A MultiAgent System for Retrieving Bioinformatics Publications from Web Sources

A. Addis, A. Manconi, M. Saba, and <u>E. Vargiu</u>

Intelligent Agents and Soft-Computing Group

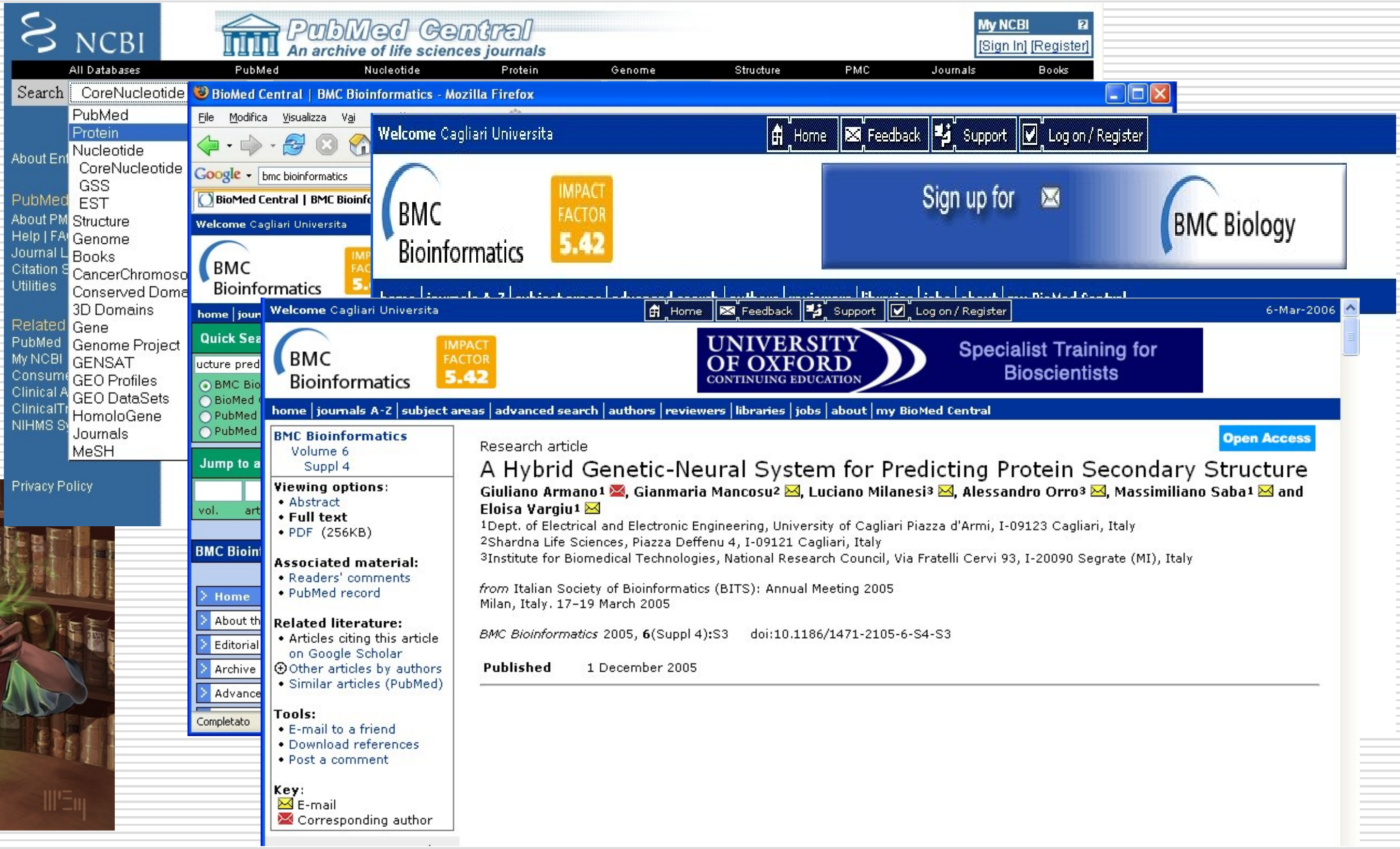
DIEE – University of Cagliari (Italy)

Outline

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

Introduction

Motivations



BMC Bioinformatics
Volume 6
Suppl 4

Viewing options:
- Abstract
- **Full text**
- PDF (256KB)

