ABSTRACT

The Web is proliferated with huge amount of data through various data sources. To retrieve useful information from the Web, data has to be collected from different data sources. The current search engines, retrieve data from multiple data sources and provide users with mere documents. The semantic web, an extension of the current web, facilitates the machines to process and interpret the data intelligently, rather than providing mere documents. The advent of XML paved a way for the standardization of data representation in the web. Semantic data integration is the process of interrelating data from diverse data sources that enable effective integration and reuse of information.

Ontology, which is developed by several domain experts is a formal specification of domain knowledge and it plays an imperative role in the semantic web. Ontology is the key technology in semantic web and is extensively used in data integration systems as they provide an explicit specification of domain knowledge. Ontologies are widely used in semantic data integration as they capture the domain knowledge. As the ontologies are developed by several domain experts, there exist structural, lexical and semantic differences in the terminologies used. Ontology mapping is the initial step in the integration of ontologies. It is the process of identifying the correspondence between the terms in the ontology. The alignment generated from the mapping process is useful in interpreting the semantics of the terminologies used in different data sources.

This research work aims to improve the quality of mapping by addressing a few critical issues. The quality of the mapping process is improved by using multiple strategies for computing the degree of

relatedness of the given terms. The similarity measure focuses on lexical, structural and semantic information represented in the ontologies. The performance of the proposed methods is evaluated using the precision, recall and f-measure metrics.

A Multi-strategy approach is used in this research for identifying the similarity between the given terminologies. MSFuzzy system, a rule based system for ontology mapping is proposed. The lexical and semantic similarity measures are used to compute the similarity value and the generated crisp set of values are fuzzified using the fuzzy inference engine. Based on the fuzzified values, the terms in the ontologies are mapped. This simple rule based system is found to improve the execution time; however it does not improve the quality of mappings as only a few distance measures are employed.

MS_SVMBoost algorithm, a hybrid statistical machine learning ensemble model is proposed to automate the process of ontology mapping. This algorithm uses multiple similarity measures as features. These features are used by SVM for generating the model and the performance of the system is better compared to the rule based system. However, the number of misclassified instances still exists and to further improve the performance the Boosting algorithm is incorporated. The simulation results indicate that the performance of this system is efficient compared to the existing systems.

Hybrid MS_ELM_LDA, a neural network based model to automate the process of ontology mapping is proposed. The learning model is trained using the similarity measures computed. The dimension of the data is reduced using the Linear Discriminant Analysis (LDA). The data after reducing the dimension is given as input to the Extreme Learning Machine (ELM), a feed-forward based neural network which is used to train the model

for mapping the terms in the ontology. The results obtained show that the proposed hybrid MS_ELM_LDA provides substantial performance improvement over the existing methods.

The biomedical ontologies developed are very large and thus require large number of comparisions to identify the correspondence between the mapping process. As the ontology mapping tasks are compute intensive, the MSMR_RF_CW system partitions the ontologies and further the candidate pair of ontologies are identified using the cosine similarity measure and then taken for deep analysis to identify the correspondence between the concepts in the ontology. Ontology partitioning is done using the cluster-walktrap algorithm. The most similar sub-ontologies are identified and the candidate sub-ontologies are subjected to the multi-strategy similarity computation. Random Forests, an ensemble based tree model is used to identify the mapping between the terminologies. The randomness associated in identifying the number of instances and the attributes generates different models and the collective decision of different models are considered for identifying whether the concepts are similar or not.

To further enhance the performance of the mapping system, a distributed environment is chosen and a MapReduce based programming model is used to parallelize the task. The similarity computations done are implemented as MapReduce tasks. The similarity score computed are aggregated and fed as features to the learning model employed. The MSMR_RF_CW MapReduce based distributed model shows substantial performance improvement over the existing systems. The benchmark datasets published by OAEI are used to evaluate the performance of the proposed system. The results obtained using MSMR_RF_CW show substantial improvement in performance when compared to the existing methods and all the other proposed techniques.