Textual information processing at the SWORD group

Fred Freitas
Centro de Informática
Universidade Federal de Pernambuco - Brazil
fred@cin.ufpe.br

Core axes

- Information gathering systems on restricted Web domains
  - MasterWeb and AGATHE
  - Extraction system WEEPAIES
- NLP-based extraction using GATE
- Ontology-Based Information Extraction (OBIE)
  - Snippet/metrics-guided OBIE
- Ontology population using ILP and deeper NLP
- Framework for blog crawler development

Core axes

- Information gathering systems on restricted Web domains
  - MasterWeb and AGATHE
  - Extraction system WEEPAIES
- Ontology-Based Information Extraction (OBIE)
  - Snippet/metrics-guided OBIE
- Ontology population using ILP and deeper NLP

An Ontology-based Information Gathering System

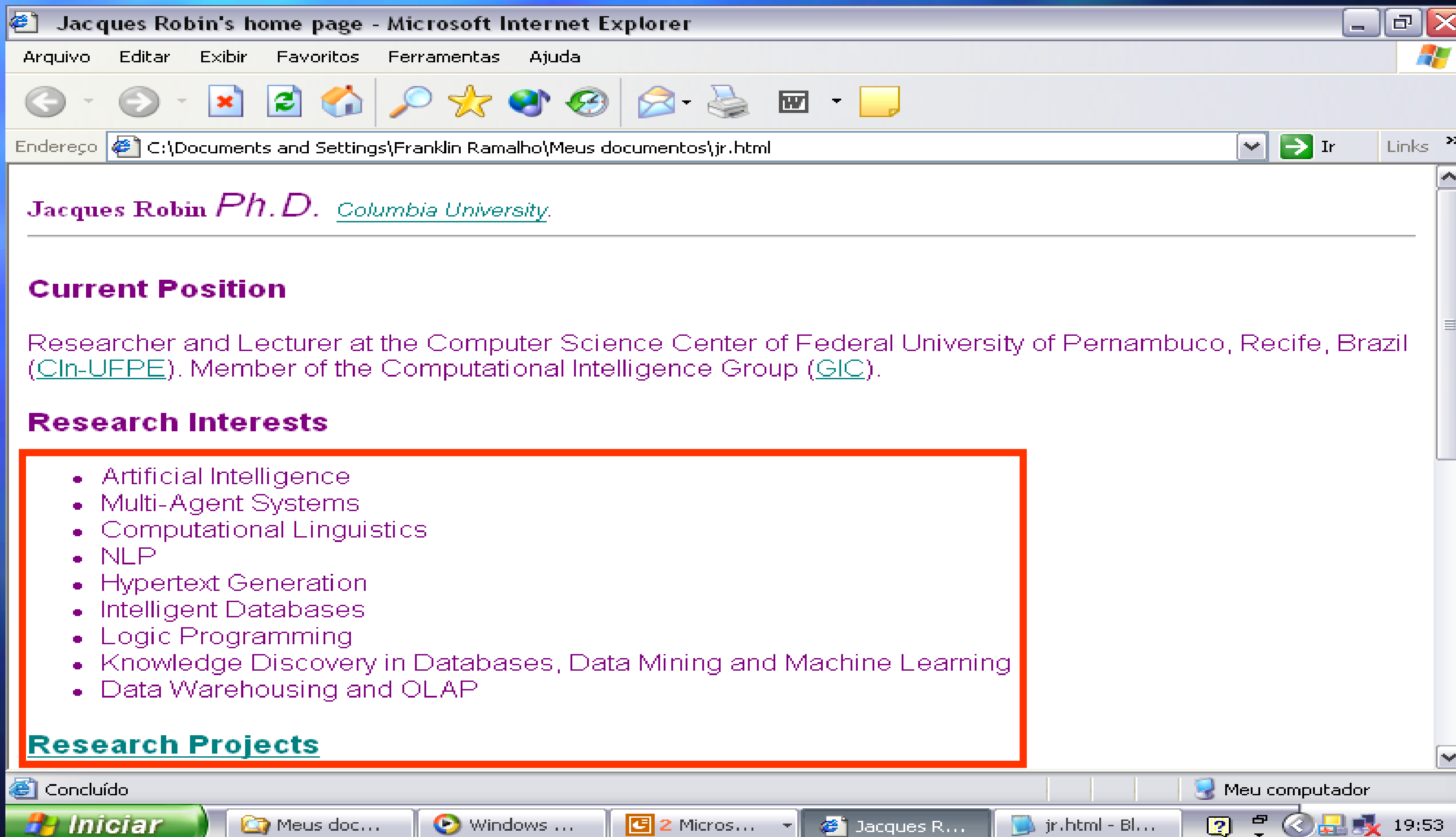
Fred Freitas
Universidade Federal de Pernambuco - Brazil
fred@cin.ufpe.br

