ISSN: 2320-1363

Evolution of Record linkage Using Proposition Logic and Generational Evolutionary Algorithm

P V KishoreKumar *1, Battula LakshmanaRao *2, B Venkatreddy *3

M.Tech student, KITS ENGINEERING COLLEGE, East Godavari(dt),AP, India

Assistant professor KITS ENGINEERING COLLEGE, East Godavari (dt), AP, India

Associate professor KITS ENGINEERING COLLEGE, East Godavari (dt), AP, India

ABSTRACT

In multiple databases redundant and noisy records problems are available. Data cleaning is major approach for removing the noisy and redundant data. All databases convert as a quality databases. Previously some approaches are detects the duplicates. Those approaches are statistical record linkage and decision tree [1][2][3]. These all approaches are require more number of training phase and testing phases for removing the duplicates. It's not gives the accurate duplicate detection. This approach utilizes more amount of operational cost.

In this paper we introduce new approaches and possible to reduces the operational cost also. These new approaches are gives the accurate solution in extraction of records from multiple databases. First record linkage related statistical approach removes the duplicates using pair matching of records. Pair's matching approach show the result as unique records content. Here we use the less number of training phases approach. These unique records filter as a truth records or guaranteed unique records we find out using propositional logic. Its reduces the irrelevant records and show the result as a minimized records content. Combine of all records show it's as fitness records using generational evolutionary approach. These results are not fixed, every time adding the new records and show the results as a refinement manner. These are good performance records [3][4].

KEYWORDS: Data mining, Propositional Logic, Record Linkage, evolutionary approach.

I.INTRODUCTION

All government organizations are maintaining the different number of web databases. In present web databases any user forward the query large amount of data is displayed here. Same data it may chance to present in two or more number of databases. All researchers are interested in this field. Same real world entities detection is the major approach in multiple databases. This complete detection is possible with the help of record linkage. Record linkage is one good process technique in detection of duplicates.

Present record matching techniques apply into different kinds of query results. Those query results are positive, negative and both. In all types of query results for removing the duplicates it's very complex. It takes more amount of training cost. After some days some new predefined rules are implemented for extracting the high quality databases maintainace. Its gives partial optimal solution results with high operational cost.

To overcome such problems, issues and risks propose the new method that is called proposition logic [7][8]. This is one of the good unsupervised classification techniques. Here we perform the classification two or more number of times in detection of duplicates. It's gives the truth records. Those results are minimized results with minimized cost. All minimized results we arranges with generational evolutionary approach. Its gives good fitness and effective results records with less operational cost.

II. RELATED WORK

This complete paper related to data integration and data cleaning. Resolve the distributed databases problems and create quality databases environment. Creation of quality databases using many number data mining many approaches. Every previous approach some limitations and risks are present here. Every approach itself calculates the overhead and operational cost [1][2][3][4].

First and starting time of creation databases there is no importance to constraints. Whenever constraints are missing data entry problems are generated. Data entry problems are gives the ambiguity solutions. It's not possible to detect all duplicates. It's show less amount of statistical performance in duplicate detection [1]. These problems we observed in many number of real time applications. All real times applications identify the problems like high amount of operational cost utilization in detection of all duplicates.

After some number days next approach to detects the duplicates in distributed databases. That new approach is called as a merge list. Merge list we apply the approximate query processing environment. Compare to previous [2] approach this present approach gives the best results. It's not gives the best performance in detection of duplicates.

Next approach is called N-grams. It is completely incremental string environment approach. After enter one string it's not detects all duplicates, in first

ISSN: 2320-1363

string add another string. Using two strings start the detection of duplicates [3]. This is process we continue until to detect all duplicates. In this approach we spend more amounts of operational cost utilization and overhead problem.

After some number of days for reducing the operational cost we introduce new format that is called as a distance based query. Distance based query also related to approximate query processing and remove the duplicates. No need to detect all duplicates in all distances. User required area itself start the detection of duplicates. User gets the less features, he will not satisfy. Distance based metrics related duplicates it's not gives the better results [3][4][2].

Next approach related to probabilistic with statistical approach and naive Bayesian with statistical approach. These two approaches works based on similarity functions. Every record of similar records we find out here. Count the similar records and allocates as a weights. In all types of weights apply the threshold and categorizes the records. Below threshold records considers as a non duplicate and above threshold information consider as a duplicate records. Using these two approaches reduces the some operational cost [2][3][4][1].

