EXPLORING RESEARCH AND INNOVATIONS

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459,ISO 9001:2008Certified Journal, Volume 3, Issue 5, May 2013)

A Detection of Duplicate Records from Multiple Web Databases using pattern matching in UDD

Dewendra Bharambe[1], Susheel Jain[2], Anurag Jain[3]

[1,2,3]Computer Science Dept., Radharaman Institute of Technology & Science, Bhopal (M.P.) India

***Abstract*—Record matching refers to the task of finding entries that refer to the same entity in two or more files, is a vital process in data integration. Most of the supervised record matching methods require training data provided by users. Such methods can not apply for web database scenario, where query results dynamically generated.**

In existing system, an unsupervised record matching method effectively identifies the duplicates from query result records of multiple web databases by identifying the duplicate and non duplicate set in the source and from that non duplicate set again searches for the existence of duplication. Then use two co-operative classifiers from the non duplicate set, they are Weighted Component Similarity Summing (WCSS) Classifier and Support Vector Machine (SVM) classifier. These two classifiers can be used to identify the query results iteratively from multiple web databases.

In this paper we modify record matching algorithm with genetic algorithm. The genetic programming is time consuming so we proposed UDD with genetic programming. A performance evaluation for accuracy is done for the dataset with duplicates using UDD and UDD with Genetic algorithm.

***Keywords*— Duplicate detection; Data deduplication; UDD; SVM; WCSS; Genetic algorithm; Pattern matching.**

I. INTRODUCTION

Web databases are the databases that produce the results dynamically in response to user queries. So the data repository from multiple heterogeneous sources contains record replicas and duplicates. A record refers to the same real world entity or object is called duplicate records. The dirty data means data with replicas and duplicates as well as no standardized representation. Since duplicate records occupies more space and even increase the access time. In the repository consisting of dirty data problems occurs like constraints quality, decreasing the performance increasing operation of cost. Thus identifying the duplicate records by analyzing and matching records from different data sources is increasingly important task

Record matching can be done by supervised learning where training dataset is required beforehand. In the web databases, the result records are obtained through online queries. They are query dependant and thus, supervised learning is in appropriate. The representative training set in supervised learning cannot be applicable for the web results that are generated on-the-fly. For each new query, depending on the results returned, the field weights should probably change too, which makes supervised learning based method even less applicable. Hence, we use a unsupervised technique named Unsupervised Duplicate Detection (UDD) which uses two classifiers for record matching and duplicate detection. This eliminates the user preference problem in supervised learning.

In this paper we modify record matching UDD algorithm with genetic algorithm by optimizing the pattern selection. The Genetic programming approach finds a proper combination of the best pieces of evidence, creates the heuristic function from a small representative portion which is then applied to the rest of repository area. Genetic programming approach combines several different pieces of evidence extracted from a data content to produce a heuristic function for identifying the duplicate records. In UDD genetic algorithm used for generation of fixed pattern for duplicate matching of valid class, the stored pattern of genetic provide the feature selection process and found the feature dissimilar in record data process recall stored pattern in classification.

II. EXISTING RECORD MATCHING TECHNIQUE

The problem of duplicate detection becomes more complicated because the record consists of multiple fields. The probabilistic approaches and supervised machine learning techniques depends on training data. As well as techniques depend on domain knowledge are generic distance metrics to match records.

A) Distance Based Technique

One way to avoid training data is to use a distance metric which does not require tuning through training data. Without the need of training data similar record can be match by using distance metric and an appropriate matching threshold. 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.

Chaudhuri et al. [2] proposed a new framework for distance-based duplicate detection, observing that the distance thresholds for detection real duplicate entries are different form each database tuple. To detect the appropriate thresholds, Chaudhuri et al. observed that entries that correspond to the same real world object but have different representation in the database tend to have small distances from each other (compact set property), to have only a small number of other neighbors within a small distance.

B) Supervised and Semi-supervised Learning Techniques

The supervised learning systems rely on the existence of training data in the form of record pairs, relabeled as matching or not. One set of supervised learning techniques treat each record pair (a, b) independently, similar to the probabilistic techniques.

A well-known CART algorithm [5] generates classification and regression trees. A linear discriminant algorithm, which generates a linear combination of the parameters for separating the data according to their classes, and a "vector quantization" approach which is a generalization of the nearest neighbor algorithm. The transitivity assumption can sometimes result in inconsistent decisions.

C) Active Learning Technique

A learning based deduplication system that allows automatic construction of the deduplication function by finding the challenging training pair interactively. In this method the learner is automated to do the difficult task 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.

Sunita Sarawagi and Anuradha Bhamidipaty [3] proposed an interactive learning based deduplication system called Active Learning led Interactive Alias Suppression (ALIAS).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-based system is not appropriate in some places because it always requires user provided training data for creating the matching models.