Guilherme Bittencourt
Departamento de Automação e Sistemas
Universidade Federal de Santa Catarina – Brazil
gb@das.ufsc.br

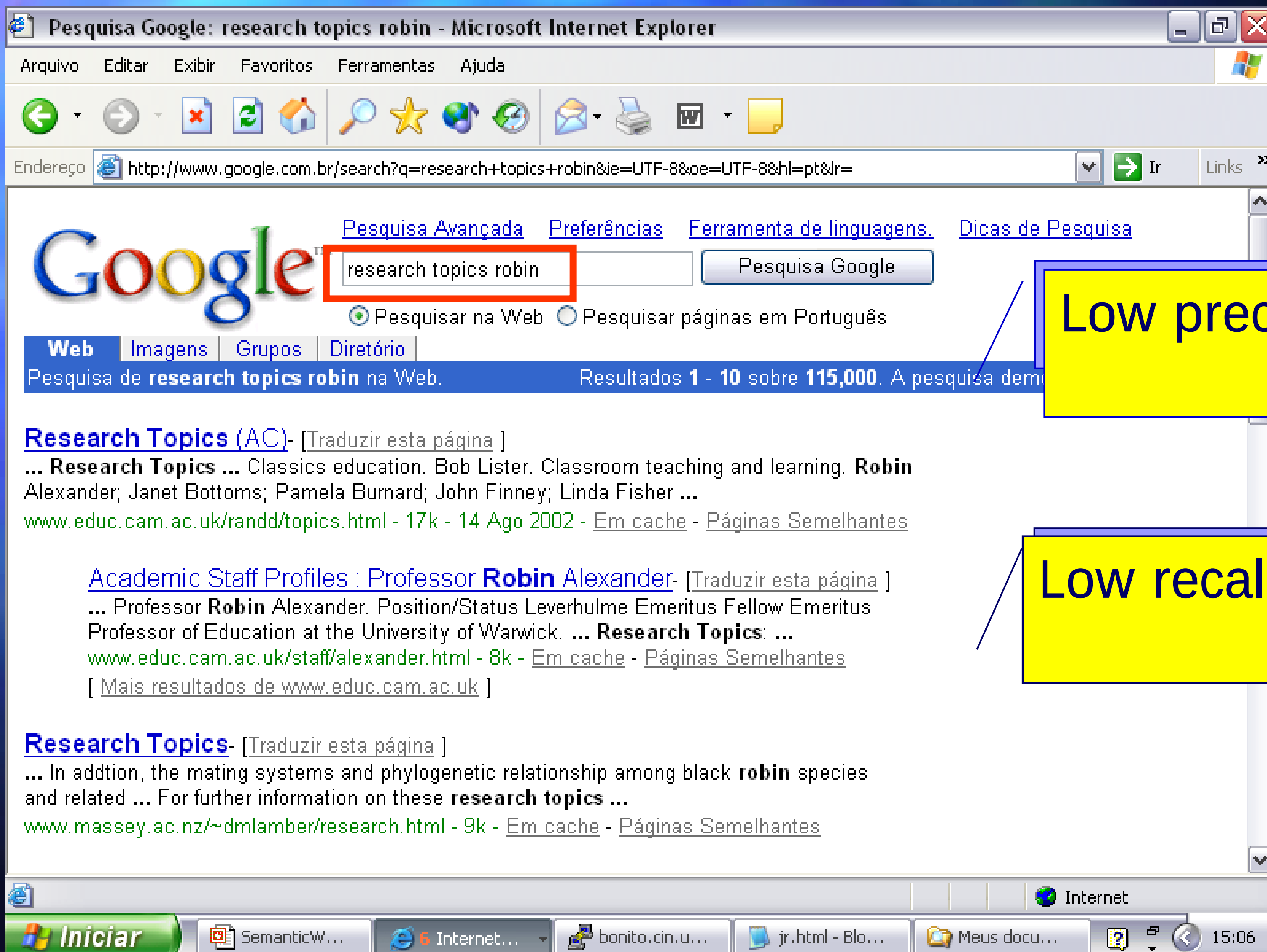
Traditional IR systems

- Users are allowed only to perform statistically lexical-based searches
- Several problems:
  - Lack of context
  - Linguistic problems: polysemy, figurative language, coreference...
    - Any language was formed as a result of a long, ever-evolving, constructive process, that takes into account mutual understanding among humans
  - Many other usages for texts beyond retrieval
    - Only retrieval clearly doesn’t suffice!

Searching for Prof. Robin's research topics



Search engines get puzzled...

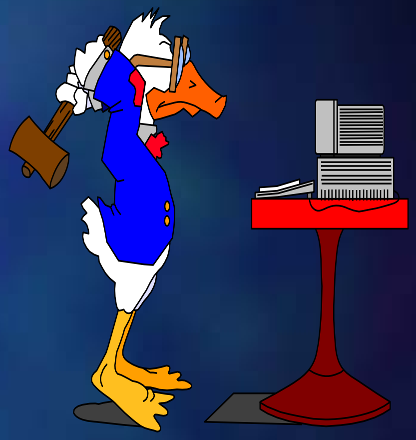


Low precision

Low recall

Why doesn't IR suffice?

- Main reason of problems
  - Lack of context
- Consequences: Users are burdened with the (hard) work of interpreting, filtering, combining, finding answers from search engine results
- How do we benefit from computer power for text processing??



Motivation: How do we agreggate context to the Web?

Possible solutions to provide contexts

- More intelligent systems
  - Intelligent Agents
  - Cooperative information gathering [Oates et al 94]: distribuition, cooperation and communication about page semantics
  - Domain restrictions
- More intelligence in the Web: Semantic Web!
  - Languages and standards that allow page definition with clear and formal semantics
  - Agents could reason and communicate using this semantics

⇒ Ontologies are fundamental to both solutions!

With ontologies, page processing gains associated context

Ontology

Person
Employee :: Person