Associated material:
- Readers' comments
- PubMed record

Related literature:
- Articles citing this article on Google Scholar
- Other articles by authors
- Similar articles (PubMed)

Tools:
- E-mail to a friend
- Download references
- Post a comment

Key:
✉ E-mail
✉ Corresponding author

Research article

A Hybrid Genetic-Neural System for Predicting Protein Secondary Structure

Giuliano Armano[1] ✉, Gianmaria Mancosu[2] ✉, Luciano Milanesi[3] ✉, Alessandro Orro[3] ✉, Massimiliano Saba[1] ✉ and Eloisa Vargiu[1] ✉

[1]Dept. of Electrical and Electronic Engineering, University of Cagliari Piazza d'Armi, I-09123 Cagliari, Italy
[2]Shardna Life Sciences, Piazza Deffenu 4, I-09121 Cagliari, Italy
[3]Institute for Biomedical Technologies, National Research Council, Via Fratelli Cervi 93, I-20090 Segrate (MI), Italy

from Italian Society of Bioinformatics (BITS): Annual Meeting 2005
Milan, Italy. 17–19 March 2005

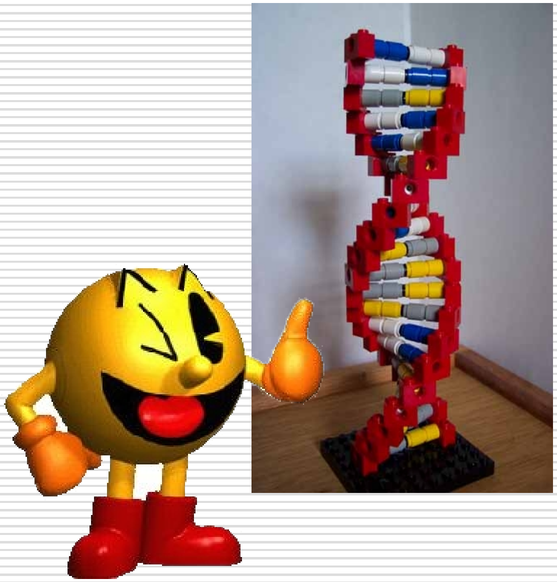
BMC Bioinformatics 2005, **6**(Suppl 4):S3   doi:10.1186/1471-2105-6-S4-S3

