

BioDARA : Data summarization approach to extracting Bio-medical structuring information

Abstract

Problem statement: Due to the ever growing amount of biomedical datasets stored in multiple tables, Information Extraction (IE) from these datasets is increasingly recognized as one of the crucial technologies in bioinformatics. However, for IE to be practically applicable, adaptability of a system is crucial, considering extremely diverse demands in biomedical IE application. One should be able to extract a set of hidden patterns from these biomedical datasets at low cost. Approach: In this study, a new method is proposed, called Bio-medical Data Aggregation for Relational Attributes (BioDARA), for automatic structuring information extraction for biomedical datasets. BioDARA summarizes biomedical data stored in multiple tables in order to facilitate data modeling efforts in a multi-relational setting. BioDARA has the advantages or capabilities to transform biomedical data stored in multiple tables or databases into a Vector Space model, summarize biomedical data using the Information Retrieval theory and finally extract frequent patterns that describe the characteristics of these biomedical datasets. Results: the results show that data summarization performed by DARA, can be beneficial in summarizing biomedical datasets in a complex multi-relational environment, in which biomedical datasets are stored in a multi-level of one-to-many relationships and also in the case of datasets stored in more than one one-to-many relationships with non-target tables. Conclusion: This study concludes that data summarization performed by BioDARA, can be beneficial in summarizing biomedical datasets in a complex multi-relational environment, in which biomedical datasets are stored in a multi-level of one-to-many relationships.