Automatic Extraction and Classification of Bioinformatics Publications through a MultiAgent System

Armano G, Manconi A, Vargiu E

Deparment of Electrical and Electronic Engineering, University of Cagliari, Cagliari

Motivation

A growing amounts of information are currently being generated and stored in the World Wide Web (WWW). Digital archives, like PubMed Central, or online journals, like BMC Bioinformatics, are more and more searched for by bioinformatics researchers to download papers relevant to their scientific interests. These services provide search and browsing facilities based on the papers' list of references. However, for a researcher, especially for a beginner, it is still very hard to determine which papers are in fact of-interest without an explicit classification of the relevant topics s/he is involved in. In our view, personalization and effective information-filtering techniques are primary features to be provided. In fact, beyond conventional search engines, users need specific tools and methods for an effective use of all the available scientific resources. In this work we present a multiagent system explicitly devoted to extract information from heterogeneous sources, and classifying them using text categorization techniques.

Methods

Searching for publications involves two main activities: information extraction and text categorization. To this end we devised a multiagent system upon the generic PACMAS architecture. PACMAS, which stands for Personalized, Adaptive, and Cooperative MultiAgent Systems, is a multiagent architecture aimed at retrieving, filtering and managing information among different and heterogeneous information sources. PACMAS agents are autonomous and flexible, and can be personalized, adaptive and cooperative, depending on the given application. The overall architecture encompasses four main levels (i.e., information, filter, task, and interface), each being associated to a specific role. The communication between adjacent levels is achieved through suitable middle agents, which form a corresponding mid-span level. At the information level, agents play the role of wrappers, each one being associated to a different information source. In the current implementation, agents wrap information sources that provide scientific publications; i.e., BMC Bioinformatics site and PubMed web services. Furthermore, an ad-hoc agent wraps a suitable taxonomy extracted from the TAMBIS ontology. At the filter level, a population of agents is devoted to manipulate the information through suitable filtering strategies according to classical text categorization techniques. In particular, a set of filter agents removes all non-informative words such as prepositions, conjunctions, pronouns and very common verbs by using a standard stop-word list. A set of filter agents performs a stemming algorithm to remove the most common morphological and inflexional suffixes from all the words. For each class, a set of filter agents selects the features relevant to the classification task according to the information gain method. After selecting the terms, for each document a feature vector is generated, whose elements are the feature values of each term. The adopted feature value is the TF (Term Frequency) x IDF (Inverse Document Frequency) measure. At the task level, a population of agents has been developed, each of them embedding a classifier. In the current implementation, each agent embed a kNN classifier Task agents have be trained in order to recognize a specific class, each class being an item of the adopted taxonomy. Given a document in the test set, each agent, through its embedded classifier, identify the category of the input document. Task agents are also devoted to measure the classification accuracy according to the confusion matrix. At the interface level, agents are aimed at interacting with the user. Interface agents are also devoted to handle user profile and propagate it through the middle agents. At least in principle an interface agent might also adapt to the changes that occur in the preferences and interests of the corresponding user through a suitable feedback mechanism.

Results

To evaluate the effectiveness of the system, a suitable training dataset is required. To this end, an online support (http://iasc.diee.unica.it/bioclassifier/index.jsp) has been provided to classify and subsequently collect articles according to the adopted taxonomy. Currently, preliminary tests has been performed using a small number of publications classified by an expert of the domain according to the first level of the proposed taxonomy. Let us pointing out that, particular care has been taken in limiting the phenomenon of false negatives (FN), which --nevertheless-- had a limited impact on the percent of false positives (FP). In particular, the ratio FN/(FN+FP) has been kept under 25% by weighting positive prototypes with an additional factor of 1.05 with respect to negative ones. This preliminary experimental results are encouraging and point to the validity of the proposed approach .

Contact email: vargiu@diee.unica.it