ABSTRACT

For provision of integrated data access or sharing among autonomous, heterogeneous and distributed databases automatic resolution of schema heterogeneity still remains a major bottleneck. In order to provide transparent access to remote data and to provide the sharing of information among databases, their schema heterogeneity is to be identified and resolved and then the correspondences among schemas is to be identified. Schema matching is a crucial process in this data integration process. A schema matching process accepts two or more schemas as input and produces a mapping between the input schemas which correspond semantically to each other. There are numerous technical challenges with automated schema matching and integration tools which are an active area of research still. This research work proposes to improve the automated schema matching by developing a schema matching technique using Efficient Centroid Based Clustering Algorithm (ECBCA) and an optimal Hybrid Hill Climbing approach with Genetic Algorithm (HHCGA).

An Efficient Centroid Based Clustering Algorithm is proposed to improve the schema matching automation by using the existing K-Means clustering algorithm enhancement in centroid section. 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. In the proposed ECBCA schema matching automation algorithm, the global
optimum is achieved through multiple splitting using the traditional k means clustering algorithm was improved by means of the centroids selection. And also the number of clusters is automatically generated so we can get the best result compared with the traditional k means clustering algorithm. From the automatic assigning process we can get the correct matched data so that the effectiveness of the proposed technique is improved. Moreover proposed ECBCA technique shows more than one matched data compared to the traditional k means clustering algorithm. In the proposed method, distance is inversely proportional to cluster size and vice versa. The performance metrics are evaluated by the experiments which use the census and medical lab real time datasets. Proposed ECBCA improves the Precision, Recall, Computation Time, Sensitivity and Specificity performance metrics by 14.56 %, 16.48 %, 32.65 %, 6.27 % and 32.29 % respectively over the existing K-Means clustering algorithm for census dataset.

An optimal Hybrid Hill Climbing approach with Genetic Algorithm (HHCGA) is proposed to improve the schema matching automation by integrating the Hill Climbing (HC) approach with the Genetic Algorithm (GA). It has two main phases, (i.e.) 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. Proposed HHCGA improves the Precision, Recall, Computation Time, Sensitivity and Specificity performance metrics by 57.86 %, 29.85 %, 48.72 %, 9.11% and 47.81 % over the existing Kang-Naughton Mutual Information (KNMI) schema matching algorithm.
Among the two proposed ECBCA and HHCGA schema matching algorithms, HHCGA outperforms the other due to the effective integration of hill climbing approach with genetic algorithm will have high precision of rejecting non-matching attributes and high precision of finding attribute matches.