An Improving Genetic Programming Approach Based Deduplication Using KFINDMR

P.Shanmugavadivu[#1], N.Baskar[*2]

[#] *Department of Computer Engineering, Bharathiar University*
Sri Ramakrishna Polytechnic College, Coimbatore-641 022, India

[*] *Department of Computer Science, Bharathiar University*
Sri Ramakrishna Mission Vidyalaya CAS, Coimbatore – 641 020, India

***Abstract*—The record deduplication is the task of identifying, in a data repository, records that refer to the same real world entity or object in spite of misspelling words, types, different writing styles or even different schema representations or data types. In existing system aims at providing Unsupervised Duplication Detection (UDD) method which can be used to identify and remove the duplicate records from different data sources. Starting from the non duplicate set, the two cooperating classifiers, a Weighted Component Similarity Summing Classifier (WCSS) and Support Vector Machine (SVM) are used to iteratively identify the duplicate records from the non duplicate record and present a genetic programming (GP) approach to record deduplication. Their GP-based approach is also able to automatically find effective deduplication functions. The genetic programming approach is time consuming task so we propose new algorithm KFINDMR (KFIND using Most Represented data samples) to find the most represented data samples to improve the accuracy of the classifier. The proposed system 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 and find the optimization solution to deduplication of records or data samples.**

***Keywords*—Extracting data, identifying duplication, deduplication, genetic programming.**

I.INTRODUCTION

Finding duplicate records in those records collected from several data sets are increasingly important tasks. Data linkage and deduplication can be used to improve data quality and integrity, which helps to re-use of existing data sources for new studies, and to reduce costs and efforts in obtaining data. Traditional methods for collecting duplicate records are time consuming and expensive survey methods.

The capacity of an organization 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., questions the overall speed or performance of data management systems. The existence of “dirty” data in the repositories leads to potential problems.

They are:

- degrading the performance
- constraints quality
- increases operational cost

These problems can be avoided by removing “dirty data “from the data source. The dirty data is the data

With replicas, with no standardized representation, etc. It requires technical efforts to manage them. By

Avoiding them, the overall speed and performance will be increased.

Using deduplication has two big advantages over a normal file system:

- Reduced Storage Allocation - Deduplication can reduce storage needs by up to 90%-95% for files such Virtual Machine Disk (VMDK) and backups. Basically situations where you are storing a lot of redundant data can see huge benefits.
- Efficient Volume Replication - Since only unique data is written disk, only those blocks need to be replicated. This can reduce traffic for replicating data by 90%-95% depending on the application.

The data mining techniques can applied when the data is available in proper format. To obtain this, information from various sources and repositories is to be structured. Many of the available web data are in unstructured form. This unstructured information cannot be omitted because it contains valuable information. Thus this data is to be integrated to structured database to enable mining activities. This can be done using high performing models, Conditional Random Fields, semi-markov models and matching [3]**.**

The web data can be queried like databases by either actually or virtually extracting information in web pages using wrappers. Once information is extracted, it can be queried like standard query languages. But the problem with the wrappers is, it has very much dependency on the structure of the source documents. Although the documents contain the same or similar information, it is therefore difficult to apply wrapper dependent source documents to documents with different formats. Thus an alternative semi automatic approach Data Extraction Group (DEG) [9] is proposed based on ontology relationships.

Information extraction on the structured data is done using several methodologies which are fully automatic. First comes the Multifactor Dimensionality Reduction (MDR) approach [5] which uses the edit distance between data segments (called generalized nodes). It traverses the Document Object Model (DOM) tree of data in pre-order. As it does not address the nested data, NET approach is proposed [6]**.** This traverses in bottom up manner and hence it is time consuming process since each traversal requires a full scan till root. Top down scan can be stopped at a point when the required information is reached thus avoiding full scan. Both the methods use pair wise similarity matching which fails when the structure is too complicated.

An alternative method to eliminate the pair wise matching is to use tag path clustering [2]. This new method for record extraction that captures a list of objects in a more robust way based on a holistic analysis of a Web page. The method focuses on how a distinct tag path appears repeatedly in the Document Object Model (DOM) tree of the Web document.