AcademicStaff :: Employee
Researcher :: AcademicStaff
PhDStudent :: Researcher
Employee[
    affiliation : Organization;
    worksAtProject : Project;
    headOf : Project;
    headOfGroup : ResearchGroup].
AcademicStaff[
    supervises :PhDStudent].
Researcher[
    researchInterest : ResearchTopic;
    memberOf : ResearchGroup;
    cooperatesWith : Researcher].

An ontology-based CIG system

Cooperative Information Gathering

[Oates, Prasad & Lesser 94]

- Proposed cooperative multi-agents systems “to integrate and evolve consistent clusters of high quality information (...)”
- DPS and knowledge-based solutions were encouraged
- Suggested domain models
  - Nowadays ontologies play this role
- Suggested implicitly task integration at agent level
  - An agent can search and process information

Problems in CIG practice

- Few systems integrate text-related tasks at agent level
  - Many systems only divide the tasks among agents
  - Lack of *semantic* cooperation of information in CIG
  - Agents can be experts on *specific* information
- Semantic cooperation is particularly suitable to Web extractor agents
  - Cooperation is neglected in IE systems
  - However, some classes processed by them form clusters (e.g. Science, Tourism)

Proposal: An Architecture for CIG

- Two design requirements:
- A Web vision
  - Support to accurate identification of specific information
  - It should couple a vision for contents (classes, attributes, etc) to a functional vision (pages can be lists, messages, class instances, garbage, etc)
- Ontologies
  - Enable cooperation
  - Provide detailed domain models useful for processing clusters formed by page classes

