ABSTRACT

Schema matching is the task of finding semantic correspondences between elements of two schemas. It takes two schemas as input and returns a mapping that identifies corresponding elements in the two schemas. Schema matching is an important and vital step in many schema and data translation and integration applications, such as integration of Web data sources, data warehousing, XML message mapping etc. Most of the schema matching approaches employ a combination of linguistic and structural matching approaches to identify the semantic correspondences.

The linguistic and structural matching is done using a number of individual matchers. Each matcher performs the matching process using a particular schema matching technique. The more the number of individual matchers used, the better is the accuracy of the system. Most of the systems for matching schemas operate based on pre defined match rules. Very few systems are capable of learning to match schemas based given examples. The match accuracy of these systems are low compared to the rule based systems.

In this thesis, we have proposed two schema matching systems. The first system HMAT – A Hybrid Matching System for XML Schemas – is a simple path based matching system that operates on schema trees. This system has been designed with minimum number of individual matchers compared to other similar systems in the literature. The accuracy of this system is also quite high compared to other similar systems. The second system uses a decision tree based approach to match two XML schemas. Experimental results reveal that the proposed approach is efficient in learning the match characteristics and reliable in predicting the similarity between two XML schemas.