Deduplication [9] is a key operation in integrating data from multiple sources. The main challenge in this task is designing a function that can resolve when a pair of records refers to the same entity in spite of various data inconsistencies. Most existing systems use hand-coded functions. One way to overcome the tedium of hand-coding is to train a classier to distinguish between duplicates and non-duplicates. The success of this method critically hinges on being able to provide a covering and challenging set of training pairs that bring out the subtlety of the deduplication function. This is non-trivial because it requires manually searching for various data inconsistencies between any two records spread apart in large lists.

Then to overcome this kind of disadvantage, a learning-based deduplication system that uses a novel method of interactively discovering challenging training pairs using a method called Active Learning came into existence**[**8], [1]. The Active Learning is done on real-life datasets which shows significantly reduced number of instances needed to be achieved for high accuracy. Even active learning techniques require some training data or some human effort to create the matching models. In the absence of such training data or the ability to get human input, supervised and active learning techniques are not appropriate. One way of avoiding the need for training data is to define a distance metric [1] for records which does not need tuning through training data. Using the distance metric and an appropriate matching threshold, it is possible to match similar records without the need for training.

Other several techniques exist like creating chunks on backups for deduplication process [4]. The deduplication module partitions a file into chunks, generates the respective summary information, which we call a fingerprint, and looks up Fingerprint Table to determine if the respective chunk already exists. If it does not exist, the fingerprint value is inserted into Fingerprint Table. Chunking and fingerprint management is the key technical constituents which governs the overall deduplication performance. Recent record deduplication approach deals with combining several different pieces of evidence extracted from the data content to produce a deduplication function that is able to identify whether two or more entries in a repository are replicas or not [7].This approach is known as genetic programming (GP) approach , which finds a proper combination of the best pieces of evidence, thus yielding a deduplication function from a small representative portion, which is then applied to rest of the repository area. Genetic programming approach combines several different pieces of evidence extracted from the data content to produce a deduplication function that is able to identify whether two or more entries in a repository are replicas or not. Since record deduplication is a time consuming task even for small repositories, their aim is to foster a method that finds a proper combination of the best pieces of evidence, thus yielding a deduplication function that maximizes performance using a small representative portion of the corresponding data for training purposes.

Our proposed system has a new algorithm for record deduplication. The proposed system 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 .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. The proposed system repeats these steps until found the most represented data samples is selected. Finally the deduplicated record is removed. we apply our proposed algorithm to genetic programming approach. 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 and find the duplicate records.

II. RELATED WORK

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 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 spite of their format, structure and size variations.

The former method is enhanced by a semi automatic extraction method using DEG [9].
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 XML 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 OntologEditor to the generated ontology. The OntologEditor 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.

In the following section different methods for deduplication are discussed. Once the data are extracted effectively, there is a need to store them in perfect format. Deduplication is the task of identifying, in a data repository, records that refer to the same real world entity or object in spite of misspelling words, typos, different writing styles or even different schema representations or data types. Hence deduplication technique is applied to extracted data which contains valuable information. This makes them resistant to inconsistencies. Various deduplication methods and the values and drawbacks are discussed in this paper.

III. MODELLING DEDUPLICATION AND ITS ANALYSIS

Data Cleaning is a time consuming process because of its lengthy activities. Since data preparation is done from multiple sources, there precedes data redundancies which brings problem in data storage capacity, processing capacity and also manual vagueness to maintainability. Deduplication is a specialized data compression technique for eliminating coarse-grained redundant data. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent across a link. There are multiple techniques for improving the efficiency and scalability of approximate duplicate detection algorithms.