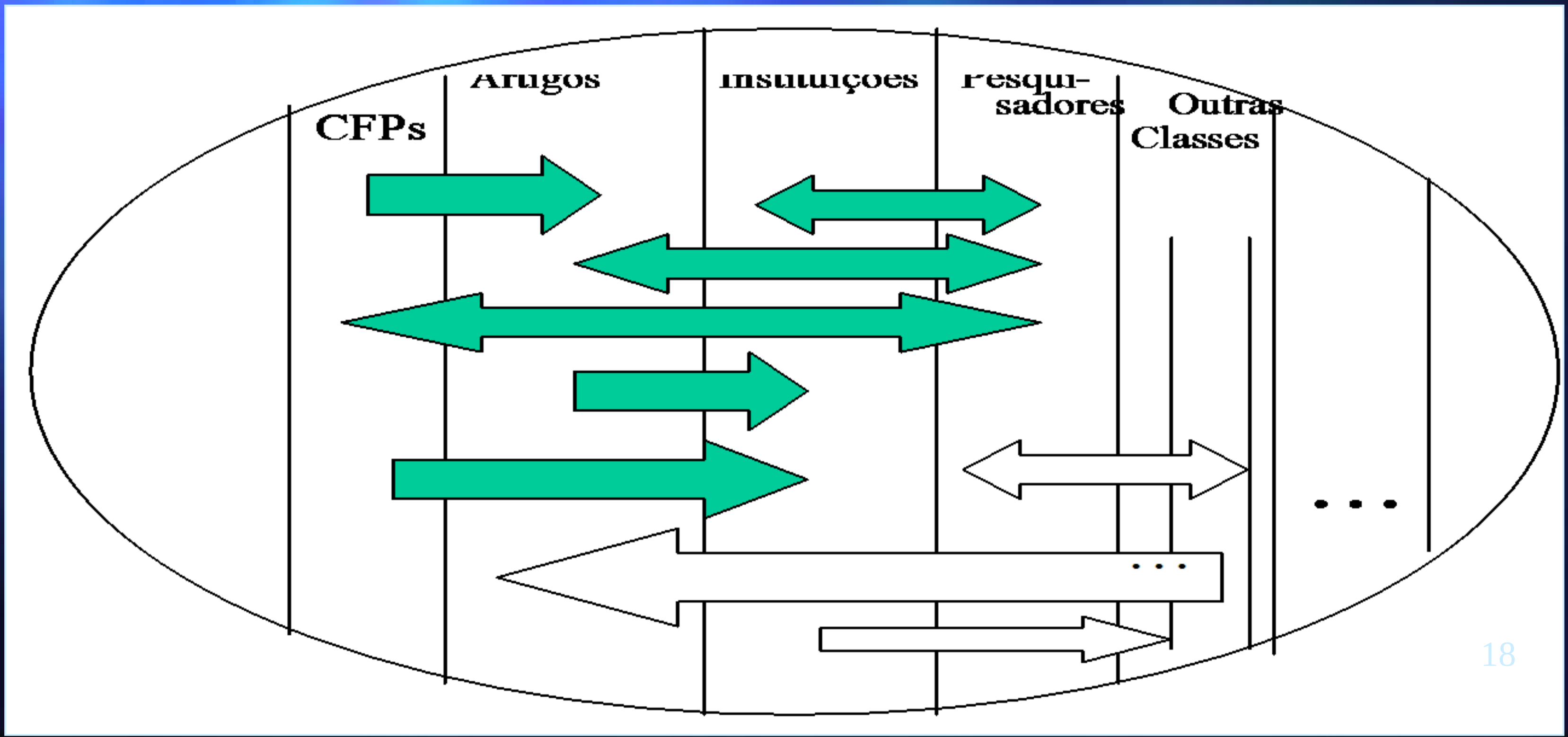
A Web vision for CIG

Vision by Contents: Page Classes

- Seek for pages that are class instances
  - Scientific article, call for papers, researcher's page, ...
- Slot are discriminators
  - Slots in an article: author, title, affiliation, abstract,...
  - Extraction and categorization are complimentary tasks

