ISSN:0975-9646

A Novel Approach for Record Deduplication using Hidden Markov Model

[1]R.Parimala devi & [2] DR.V.Thigarasu

1. *Research Scholar, Department of Computer Science, Karpagam University, Coimbatore, Tamilnadu, India.*
2. *Associate Professor, Department of Computer Science, Gobi Arts and Science College, Gobichettipalayam, Erode, Tamilnadu, India*

Abstract One of the challenging research areas in data mining is record deduplication. In most of the organizations the storage systems having duplicate copies of several pieces of data. The dedicated data compression method is data deduplication which is used for remove the duplicate copies of repeating data. Previous research used genetic programming based record deduplication which combined various pieces of evidence extracted from the data content. However the true positive level of the system will be low. Therfore, the performance of the record deduplication system is degrades .To solve this problem we are propopsing the Hidden markov model based record deduplication method. In a HMM model the records with different attributes are called states and a similarity functions among the couple of records are called transition. The data records attribute information of are cleaned, standardised and implemented through a hidden Markov models (HMMs). Evaluating the performance of the system is performed using Restaurants data set and Cora Bibliographic data set. The result obtained from the HMM based results the duplicate and non-duplicate records of datas. The system improves true positive level of the system.

Key words : Record deduplication, Hidden Markov Model, Genetic Programming

INTRODUCTION

The data sets to be integrated may contain data on the same real-world entities. In order to combine two or more data sets in a significant way, it is essential to identify representations belonging to the identical real-world entity. Therefore, duplicate detection is a significant component in an integration process. Due to deficiency in data collection, data modeling or data management, real-life data is often incorrect and/or incomplete. This principally hinders duplicate detection. Therefore, duplicate detection methods have to be designed for accurately handling dissimilarities due to typos, data missing, data obsolescence or misspellings.

Duplicate detection is the trouble of identifying many representations of a same real-world object. It is a crucial task in data cleansing and has applications in scenarios such as data integration, customer relationship management, and personal information management. To detect and remove duplicate records is a key step for data cleaning and also a significant problem for improving quality of data. Duplicate records are the records that signify the same entity in the real world while are not identified by DBMS due to various data format or misspell. The reason of duplicate record detection is to match, merge and remove the redundant database records that signify the same entity while with various data expression.

Deduplication is a task of recognizing the duplicate data in a warehouse that refer to the same real world entity or object and systematically substitutes the reference pointers for the redundant blocks. It is called as storage capacity optimization. Dirty data is classified in various classes.

(1) Performance degradation: As extra useless data demand extra processing, extra time is necessary to answer simple user queries.

(2) Quality loss: The presence of replicas and other inconsistencies direct to alteration in reports and misleading conclusions based on the existing data information

(3) Increasing operational costs: Because of the extra volume of useless data, investments are required on more storage media and extra computational processing power to keep the response time levels acceptable.

The trouble of detecting and eliminating duplicate entries in a repository is usually known as record deduplication. More specifically, record deduplication is the task of recognizing, in a data repository, records that refer to the same real world entity or object in spite of misspelling words, typos, various writing styles or even various schema representations or data types. Thus, there have huge investments from private and government associations for developing techniques for eliminating replicas from data repositories. This is due to the fact that clean not only allow the retrieval of greater quality information but also lead to a extra concise data representation and to potential investments in computational time and also resources to process this data

PREVIOUS RESEARCH

A genetic programming (GP) technique is used to record deduplication. In this technique combines many various pieces of evidence extracted from the data content to create a deduplication function that is able to identify the two or more entries in a repository are replicas or not.

The main role of this paper is a GP-based approach to record deduplication that

- Outperforms previous state-of-the-art machine learning based technique found in the novel
- Provides results less computationally intensive, since it recommends deduplication functions that develop the available evidence additional efficiently.

- Frees the user from the burden of choosing how to mingle similarity functions and repository attributes. This distinguishes our approach from all existing technique, since they need user-provided settings
- Frees the user from the burden of choosing the replica classification boundary value, since it is able to automatically select the deduplication functions that enhanced fit this deduplication parameter.

