Minor Research Project - Summary

Design & Development of Efficient Data Cleaning Algorithm for Biological Database

Dr.V.Bhuvaneswari, Asst. Prof, Dept. of Computer Applications

The objective of the work is to clean the biological artifacts that exeunt in the genomic and proteomic databases in xml form A Framework is proposed to clean the artifacts using the data mining clustering technique. The pre-processing phase the attributes are retrieved from xml documents using xquery. The clustering algorithm is applied in two different approaches one to group documents syntactically and then group documents semantically. The documents are filtered based on the path length using xquery to group syntactically. The semantic similarity is done based on keyword based approach for structurally similar documents. The rules are inferred for the clustered documents based on keywords using association rule mining. The duplicate documents are verified using the keyword based approach , and it is compared with BLAST approach.

The Dataset used in experiment is downloaded from SwissProt for human taxonomy and E.Coli in xml format and also used the GO Ontology recent download 2009 to verify the clusters generated based on the functionality of genes described in the second approach. The framework is validated using the precision and recall measures. The rules are inferred using the association rule mining technique. It is inferred from the study that GO terms groups similar documents which has high content similarity in terms of functionality that the first approach.