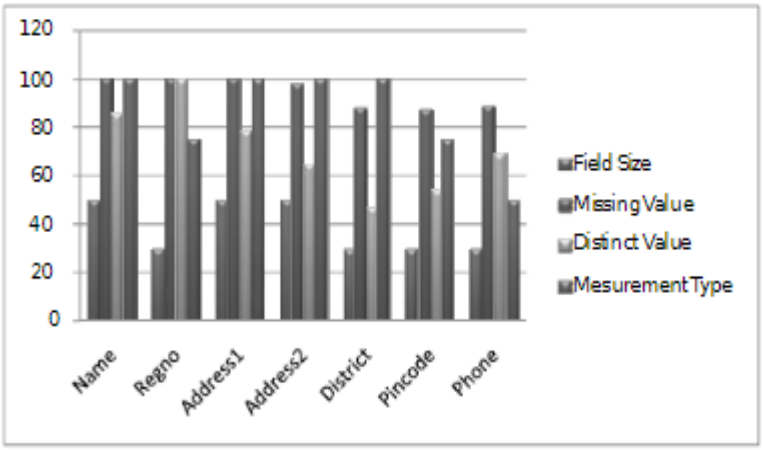

Figure: Attribute selection

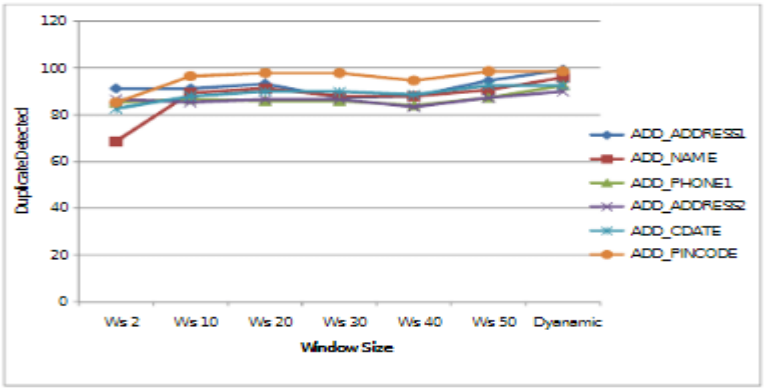

Figure: Attribute Vs Duplicate detected with varying window size.

Duplicates Vs No. of Attributes Efficiency of the record matching algorithm mainly depends on the selection of the attributes. The selection of combination of multiple attribute will be more useful to identify exact and inexact duplicates.

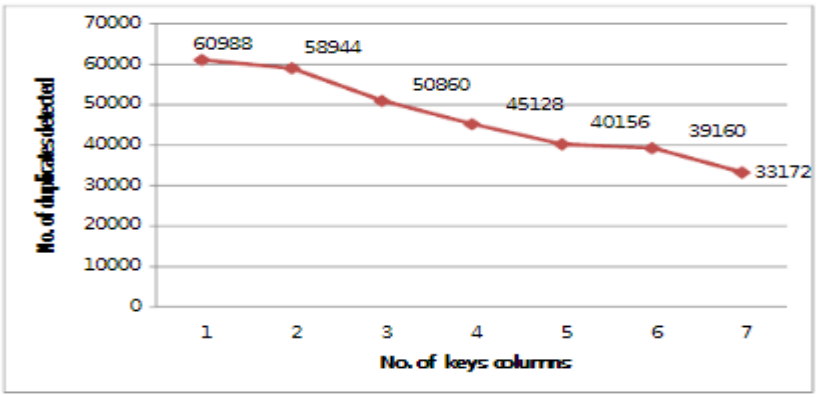

Figure: Duplicate detected Vs No. of attribute selected.

Table: Key columns and no. of duplicate detected

No. of Columns	Key Columns Selected	No. of duplicate detected
1	ADD_ADDRESS1	60988
2	ADD_ADDRESS1 ADD_NAME	58944
3	ADD_ADDRESS1 ADD_NAME ADD_PHONE1	50860
4	ADD_ADDRESS1 ADD_NAME ADD_PHONE1 ADD_CDATE	45128
5	ADD_ADDRESS1 ADD_NAME ADD_PHONE1 ADD_CDATE ADD_DEL	40156
6	ADD_ADDRESS1 ADD_NAME ADD_PHONE1 ADD_CDATE ADD_DEL ADD_PARENTTYPE	39160
7	ADD_ADDRESS1 ADD_NAME ADD_PHONE1 ADD_CDATE ADD_DEL ADD_PARENTTYPE ADD_PINCODE	33172

Duplicates Vs Size of dataset and attributes: Number of blocks are varied based on the size of dataset and number of duplicates. Basically, size of block depends on the number of duplicates available in the data set. Attribute address1 identifies more duplicates because attribute address1 has more high distinct values than other attributes. As a result, high power attributes identify high amount of duplicates than other attributes. At the same time, identification of duplicates varies for different sizes of data.

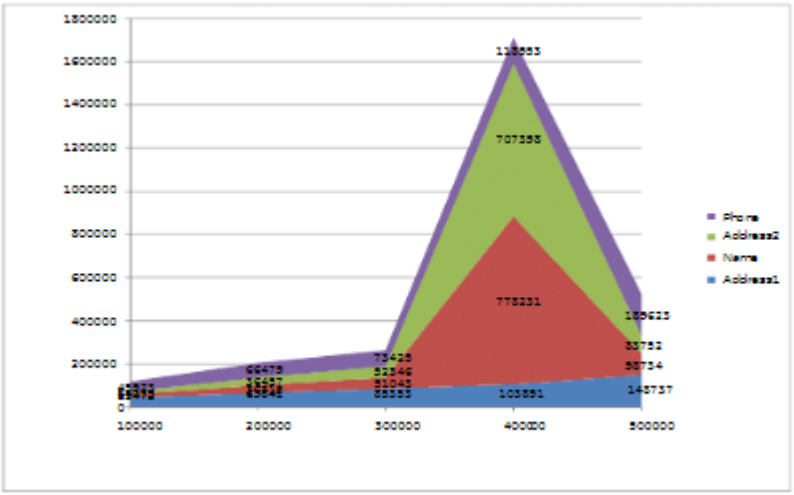

Figure: Duplicates Vs Size of dataset and attributes.