Published 1 December 2005

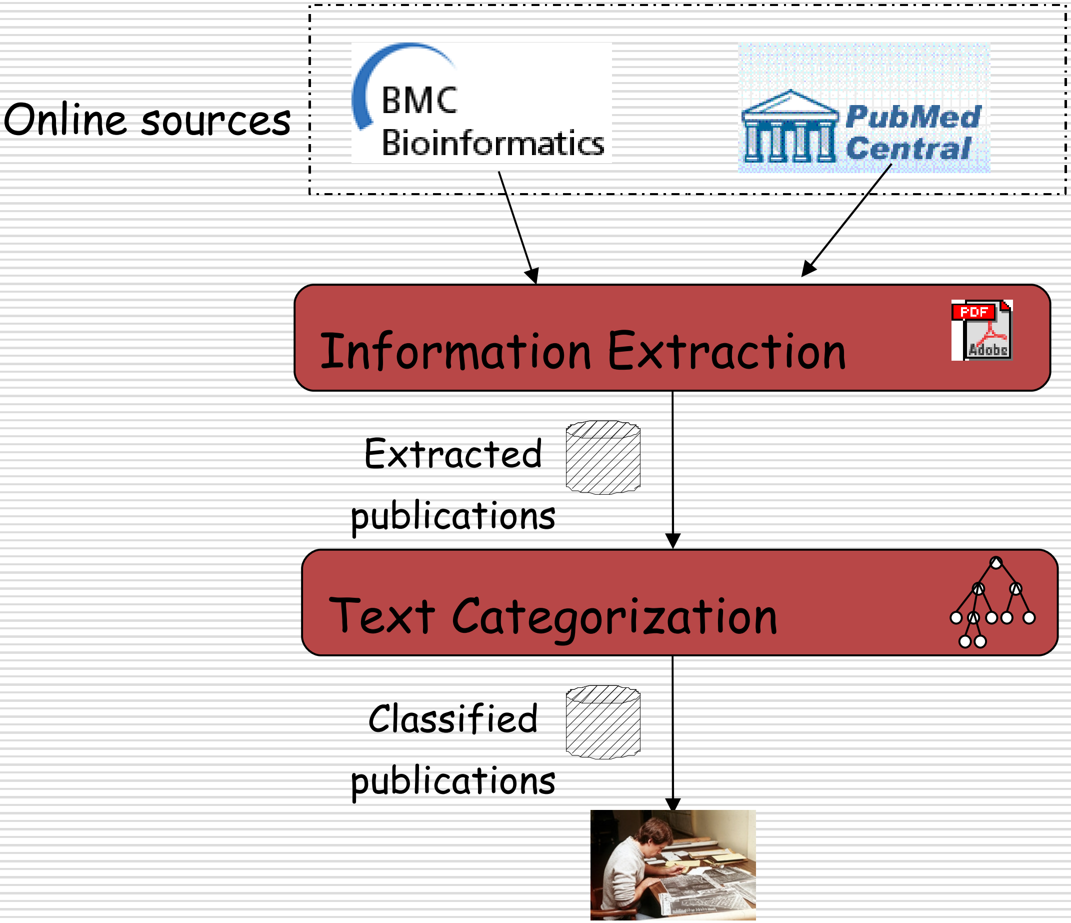
Motivations

- 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

BMC Bioinformatics

PubMed Central

Information Extraction