During the evolutionary procedure, the individuals are handled and customized by genetic operations in a repeated way. The genetic operations are crossover, reproduction and mutation [6].

Reproduction

Reproduction is a process of copy of individuals without any modification. Generally, this operator is used to execute an elitist strategy that is adopted to keep the genetic code of the fittest individuals across the changes in the generations. If a excellent individual is found in the previous generations, it will not be lost during the process.

Crossover

The process of crossover permit genetic content which is sub trees swap among two parents tree, in a operation that can produce two or more children. Genetic programming evolutionary operation, couple of parent trees are selected based on a pairing strategy and then, a random sub tree is selected in every parent tree. Child trees are the result from the exchange of the selected subtrees among the parents tree .

Mutation

Keeping a minimum diversity level of individuals is a role of mutation operation in the population, thus avoiding premature convergence. Every solution tree output from the crossover operation has an equal chance of suffering a mutation operation. Genetic Programming tree representation, a random node was selected and the corresponding sub tree is put back by a new randomly created sub tree. Genetic Programming evolutionary operation is guided by a creational evolutionary algorithm. In that technique each piece of evidence (or simply "evidence") E is a couple <attribute; similarity function> that represents the need of a specific similarity function over the values of a specific attribute found in the data being calculated. At the final stage entire number of correct and incorrect replicas is determined.

In that Genetic Programming-based technique is used to record deduplication. That technique is able to automatically propose deduplication functions depends on evidence present in the data repositories. The recommended functions properly merge the most excellent evidence available in order to recognize if two or more distinct record entries are replicas.

PROPOSED METHODOLOGY

In Genetic programming technique joined more than a few different pieces of evidence extracted from the data content and generates the deduplication function. In that process accuracy level of the technique will be low. Record Deduplication using GP, works to find the replica records only in local repository and not in all records, when matched to other optimization it becomes less efficient . To overcome the issues of genetic programming approach we use Hidden Markov Models (HMM) based record deduplication.

An HMM is defined by the probabilistic finite state machine constructed based on the set of hidden or unobserved states, transition edges connecting these states and a fixed dictionary of distinct observation output. Each and every edge is connected with a transition probability, and each state produce observation output from the dictionary with a definite probability distribution.

The states are represented as records with various attributes and transition as are defined as similarity function between a couple of records. Attribute information of data records such as author names, year, title, venue, pages and other information of records are cleaned and standardised and implemented through a hidden Markov models (HMMs). To perform this, the training of HMM data is done from the same data sets. The result obtained from the HMM based results the duplicate and non-duplicate records of datas.

Normally solved problems are:

1. Matching the most likely system to a series of observations -evaluation, solved using the forward algorithm;
2. Calculating the hidden sequence most likely to have produced a series of observations - decoding, solved using the Viterbi algorithm.
3. Calculating the model parameters most likely to have produced a sequence of observations - learning, solved using the forward-backward algorithm.

More specifically, record deduplication is the task of recognizing, 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 various schema representations or data types. Thus, there have beenlarge investments from private and government associations for developing techniques for eliminating replicas from data repositories. This is due to the fact that clean and replica-free repositories not only allow the retrieval of higher quality information but also lead to a more concise data representation and to potential savings in computational time and resources to process this data. The hidden markov model is used for record duplication detection the method of record deduplication mentioned in below.

The data cleaning phase recognized records that was invalid for linkage and performed corrections in the name field, preparing it for the subsequent standardization phases [9]. Standardization of the form included some corrections and/or substitutions of some spelling variations according to a standard established for representing the name's form: capitalization of the letters; elimination of accent marks; removal of spaces at the beginning and end of the name; removal of double spaces; removal of prepositions; and removal of punctuation marks.

The name standardization phase makes "dictionary" tables. These tables consisted of two fields, current_name and correct_name. That functioned are when a term from the name was found in the current_name table, the term was corrected according to the correct_ name field. For example, this technique can replace all the variations for the surname “GONCALVES”, such as “GONCAVES”, “GONEALVES”, “GONCAOLVES”. Three such tables were make, for given names (dic_name), surnames (dic_surname), and suffixes (dic_suffix)[10].

