A Text Based Clustering Scheme With Genetic Programming To Eliminate Replicas

M. Yuvaraju
Assistant Professor
Department of IMER
Anna University, Regional Center
Coimbatore-641047,India

S. R. Nivedita
PG Scholar
Department of Computer Science and Engineering
Anna University, Regional Center
Coimbatore-641047, India

Abstract— *Duplicate data entries in repositories are a major problem for data administrators which lead to inconsistency. Therefore the quality of services is affected by the existence of duplicates in repositories. The need for clean and replica free repositories is to have high quality and concise data. The existing system uses less evidence to find out the duplicates and they are statically allocated. So we propose an extended Genetic programming based approach that automatically uses more evidences which are dynamically allocated. In this approach several different pieces of evidence extracted from the data content are combined to find a de-duplication function which is able to identify whether two entries are replicas or not in a repository. The genetic programming approach is capable of adapting these functions to a given fixed replica identification boundary automatically, allowing the user to choose and tune the parameter thereby freeing the burden of the user.*

Keywords— Replica, Record De duplication, Data Storage, Evidence, Repository

1. Introduction

The occurrence of dirty data in the repositories leads to performance degradation, quality loss and increasing operational cost. To overcome the occurrence of dirty data, we go for record deduplication which is the task of identifying same records. The main challenge in this task is finding a function that can resolve when two records refer to the same entity inspite of errors and inconsistencies in data.We present a genetic programming approach that uses more evidence extracted from the data or records to find out whether two or more entries in a repositories are replicas or not. These evidences are combined to produce a de-duplication function that identifies the distinct properties in records to find out which records are replicas. Genetic programming approach is capable of finding solutions to the problems without considering the large space of the repositories. Record de-duplication is to compare if the records refer to the same real world entity.

2. Related Works

A. A Genetic Programming approach for record de-duplication

We use algorithm based on various factors such as word matching, phrase matching, and file name, type, size and so. There are two categories
(a) Ad hoc or domain knowledge approaches: This approach is based on the domain knowledge.
(b)Training based approaches: This approach is based on some sort of training supervised or semi supervised. Probabilistic and machine learning approaches come under this category.

Domain Knowledge approach: In this approach we use a matching algorithm, for a record in a repository that matches another record using the similarity function. This depends on the user defined minimum similarity threshold. This approach selects records of high weight tokens than the low weight tokens.

Probabilistic approach: In this approach, two boundary values are used to classify a pair of records as being replicas or not. The boundary values are classified as positive boundary values and negative boundary values.

- Positive boundary value: If the similarity value lies above this boundary, then the records are replicas.
- Negative boundary value: If the similarity value lies below this boundary, then the records are not replicas.

Machine learning approach: In machine learning approach, we consider a set of record pairs and find the similarity between these record pairs which is the probability of finding the rank and score of best alignment between these pairs. Based on the probability value, higher the probability, bigger the similarity between the attributes. In this approach we generate individual ranking for each field based on the scores.

A. Adaptive Duplicate Detection using learnable String Similarity Measures

The problem of identifying duplicate records in database is an essential step. A framework for improving duplicate detection using textual similarity. Since variations and representations occur from abbreviation, misspellings and typographical errors, text distance functions for each field is employed [4].

B. Eliminating Fuzzy duplicates in data warehouses

One of the important data cleaning problems is the problem of eliminating duplicates which detect multiple tuples that refer to the same real world entity. Standard textual similarity functions between multiple attributes tuples were the solution to this problem. But this resulted in unexpected output and fake positives. Therefore here an algorithm is developed to overcome this problem. This algorithm deals with hierarchies. This hierarchical procedure produces a high quality and it works on the real datasets. Data that arrive at data warehouse usually contain errors, spelling mistakes, and wrong conventions. Therefore, large amount of time and money is spent on data cleaning.

C. Robust and Efficient Fuzzy Match for Online Data Cleaning

Data warehouses receive data tuples from external sources .So the data warehouse should be checked to produce high data quality. For example, student name and roll number fields in attendance record must match the pre- recorded student name and roll number fields in the product reference relation.. the major task is to put an efficient and accurate fuzzy match operation in action to produce a clean tuple. This produces a new similarity function that overcomes the existing similarity function thereby developing an efficient fuzzy matching algorithm which proves the effectiveness of matching by implementing on real datasets [7].