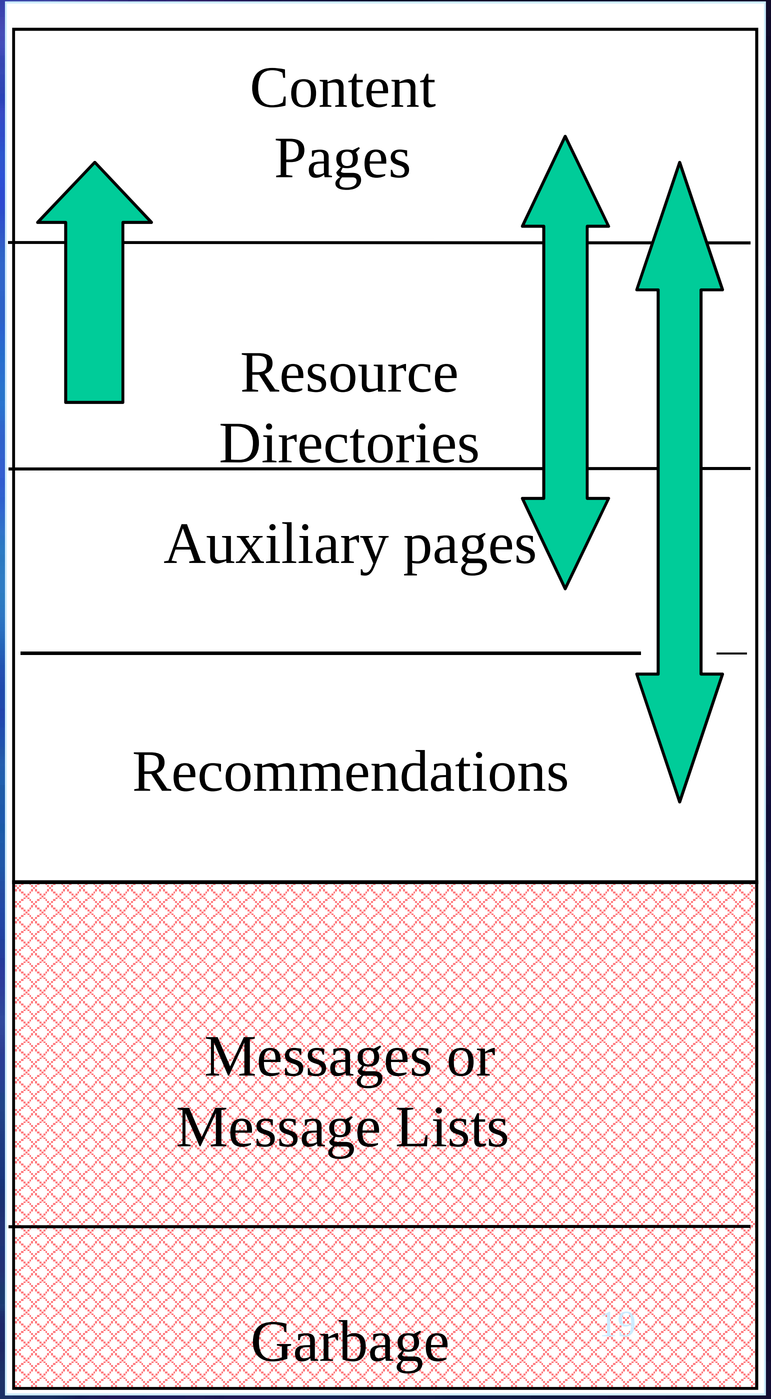
Vision by Contents: Clusters of Page Classes

- *Hypothesis:* Most links in classes' pages point to pages containing data from a few other classes
- Interrelated classes form a cluster about a domain
- Class Relations
  - Extraction and search can be viewed as complimentary