Next approach is called as a rule based prediction. This rule based prediction also works based similarity record weight. Using different kinds of rules different clusters we generate here. In total number of clusters which cluster is best cluster no user is not identify the effectively. This is also one good information retrieval techniques under detection of duplicates. Its level by level approach in detection of duplication. Here also we spend the more amount of operational cost here [6]. The above all approaches are related supervised clustering. These all approaches are training approaches.

After completion of training approaches we choose the different number of testing approaches. Combine of training and testing approaches are called unsupervised clustering. Here in training approach only detection of all duplicates we spend more amount of operational cost. Second time classification detects the number duplicates and gives the results high duplicate ratio. Unsupervised is the one good learning approaches in detection of duplicates here.

The above all problems and limitations we overcome in new approach. That new approach we discuss in following contents.

III.PROBLEM STATEMENT

Previously decision trees are used. Decision tree starts align records based on conditions or constraints. This is one of the good classification approach. Classification is works as a training approach. Many times perform the classifications approach in present record linkage with decision trees. These approaches are requires more number of training phases. In classification time we spend more amount of operational cost.

In this paper we propose new statistical record linkage with new generational evolutionary approach. Using these approaches reduces the number of training phases and automatically reduces the operational cost. Using normal query process extracts the records. In these displayed records some records are available as a duplicate. We remove duplicate records using record linkage technique. Unique records we align with the help of different kinds of prepositions in generational evolutionary algorithm. Proposition related results show as a better and effectiveness results.

IV.SYSTEM MODEL

4.1 OUTLINE OF RECORD LINKAGE:

User forwards the query for retrieving of results here. Results are extracted from multiple databases. It may chance same results are present in two or more number of databases. Those results are duplicate results or records. Using cleaning approaches remove the duplicate records of content. All records we display as a unique records. We align all unique records as a index format. Index based records are aligns using similarity function. Similarity function starts the calculation of each and every record weight based on matching environment process. It is one of the evolutionary processes. It's not gives the proper alignment result, no user is not satisfied with results.

Similarity function related approach completely statistical approach. In statistical approach applies the probability related mechanism.

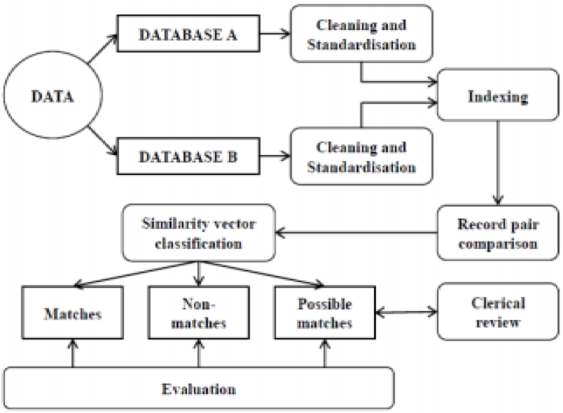

Fig4. 1: Outline related to record linkage

In this present record linkage whenever to start the calculation of statistical performance or statistical weight, there is no communication in between of one record to another record. This is called as a correlation problem. All problems and

ISSN: 2320-1363

risks we overcome using proposition related generation evolutionary approach.

4.2 proposition related generational evolutionary approach

Using similarity function we identify the unique records weight. In similarity record weight it may chance to present the duplicates, no author is not gave the guarantee there is no duplicates. We give guarantee to remove the maximized duplicates using proposition logic. Proposition logic gives truth records in alignment with understandable content. It's show high duplicate ratio in detection of duplicate records.

These kinds of truth results we show with the help of different experiments. These experiments give the good proper alignment and performance results.

V. IMPLEMENTATION APPROACH WITH PREPOSITION LOGIC

We use the proposition logic for mapping of records. Using propositional logic aligns the records with fewer amounts of operational cost and time consuming. Propositional logic works as good inference rule logic. Using inference rules formulate the records with good relationship. Relationship of records creates as a structure.

Relationship generate with the help of different number of symbols and operators. Operators are use in between two or more number of records. This is called one verification approach in alignment of records. Verification approach gives the result as true or false information. This is one of the good mathematical component procedures in relationship of records. We increase the efficiency in record alignment.

