

A method for identifying the interesting abstracts from a relevant collection

Theodoros Soldatos, PhD Student

EMBL Heidelberg, Meyerhofstraße 1, 69117 Heidelberg, Germany
soldatos@embl.de

In this paper we develop a system to extract interesting literature (abstracts) based on a training that uses a predefined set of interesting and non-interesting abstracts. A classifier (support vector machine) learns how to perform the separation using the terms found in the abstracts. The user can select the terms to be used, changing thus from case to case the concept or the context that the classifier should learn. The text mining implementation is based on the AKS2 system (from Bioalma bioinformatics company, <http://www.bioalma.com>), which is a biological knowledge system that manages biological terms and other information extracted directly from the scientific literature. With Arena Text one can collect more interesting abstracts about a subject and expand the relevant information. Arena Text can also be trained to identify the part of a collection that refers to specific context requirements. Furthermore, Arena Text allows the connection of different abstract collections based on a context that they share. With this way the knowledge about biological questions, phenomena, mechanisms, procedures that have been studied well in specific contexts (e.g. diseases, organisms) can be applied to others that are not well studied. Results from two different biological tasks are presented. The performance of the system is evaluated with a wide range of information theoretic measures.