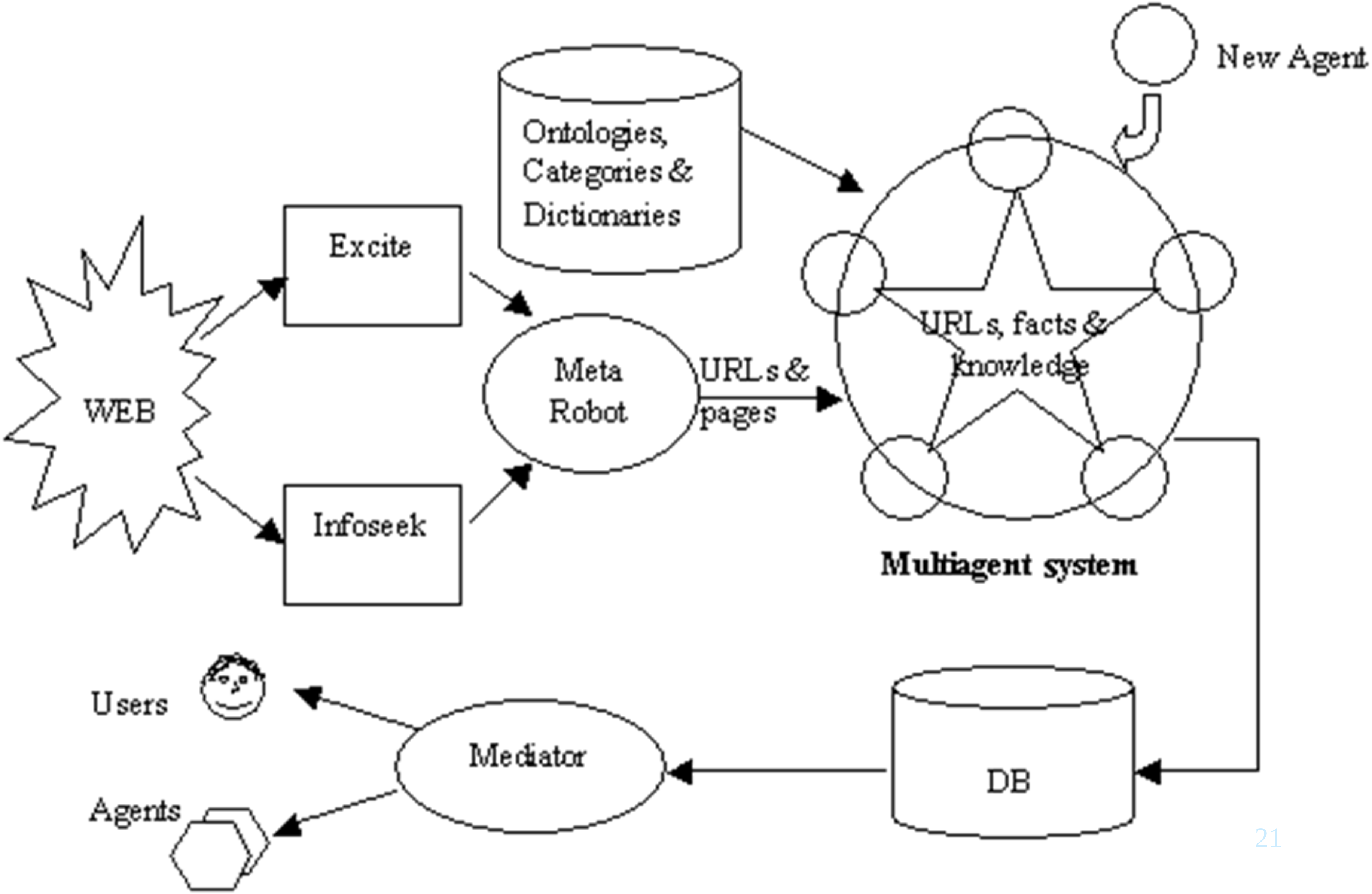
Vision by Functionality

- Inspired on [Pirolli 95]
- Divides pages by the role played in linkage and information storage
- Classes:
  - Content pages
  - Auxiliary pages
  - Resource directories (lists)
  - Messages and messages lists
  - Recommendations pages (other classes' pages)
  - Garbage

