ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

A Novel Approach Using Genetic Algorithm and Hidden Markov Model for Optimising Record Deduplication

[I]R.Parimala Devi, [II]Dr. V.Thigarasu

[I]Research Scholar, Dept. of Computer Science, Karpagam University, Coimbatore, Tamilnadu, India.
[II]Associate Professor, Dept.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 th;e 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

HMModel, Genetic Algorithm, Deduplication

I. Genetic Algorithm

In the domain of computer science of artificial intelligence, a genetic algorithm (GA) is a search heuristic that mimics the method of natural selection. This heuristic also known as metaheuristic. That is routinely used to create benefit explanation for search trouble and optimization.

GAs are stochastic search procedure based on the mechanism of natural selection and natural genetics to imitate living beings for reduce those complicated problems with high difficulty and or undesirable structure

Nature of GAs

- It employment among a coding of result set, not the result themselves
- Genetic algorithms search from a population of result, not a single result
- It use payoff information fitness function, not derivatives or other auxiliary knowledge
- Genetic algorithms use probability conversion rules,not deterministic rules
- GA exploit the best result while exploring the search space

Advantages of GA

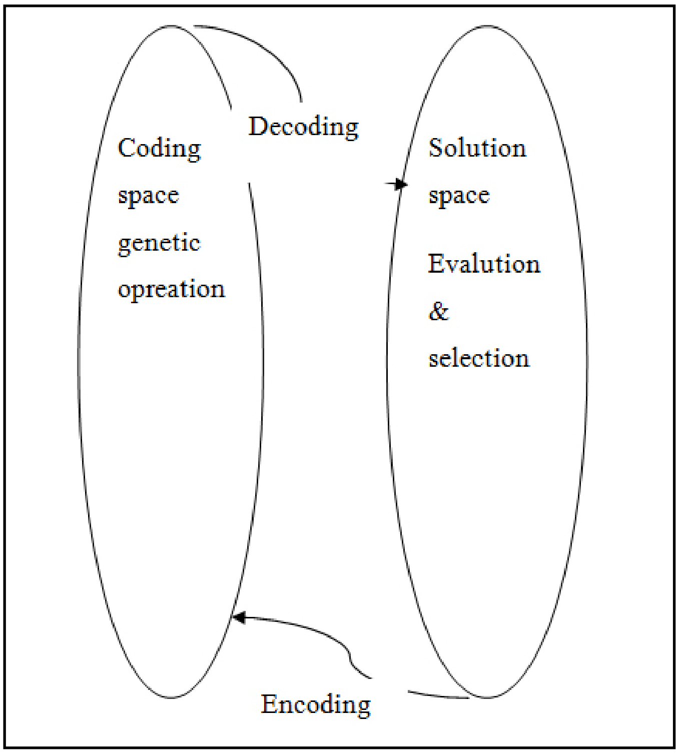
- Genetic algorithms do not have a large amount mathematical needs about the optimization problems
- The ergodicity of evolution operators formulate genetic algorithms very useful at performing global search in probability
- Genetic algorithms give us a great flexibility to hybridize with domain dependent heuristics to create an efficient presentation for a specific problem

Key components of GA

- Encoding / Decoding
- Crossover / Mutation
- Selection

Encoding / Decoding

How to encode a solution of the problem into a chromosome is a key to success



Genetic Operations Crossover Mutation

- The genetic procedure mimic the process of heredity of genes to create new offspring at every invention
- Crossover operates on two chromosomes at a time and produce offspring by mingle both chromosomes features
- The crossover rate defined as the ratio of the number of offspring produced in every generation to the population size
- A higher crossover rate allows exploration of more of the solution space at the cost of computations
- Effectiveness and feasibility problem
- Mutation is a background operator which generates spontaneous random changes in different chromosomes

1. replacing the genes lost from the population during the selection procedure so they can be tried in a fresh context
2. or giving the genes that were not here in the original population.

Evolution Operation Selection

The theory following genetic algorithms is fundamentally Darwinian natural selection. It convey a genetic algorithm search toward promising regions in the search space.

Basic Genetic Algorithms structure

```
Beginning (1)
        t = 0
   Initialization P(t)
   evalution P(t)
While (the stop contition is not verified) do
  Beginning (2)
        t = t + 1
        selection P'(t) from P(t-1)
      P''(t) ← crossover P'(t)
    P'''(t)  ← mutation P''(t)
   P(t) ← replacement (P(t-1),P'''(t))
```

II. Hidden Markov Model

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.

A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. An HMM can be presented as the simplest dynamic Bayesian network.

A statistical tool used for modeling generative sequences characterized by a set of observable sequences. The HMM framework can be used to model stochastic processes where

1. In a Markov process non-observable state of the system is governed
2. The observable sequences of system have an underlying probabilistic dependence.

Hidden Markov Model

HMM Model Parameters

$$\lambda = (\pi,A,B)$$

Three Basic Problems in HMMs

- Given a set of observation sequences O= and the HMM

  parameters **$\lambda = (\pi,A,B)$,** computing the probability **P().**
- Givenasetofobservationsequences O=andtheHMMparameters **$\lambda = (\pi,A,B)$,** ,computing the optimal state sequences
- Given a set of observation sequences O= adjusting the HMM parameters **$\lambda = (\pi,A,B)$** to maximize the probability **P()**

III. Applications

HMMs can be applied in many fields where the goal is to recover a data sequence that is not immediately observable .Applications include: Single Molecule Kinetic analysis, Cryptanalysis ,Speech recognition and Speech synthesis, Part-of-speech tagging, Document Separation in scanning solution, Machine translation and Partial discharge, Gene prediction, Alignment of bio-sequences, Time Series Analysis and Human Activity recognition and Protein folding

IV. 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 accleration coefficent. 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.