International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 2, February- 2013

Optimizing GP Approach Using KFINDMRD Algorithm For Record Deduplication

[1]D.Saradha Priyadharshini, [2]Linda Joseph

Abstract - Deduplication is the key operation in data integration from multiple data sources. To achieve higher quality information removal of replicas within the repository is required**.** The accuracy in genetic programming is so less KFINDMRD(KFIND using Most Represented Data samples) is proposed to improve the accuracy of the classifier. The proposed system removes the duplicate dataset samples in the system and find the optimization solution to the deduplication of records or data samples in an efficient manner. In addition, the efficiency of the GP training phase is improved by giving the selected samples from the KFINDMRD algorithm. This deduplication process is carried out by the admin before storing the data within database so that the user can access data with good quality of information without the presence of duplicates. Secured login will be given to the users for accessing the data after the completion of the deduplication process. A performance evaluation for accuracy is done for the dataset with duplicates using Genetic programming and Genetic programming with KFINDMRD algorithm.

.

Keywords- deduplication, genetic programming, KFINDMRD,.

I. INTRODUCTION

During the time of integrating data from multiple heterogeneous sources, record replicas and duplicates will occur. In a data repository, a record that refers to the same real world entity or object is referred as duplicate records. Due to the duplicate records and dirty data, many problems will occur like performance degradation, quality loss and increasing operational costs [1]. The capacity of an organization is to provide useful services to its users is proportional to how well the data is handled by its systems. To keep repositories with "dirty" data i.e., data with replicas, with no Standardized representation, etc. Having duplicate records occupies more space and even increases the access time. Thus there is a need to eliminate duplicate records. This sounds to be simple but requires an tedious work since duplicate records do not share a common key and/or they contain errors that make duplicate matching a difficult task. Deduplication is one such tasks which fastens the search and help producing efficient results. The deduplication process can also be called as merge-purge[4] and instance identification[7] in the database community. More number of industries and systems depends the accuracy of databases to carry out operations. Dirty data—inaccurate, redundant, or outdated information—has been the bane of many organizations struggling to use information for higher performance. So, an error free database is required for reteriving high quality of information. The replica free not only allow the retrieval of higher quality information but also lead to more concise data and to potential savings in computational time and resources to process this data information but also lead to more concise data and to potential savings in computational time and resources to process this data. The error free database can be obtained by removing the dirty data present within it. Presence of dirty data leads to degrading the performance, constraints quality, increases operational cost. With the increase in size of the database the problem intensifies taking into account the huge amount of computational resource required for examination and removal of duplicate records. Duplicates can occur out of numerous scenarios, for instance when a large database is updated by an external source and registry numbers are not accessible.

[1] Department of Computer science and engineering , Hindustan university, Chennai.
E-mail: saradhapriyadharshini@gmail.com
[2]Assistant professor Department of . Computer science and engineering , Hindustan university, Chennai.

II.RELATED WORK

Moreover replica of documents is made for Optical Character Recognition (OCR) documents. This lead to inconsistencies among the data stored in repositories. To solve these issues, information from unstructured data is to be extracted and stored in databases with perfect structure. This enables user to obtain information retrieval with increased speed and accuracy. The common problems met are:

1) The existing structured databases of entities are organized very differently from labeled unstructured text.

2)There is significant format variation in the names of entities in the database and the unstructured text.

3)In most cases the database will be large whereas labeled text data will be small. Features designed from the databases should be efficient to apply and should not dominate features that capture contextual words and positional information from the limited labeled data.

To address these issues, the data integration system [3] is designed. This system uses Semi-Markov models for extracting information from structured data and labeled unstructured data in There arise so many problems when data collected from different sources are to be used since these data uses different styles and standards. Moreover replica of documents is made for Optical Character Recognition (OCR) documents. This lead to inconsistencies among the data stored in repositories. The problem becomes more complicated when

a user needs to obtain user-specified information from huge volume of data stored in large databases like repositories. To solve these issues, information from unstructured data is to be extracted and stored in databases with perfect structure. This enables user to obtain information retrieval with increased speed and accuracy. The common problems met are:

1)The existing spite of their format, structure and size variations.The former method is enhanced by a semi automatic extraction method using DEG [6].