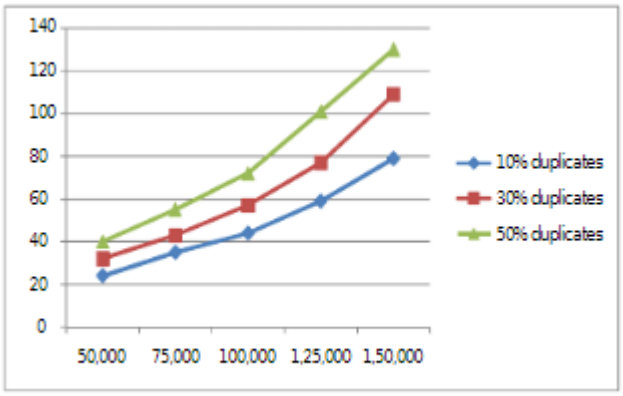

Figure: Time Vs Database size and % of Duplicates

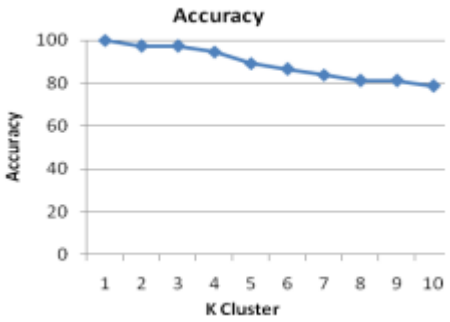

Figure. Graph between k-cluster and accuracy

We need to identify partially duplicated records which may occur in different groups. We cannot compare partially duplicated records which are present in different groups. In Fig, the graph represents the accuracy of partially duplicated records, which decreases by increasing the number of clusters.

CONCLUSION:

Deduplication and data linkage are important tasks in the pre-processing step for many data mining projects. It is important to improve data quality before data is loaded into data warehouse. In this research wok, a framework is designed to clean duplicate data for improving data quality and also to support any subject oriented data. This framework is useful to develop a powerful data cleaning tool by using the existing data cleaning techniques in a sequential order. The new method offers more accuracy dis-similarity measure for each cluster data without manual intervention at the time of duplicate deduction. If we apply the propose deduplication algorithm with this new method, surely it will reduce the total time consumption as well as avoid the unwanted record comparison.

The framework is mainly developed to increase the speed of the duplicate data detection and elimination process and to increase the quality of the data by identifying true duplicates and strict enough to keep out false-positives.

REFERENCES:

1. Lavanya Pamulaparty, Dr. C.V Guru Rao, Dr. M. Sreenivasa Rao (2014), "A Near-Duplicate Detection Algorithm To Facilitate Document Clustering", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.4, No.6, PP: 39-49
2. Bilal Khan, Azhar Rauf, Huma Javed, Shah Khusro, and Huma Javed (2012), "Removing Fully and Partially Duplicated Records through K-Means Clustering", IACSIT International Journal of Engineering and Technology, Vol. 4, No. 6, PP: 750-755
3. Lohman, G., J. Champlin, and P. Sohn (2005), "Quickly Finding Known Software Problems via Automated Symptom Matching", In Proceedings of the Second International Conference on Automatic Computing (ICAC '05). IEEE Computer Society, Washington, DC, USA, 101-110
4. Manikandan.G, Rajendiran.P, Kamarasan.M, SowndaryaShekar (2011), "Performance Analysis of Clustering Algorithms in Outlier Detection Based on Statistical Models and Spatial Proximity", International Journal of Computer Science and Information Technologies, ISSN: 0975-9646, Vol. 2 (4) , PP: 1747-1749
5. Zeng, J., L. Gong, Q. Wang and C. Wu, 2009. Hierarchical clustering for topic analysis based on variable feature selection. Proceedings of the 6th International Conference on Fuzzy Systems and Knowledge Discovery, Aug. 14-16, EEE Xplore Press, Tianjin, pp: 477-481.
6. Marrakchi, Z., H. Mrabet and H. Mehrez, 2005. Hierarchical FPGA clustering based on multilevel partitioning approach to improve routability and reduce power dissipation. Proceedings of the International Conference on Reconfigurable Computing and FPGAs, Sept. 28-30, IEEE Xplore Press, Puebla City, pp: 25-28.

Vol 11, Issue 3, MARCH/2020
ISSN NO:0377-9254

7. Velmurugan, T. and T. Santhanam, 2011. A survey of partition based clustering algorithms in data mining: An experimental approach. Inform. Technol. J., 10: 478-484.
8. R. Arora, P. Pahwa, and S. Bansal, "Alliance Rules for Data Warehouse Cleansing,"in Proceeding of International Conference of Signal Processing Systems, IEEE, 2009.
9. T. E. Ohanekwu and C. I. Ezeife, "A Token-Based Data Cleaning Technique for Data Warehouse System," University of Windsor, in Proceeding of 9th International Conference on Database Theory ICDT, Siena, Italy 2003.
10. A. K. Elmagrmid, P. G. Ipeirotis, and V. S. Verykois, "Duplicated Record Detection: A Survey," IEEE Transaction on Knowledge and Data Engineering, vol. 19, no. 1, IEEE, 2007.