Abstract

Normally setting the threshold is an important issue in applications where the similarity functions are used and it relies more on human intervention. The proposed work addressed two issues: first to find the optimal equation using Genetic Algorithm (GA) and next it adopts an intelligence algorithm, Particle Swarm Optimization (PSO) to get the optimal threshold to detect the duplicate records more accurately and also it reduces human intervention. Restaurant and CORA data repository are used to analyze the proposed algorithm and the performance of the proposed algorithm is compared against marlin method and the genetic programming with the help of evaluation metrics.

References

Automatic Threshold Selection using PSO for GA based Duplicate Record Detection

- Arthur D. Chapman, Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data, version 1.0, Report for the Global Biodiversity Information Facility, Copenhagen, 2005.
- A. E. Monge and C. P. Elkan, The field matching problem: Algorithms and
Automatic Threshold Selection using PSO for GA based Duplicate Record Detection

applications; SIGMOD workshop on research issues on knowledge discovery and data mining, pp. 267-270, 1996.

**Index Terms**

<table>
<thead>
<tr>
<th>Computer Science</th>
<th>Information Sciences</th>
</tr>
</thead>
</table>

**Keywords**

GA  PSO  Similarity metrics  Threshold