An Efficient Data Cleansing by Duplication Record Detection Algorithm

R Janardhan Naidu[1] Ms. I.Madhavi Latha[2]
[1]Student [2]Assistant Professor
[1,2]Department of Computer Applications
[1,2]KMM Institute of PG Studies, Tirupati, India

Abstract— Many industries and businesses have huge amount of data stored in different databases. In this fast world, it is necessary that data operations on the database are carried out smoothly and efficiently. However, to access the useful information that can help in decision making for industries and businesses, it is necessary to integrate large dataset. In existing system, DCS++ algorithm is used. It is very difficult to analyzed or understand. To improve this limitation and improve their performance we are use proposed system. In this proposed system, we are use Record Detection Algorithm. In record detection algorithms are classified into three types. They are knowledge based techniques, probabilistic techniques, and empirical techniques. Knowledge based algorithms demand training and the use of that training and reasoning skills in order to perform detection. Probabilistic algorithms are based on geometric and probability methods that are Bayesian networks, anticipation maximization and data clustering. Empirical algorithms consist on sorting, blocking and windowing methods. Here mainly blocking and windowing is used. By this accuracy are increases and efficiency and the performance of this system will be increased?

Key words: Duplication Records Detection Algorithm, DCS++, Windowing, Blocking

I. INTRODUCTION

Now-a-days, the digital economy is totally dependent on the databases. Many industries and businesses have huge amount of data stored in different databases. In this fast world, it is Necessary that data operations on the database are carried out smoothly and efficiently. However, to access the useful information that can help in decision making for industries and businesses, it is necessary to integrate large dataset. When data is integrated from different sources then it contains a huge part of dirty data. This dirty data contain mistakes in record values, duplication in records, spelling mistakes, null or illegal values, disobedience referential integrity and inconsistency in records. Quality assurance of data is necessary for fast retrieval of data, quick and smooth data processing, and right decision making. Business organizations are paying high attention towards data quality because dirty data can effect important decisions in businesses. In addition, cleansed data can improve the production because of quality decisions . Data cleansing is performed to get cleansed and quality data. Therefore, Data cleaning is important for business industry. The available data cleaning methods are not time and cost effective.

Duplication in data is one of the most important issues of Data cleaning. When data is gathered from different source then due to mistakes in spells or difference or inconsistency of format may cause presence of duplicate records in data. Extraction of knowledge from huge databases is known as data mining . Duplicate record deduction and data redundancy control are also hot topics of data mining and data integration With the increase of Quality data demand, many logical and statistical methods have been provided to resolve the problem. In this regard, there are three basic techniques of Duplicate records detection which are knowledge-based techniques, probabilistic techniques and empirical techniques. Many algorithms have been proposed under those techniques but all of them somehow lack in one of these parameters which are time efficiency, cost effectiveness, space consumption and accuracy . Duplication record detection is a very diverse field so this decision was made that one of its basic technique will be chosen and then focus will be on algorithms which lie within that technique. It was decided to select empirical technique and compared all the algorithms under this category. After comparison, most effective algorithm will be selected and improved accordingly. The objectives of this research study are as follows: To study the algorithms of duplication records detection. To perform comparative analysis of duplicate records detection algorithms lying under the empirical technique .To implement the best selected algorithm after performing comparative analysis .To suggest improvement in the selected algorithm.

II. RELATIVE STUDY

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, ten next steps are to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system.

A. "An Efficient way to find Frequent Pattern with,"

The purpose of this research was to review, analyze and compare algorithms lying under the empirical technique in order to suggest the most effective algorithm in terms of efficiency and accuracy. The research process was initiated by collecting the relevant research papers with the query of "duplication record detection" from IEEE database. After that, papers were categorized on the basis of different techniques proposed in the literature. In this research, the focus was made on empirical technique. The papers lying under this technique were further analyzed in order to come up with the algorithms. Finally, the comparison was performed in order to come up with the best algorithm i.e. DCS++. The selected algorithm was critically analyzed in order to improve its working. On the basis of limitations of

selected algorithm, variation in algorithm was proposed and validated by developed prototype.