A. Active-Learning Techniques

The main task that must be carried out is to project a function that must be able to resolve when a pair of records refers to the same entity in spite of various data inconsistencies. The earlier function to resolve was hand coded function where requires manually searching for various data inconsistencies between any two records spread apart in large lists, which the task very non-trivial and challenging. Learning-based deduplication system was introduced which discovered challenging training pairs using method called Active learning [8].The designed technique is a learning based deduplication system that allows automatic construction of the deduplication function by using a novel method of interactively discovering challenging training pairs. In this method the learner is automated to do the difficult task of of bringing together the potentially confusing record pairs. So the user has to only perform the easy task of labeling the selected pairs of records as duplicate or not. The system for deduplication consist of three primary inputs they are

a) Database of records (D) The original set D of records in which duplicates need to be detected.
b) Initial training pairs (L) An optional small(less than ten) seed L of training records r_1, r_2 arranged in pairs of duplicates or non-duplicates.
c) Similarity functions (F) A set F of n_f functions each of which computes a similarity match between two records based on any subset of d attributes.

Finally a function is provided as an output of this system.

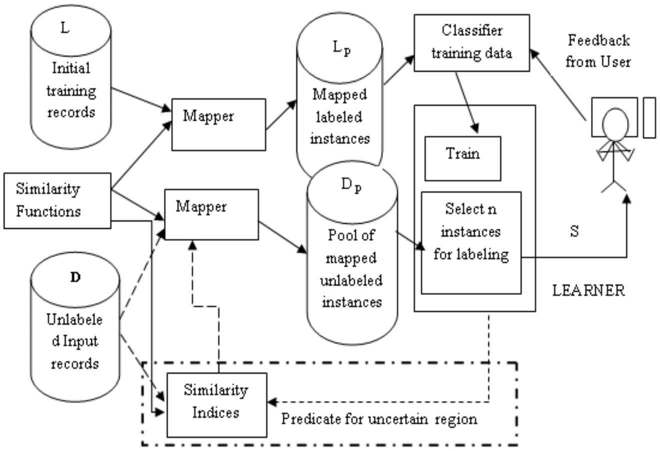

Fig.1 Overall design and working of Active- learning- based technique [8]

The main idea behind this system is that most duplicate and nonduplicate pairs are clearly distinct. The system starts with small subsets of pairs of records designed for training which have been characterized as either matched or unique. This initial set of labeled data forms the training data for a preliminary classifier. In the sequel, the initial classifier is used for predicting the status of unlabeled pairs of records. The initial classifier will make clear determinations on some unlabeled instances but lack determination on most. The goal is to seek out from the unlabeled data pool those instances which, when labeled, will improve the accuracy of the classifier at the fastest possible rate. Pairs whose status is difficult to determine serve to strengthen the integrity of the learner. Conversely, instances in which the learner can easily predict the status of the pairs do not have much effect on the learner. Using this technique, Active-learning-based system can quickly learn the peculiarities of a data set and rapidly detect duplicates using only a small number of training data **[1]**.Active-learning-based system is not appropriate in some places because it always requires some training data or some human effort to create the matching models.

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. An alternative approach of creating the distance metric that is based on ranked list merging. Here the idea is to compare only one field using matching algorithm and find out the best matches and rank them according to the similarities, where the best match catches the top position in rank. Finally, one of the problems of the distance-based techniques is the need to define the appropriate value for the matching threshold. In the presence of training data, it is possible to find the appropriate threshold value. However, this would nullify the major advantage of distance-based techniques, which is the ability to operate without training data.

C. Deduplication Techniques for Backup Operation

Now-a-days having backups are the safer side of loses of data in data repository but in addition to it there arise a problem of replication while duplicating the repository [4]. So to avoid this duplication module partitions a file into chunks, generates the respective summary information, which is knows a fingerprint. There is table called looks up Fingerprint Table to determine if the respective chunk already exists. If it does not exist, the fingerprint value is inserted into Fingerprint Table. Chunking and fingerprint management is the key technical constituents which governs the overall deduplication performance. There are a number of ways for chunking, e.g., variable size chunking, fixed size chunking, or mixture of both. There are a number of ways to managing fingerprints. Legacy index structure, e.g., B+ tree, and hashing does not fit for deduplication workload. Chunking and fingerprint management is the key technical constituents which governs the overall deduplication performance.