A HMM consists of four part .That is (1) a set of hidden states *S*; (2) a probability of transition $P[s'|s]$ between hidden states s *e* $s' \in S$; (3) a set of symbols (observations) *T* emitted by the hidden states; (4) a probability distribution of symbol emissions for each hidden state.

HMM Training

Training of an HMM is an offline method. We use Baum-Welch algorithm to train an HMM. Baum-Welch algorithm uses observation attributes created at the end of method. At the end of training phase we get an HMM corresponding to each cardholder. Baum-Welch algorithm is as follow [11]:

Particular observation sequence is $\mathrm{O}_1, \mathrm{O}_2, \dots \mathrm{O}_\mathrm{T}$. Initialization: set λ = (π,A,B) with random initial conditions. The algorithm updates the parameters of λ iteratively until convergence, following the procedure below:

The forward procedure: We define: $\mathrm{P}(\ \mathrm{O}_1, \mathrm{O}_2, \dots \mathrm{O}_\mathrm{T}, \mathrm{S}_\mathrm{t} = \frac{\mathrm{i}}{\lambda}$)which is the probability of seeing the partial sequence $\mathrm{O}_1, \mathrm{O}_2, \dots \mathrm{O}_\mathrm{T}$ and ending up in state *i* at time *t*. We can efficiently calculate $\propto_\mathrm{i}\ (\mathrm{t})$ recursively as

$$\propto_\mathrm{i}\ (\mathrm{t})\ = \pi_\mathrm{i} \mathrm{b}_\mathrm{i}(\mathrm{O}_\mathrm{i})$$

$$\propto_\mathrm{j}\ (\mathrm{t}+1) = \mathrm{b}_\mathrm{i}(\mathrm{O}_{\mathrm{t}+1})\ \sum_{i=1}^{N} \propto_i\ (t) \,.\, a_{ji}$$

Testing

Let initial series of observation attributes of length *R* up to time *t* is $\mathrm{O}_1, \mathrm{O}_2, \dots \mathrm{O}_\mathrm{R}$. In our implementation we have taken 50 as length of series. We determine the probability of acceptance of this series by HMM, let$\propto_1$ be the probability of acceptance [12].

$$\propto_1 = \quad \mathrm{P}(\mathrm{O}_1, \mathrm{O}_2, \dots \mathrm{O}_\mathrm{R} | \lambda)$$

At time t+1 sequence is $\mathrm{O}_1, \mathrm{O}_2, \dots \mathrm{O}_{\mathrm{R}+1}$, let $\propto_2$ be the probability of acceptance of this sequence

$$\propto_2 = \quad \mathrm{P}(\mathrm{O}_1, \mathrm{O}_2, \dots \mathrm{O}_{\mathrm{R}+1} | \lambda)$$

Let $\Delta\propto\ = \propto_1 - \propto_2$, $\Delta\propto > 0$, it mean new series is accepted by an HMM with minimum probability, and it could be a fraud. The new added transaction is calculated to be fraudulent if percentage change in probability is above threshold, that is

$$\text{Threshold} \leq \Delta\propto | \propto_1$$

The threshold value can be well-read empirically and Baum-Welch algorithm determines it automatically. If O_{R+1} is malicious, the issuing bank does not approve the transaction, and the FDS discards the symbol. Otherwise, O_{R+1} is added in the sequence permanently, and the new series is used as the base sequence for calculating the validity of the next transaction [13].

The training and refinement phases want to achieve the best fit of the initial model to the real data. The sequence of observations used to construct this fit. That is called a “training sequence”, since it is utilised to train the HMM. For each phase, another random sequence of a thousand records was selected from Cora Bibliographic data set and Restaurants data set producing the corresponding identification symbols. The Baum-Welch algorithm 13 was used to adjust the initial model’s parameters. The algorithm is a technique of iterative re-estimation which creates a series of observations with maximum probability than the existing method. Repetition of the methods was done, and the Kullback-Leibler divergences 14 among the two models were determined; the iterations were interrupted when divergence among two consecutive methods dropped below 10-5.