B. "Efficient Exact Similarity Searches using Multiple,"

Similarity searches are essential in many applications including data cleaning and near duplicate detection. Many similarity search algorithms first generate candidate records, and then identify true matches among them. A major focus of those algorithms has been on how to reduce the number of candidate records in the early stage of similarity query processing. One of the most commonly used techniques to reduce the candidate size is the prefix filtering principle, which exploits the document frequency ordering of tokens. In this paper, we propose a novel partitioning technique that considers multiple token orderings based on token co-occurrence statistics. Experimental results show that the proposed technique is effective in reducing the number of candidate records and as a result improves the performance of existing algorithms significantly

C. "Scalable Iterative Graph Duplicate Detection,"

Duplicate detection determines different representations of real-world objects in a database. Recent research has considered the use of relationships among object representations to improve duplicate detection. In the general case where relationships form a graph, research has mainly focused on duplicate detection quality/effectiveness. Scalability has been neglected so far, even though it is crucial for large real-world duplicate detection tasks. We scale-up duplicate detection in graph data (DDG) to large amounts of data and pair wise comparisons, using the support of a relational database management system. To this end, we first present a framework that generalizes the DDG process. We then present algorithms to scale DDG in space (amount of data processed with bounded main memory) and in time. Finally, we extend our framework to allow batched and parallel DDG, thus further improving efficiency. Experiments on data of up to two orders of magnitude larger than data considered so far in DDG show that our methods achieve the goal of scaling DDG to large volumes of data.

D. "Adaptive Windows for Duplicate Detection,"

Duplicate detection is the task of identifying all groups of records within a data set that represent the same realworld entity, respectively. This task is difficult, because (i) representations might differ slightly, so some similarity measure must be defined to compare pairs of records and (ii) data sets might have a high volume making a pair-wise comparison of all records infeasible. To tackle the second problem, many algorithms have been suggested that partition the data set and compare all record pairs only within each partition. One well-known such approach is the Sorted Neighborhood Method (SNM), which sorts the data according to some key and then advances a window over the data comparing only records that appear within the same window. We propose with the Duplicate Count Strategy (DCS) a variation of SNM that uses a varying window size. It is based on the intuition that there might be regions of high similarity suggesting a larger window size and regions of lower similarity suggesting a smaller window size. Next to the basic variant of DCS, we also propose and thoroughly evaluate a variant called DCS++ which is provably better than the original SNM in terms of efficiency (same results with fewer comparisons).

E. "Approximate Record Matching Using Hash Grams,"

Data Linkage is an important step that can provide valuable insights for evidence-based decision making, especially for crucial events. Performing sensible queries across heterogeneous databases containing millions of records is a complex task that requires a complete understanding of each contributing database's schema to define the structure of its information. The key aim is to approximate the structure and content of the induced data into a concise synopsis in order to extract and link meaningful data-driven facts. We identify such problems as four major research issues in Data Linkage: associated costs in pairwise matching, record matching overheads, semantic flow of information restrictions, and single order classification limitations. In this paper, we give a literature review of research in Data Linkage. The purpose for this review is to establish a basic understanding of Data Linkage, and to discuss the background in the Data Linkage research domain. Particularly, we focus on the literature related to the recent advancements in Approximate Matching algorithms at Attribute Level and Structure Level. Their efficiency, functionality and limitations are critically analysed and open-ended problems have been exposed.

III. PROPOSED ALGORITHM

In this proposed system, we are use Record Detection Algorithm. In record detection algorithms are classified into three types. They are knowledge based techniques, probabilistic techniques, and empirical techniques. Knowledge based algorithms demand training and the use of that training and reasoning skills in order to perform detection. Probabilistic algorithms are based on geometric and probability methods that are Bayesian networks, anticipation maximization and data clustering. Empirical algorithms consist on sorting, blocking and windowing methods. Here mainly blocking and windowing is used. By this accuracy is increases and efficiency and the performance of this system will be increased.

IV. ALGORITHM

A. Duplicate record detection algorithm

Duplicate record detection algorithm is a data mining technique algorithm for predicting the duplicate records in the dataset. In this paper we are calculating the no of times that a particular word is repeated in the dataset and we can also search the particular word and view the no of times that particular word is repeated.

Duplication record detection is a very diverse field so this decision was made that one of its basic technique will be chosen and then focus will be on algorithms which lie within that technique. It was decided to select empirical technique and compared all the algorithms under this category. After comparison, most effective algorithm will be selected and improved accordingly. The objectives of this research study are as follows:

1) To study the algorithms of duplication records detection

2) To perform comparative analysis of duplicate records detection algorithms lying under the empirical technique
3) To implement the best selected algorithm after performing comparative analysis
4) To suggest improvement in the selected algorithm

