A MultiAgent System for Retrieving Bioinformatics Publications from Web Sources

A. Addis, A. Manconi, M. Saba, and E. Vargiu

Intelligent Agents and Soft-Computing Group

DIEE – University of Cagliari (Italy)
Outline

- Introduction
- The Proposed MAS
- Experimental Results
- Conclusions and Future Work
Introduction
Motivations
Motivations

- Support the user through an automated system, able to:
  - *Retrieve* and *extract* information from heterogeneous sources
  - *Select* the contents really deemed relevant for the user, according to her/his *personal* interests
The Proposed MAS
Retrieving Bioinformatics Publications: main activities

Online sources

Information Extraction

Extracted publications

Text Categorization

Classified publications
The Proposed Approach

- A multiagent system able to:
  - *take* into account user’s needs and preferences (Personalization)
  - *adapt* to changes occurring in the environment (Adaptation)
  - *interact* with other agents and the user (Cooperation)
Implementation: The PACMAS Architecture

- A multiagent architecture designed to support the development of applications aimed at:
  - *Retrieving* heterogeneous data spread among different sources
  - *Filtering* and organizing them to personal interests explicitly stated by each user
  - Providing adaptation techniques to improve and refine user profile
Implementation: The PACMAS Architecture

Information Sources

- Information Level
- Filter Level
- Task Level
- Interface Level

Mid-span Levels
Retrieving Bioinformatics Publications: main activities

Online sources

Text Categorization

Information Extraction

Performed by agents belonging to the Information Level

Extracted publications

Classified publications
Information Extraction

- At the information level:
  - An agent wraps the BMC Bioinformatics site
  - An agent wraps the PMC web service
  - An agent wraps the adopted taxonomy
Information Extraction: BMC

- RSS is a family of web feed formats providing web contents and other metadata.
- An information agent is aimed at extracting information from a corresponding structured RSS source.
Information Extraction: PMC

- WSIG is a JADE add-on providing support for bidirectional interactions between web services and JADE agents (and JADE agent services from web service clients)

- An information agent is aimed at interacting with a corresponding web service using WSIG
Retrieving Bioinformatics Publications: main activities

Online sources

Information Extraction

Extracted publications

Text Categorization

Classified publications

Performed by agents belonging to the Filter and the Task Level
Text Categorization step by step

I. Disregarding *stop words*
II. Applying the *stemming algorithm*
III. Creating the *bag of words*
IV. *Creating the vocabulary*
V. Applying a *feature selection* technique
VI. Creating the *feature vector*
VII. *Classifying* the resulting document according to a predefined *taxonomy*
Text Categorization: the adopted taxonomy

Filter Agents

- At the filter level, agents:
  - remove all non-informative words by using a stop-word list
  - remove the most common morphological and inflexional suffixes by using a stemming algorithm
  - select the relevant features by using the information gain method
  - generate for each document a feature vector
Task Agents

- At the task level, agents:
  - embody a wkNN classifier
  - are trained to recognize a specific class, each class being an item of the adopted taxonomy
  - measure the classification accuracy
Interface Agent(s)
Experimental Results
Experimental Results

- Several tests have been performed, aimed at highlighting –and getting information about– the validity of the approach.
- We estimated the (normalized) confusion matrix for each classifier belonging to one of the two highest levels of the taxonomy.
Experimental Results

- Tests have been conducted using selected publications extracted from the BMC Bioinformatics site and the PubMed Central digital archive.
- Publications have been classified by an expert of the domain according to the first two levels of the proposed taxonomy.
Experimental Results

- For each item of the first and second level of the taxonomy:
  - a set of about 80-100 articles has been selected to the training phase
  - a set of about 200-300 articles have been used to the test phase
## Experimental Results

<table>
<thead>
<tr>
<th>Category</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Macromolecular Structure</em></td>
<td>0.95</td>
<td>1</td>
<td>0.9</td>
</tr>
<tr>
<td><em>Biological Structure</em></td>
<td>0.86</td>
<td>0.92</td>
<td>0.79</td>
</tr>
<tr>
<td><em>Chemical Structure</em></td>
<td>0.9</td>
<td>0.97</td>
<td>0.83</td>
</tr>
<tr>
<td><em>Molecular Compound Structure</em></td>
<td>0.87</td>
<td>1</td>
<td>0.74</td>
</tr>
<tr>
<td><em>Part of Physical Structure</em></td>
<td>0.86</td>
<td>1</td>
<td>0.71</td>
</tr>
<tr>
<td><em>Molecular Structure</em></td>
<td>0.87</td>
<td>1</td>
<td>0.74</td>
</tr>
<tr>
<td><em>Physical Organisation</em></td>
<td>0.87</td>
<td>1</td>
<td>0.74</td>
</tr>
<tr>
<td><em>Physical Space</em></td>
<td>0.88</td>
<td>1</td>
<td>0.76</td>
</tr>
</tbody>
</table>
Conclusions and Future Work
Conclusions

- We presented a system aimed at
  - retrieving publications from bioinformatics sources
  - classifying them using suitable machine learning techniques
- The system has been built upon PACMAS, a support for implementing Personalized, Adaptive, and Cooperative MultiAgent Systems
Future Work

- To implement...
  - more sophisticated classification algorithms
  - automatic composition of categories
  - suitable feedback mechanisms
That’s all folks!

Any question?