3. Dataset Creation

This is the initial step of the record linkage where data are stored in data storage. End user, administrator and website logger are involved in this phase. Duplicates are detected from the original set D of records. The goal is to find the pairs in the cross product D-D that can be labeled as duplicates.

4. Genetic Operation for Selected Item

The function set is the collection of statements, operators and basic or user defined functions that can be used to terminal values. It produces better individuals in the future generations. The various genetic operations include reproduction, cross over and mutation. Using these operations such as cross over the cross product of two records are compared to find out if the refer to same real world entity or not. That is to check if the records are duplicate or not.

5. Evidence Extraction

Here several different piece of evidence extracted from various contents produce a de-duplication function which identifies whether the entries are replicas or not. It combines best piece of evidence to maximize the performance. The evidence forms the tree structure.

6. De-duplication Identification

Tree input is the evidence instances and output is the real number value. This value is compared against the replica identification value. It is an estimation of the similarity between the records being processed to check whether the records are replicas or not. This comparison is done for all candidates record pairs and the total number of correct and incorrect identified replicas is computed.

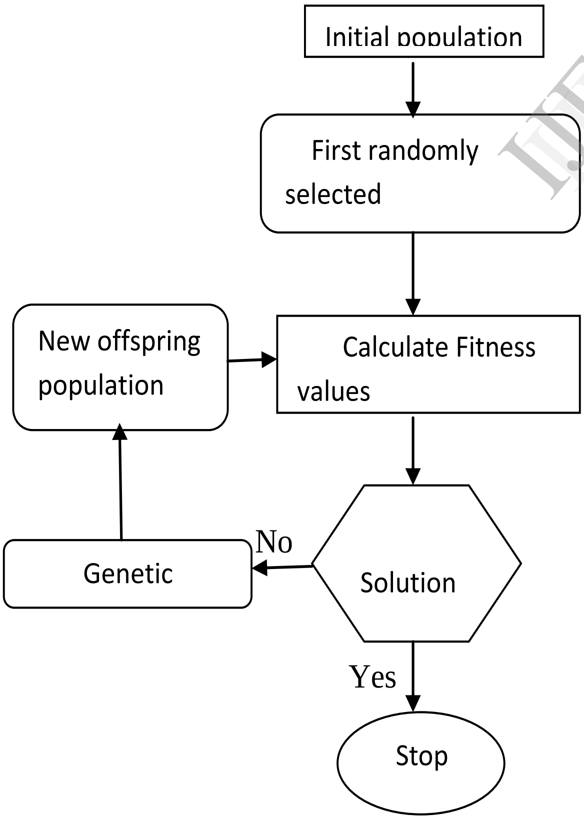

Fig1.Genetic programming architecture diagram

De-duplication Function

A set of nf functions each of which computes a similarity match between two records r1; r2 based on any subset of d attributes.

Database records –D

Training pairs –L

De-duplication function –F

1. Input L,D,F.
2. Create pairs Lp from the labeled data L and F.
3. Create pairs Dp from the unlabeled data D and F.
4. Initial training set T = Lp.
5. Loop until evidence identification

Training C using T.

Use C to select a set S of n instances from Dp for labeling.

If S is empty , exit loop.

Collect evidence on the label of S.

Add S to T and remove S from Dp.

6. Output classifier C.

7. Duplicate Elimination

Duplicate elimination depends upon the evidence extraction phase which retrieves result without any duplicate values. Elimination of duplicate results provides a better evidence for further process. The extended genetic programming approach identifies duplicates among records from multiple databases.

8. Reports

The reports are verified and controlled by the administrator. The searching efficiency and the download ratio is verified and maintained. So the administrator can view all the data of the users and web loggers.

9. Genetic Programming

A. Basic Concepts

Genetic Programming is a best evolutionary programming technique that inherits the properties of genetic operations. It is used for optimization problem and for the multi-objective problem. It is well known for its best performance in searching large spaces and as well as its capability to operate over the population of individuals. It not only creates new solutions but also allows new combination of features.