5.1 Basic concepts related to proposition Logic in record alignment

In basic concepts we follow some steps. Those steps are

1. Collection of similar records with good primitive symbols and operators.

2. Align the records with good meaningful representation

3. Apply the different inference rules in implementation process.

5.2 Identification of Truth records with Different Inference Rules

5.2.1 Apply the conjunction in between of Different Records

In between of two or more number of records apply the conjunction operation identifies the relationship records. These type results are deeper results and display highest features of content or vectors of content. Deep web data extraction is

ISSN: 2320-1363

possible with the help of "AND" operation. It does not allow any false records information. Its show the minimization records in alignment. That's why we spend the less amount of operational cost in implementation process.

5.2.2 Apply the disjunction in between of Different Records

In between of two or more number of records apply the disjunction operation. Disjunction operation performs based on "OR". These operation related results also we discard in implementation of results extraction process. These types of conditions related results we reduce in extraction. Its gives the solution with less operational cost specification process.

VI.TEST ACCURACY IN DUPLICATE DETECTION

Record linkage and proposition logic reduces the number of duplicates records and provides the truth records. Compare to previous all approaches its show the good results. Those test accuracy operations are called Precision, Recall and F-measure.

Precision= Number of Correctly Identified Duplicate Pairs / Number of Identified Duplicate pairs

Recall= Number of Correctly Identified Duplicate Pairs / Number of True Duplicate Pairs

F-Measure = 2*P*R / P+R

These three operations are provides the good metric results in maximization of duplicate detection. It's providing the meaningful in alignment with different proposition logic operations. It's give high accuracy with less amount of operational cost.

VII. PERFORMANCE EVOLUTION

Different approaches related duplicate detection works based on operational cost. Compare to all previous concepts, present concepts show the less amount operation cost and overhead results. Those results show into output with help of experiment.

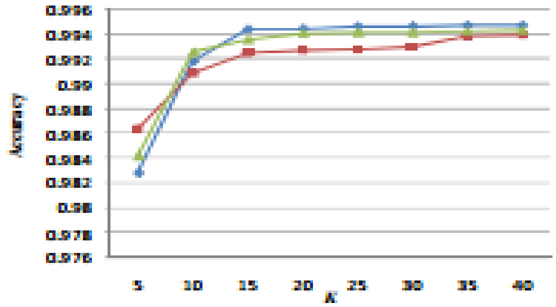

Fig2: Test Accuracy evolution with operational cost

In X-axis consider the operational cost and y-axis considers accuracy. Here in graph three iterations are present here. Every iteration automatically increases the accuracy. Those performance levels store in each and every line.

VIII. CONCLUSION

We test different approaches in record linkage environment. We have done the analysis of operational cost with different approaches. Previous all approaches utilizes the more amount of operational cost. Now in this paper we discuss about different approaches like propositional logic and generational evolutionary approach. Its gives good accuracy duplicates detection with less amount of operational cost. It's show the good optimal solution.

IX. REFERENCE PAPERS

[1] Duplicate Record Detection: A Survey, Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios, 2006

[2] An Iterative Two-Party Protocol for Scalable Privacy-Preserving Record Linkage, 2012 Dinusha Vatsalan and Peter Christen

[3] A Comparison of Fast Blocking Methods for Record Linkage, Rohan Baxter, Peter Christen, Tim Churches, 2003

[4] Towards Duplicate Detection for Situation Awareness Based on Spatio-Temporal Relations, Norbert Baumgartner1, Wolfgang Gottesheim2, Stefan Mitsch2, Werner Retschitzegger2, and Wieland Schwinger2, 2010

[5] A Comparative Study of Duplicate Record Detection Techniques, Osama Helmi Akel, 2012

[6] Improving the Accuracy of Duplicate Detection at Scale, Aubrey Barnard and Allison Terrell,2009.

[7] A manual and introductory tutorial, David Martin, James Procter, Andrew Waterhouse, Saif Shehata and Geoff Barton, 2013

[8] Formal Veri_cation of Data Provenance Records, Szymon Klarman1, Stefan Schlobach1, Luciano Sera_ni2,2011

[9] Efficient Record Linkage in Large Data Sets, Liang Jin, Chen Li, and Sharad Mehrotra

[10] Efficient Private Record Linkage, Mohamed Yakout, Mikhail J. Atallah, Ahmed Elmagarmid, 2013