There are three compartments in the above process which composed of

a) Development of a novel chunking method called context-aware chunking. To reduce the computational overhead the exploitation of the algebraic nature of the modulo arithmetic and development of incremental Modulo-K algorithm is done. The chunking can be either type may be fixed size or variable size which corresponds to the file type.

b) Development of an efficient fingerprint management scheme called Least Recently Used (LRU) based index partitioning. Where tablets are formed by partitioning fingerprint table into smaller sized tables called tablets. When inserting a sequence of fingerprints to a tablet than to a large size table, placing a sequence of fingerprint values onto the disk will give clustered manner representation. Access history of the tablets as the LRU list is maintained to reduce the overhead of fingerprint table.

c) In third step performance is studied between the chunking and fingerprint lookup overheads.

Finally the efficiency of the deduplication is measured in two aspects. They are backup speed and the size of resulting deduplicated backup. For deduplication backup, a number of factors exist to optimize the performance: the false positive rate of the Bloom filter, the deduplication ratio, the chunking speed, the fingerprint lookup speed, etc. For optimizing deduplication performance, particular care needs to be taken to orchestrate the various factors involved in the entire deduplication process. In appropriate optimization may result in an unacceptable penalty to the overall backup performance.

D. Unsupervised Duplicate Detection

UDD [12] can effectively identify duplicates from the query result records of multiple Web databases for a given query it uses two classifiers. The WCSS classifier act as the weak classifier which is used to identify “strong” positive examples and an SVM [10], [11] and [12] classifier acts as the second classifier. First, each field’s weight is set according to its “relative distance,” i.e., dissimilarity, among records from the approximated negative training set. Then, the first classifier utilizes the weights set to match records from different data sources. Next, with the matched records being a positive set and the nonduplicate records in the negative set, the second classifier further identifies new duplicates. Finally, all the identified duplicates and nonduplicates are used to adjust the field weights set in the first step and a new iteration begins by again employing the first classifier to identify new duplicates. The iteration stops when no new duplicates can be identified. This method is well suited for only web based data but still it requires an initial approximated training set to assign weight. Compared to the existing work, UDD (Unsupervised Deduplication Detection) is specifically designed for the Web database scenario..Moreover, UDD focuses on studying and addressing the field weight assignment issue rather than on the similarity measure.
UDD identifies duplicates as follows:
WCSS classifier weight is assigned to each set of the field according to the relative distance that is the dissimilarity, among records from the approximated negative training set and the WCSS classifier, which utilizes the weights set in the first step of the UDD algorithm .It is used to match records from different data sources. then matched records being a positive set and the non duplicate records in the negative set, the SVM classifier further identifies new duplicates. Finally, all the identified duplicates and non duplicates are used to adjust the field weights set in the first step and a new iteration begins by again employing the first classifier to identify new duplication.

Input : Potential duplicate vector set P
Non-duplicate vector set N

Output: Duplicate Vector set D

C_1: a classification algorithm with adjustable parameters W that identifies duplicate vector pairs from P

C_2: a supervised classifier

Algorithm:

```
1.  $D=\emptyset$
2.  Set the parameters W of $C_1$ according to N
3.  Use $C_1$ to get a set of duplicate vector pairs $d_1$ from P
4.  Use $C_1$ to get  a set duplicate vector pairs f from N
5.  $P=P - d_1$
6.  While $|d_1| \neq 0$
7.  $N'=N-f$
8.  $D=D+d_1+f$
9.  Train $C_2$ using D and N'
10. Classify p using $C_2$ and get a set of newly identified duplicate vector pairs $d_2$
11. $P=P - d_2$
12. $D=D+d_2$
13. Adjust the parameters W of $C_1$ according to N' and D
14. Use $C_1$ to get a new set of duplicate vector pairs $d_1$ from P
15. Use $C_1$ to get a new set of duplicate vector pairs f from N
16. N=N'
17. Return D
```

Fig 2: Algorithm for UDD Duplicate Detection [12]

