CHAPTER 7

CONCLUSION AND SCOPE FOR FUTURE WORK

There are several approaches to improve the performance of the duplicate detection algorithm. This research made one such attempt that exploited the natural and biological phenomenon to bring about invention of new algorithm and performance amelioration in duplicate detection algorithm. The work in this thesis presented the application of bio-inspired concept in duplicate detection algorithm. The main goal of this research work was to develop a dis-similarity calculation algorithm that supported mixed attribute and multiple tokens. In general, a data set contains both nominal and continuous attribute. There is a need to develop a dis-similarity calculation algorithm for various kinds of attribute.

7.1 CONCLUSION

A dis-similarity calculation algorithm was developed by using entropy and information gain method for discovering the dis-similarity value for various kinds of attribute and various kinds of tokens. The truth table construction was employed to represent the character into numeric phenomenon value for various kinds of attribute. Entropy was used as a measure of how “mixed up” an attribute is and also here it is considered purity and impurity of a variable.
The quality of algorithm was measured by using rule confidence threshold value. Two different algorithms were employed to exploit and explore the search space efficiently and to find the optimal exploring rate for each data set. ADTree_Group Induction Algorithm() procedure was used to find group in the given dataset and DS_Calculation() procedure was used to find dis-similarity value. The proposed algorithm produced better results for various kinds of attribute than that of the existing algorithm.

A group induction algorithm (ADTree_groupInductionAlgorithm()) was developed by coalescing the merits and features of Hierarchical Clustering, Group Detection, and Multi-Level Clustering technique. The algorithm consists of a two-step process. In the first phase, a precondition was created by using the standardization and hygiene concept. In the second phase the continuous attribute was added with vector by using the excellence of short name index and if short name index did not exist, then a new vector by using short name as index was created.

This technique improved the group object accuracy and iteration capability of duplicate detection. This group induction algorithm executed the two phases in sequential order one by one. The group induction algorithm produced better results for various attribute than the without grouping algorithm for duplicate detection. So here is conclusive evidence that the group induction algorithm produces a good decision making system for real life application.

Dis-Similarity calculation algorithm (DS_Calculation) was also developed to identify the dis-similarity percentage between pair of string. It is integrated with the group induction algorithm to improve the performance and scalability of both algorithms. It has produced a more accurate result when the
attribute had more number of characters. This approach not only reduces the elapsed time but it also generates more accurate result and minimizes the number of iteration and the number of comparison. It is also simple to make a good decision. As the various combination values were defined to the parameter, several interesting phenomena were observed. From the observations, it is found that the number of comparison and number of iteration are fully depending on the characteristics of the data set. The scalability of the algorithm was measured by using various sector dataset. The result clearly indicated that the algorithm produced more accurate result for large data set with minimal time consumption.

In this research, we have proposed group detection algorithm and dis-similarity calculation algorithm that take into account entity relationships for the purpose of grouping similar entities in linked environments. We show how these measures are useful for the important data mining and data warehousing tasks of data de-duplication and group detection.

We argue that the two problems are quite related and use a unified generative model for link-data to evaluate our approaches. We present comparisons of the different distance measures for varying data characteristics and results that show significant improvements over algorithms based on attributes.

7.2 SCOPE FOR FUTURE WORK

The present study can be extended in the following direction:

The record hiding concept may be adopted to hide sensitive data to maintain the privacy of data. Future work will involve looking into ways to improve the scalability and to combine different de-duplication approaches
into a cloud computing system. The algorithm may be extended to handle missing values in a natural way during grouping. To solve uncertainty grouping problem, the association rule mining may be integrated into grouping algorithm. This algorithm may be extended to grouping and to identify duplicate in web document and image text.

A present work won’t support for encrypted or image kind of text. Future we may extend this work for support encrypted and image kind of text.