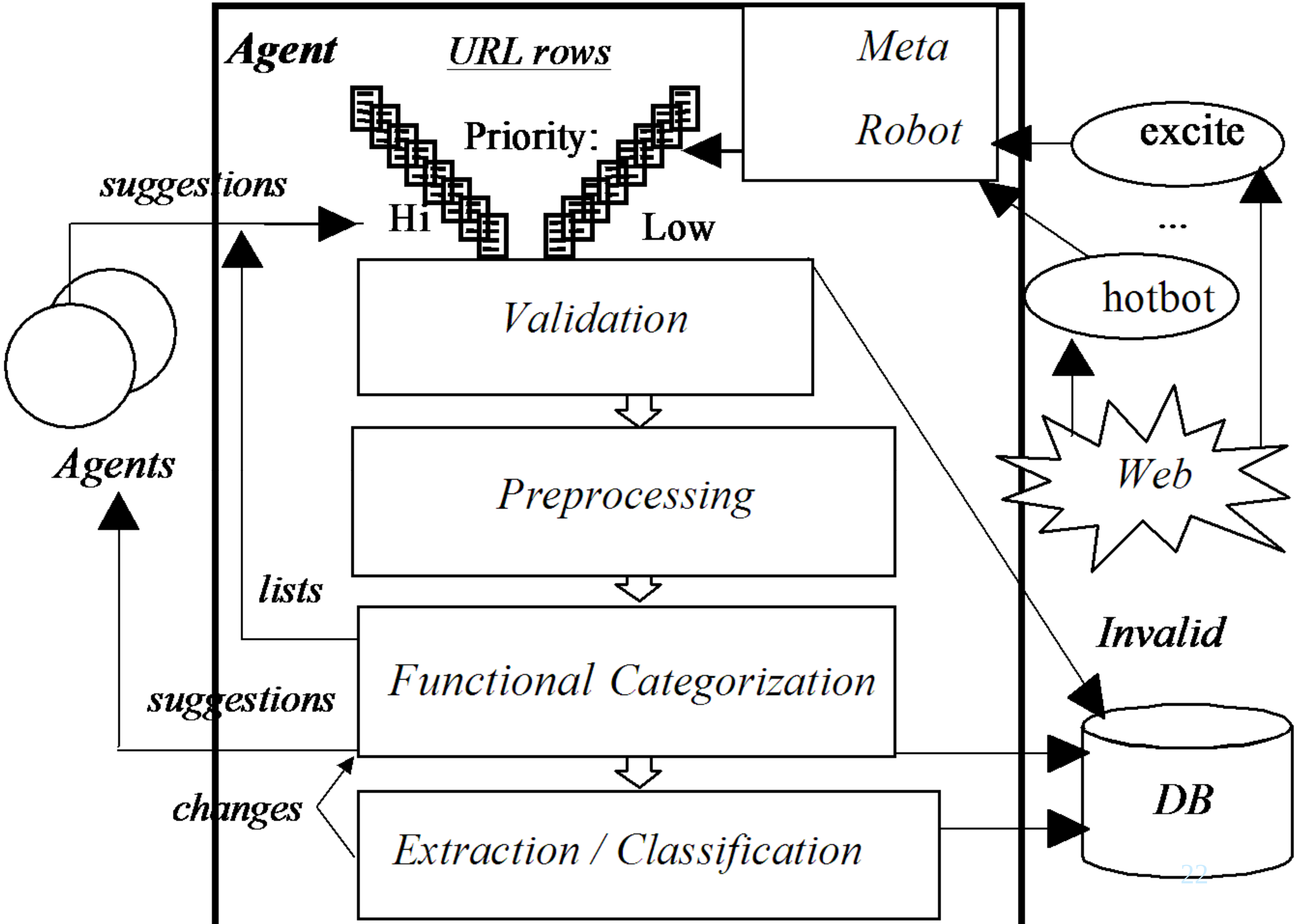


Proposal of a CIG Architecture

Proposed Architecture



Agents' tasks



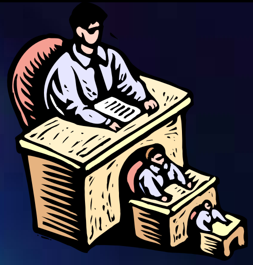
Agents' knowledge

- Cluster (domain) ontology
  - Comprehensive as possible
- Web ontology
  - Pages, URLs, anchors, ...
  - Protocol data (HTTP, FTP,...)
  - Page elements (links, tags contents, e-mails addresses, ...)
  - IR representations (terms, frequencies, centroid, ...)
  - NLP representations

Agents' knowledge (cont.)

