CHAPTER 3

PROBLEM STATEMENT AND OBJECTIVE

3.1 INTRODUCTION

The problem taken up is to improve the performance of detecting duplicate record. The performance of entropy based duplicate detection could be enhanced through various means. It could be in the form of predictive accuracy, comprehensibility, speed, and scalability. This research concentrates on the performance enhancement of duplicate detection through Multi-Level clustering technique and implemented it by using decision tree framework.

3.2 PROBLEM DEFINITION

The goal of this work is to identify groups of similar entities in the presence of linked environment and the searching methods should reduce the number of unwanted comparisons during de-duplication. It will maximize the performance of data de-duplication. But de-duplication and group detection of similar entities have mostly been dealt with separately and as an unrelated problem. We argue that the two problems occur in most real world application and this calls complete frame work for addressing them. In order to achieve this goal, in this paper I propose a new technique. First one is a clustering algorithm which will overcome the existing clustering dis-advantage, that may be either partition or hierarchical. Second one is de-duplication algorithm which will
produce the dis-similarity percentage of the pair of string in each clustered group.

In this thesis I give an investigation of the viability of a Multi-Level Clustering based de-duplication algorithm compared within the group. I show that duplicate data does exist in cluster groups and an entropy-based de-duplication algorithm is a viable option for de-duplication. I offer the design and analysis of different algorithms used to perform data de-duplication. I show through the use of different data sets that it is possible to perform de-duplication on archival data, static primary data and dynamic primary data. Finally I present a full scale implementation using Multi-Level Clustering algorithms, de-duplication algorithm and report the algorithm’s impact on de-duplication activity.

3.3 OBJECTIVES

The aim of the research work is to enhance the performance of the existing classification technique. The objectives of this research are:

- Enhancing the performance of the classification algorithm that supports the mixed attributes and generates a group with most closely related object.

- Explore the search space effectively by creating a group between Intensification and diversification.

- Developing a hybrid algorithm by combining the advantages of classification based algorithm and entropy technique for duplicate detection.
• Compare the performance of the algorithm with existing algorithm, both experimentally and statistically with respect to
  • Performance
  • Reduce Number of record comparison
  • Memory Usage
  • Accuracy