It follows the following three steps.

1) To gather the necessary knowledge and then transform them into useable form. The knowledge can be obtained from any source such as encyclopedia, a traditional relational database, a general ontology like Mikrokosmos ([Mik]), etc. This needs to handle data in different formats.

2) Automatically generate an initial data-extraction ontology based on the acquired knowledge and sample target documents. Gathered knowledge is transformed into Extensible Markup Language. (XML) format and various documents are combined to produce a high level schema. This schema defines the set of attributes that may appear in generated data extraction ontology.

3) Finally user validates the initial data extraction ontology which is generated using set of validation documents. If the result is not satisfactory, user applies Ontolog Editor to the generated ontology. The Ontolog Editor provides a method of editing an Object Relationship Model (ORM) and its associated data frames and also provides debugging functionality for editing regular expressions in data frames by displaying sample text with highlighting on sample source documents .Even though Database Enhancement Gateway (DEG)method is efficient, it requires human validation. Thus a fully automatic method is proposed which uses tag path clustering[2]. Usually the list of objects is extracted from databases using pair wise similarity match. But this pair wise similarity match did not address the nested data structures or more complicated structure. Hence the tag path clustering focuses on how a distinct tag path (*i.e.*, a path from the root to a leaf in the DOM tree) appears repeatedly in the document. The occurrence of a pair of tag path patterns (called visual signals) is compared to estimate how likely these two tag paths represent the same list of objects. Comparison is done using a similarity measure which uses a similarity function which captures how likely two visual signals belong to the same data region. There are still various advanced fully automatic methods to extract information from structured and unstructured data.

III. MODELLING DEDUPLICATION AND ITS ANALYSIS

A. Indexing Based Record Deduplication and Record Linkage

Peter Christen [6] surveyed various indexing techniques for record linkage and dedupliaction. Record linkage refers to the task of identifying records in a data set that refers to the same entity across different data sources[10]. Blocking technique is used in traditional record linkage approach. Blocking key values are used to place the records into different blocks. According to this BKV, the matched records are placed in same block and non matching records into different blocks. The record linkage process has divided into two phases: Build and Retrieve. In build phase, at the time of linking two data bases, a new data structure is formed: i) Separate index data structures ii) Single data structures with common key values. The hash table data structure is also used for indexing. In retrieve phase, the retrieval of records from block and it will be paired with other records which having same index value. This resulting vector given to classification steps. There are many indexing techniques available. These techniques are mainly used to reduce the number comparison between the records. This can be achieved by removing non matching pairs from the block
[6].

B. Distance-Based Techniques

One way to avoid training data is to introduce a distance metric for records which does not need tuning through training data. Without the need of training data with help of distance metric and an appropriate matching threshold, it is possible to match similar records without the need for training. Here each record is considered as field where the distance between individual fields are measured, using the appropriate distance metric for each field, and then the weighted distance between the records are computed. But the computation part of weighted distance moves bit probabilistic and difficult.

C. Active Learning Based Deduplication

Sunita Sarawagi and Anuradha Bhamidipaty [8] proposed an interactive learning based deduplication system called Active Learning led Interactive Alias Suppression (ALIAS). This technique automatically constructs the deduplication function by interactively finding the challenging training pairs. An active learner actively picks the subset of instances. It easing the deduplication task by limiting the manual effort for inputting simple, domain specific attributes similarity functions. It interactively labeling a small number of record pairs. First they took the small subset of pair of records. Then they find the similarity between records and this initial set of labeled data creates the training data for the preliminary classifier. To improve the accuracy of classifier they selected only n instances from the pool of unlabeled data [8]. They conclude that, active learning process is practical effective and provide interactive response to the user. It is easy to interpret and efficient to apply on large datasets. Active learning requires some training data but in some real world problems the training data are not available, so active learning technique is not suitable for all the problems .