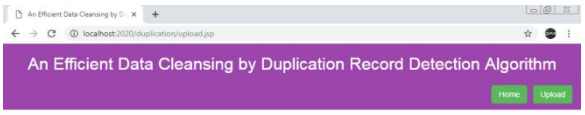
Duplicates are the records that represent the same real-world object or entries. Record matching is a state of art technique to find these duplicates. Duplicates can be of two types that are exact or mirror duplicates and approximate or near duplicates. Exact duplicate records contain the same content but on the other hand content of near duplicate records vary slightly. The records which contain syntax differences or typographical errors but represent the same real world entity are known as near duplicates. Duplication detection is used to identify the same real world entities which exist in different format or representation in database. It is very common to find some non-identical fields or records that refer the same entity. Efficient and accurate detection of duplicates is hotspot of the data mining and online analyzer

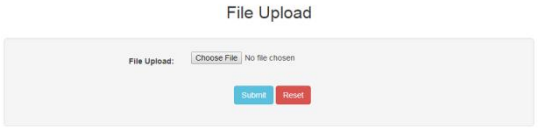
Now-a-day, duplication detection is the most popular topic in research. Duplication detection is based on two basic Stages. The first one is the outer stage in which record matching technique or duplication record matching technique is applied. The second one is the inner stage that is based on field matching techniques. Duplication record detection algorithms are divided in three type's i.e. Knowledge- based techniques, probabilistic techniques, and empirical techniques. Empirical algorithms consist on sorting, blocking and windowing methods. Knowledge based algorithms demand training and the use of that training and reasoning skills in order to perform detection. Probabilistic algorithms are based on statistical and probability methods that are Bayesian networks, expectation maximization and data clustering. In this research study, focus is on empirical algorithms.

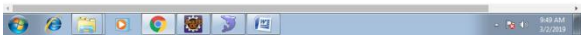
V. SCREEN SHOTS

A. *Home page:*

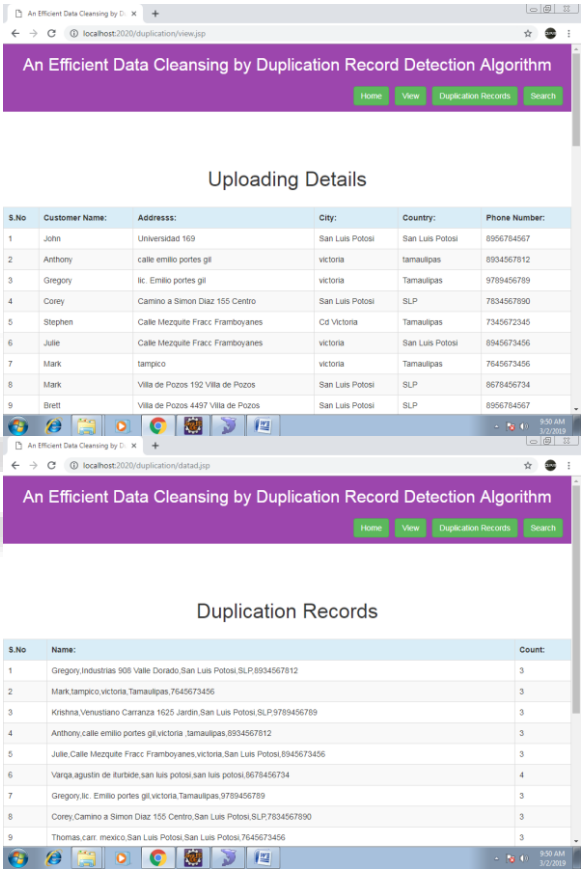
B. *File upload*



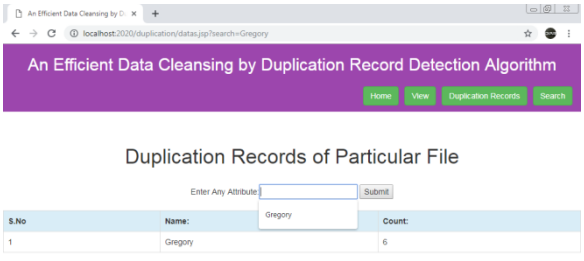




C. *Uploading details*



D. *Search*



VI. CONCLUSION

In proposed system Record Detection Algorithm is used. It is classified into three types they are knowledge based technique, probabilistic technique and empirical technique. Mainly here empirical technique is used. It is domain independent but input dependency is there. The algorithm provides almost similar results than of DCS++ in terms of accuracy excluding some cases where accuracy of proposed algorithm is higher. On the other hand, efficiency of proposed algorithm is equal or higher in some cases.