The hidden Markov model’s conformity was evaluated by the proportion of hits in the sequence of states created by the names of the test samples. This study adopted the terminology proposed by Müller & Buttner 15, which defines conformity as the contract between two observations when one is taken as the reference or standard, and consistency as the agreement among two observations when neither can be taken as the reference. In order to estimate the application of name segmentation via HMM in record linkage of Restaurants data set and Cora Bibliographic data set, 20 thousand records was randomly selected from every respective database. The fields selected for record linkage were: author name, title, year and etc.

EXPERIMENTAL RESULTS

Dataset description

In our experiments, we used two real data sets known as Bibliographic data set and Restaurants data set. They are commonly employed which are based on real data gathered from the web. The Cora Bibliographic data set is a first real data set. That is collection of 1,295 distinct citations to computer science papers of 122 taken from the Cora research paper search engine. These citations were split into multiple attributes (authornames, year, title, venue, and pages and other info) by an information extraction method. Restaurants data set is a second real data set; it contains 864 entries of restaurant names and additional information, including 112 duplicates that were obtained by integrating records from Fodor and Zagat’s guidebooks. We used the following attributes from this data set: (restaurant) name, address, city, and specialty.

Performance evaluation

In our experiment, we are analyze and compare the performance of record deduplication systems such as Hidden markov model (HMM) based record deduplication and genetic programming based record deduplication. The performance of three parameters such as accuracy rate, precision and recall in the HMM based record duplication detection is better than GP based record duplication detection.

Accuracy rate

The Accuracy of the system is calculated with the values of the True Negative, True Positive, False Positive, False negative actual class and predicted class outcome it is defined as follows,

Accuracy

$$= \frac{True\ positive + True\ negative}{True\ positive + True\ negative + False\ positve + False\ negative}$$

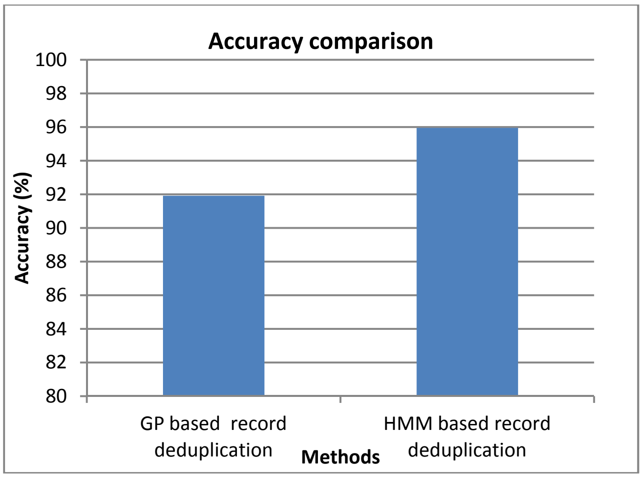

Fig.1. Accuracy comparison

In this graph, x axis will be the two approaches of record deduplication and y axis will be accuracy in %. From the graph see that, accuracy of the system is reduced in Genetic programming than our proposed hidden markov model based record deduplication. From this graph, we can say that the accuracy of record deduplication approach is increased, which will be the best one

Precision

Precision value is determined based on the retrieval of information at true positive prediction, false positive. In healthcare data precision is determined the percentage of positive outcome returned that are relevant.

Precision =TP/ (TP+FP)

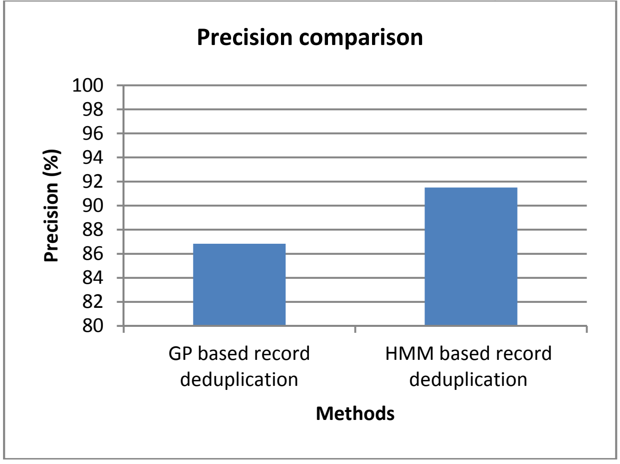

Fig.2. Precision comparison

Compare the methods of genetic programming and hidden Markov model deduplication. In this graph, x axis will be the two approaches of record deduplication and y axis will be precision in %. Hmm based record deduplication has high precision compare to another one.

