International Journal of Computer Science and Engineering (IJCSE)
ISSN(P): 2278-9960; ISSN(E): 2278-9979
Vol. 3, Issue 5, Sep 2014, 61-66


International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

# AN EXTENDED RANGE OF USE OF GENETIC PROGRAMMING APPROACH TO RECORD DEDUPLICATION

**SUPRIYA THATAVARTHI[1], GURU RAMANADHA BABU THOTA[2] & VIJAY SOWPATI[3]**

[1]M.Tech Student, Department of PG (CSE), Loyola Institute of Technology and Management, Sathenapalli, Guntur Affiliated to Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh, India

[2]Assistant Professor, Department of CSE, Loyola Institute of Technology and Management, Sathenapalli, Guntur, Andhra Pradesh, India

[3]Assistant Professor, Department of CSE, Satyam Learning Campus Institute of Engg. & Tech., Hyderabad, Telangana, India

## ABSTRACT

The task of recognizing, in a data warehouse, account that pass on to the matching real world entity regardless of misspelling words, kinds, special writing styles or even unusual schema versions or data types is called as the record deduplication. In presented research [1] they offered a genetic programming (GP) approach to record deduplication. [2] Their approach combines several different parts of substantiation extracted from the data content to generate a deduplication purpose that is capable to recognize whether two or more entries in a depository are duplications or not. Because record deduplication is a time intense task even for undersized repositories, their aspire is to promote a method that discovers a proper arrangement of the best pieces of confirmation, consequently compliant a deduplication function that maximizes performance using a small representative portion of the corresponding data for preparation purposes also the optimization of process is less. Our research deals these issues with a novel technique called modified bat algorithm for record duplication. The incentive behind is to generate a flexible and effective method that employs Data Mining algorithms. The structure distributes many similarities with evolutionary computation techniques such as Genetic programming approach [1]. this scheme is initialized with an inhabitant of random solutions and explores for optima by updating bat inventions. Nevertheless, disparate GP, modified bat has no development operators such as crossover and mutation. We also compare the proposed algorithm with other existing algorithms, together with GP from the experimental results.

**KEYWORDS:** Database Integration, Data Sets, Data Ware House, Evolutionary Computing and Genetic Algorithms