D. Unsupervised Duplicate Detection (UDD) Algorithm

Weifeng Su et al. [9] proposed an unsupervised,online record matching method called Unsupervised Duplicate Detection (UDD) algorithm. There are two classifiers in UDD for iteratively identify the duplicate records. The duplicate records from the same source are removed using the exact matching method. In this method relative distance of each field of the records are calculated and according to this value,

field's weight will be assigned. After that Weighted Component Similarity Summing(WCSS) Classifier utilizes this weight set for matching the records from various data sources. It places the duplicate records in the positive set and non duplicate records in the negative set. The SVM classifier again identifies duplicates from the positive set. These two classifiers iteratively working together and identify the duplicate records in efficient manner. The iteration stops when new duplicates
cannot be found. This algorithm mainly used in the web databases because UDD does not require human labeled training data from the user. So it solves the online duplicate detection problem where the query results are generated on the-fly [9].In SVM based record deduplication only the concrete implementation has been done. However sometimes it an initial approximated training set to assign weight.

E. Genetic Programming

Several researches havebeen done in field of deduplication. Recently,deduplication in distributed manner has fascinated lots of researchers due to the demand of scalability and efficiency. Here, we reviewed the recently done works in the literature for deduplication and the different approaches used for it. Moises *et al*. [5] have proposed a genetic programming based approach to record deduplication that prepares a deduplication function to identify the replicate pairs on the basis of evidence extracted from the data content. They have also shown experimentally that their approach better than existing state of-the-art method which has been proposed in literature. Moreover, the devised functions take less time computationally because they used less evidence. In addition, the genetic programming approach was capable of automatically limiting these functions to a given fixed replica identification boundary, which frees the user from the load of choice and tunning this parameter.

IV. MODELLING PROPOSED SYSTEM

The proposed system of new algorithm KFINDMR (KFIND using Most Represented data samples) calculates the mean value of the most represented data samples in centroid of the record members; it selects the first most represented data sample that closest to the mean value calculates the minimum distance. The system Remove the duplicate dataset samples in the system which is less than the mean value and obtain new dataset samples, calculates the centroid of the dataset. It selects the data samples whose value is the most similar to the $\boldsymbol{C}p-1$ as the second most represented sample training sample. Repeat the steps until required number of most represented data sample is selected. This Proposed algorithm is used along with the genetic programming to improve the accuracy . We can easily find the deduplication records.

The steps involved in the proposed algorithm are:

1. Compute $\boldsymbol{C}_p=(1/p)\ \sum_{pi=1}\mathbf{e}_i$, i.e., the centroid of the record members $\boldsymbol{S}=\boldsymbol{S}=\{\mathbf{e}_i\}_{pi=1.}$
2. Select a first most represented sample that corresponds to the sample is closest to $\boldsymbol{cp}$ using $\mathbf{t}1 = \text{argmin}\ j\{Dist(\mathbf{r}j, \boldsymbol{C}p)\}$
3. For each of the end members in the member set $\boldsymbol{S}$ do:

4.1 Remove from $\boldsymbol{S}$ the members which is less similar to $\boldsymbol{C}p$,thus obtaining a new member set $\{\mathbf{e}i\}_{p-1\ i=1}$

4.2 Calculate the centroid of the set $\{\mathbf{e}i\}_{p-1\ i=1}\ \boldsymbol{C}_{p-1}(1/p-1)\ \sum_{p-1i=1}\mathbf{e}_i$.

4.3 Select the data sample whose value is the most similar to $\boldsymbol{C}p-1$ as the second most represented sample.

4.4 repeat from step 3 until the duplicate samples or records removed.

IV. EXPERIMENTAL RESULTS

In our experiments, a real data sets commonly employed for evaluating record deduplication approaches[1], which are based on real data gathered from the web. This section, we present and discuss the results of the Experiments performed to evaluate our proposed algorithm to record deduplication. The California restaurant dataset are used to found the duplicate records.