Recall

Recall value is determined based on the retrieval of information at true positive prediction, false negative. Recall in this context is also referred to as the True Positive Rate. In that process the fraction of relevant instances that are retrieved.

Recall =TP / (TP+FN)

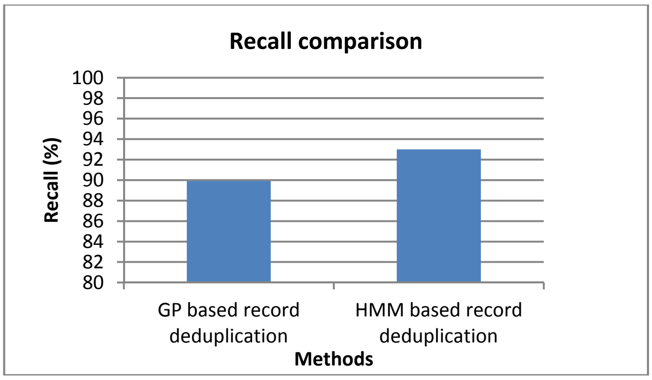

Fig.3. Recall comparison

In this graph, x axis will be the two approaches of record deduplication and y axis will be recall in %. From the graph see that, recall of the system is GP systems than our proposed Hidden markov model based record deduplication. From this graph, we can say that the recall of record deduplication approach is increased, which will be the best one.

CONCLUSION

Identifying and handling replicas is important to guarantee the quality of the information made available by data intensive methods they are digital libraries and also e-commerce brokers. These methods rely on consistent data to offer high-quality services, and may be affected by the existence of duplicates or near-duplicate entries in their repositories. Thus the reason the hidden markov model used for record duplication detection. Hidden markov model based record deduplication attribute information of data records are standardised and achieved. The performance of the system is maximised. Experiment with datasets such as Restaurants data set and Cora Bibliographic data set are evaluated. The result obtained from the HMM based results the duplicate and non-duplicate records of datas. The parameters of accuracy, precision and recall are better performance compare to the existing GP method.

REFERENCE

1. Weifeng Su, Jiying wang, Frederick H Lochovsky., Record matching over query results from multiple web databases. IEEE Transcations on Knowledge and Data Engineering, Vol 22, 578 – 588, 2010.
2. Li Yi and Kang Wandi, A new genetic programming algorithm for building decision tree. Procedia Engineering, Vol. 15, 3658 – 3662, 2011.
3. Prabhat Srivastava and Margaret O Mahony, Amodel for development of optimisied feeder routes and coordinated schedules – A genetic algorithms approach. Transport Policy, Vol.13, 413 – 425, 2006.
4. Brandye M. Smith,and Paul J Gemperline. Wavelength selection and optimisation of pattern recognition methods using the genetic algorithm. Analytica Chimica Acta, Vol.423, 167 -177, 2000.
5. Brain Carse , Terence C Fogarty. Evolving fuzzy rule based controllers using genetic algorithms. Fuzzy Sets and Systems, Vol.80, 273 – 293, 1996.
6. Parimala devi and Thigarasu. A genetic programming approach for record dedepulication. Int. J. Computer Sci. Information Technologies, Vol.5, 2895 – 2898, 2014.
7. Praveen kumar, Sankar kumar paul. Multiobjective PSO with time variant inertia and acclerationcoefficent. Information Sciences, Vol.177, 5033 – 5049, 2007.
8. Dawei Zhou, Xiang Gao et al., Randomisation in PSO for global search ablity. Expert Systems with Applications, Vol.38, 15356 – 64, 2011.