E. Genetic Programming Approach for Deduplication

The data gathering is done from multiple sources to make data repository. Data repository at that stage is said to contain “dirty data”. The data with no standard representation and presents of replicas is said to be dirty data. Due to this kind of contamination usage of such repository faces few problems. They are 1) performance degradation—as additional useless data demand more processing, more time is required to answer simple user queries; 2) quality loss—the presence of replicas and other inconsistencies leads to distortions in reports and misleading conclusions based on the existing data; 3) increasing operational costs—because of the additional volume of useless data, investments are required on more storage media and extra computational processing power to keep the response time levels acceptable. The problem of detecting and removing duplicate entries in a repository is generally known as record deduplication

To deal with the above problem approach based on Genetic programming is used. This approach combines several different pieces of evidence extracted from the data content to produce a deduplication function that is able to identify whether two or more entries in a repository are replicas or not [7].Record deduplication is a kind of time consuming process so the aim is to make out duplication function for small repository and resulting function is applied to other areas. The resulting function should be able to efficiently maximize the identification of record replicas while avoiding making mistakes during the process. Genetic Programming is one of the best known evolutionary programming techniques.

During the evolutionary process, the individuals are handled and modified by genetic operations such as reproduction, crossover, and mutation, in an iterative way that is expected to spawn better individuals (solutions to the proposed problem) in the subsequent generations. The steps of Genetic algorithm are the following:

1. Initialize the population (with random or user provided individuals).
2. Evaluate all individuals in the present population, assigning a numeric rating or fitness value to each one.
3. If the termination criterion is fulfilled, then execute
4. The last step. Otherwise continue.
5. Reproduce the best n individuals into the next generation population.
6. Select m individuals that will compose the next generation with the best parents.
7. Apply the genetic operations to all individuals selected. Their offspring will compose the next population. Replace the existing generation by the generated population and go back to Step 2.
8. Present the best individual(s) in the population as the output of the evolutionary process.

The evaluation at Step 2 is done by assigning to an individual a value that measures how suitable that individual is to the proposed problem. In our GP experimental environment, individuals are evaluated on how well they learn to predict good answers to a given problem, using the set of functions and terminals available. The resulting value is also called raw fitness and the evaluation functions are called fitness functions. The results are represented in tree format in this case, the rule is that each possible solution found is placed in the tree and evolutionary operation is applied for each tree. The fitness function is the GP component that is responsible for evaluating the generated individuals along the evolutionary process. If the fitness function is badly chosen or designed, it will surely fail in finding a good individual.

Using GP [13] approach three set of experiments are done with different conditions (a) GP was used to find the best combination function for previously user-selected evidence (b) GP was used to find the best combination function with automatically selected evidence (c) GP was tested with different replica identification boundaries. The boundary decides whether the pair is replica or not. This method is able to automatically suggest deduplication functions based on evidence present in the data repositories. The suggested functions properly combine the best evidence available in order to identify whether two or more distinct record entries are replicas.

As the result of GP [13] approach following criteria must be satisfied: outperforms an existing state-of-the-art machine learning based method, provides solutions less computationally intensive, frees the user from the burden of choosing how to combine similarity functions and repository attributes, frees the user from the burden of choosing the replica identification boundary value, since it is able to automatically select the deduplication functions that better fit this deduplication parameter.

F. Genetic Algorithm with Most Represented Data Samples

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. We can easily find the deduplication records.

The steps involved in the proposed algorithm are:

1. Compute $\boldsymbol{C}_p=(1/p)\ \Sigma^p{}_{i=1}\ \mathbf{e}_i$, i.e., the centroid of the record members $\boldsymbol{S} = \boldsymbol{S} = \{\mathbf{e}_i\}^p{}_{i=1}$

2. Select a first most represented sample that corresponds to the sample is closest to ***cp*** using **t**1 = argmin $j\{Dist(\mathbf{r}j, \boldsymbol{C}p)\}$

3. For each of the end members in the member set ***S*** do:

4.1 Remove from ***S*** the members which is less similar to ***Cp***, thus obtaining a new member set

$\{\mathbf{e}i\}^{p-1}{}_{i=1}$

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

4.3 Select the data sample whose value is the most similar to ***C****p*−1 as the second most represented sample training sample

4.4 repeat from step 3 until required number of most represented data sample

IV. EXPERIMENTAL RESULTS

In this section, we present and discuss the results of the Experiments performed to evaluate our proposed algorithm to record deduplication. In our experiments, we used Cora dataset to found the duplicate records.