The advantages of genetic programming are:

- Genetic programming focuses on the concepts and interprets the problem as a computer program.
- It is applicable to symbolic regression problem since the representation structures are variable.
- It is capable of finding out the dependent and independent variables, and establishes the relationships between them.

B. Genetic Operations

Genetic programming makes use of length free data structures called individuals; they represent a single solution to the problem. In this method, we make use of tree representation. This tree representation includes set of terminals and functions which is to be defined. Terminals include input, constants or zero arguments which forms the termination of a tree. It is also called as tree leaves. Function set includes the collection of operators, statements and uses defined functions for manipulations of terminal values.

The various Genetic operations are reproduction, crossover and mutation. Reproduction is defined as copy of the individual without any modifications in them. Crossover is the exchange of two parents which produces two or more children. Crossover produces child trees which is result of swap of selected sub trees of the parents. Mutation operation is selecting a random node and the created sub tree for that node is replaced for the corresponding sub tree.

C. Generational Evolutionary Algorithm

Generational Evolutionary Algorithm is used in Genetic programming evolutionary process. The various steps in the algorithm are:

- Collect the population
- Assign numeric rating for the individuals in the population.
- Execute last step if the termination criteria are fulfilled.
- Select and reproduce the best individuals.
- The individuals are applied with the genetic operation that will produce the offspring of the next generation.

Genetic programming evaluates individuals on how they predict good answers to the problem. The final value is called raw fitness and the evaluation function is known as fitness function. Genetic programming uses a tree based representation as it the natural representation of the function. The requirements are:

- The solution to the problem is presented by tree structure.
- The operation for every individual tree terminates with a valid tree.
- Automatically every individual tree is evaluated.

Algorithm

```
1  Let DB be the set of records to de-duplicate;
2  Let sim be a record similarity function;
/*Training Phase*/
3  Generate a set of pairs P = p1 ::: pn from DB
```

```
Compute sim(p) for each pair p ∈ P;
R ← rank P according to sim(p) values;
Let T be the top k pairs in R;
Let B be the bottom k pairs in R;
User labels each pair p ∈ T ∪ B with Lp ∈
   {T ; F };
Compute a weight WP for each p ∈ P;
/*Evolution process*/
Gen0 ← Generate m functions;
Evaluate (Gen0,T∪B);
Compute a weight Wf for each function
   f ∈ Gen0;
Com0 ← ϕ;
for i=0 to a predefined number of generations do
   Comi ← top C functions in Geni;
   for each pair p∈ P do
      Comi labels p with Lp ∈ {T;F;D};
      if Lp=D then user labels p with Lp ∈
         {T; F}
   /*Genetic process implementation*/
   Update WP for each p∈ P;
   Geni + 1 ← Crossover & Mutation ( Geni –
      Comi);
   Geni + 1 ← Geni +1 ∪ Comi;
   Evaluate (Geni +1,P);
   Compute Wf for each function f ∈ Geni+1;
   Comi ← top C functions in Geni;
   for each pair p∈P do
      Comi labels p with Lp ∈ {T;F}
```

10. Result

In this paper, website is created in which the web loggers can register and login into the website to post and upload the books. End users can login and search for the books. This work is an real time online application. It is a priority based search scheme which identifies the replicas using text analysis. It retrieves the result whether the records are replicas or not by comparing the records Based on clustering, word analysis and frequency count. Apart from comparing the external properties such as file type, file name and size of file , it also takes into account the internal properties such as line count, frequency which is number of occurrences of a particular word in a record. By comparing the frequency of the words in each record, it returns whether the records are duplicate or not. Administrators can view the report, view the rank – numbers of searches made for the particular book, view the website and the books in the corresponding website.

The graph for the proposed work shows that the time consumption is less as it uses the priority, accuracy and improves efficiency. Theoretically, for the existing work time consumption is more as it consider all the properties.