Extracted publications

Text Categorization

Classified publications

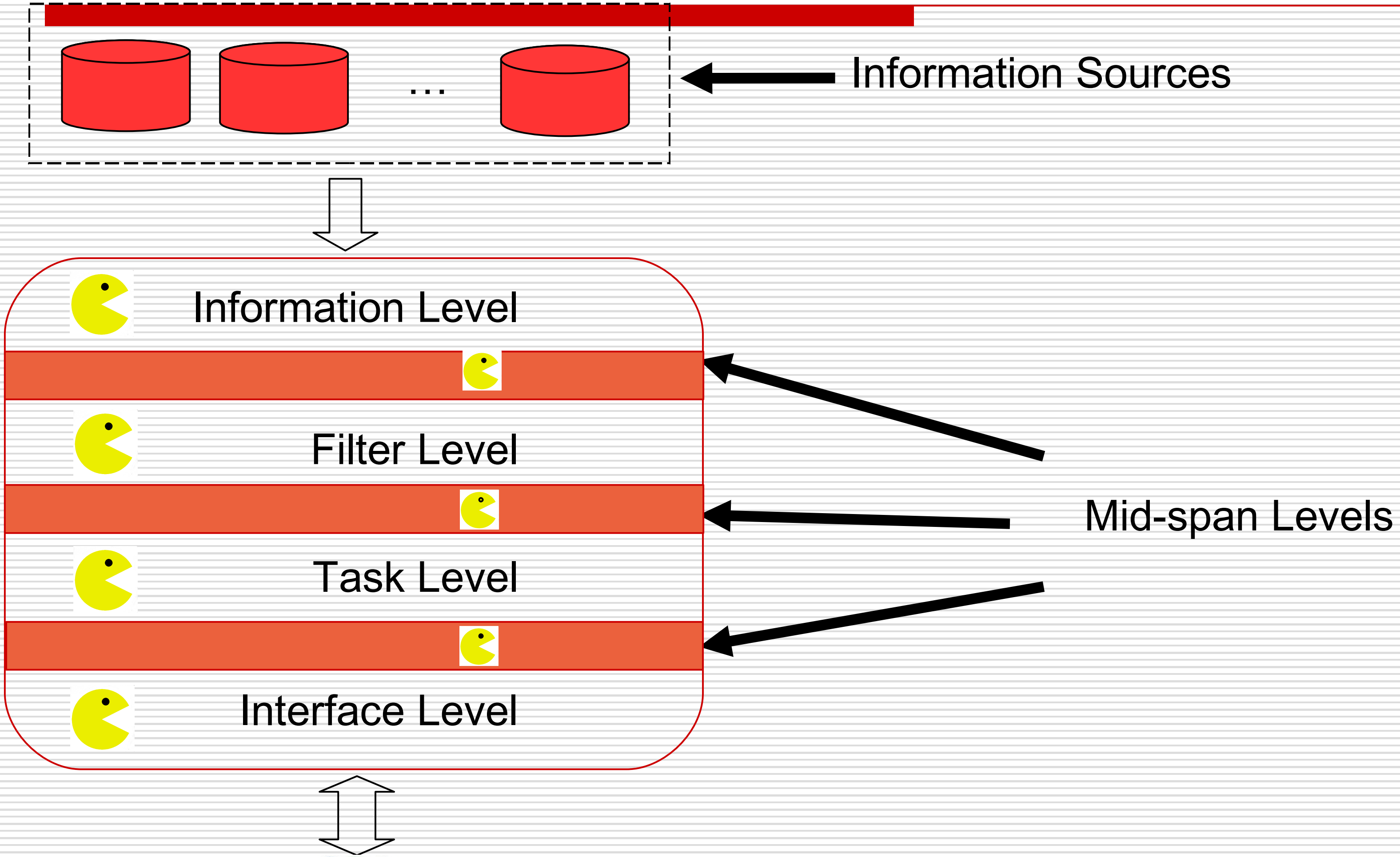
The Proposed Approach

- 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)