D) Indexing Based Record Matching Technique

Peter Christen [4] surveyed various indexing techniques for record linkage and deduplication. Record linkage refers to the task of identifying records in a data set that refers to the same entity across different data sources. 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

III. PROPOSED SYSTEM

In proposed system we modify record matching algorithm with genetic algorithm. Genetic algorithm is heuristic function; the nature of genetic algorithm is single objective for optimization of given problem.

In UDD genetic algorithm used for generation of fixed pattern for duplicate matching of valid class, the stored pattern of genetic provide the feature selection process and found the feature dissimilar in record data process recall stored pattern in classification. The process of UDD proceeds with graph traversing technique in follower of clustering and classification. In the process of UDD the graph points of number of pattern point selection executed by genetic algorithm. The process of modified UDD with stored pattern describe here.

Algorithm

```
1.  Input: G = (V, E) ←empty //initialize graph
2.  NP_list ← K-means (N_list, $K_v$)
3.  Input NP_list X , the clustering number cn, population scale XN , crossover probability cP, mutation probability mP, Pattern probability vP, stop conditions cS ;
4.  Code the chromosome in real number and initialize population A(i),i = 0 at random;
5.  Calculate the fitness of each individual in the current instant;
6.  UDD clustering creates stored pattern for classification, which means find dissimilar feature cluster. Hence the fitness function of algorithm is determined by f(x).
7.  F(x) = {(α +2β)-αi, αi<β+2α
              0,      αi≥αi+2β
         I=1, 2,……………………………..,N
8.  Judge the termination conditions. If the termination conditions are satisfied, then turn to step 9, otherwise, turn to step 10;
9.  Decode to find and calculate the optimal clustering and pattern matrixes. And set the optimal clustering for classification.
10. Do the parallel crossover and mutation operation on population A(i), then we can get population B(i), C(i) respectively;
11. Carry out the genetic selection on the instant composed of population A(i), B(i), C(i) and population D(i) is got;
12. Take the UDD optimization on population D(i) and generate the next generation A(i +1) . Then turn to step
13. for h ∈ A(i+1) do
14. h.nn ← Nearest-neighbor (A(i+1)- {h})
15. h.sc ← Compute-SC (h, h.nn)
16. V←V ∪ {h}
17. V←V ∪ {h.nn}
18. if h.sc <$th_{sc}$ then
19. E←E ∪ {(h,h.nn)}
20. end if
21. end for
```

Unsupervised Duplicate Detection (UDD) Algorithm

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

Output: Duplicate Vector set E
C_1: a classification algorithm with adjustable parameter W that identifies duplicate vector pairs from P
C_2:a supervised classifier,SVM

Algorithm

```
1.  $E = \emptyset$
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 of 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 E
```

UDD Algorithm Overview

UDD [1] identifies the duplicate records by using two classifiers iteratively. The duplicate records from the same source are removed using the exact matching method. At first each field's weight is set based its relative distance is dissimilarity among records from the approximated negative training set.

Then the WCSS classifier that utilizes the weight set is used to match records from various data sources. Next the matched records acting as positive set and Non-duplicate records as negative sets. After the SVM classifier again identifies the duplicates from positive set. At last all the identified duplicates and non-duplicates are used to adjust the field weights set in first step and new iteration begins by employing WCSS to identify new duplicates. The iteration stops when no new duplicates can be identified.

IV. Experiments And Result

In this section we present and discuss the result of the experiments performance to evaluate our proposed system for detection of duplicate record.

A. Dataset

We use three dataset to test our system. This method should be compared to existing techniques and approached to run out experiments with Cora dataset [6], collection of research paper citations provided by Cora citation matching. The citations were divided into multiple attributes (author, name, year etc) Likewise we used similar dataset like Book-full and Movies for testing.

B. Evaluation Metric

In many duplicate detection approaches the overall performance should be calculated using recall, precision and F-measure which are defined as:

$$\text{Recall} = \frac{\text{Noofcorrectlyidentifiedduplicatepairs}}{\text{No. oftrueduplicatepairs}}$$

$$\text{Precision} = \frac{\text{No. ofCorrectlyidentifiedduplicatepairs}}{\text{No. ofallidentifiedduplicatepairs}}$$

$$\text{F} - \text{measure} = \frac{2 * \text{precision} * \text{recall}}{(\text{precision} \ + \text{recal})}$$

C. Result

After running proposed system on different datasets we found that UDD with genetic algorithm outperforms than UDD. That is given below in the table.

