CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENT

6.1 SUMMARY

Schema matching is the process of identifying semantically related objects and also it plays a vital role the architecture of data integration. The actual process of schema matching is, the source elements are mapped with the target schema matching elements. It has a great part in enterprise information integration and also is a well-known data management research topic, mainly in building data warehouses.

In this thesis, two automated schema matching techniques such as an Efficient Centroid Based Clustering Algorithm (ECBCA) and Hybrid Hill Climbing with Genetic Algorithm (HHCGA) approach to improve the schema matching automation. The performances of the proposed schema matching algorithms are compared with the existing schema matching algorithms such as Kang-Naughton Entropy only (KNE), Kang-Naughton Mutual Information (KNMI), K-Means Clustering and modified Kang-Naughton Genetic Algorithm (KNGA).

6.2 CONTRIBUTIONS

An Efficient Centroid Based Clustering Algorithm (ECBCA) is proposed to improve the schema matching automation technique using the existing K-Means clustering algorithm. According to the proposed algorithm, clustering of schemas is achieved on the basis of their contextual similarity,
attributes of the schemas are clustered that are in the same schema cluster to find attribute correspondences between these schemas. To achieve this, the existing K-Means clustering algorithm is modified in centroid section to improve the schema matching automation.

Hybrid Hill Climbing and Genetic Algorithm (HHCGA) is proposed with the two main phases, Dependency Graph Generation and Dependency Graph Matching. For generating the Dependency Graph from the schema of table instances, the properties Entropy and Mutual Information are initially found from these schemas. Entropy property obtains for each and every element of the schema, separately, which is independent of the actual values of attributes. Mutual Information property measures by the correlation measure, Mean-Square Contingency Co-efficient between two elements of a schema. These two properties are separately computed for each schema and then the Dependency Graphs are generated separately for each schema of table instances. For these two Dependency Graphs, the matching of schema of table instances is identified using Hybrid Hill Climbing and Genetic Algorithm (HHCGA). Integrating hill climbing approach with the Genetic Algorithm will considerably minimize the search space and improves the performance metrics such as Precision, Recall etc.

The performance metrics that are evaluated for the proposed ECBCA and HHCGA schema matching algorithms are Precision and Recall, Computation time, Sensitivity, and Specificity. Real-world datasets were used from two different data domains: medical lab data and census data. The experiments run while increasing the number of attributes in two input tables to be matched. For each table width, two to 24, we iterated the measurement 30 times with randomly chosen subsets of attributes and averaged the results.

From the experimental results, it is inferred that the proposed ECBCA and HHCGA algorithms improve the schema matching automation
by considerably improving the Precision, Recall, Computation Time, Sensitivity and Specificity performance metrics. For the Census dataset, the proposed ECBCA and HHCGA improve the precision by 54.02 % and 57.86% over the existing KNMI algorithm. Similarly, when compared to the existing KNGA and K-Means algorithms, the proposed HHCGA improves precision by 11.93 % and 17.41 % respectively. In the case of proposed ECBCA, 14.56 % improvement in precision over the existing K-Means algorithm for the census data. Among the proposed ECBCA and HHCGA algorithms, HHCGA outperforms precision by 2.49 % over the ECBCA.

Average computation time for the proposed ECBCA and HHCGA are 8051 ms and 7113 ms; but the existing algorithms such as KNE, KNMI, K-Means and modified KNGA show about 9725 ms, 23298 ms 11709 ms and 8810 ms respectively for the census dataset. Similarly for the medical lab dataset, the proposed HHCGA, ECBCA and existing KNE, KNMI, K-Means and modified KNGA show about 8423 ms, 87933 ms, 10631 ms, 25372 ms, 10059 ms and 9433 ms respectively.

As the proposed HHCGA algorithm uses both hill climbing to find the local matches and GA for finding the global matches using the local matches, it considerably improves the sensitivity performance. From the experimental results, the HHCGA improves sensitivity by 13.77 %, 9.11 %, 7.33 %, 2.81 % and 1 % over the existing KNE, KNMI, K-Means, modified KNGA algorithms and the proposed ECBCA algorithm for Census dataset.

6.3 FUTURE ENHANCEMENT

The thesis proposed two different schema matching automation algorithms to improve the performance metrics such as Precision, Recall, Computation Time, Sensitivity and Specificity. To make the automation of the proposed algorithms more efficient and feasible to the practical
implementation of the network, it is suggested to further enhance the work by the following components: The work can be further enhanced by increasing the number of attributes from 24 to 100, Sample data size may be increased by 10k tuples to few 100k tuples, Genetic Algorithms proposed in this thesis may be integrated with some meta heuristic techniques such as Particle Warm Optimization, Ant colony algorithm etc, Optimized experimental results can be achieved more by increasing the total number of average, Cluster size may be increased and the experiments can also be analyzed with individual cluster.