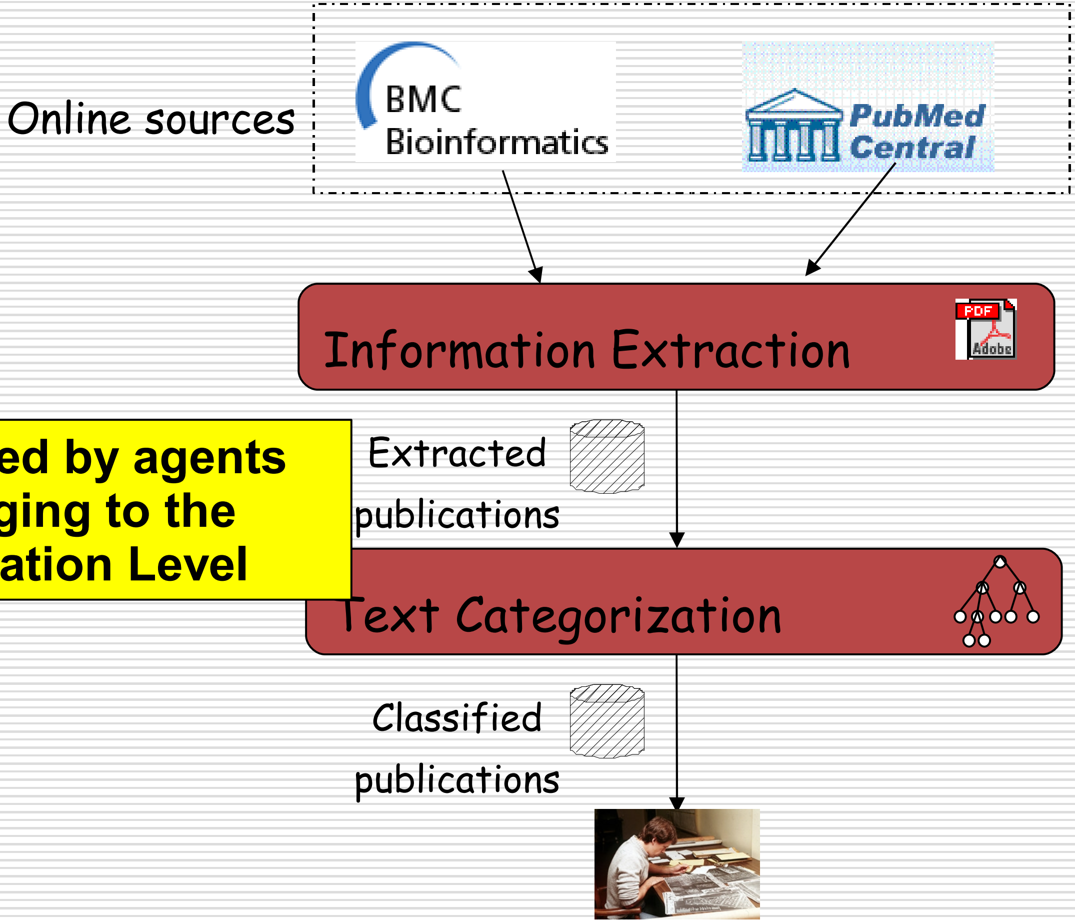
Implementation: The PACMAS Architecture

- 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



Retrieving Bioinformatics Publications: main activities



Online sources

Information Extraction

Extracted publications

Text Categorization

Classified publications

Performed by agents belonging to the Information Level

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