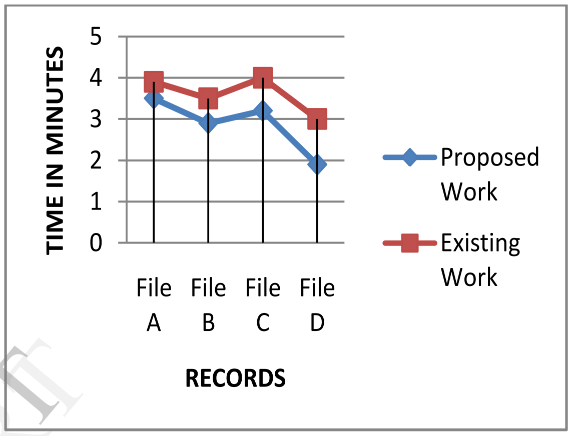

Fig2. Comparison of existing and proposed Work

11. Conclusion and Future Work

Identifying and handling replicas guarantee the quality of the information. Genetic programming approach automatically suggests the de -duplication function related to the evidence present in repositories. This approach combines the best set of evidence though it is not previously known. This approach also makes use of boundary values to find out if the record is replica or not. This approach is able to adapt the deduplication functions to different boundary valuesto classify the records are replicas or not. The comparison is done dynamically which is a real time implementation.

In our future work, record deduplication is an expensive and demanding task so we intend to improve the efficiency of Genetic programming training phase. Therefore, it is necessary to minimize the training effort without affecting the quality of the final solution.

References

[1] Moises G.de Carvalho, Alberto H.F.Laender, Macros Andre Goncalves, and Altigran S.da Silva, "A Genetic Programming Approach to Record Deduplication" IEEE Trans on Knowledge and Data Engineering,Vol 24, No 3,March 2012.

[2] H.M. de Almeida, M.A. Gonc¸alves, M. Cristo, and P. Calado, "ACombined Component Approach for Finding Collection-Adapted Ranking Functions Based on Genetic Programming," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 399-406, 2007.

[3] M.G. de Carvalho, M.A. Gonc¸alves, A.H.F. Laender, and A.S.daSilva, "Learning to Deduplicate,"Proc.Sixth ACM/ IEEE CSJointConf. Digital Libraries, pp. 41-50, 2006.

[4] 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.

[5] 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.

[6] N. Koudas, S. Sarawagi, and D. Srivastava, "Record Linkage: Similarity Measures and Algorithms," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 802-803, 2006.

[7] S. Chaudhuri, K. Ganjam, V. Ganti, and R. Motwani, "Robust and Efficient Fuzzy Match for Online Data Cleaning," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 313-324, 2003.

[8] I.P. Fellegi and A.B. Sunter, "A Theory for Record Linkage," J. Am.Statistical Assoc., vol. 66, no. 1, pp. 1183-1210, 1969.

[9] Bhattacharya and L. Getoor, "Iterative Record Linkage for Cleaning and Integration," Proc. Ninth ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery, pp. 11-18,2004.

[10] M. Wheatley, "Operation Clean Data," CIO Asia Magazine, http://www.cio-asia.com, Aug. 2004.

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

[12]I.P. Fellegi and A.B. Sunter, "A Theory for Record Linkage," J. Am.Statistical Assoc., vol. 66, no. 1, pp. 1183-1210, 1969.

[13] I. Bhattacharya and L. Getoor, "Iterative Record Linkage for Cleaning and Integration," Proc. Ninth ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery, pp. 11-18,2004.

[14] J.R. Koza, Gentic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, 1992.

[15] R. Bell and F. Dravis, "Is You Data Dirty? and Does that Matter?,"Accenture Whiter Paper, http://www.accenture.com, 2006.

[16] V.S. Verykios, G.V. Moustakides, and M.G. Elfeky, "A Bayesian Decision Model for Cost Optimal Record Matching," The Very Large Databases J., vol. 12, no. 1, pp. 28-40, 2003.

[17] W. Banzhaf, P. Nordin, R.E. Keller, and F.D. Francone, Genetic Programming - An Introduction: On the Automatic Evolution of Computer Programs and Its Applications. Morgan Kaufmann Publishers,1998.