Table 1.
Performance comparison between UDD and UDD with GA (T_{sim}=0.85)

Dataset		UDD	UDD withGA
Cora	Precision	1.3535	1.5535
	Recall	4.000	0.9915
	F-measure	30.55	2.4570
	Avg.Exe.Time	58.13	2.66
Movies	Precision	1.3535	1.5335
	Recall	13.00	4.9915
	F-measure	0.4758	0.4914
	Avg.Exe.Time	0.2329	0.2033
Book-full	Precision	1.3535	1.553
	Recall	16.000	1.991
	F-measure	1.0374	3.151
	Avg.Exe.Time	0.3779	0.2287

Table 2.
F-measure comparision of UDD and UDD with GA with Different Threshold Values

Threshold Values	UDD (F-measure)	UDD With GA (F-measure)
0.75	31.05	2.54
0.8	24.01	2.55
0.85	30.88	2.63
0.9	17.34	2.37
0.95	28.17	2.47
0.75	31.05	2.54

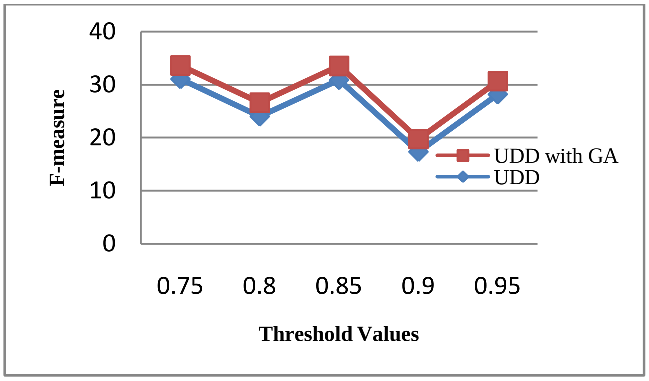

Fig 1. F-measure comparisons of UDD and UDD with GA with Different Threshold Values.

The figure shows the F-measure comparisons between UDD and UDD with Genetic Algorithm. We consider the Cora Dataset for experiment and take different threshold value in X-axis and measure the F-measure value in Y-axis. After applying different threshold values, chart shows that UDD with Genetic algorithm gives better performance than UDD.

V. CONCLUSION

Detection of duplicate records is an important problem in data management system. There are various techniques for record matching is explained with their advantages and disadvantages. The proposed algorithm modifies UDD with stored pattern. We used UDD with Genetic Algorithm for generation of fixed pattern for duplicate matching of valid class. In our system the process of UDD proceeds with graph traversing technique in follower of clustering and classification advantage of this algorithm is that if pattern is present then there is no need to search the entire text. The search is carried out on an area where the probability of successfully locating the pattern is highest. We used UDD with Genetic algorithm to provide better solution for the problem of finding out duplicate records from multiple web databases. The experimental result shows that our approach is comparable to previous work and the result shows that our alogorithm requires less time than UDD to find out the duplicates. As future work we planned to introduce a new algorithm for finding the duplicate records by using Optimization techniques.

REFERENCES

[1] W.Su, J. Wang, and Frederick H. Lochovsky, “Record Matching over Query Results from Multiple Web Databases” IEEE Transactions on knowledge and data engineering.S. Sarawagi and A. Bhamidipaty,“Interactive Deduplication Using Active Learning,” Proc. Eighth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD ’02), pp. 269-278 2002.

[2] S. Chaudhuri, V. Ganti, and R. Motwani, “Robust Identification of Fuzzy Duplicates,” Proc. 21st IEEE Int’l Conf. Data Eng. (ICDE ’05),pp. 865-876, 2005.

[3] 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.

[4] 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.

[5] O. Bennjelloun, H. Garcia-Molina, D. Menestrina, Q. Su,S.E.Whang, and J. Widom, “Swoosh: A Generic Approach to Entity Resolution,” The VLDB J., vol. 18, no. 1, pp. 255-276, 2009.

[6] A. McCallum, “Cora Citation Matching,” http://www.cs.umass.edu/~mccallum/data/cora-refs.tar.gz, 2004.

[7] 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.

[8] 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.

[9] 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.

[10] W.E. Winkler, “Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage,” Proc. Section Survey Research Methods, pp. 667-671, 1988.

BIBILOGRAPHY

Dewendra Bharambe is a scholar of M.Tech, (Computer Science Engineering), at R.I.T.S. Bhopal, under R.G.T.U. Bhopal, India. He is working as a lecturer in J.T.Mahajan College of Engineering, Faizpur.

Susheel Jain, Assistant Professor in Computer science department of R.I.T.S., Bhopal, M.P. He has done his M.Tech. in Software Engineering From Gautam Buddh Technical University,Lucknow, India.

Anurag Jain, H.O.D. of Computer science department of R.I.T.S. Bhopal, M.P. He has done his M.Tech, in Computer Science and Engineering, From Barkatullah University, Bhopal, India.

Manuscript received May 13, 2013.

Dewendra Bharambe, *Computer science department, RITS, Bhopal, M.P. (e-mail: dewendra.bharambe@gmail.com). Mobile No.09970899254*
Susheel Jain, *Assistant Professor Computer science department, RITS, Bhopal, M.P. (e-mail: jain_susheel65@yahoo.co.in).*
Anurag Jain, *H.O.DComputer science department, RITS, Bhopal, M.P. (e-mail: anurag.akjain@gmail.com).*