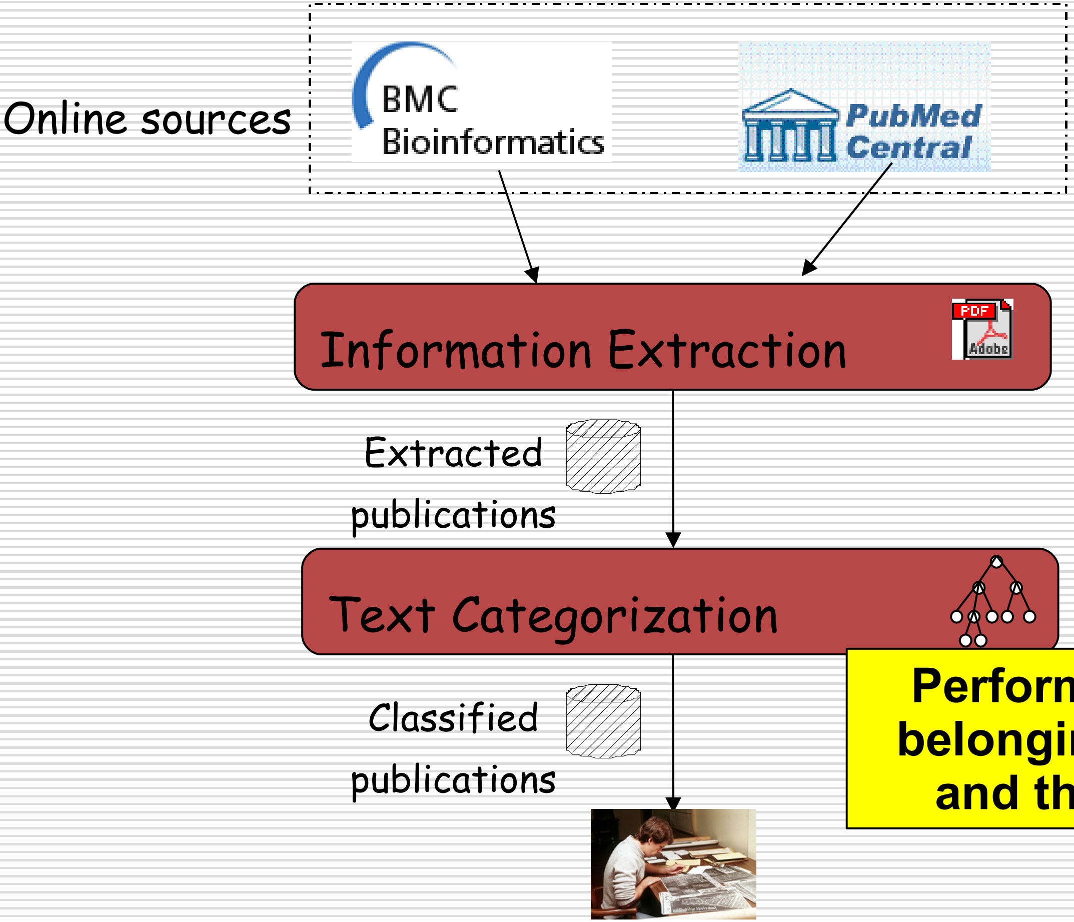
Information Extraction: BMC

- 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



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*

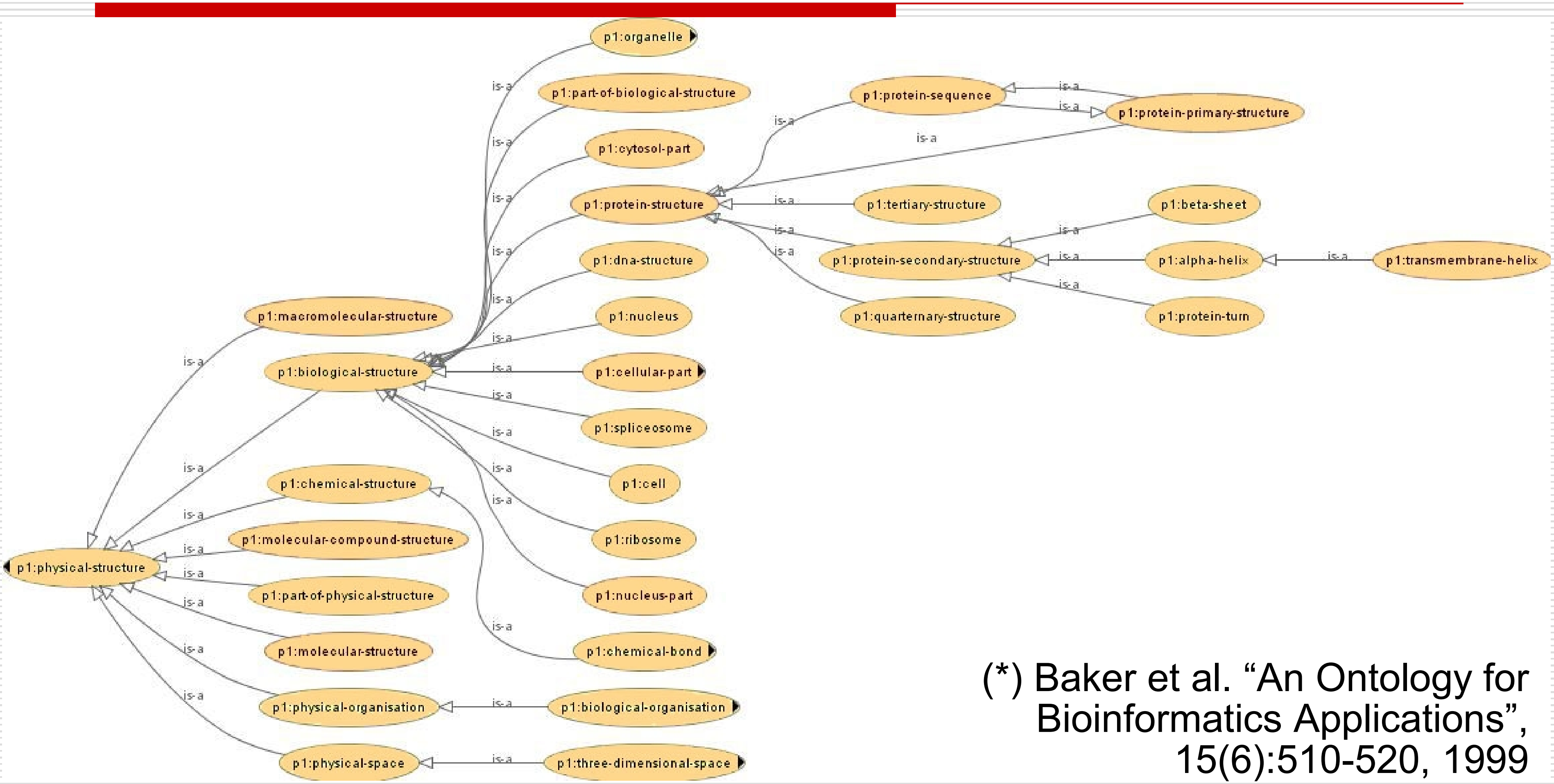
IV. *Creating the vocabulary*

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



(*) Baker et al. “An Ontology for Bioinformatics Applications”, 15(6):510-520, 1999