The first real data set, the Cora Bibliographic data set, is a collection of 1,295 distinct citations to 122 computer science papers taken from the Cora research paper search engine. These citations were divided into multiple attributes (author names, year, title, venue, and pages and other info) by an information extraction system.

Precision comparison

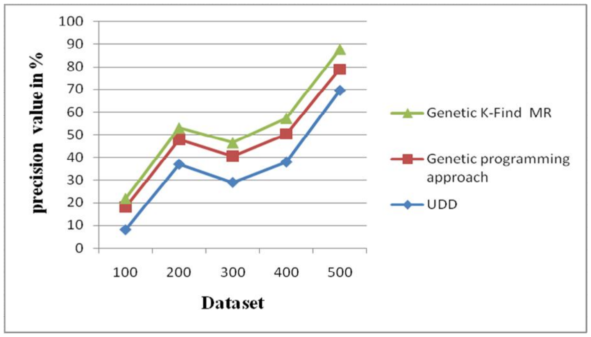

Fig. 3. Precision comparison

In this Figure 3 shows that the Precision Comparison of the system between the UDD, Genetic programming approach, Genetic programming approach for KFINDMR most relevant sample selection. We measure the precision value in % at Y-axis as algorithm and consider the Cora dataset in the X-axis. The precision value of the genetic KFINDMR is higher than the GP and the precision value of the GP is higher than the UDD. Finally our proposed algorithm achieves the higher level of the precision value rather than the other algorithm and Comparison also shown in Table1.

F measure comparison

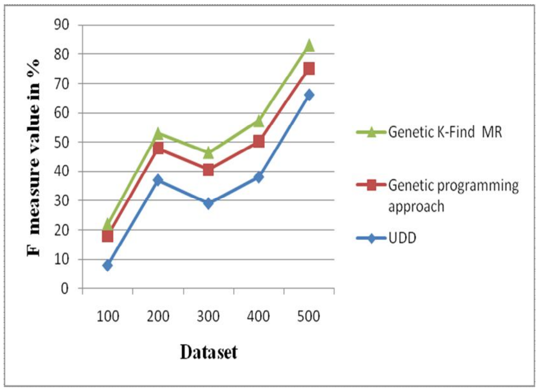

Fig. 4. F measure comparison

In this Figure 4 shows that the F measure Comparison of the system between the UDD , 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 and the F measure value of the GP is higher than the UDD. Finally our proposed algorithm achieves the higher level of the F measure value than the other algorithm and Comparison also shown in Table1.

Recall comparison

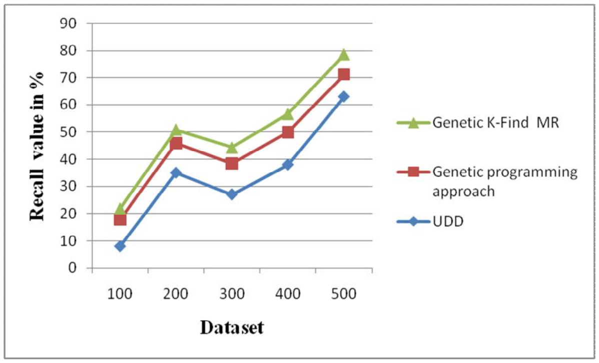

Fig. 5. Recall comparison

In this Figure 5 shows that the recall comparison of the system between the UDD, Genetic programming approach, Genetic programming approach for k- find most relevant sample selection. We measure the recall value in % at Y-axis as algorithm and consider the cora dataset in the X-axis. The recall value of the Genetic KFINDMR is higher than the GP and the precision value of the GP is higher than the UDD. Finally our proposed algorithm achieves the higher level of the Recall value than the other algorithm and Comparison also shown in Table1.