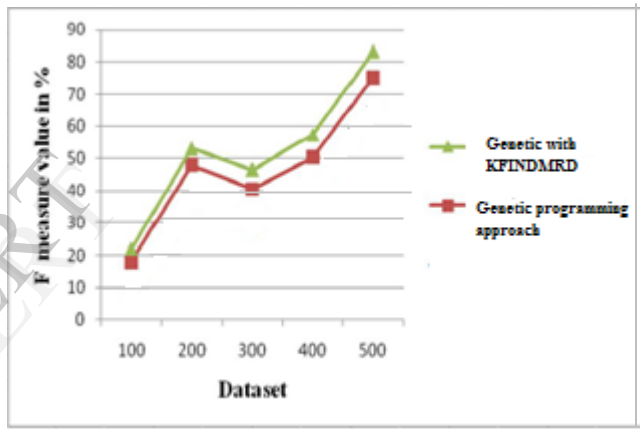

Figure:1 F MEASURE COMPARSION

In this Figure 1 shows that the F measure Comparison of the system between Genetic programming approach, Genetic programming approach for KFINDMR most relevant sample selection. We measure the F measure value in % at Y-axis as algorithm and consider the Cora dataset in the X-axis. The F measure value of the genetic KFINDMR is higher than the GP Finally the proposed new algorithm along with gp achieves the higher level of the F measure value than the Genetic programming approach.

V. CONCLUSION

Identifying and handling replicas is considered to be an important problem since it guarantees the quality of the information made available by data intensive systems. These systems rely on consistent data to offer high-quality services and may be affected by the existence of duplicates, quasi replicas, or near-duplicate entries in their repositories. So far various methods for deduplication are explained and the advantages and disadvantages of these techniques are discussed. The proposed algorithm finds the best optimization

solution to deduplication of the records. In addition, we intend to improve the efficiency of the GP training phase by selecting the most representative examples for training. Our proposed new algorithm selects the most represented data samples to improve the accuracy in finding duplicate records. This deduplication process is applied to the admin side in a distributed network. so that the user can access high quality of information without the presence of duplicates within it.

REFERENCES

[1] M. Bilenko and R.J. Mooney, "Adaptive Duplicate Detection Using Learnable String Similarity Measures," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 39-48, 2003.

[2] Gengxin Miao1 Junichi Tatemura2 Wang-Pin Hsiung2 Arsany Sawires2 Louise E. Moser11 ECE Dept., University of California,Santa Barbara, Santa Barbara, CA, 93106 2 NEC Laboratories America, 10080 N. Wolfe Rd SW3-350, Cupertino, CA, 95014, "Extracting Data Records from the Web Using Tag Path Clustering".

[3] Imran R. Mansuriimran@it.iitb.ac.in IIT Bombay ,Sunita Sarawagi sunita@it.iitb.ac.in IIT Bombay, "Integrating unstructured data into relational databases.

[4] Mauricio Antonio Hern´andez and Salvatore Joseph Stolfo. Real-world data is dirty: Data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery, 2(1):9.37, January 1998.

[5] Moises, G., D. Carvalho, H.F.A. Laender, M.A. Goncalves and A.S.D. Silva, 2011. A genetic programming approach to record deduplication. IEEE Trans. Knowl. Data Eng., 24: 399-412.2010.

[6] Peter Christen, "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication", *IEEE Trans. Knowledge and Data Eng.*, vol. 24, no. 9, pp. 1537-1555, Sept.2012.

[7] Y. Richard Wang and Stuart E. Madnick. The inter-database instance identi_cation problem in integrating autonomous systems. In Proceedings of the Fifth IEEE International Conference on Data Engineering (ICDE 1989), pages 46.55,1989

[8] Sunita Sarawagi and Anuradha Bhamidipaty. Interactive deduplication using active learning. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002), pages 269.278, 2002.

[9] Weifeng su, Jiying Wang, Frederick H. Lochovsky, " Record Matching over Query Results from Multiple Web Databases", *IEEE Trans. Knowledge and Data Eng.*, vol. 22, no. 4, pp.578-588, April 2010).

[10] Yihong Ding, A Thesis Proposal Presented to the Department of Computer Science Brigham Young University, "Semiautomatic Generation of Data-Extraction Ontologies", July 3, 2001.

D.Saradha priyadharshini was born in Tuticorin. She received the B.E degree in Computer science and Engineering in Sri Shakthi Institute of Engineering and Technology, Coimbatore, in 2010, and currently pursuing M.Tech degree in Software Engineering in Hindustan University, Chennai. Her research interests are in the area of data mining and software engineering.

Mrs. Linda Joseph was working as a assistant professor in the Department of computer science and Engineering, Hindustan University, Chennai. She published several papers in journels and conferences. Her research interest are in the area of network security and data mining.