REFERENCES

[1] Ahmed K. Elmagarmid, P., G. Ipeirotis, and Vassilios S. Verykios, "Duplicate Record Detection: A Survey," IEEE Trans. on Knowl. and Data Eng., vol. 19, pp. 1-16, 2007.
[2] P. Ying, X. Jungang, C. Zhiwang, and S. Jian, "IKMC: An Improved K-Medoids Clustering Method for NearDuplicated Records Detection," in Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference on, Wuhan, 2009, pp. 1 - 4.
[3] M. Rehman and V. Esichaikul, "DUPLICATE RECORD DETECTION FOR DATABASE CLEANSING," in Machine Vision, 2009. ICMV '09. Second International Conference on , Dubai, 2009 , pp. 333 - 338.
[4] X. Mansheng, L. Youshi, and Z. Xiaoqi, "A PROPERTY OPTIMIZATION METHOD in SUPPORT of APPROXIMATELY DUPLICATED RECORDS DETECTING," in Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on, 2009.
[5] Q. Hua, M. Xiang, and F. Sun, "An Optimal Feature Selection Method for Approximately Duplicate Records," in Information Management and Engineering (ICIME), 2010 The 2nd IEEE International Conference on, Chengdu, 2010.
[6] D. Bhalodiya, M., K. Patel, and C. Patel, "An Efficient way to Find Frequent Pattern with," in Nirma University International Conference on Engineering, 2013.
[7] L. Huang, P. Yuan, and F. Chu, "Duplicate Records Cleansing with Length Filtering and Dynamic Weighting," in Semantics, Knowledge and Grid, 2008. SKG '08. Fourth International Conference on, Beijing, 2008, pp. 95 - 102.
[8] M. Gollapalli, X. Li, I. Wood, and G. Governatori, "Approximate Record Matching Using Hash Grams," in 11th IEEE International Conference on Data Mining Workshops, 2011.
[9] Z. Wei, W. Feng, and L. Peipei, "Research on Cleaning Inaccurate Data in Production," in Service Systems and Service Management (ICSSSM), 2012 9th International Conference on, Shanghai, 2012.
[10] L. Zhe and Z. Zhi-gang, "An Algorithm of Detection Duplicate Information Based on Segment," in International Conference on Computational Aspects of Social Networks, 2010.
[11] H., H. Shahri and Z., A., A. Barforush, "Data Mining for Removing Fuzzy Duplicates Using Fuzzy Inference," in Processing NAFIPS '04. IEEE Annual Meeting of the (Volume:1), 2004.
[12] W. Su, J. Wang, and H., F. Lochovsky, "Record Matching over Query Results from Multiple Web Databases," in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2010.
[13] R. Naseem, S. Anees, M., and S. Farook, "Near Duplicate Web Page Detection With Analytic Feature Weighting," in Third International Conference on Advances in Computing and Communications, 2013.
[14] L., Wan Zhao and Wah, C. N., "Scale-Rotation Invariant Pattern Entropy for Keypoint-Based Near-Duplicate Detection," in IEEE TRANSACTIONS ON IMAGE PROCESSING, 2009.
[15] G. Beskales, A., M. Soliman, F., I. Ilyas, S.i Ben-David, and Y. Kim, "ProbClean: A Probabilistic Duplicate Detection," in Data Engineering (ICDE), 2010 IEEE 26th International Conference on, 2010.
[16] J. Kim and H. Lee, "Efficient Exact Similarity Searches using Multiple," in IEEE 28th International Conference on Data Engineering, 2012.
[17] M. Ektefa, F. Sidi, H. Ibrahim, and M.,A. Jabar, "A Threshold-based Similarity Measure for Duplicate Detection," in Open Systems (ICOS), 2011 IEEE Conference on, Langkawi, 2011, pp. 37 - 41.
[18] M. Herschel, F. Naumann, S. Szott, and M. Taubert, "Scalable Iterative Graph Duplicate Detection," in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2012.
[19] Q. Kan, Y. Yang, S. Zhen, and W. Liu, "A Unified Record Linkage Strategy for Web Service," in Third International Conference on Knowledge Discovery and Data Mining, 2010.
[20] U. Draisbach and F. Naumann, "A Generalization of Blocking and Windowing Algorithms for Duplicate Detection," in Data and Knowledge Engineering (ICDKE), 2011 International Conference on , Milan, 2011, pp. 18 - 24.
[21] A. Bilke and F. Naumann, "Schema Matching using Duplicates," in Proceedings of the 21st International Conference on Data Engineering, 2005.
[22] Q. kan, Yan, Y. g, W. Liu, and X. Liu, "An Integrated Approach for Detecting Approximate Duplicate Records," in Second Asia-Pacific Conference on Computational Intelligence and Industrial Applications, 2009.
[23] U. Draisbach, F. Naumann, S. Szott, and O. Wonneberg, "Adaptive Windows for Duplicate Detection," in 28th International Conference on Data Engineering, 2012.