TABLE I
PERFORMANCE COMPARISON BETWEEN SVM AND GENETIC

	Comparison: Before and After selection			
	Before Selection		**After selection**	
	SVM	**Genetic**	**SVM**	**Genetic**
Precision	66.554	72.183	74.125	84.254
Recall	60.552	67.265	78.102	82.637
F-Measure	63.395	69.637	76.062	83.437

V. CONCLUSIONS

The problem of identifying and handling replicas is considered important 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 of these techniques are discussed. A particular function is obtained from the list of objects and that function is used to compare with other objects. After deduplication, information retrieval will be fast and efficient.

Deduplication is a very expensive and computationally demanding task, it is important to know which cases our approach would not be the most suitable option. Thus there is a need to investigate in which situations (or scenarios) our GP-approach would not be the most adequate to use. The genetics programming approach combines several different pieces of evidence extracted from the data content and produces the deduplication function that is able to identify whether two or more entries in a repository are replicas or not is difficult task .Our proposed new algorithm meets the most representative data samples from the Cora dataset. We intend to improve the efficiency of the GP training phase by selecting the most representative examples for training.

As future work, we can design criterion is choosing a deduplication process in distributed network .We plan to introduce new algorithm to find the duplication records using some techniques such as: i) an Optimization technique which is more efficient than genetic along with deduplication techniques ii) collaborative approach from different vendor and iii)We also introduce semantic based approaches to improve record deduplication in syntactically varied duplicate records.

REFERENCES

[1] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios, “Duplicate Record Detection: A Survey”, IEEE transactions on knowledge and data engineering, vol. 19, no. 1,January 2007.

[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] Jaehong Min, Daeyoung Yoon, and Youjip Won,“Efficient Deduplication Techniques for Modern Backup Operation”, IEEE transactions on computers, vol. 60, no. 6, June 2011**.**

[5] B. Liu. Mining data records in Web pages. In Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining, pages 601-606, 2003.

[6] B. Liu and Y. Zhai. NET: System for extracting Web data from °at and nested data records. In Proceedings of the Conference on Web Information Systems Engineering, pages 487-495, 2005.

[7] Moise´s G. de Carvalho, Alberto H.F. Laender, Marcos Andre´ Gonc¸alves, and Altigran S. da Silva “A Genetic Programming Approach to Record Deduplication” IEEE Transaction on knowledge and data engineering,vol.24, No.3, March 2012.

[8] Sarawagi and A. Bhamidipaty, “Interactive Deduplication Using Active Learning,” Proc. Eighth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, pp. 269-278, 2002.

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

[10] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg, “Adaptive Name Matching in Information Integration,” IEEE Intelligent Systems, vol. 18, no. 5, pp. 16-23, Sept./Oct. 2003.

[11] 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.

[12] Weifeng Su, Jiying Wang, and Frederick H. Lochovsky,” Record Matching over Query Results from Multiple Web Databases”, Knowledge Discovery and Data Mining, VOL. 22, NO. 4, APRIL 2010.

[13] M.G. de Carvalho, A.H.F. Laender, M.A. Gonc¸alves, and A.S. da Silva, “Replica Identification Using Genetic Programming,” Proc. 23rd Ann. ACM Symp. Applied Computing (SAC), pp. 1801-1806, 2008.

[14] T.P.C. Silva, E.S. de Moura, J.M.B. Cavalcanti, A.S. da Silva, M.G. de Carvalho, and M.A. onc¸alves, “An Evolutionary Approach for Combining Different Sources of Evidence in Search Engines,” Information Systems, vol. 34, no. 2, pp. 276-289, 2009.

[15] Bilal Khan, Azhar Ranf, Sajid H.Shah and ShanKhusrso “Identification and removal of Duplicated Records” World Applied Sciences Journal 13(5):1187-1184, 2011