An Ontology-based and Distributed KDD Model for Biomedical Sources

David Perez-Rey, Alberto Anguita, Jose Crespo, PhD, Victor Maojo, MD, PhD
Biomedical Informatics Group, Artificial Intelligence Lab. Facultad de Informática, Universidad Politécnica de Madrid, Spain

Abstract

Knowledge discovery approaches in modern biomedical research usually require to access heterogeneous and remote data sources in a distributed environment. Traditional KDD models assumed a central repository, lacking mechanisms to access decentralized databases. In such distributed environment, ontologies can be used in all the KDD phases. We present here a new model of ontology-based KDD approach to improve data preprocessing from heterogeneous sources.

Introduction

Traditionally, to create new data mining algorithms has been the main challenge for informaticians in knowledge discovery processes. In biomedical research, distributed environments demand new KDD approaches. In such context, ontologies provide a structural and semantic support to manage heterogeneous data source collections.

Methods

Previous own works in ontology-based database integration and preprocessing had been combined to provide a new, global and distributed KDD model. Figure 1 shows the updated KDD methodology.

Two ontology applications are identified within this model: (i) preprocessing ontologies for instance level integration and preprocessing and (ii) virtual schemas for schema level integration. Both can be manually or semi-automatically generated (always supervised) and unified. The system allows the association of various ontologies to each source. Once incorporated into the system, these ontologies drive data transformations when users launch a query. The approach has been tested with different biomedical databases, improving final KDD results.

Conclusions

The complexity of the different data sources in genomic medicine and their semantic variability requires new approaches in KDD research. This paper deal with such heterogeneity in distributed environments providing a novel and global approach.

References


Figure 1. New model of distributed and ontology-based KDD for biomedical sources.