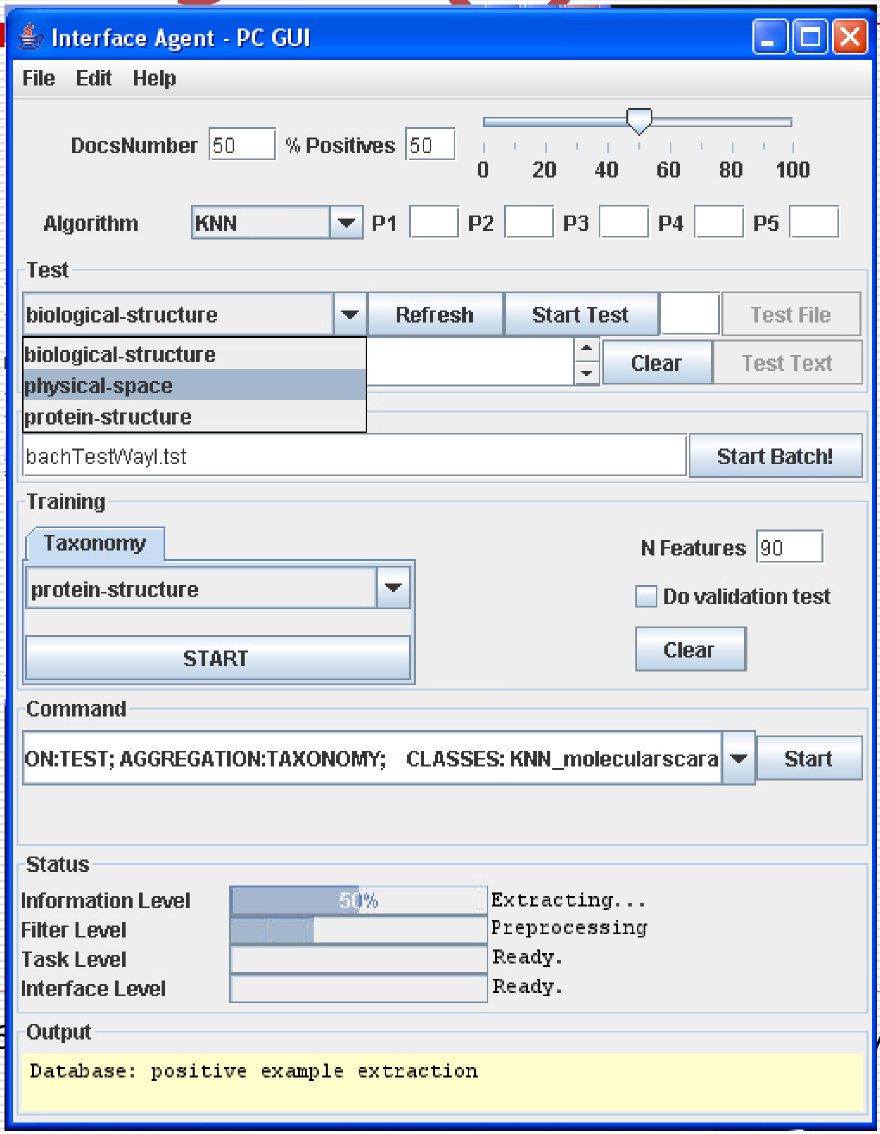
Filter Agents

- 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)



DocsNumber 50 % Positives 50

Algorithm KNN

Test

biological-structure

biological-structure
physical-space
protein-structure

bachTestWayl.tst

Training

Taxonomy

protein-structure

N Features 90

□ Do validation test

Command

ON:TEST; AGGREGATION:TAXONOMY; CLASSES: KNN_molecularscara

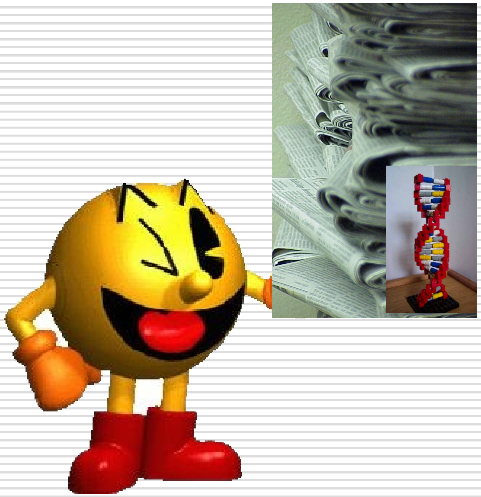
Status

```
Information Level   50%   Extracting...
Filter Level              Preprocessing
Task Level                Ready.
Interface Level           Ready.
```

Output

```
Database: positive example extraction
```

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

Category	**Accuracy**	**Precision**	**Recall**
Macromolecular Structure	0,95	1	0,9
Biological Structure	0,86	0,92	0,79
Chemical Structure	0,9	0,97	0,83
Molecular Compound Structure	0,87	1	0,74
Part of Physical Structure	0,86	1	0,71
Molecular Structure	0,87	1	0,74
Physical Organisation	0,87	1	0,74
Physical Space	0,88	1	0,76

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!