- CIG ontology
  - Templates for slot extraction, functional category identification and classification
  - Agents descriptors (identification, abilities,...)
  - Dictionaries (synonyms, keywords, ...)
  - Complex and expressive cases for recognition
- Auxiliary ontologies
  - Wordnet
  - Time and places
  - Topic-specific ontologies (e.g. Bibliographic-data, for the scientific articles agent)

Types of Reuse enabled



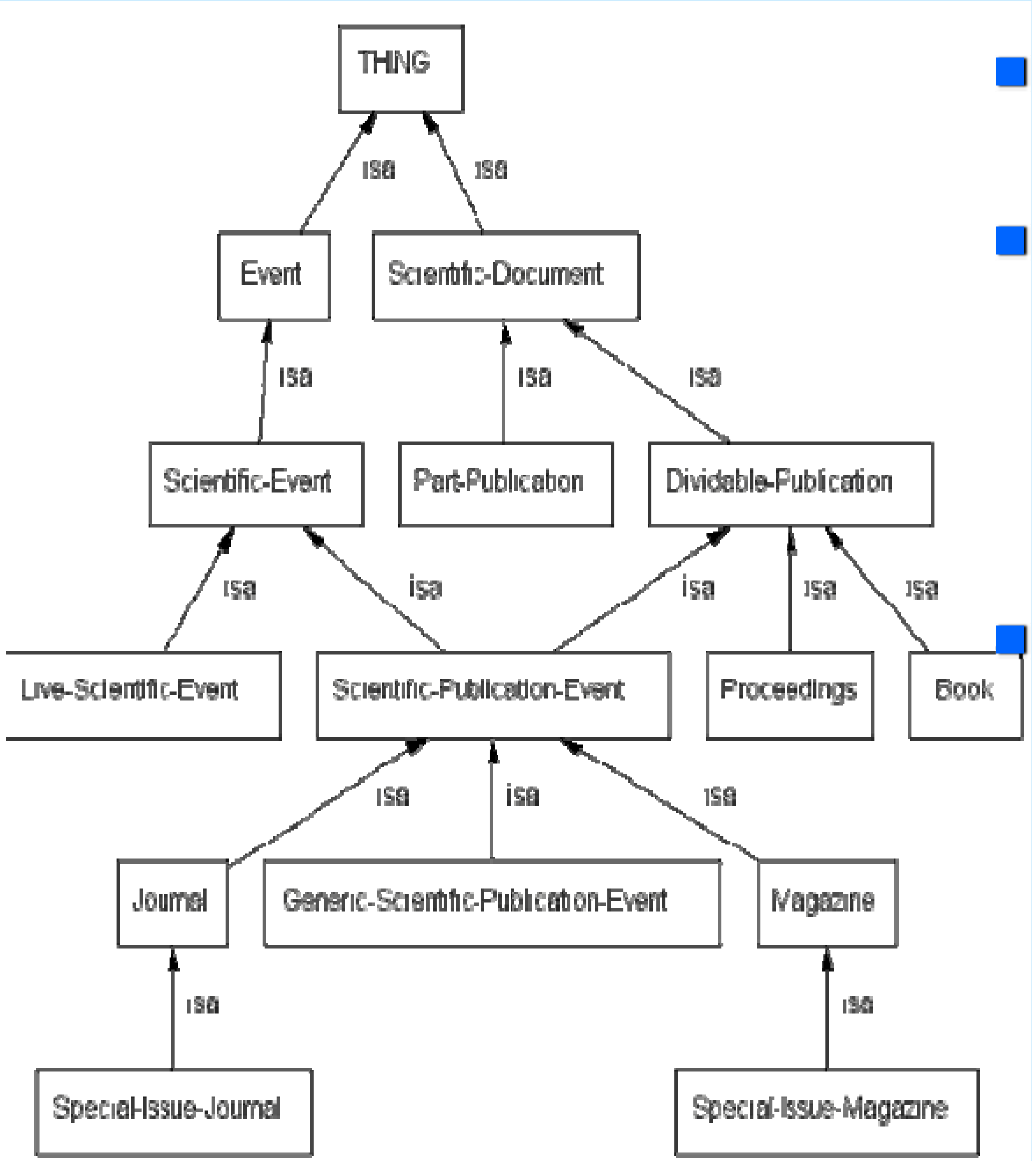
- Code
- Robots and search engines services
- DB definitions
- Knowledge
  - Agents share ontologies about the domain and the Web
  - Domain ontology can be reused, if available
  - CIG and auxiliary ontologies are also reused, but instances are agent specific
  - Most of the rules can be reused

Case Studies

Case 1: The scientific cluster

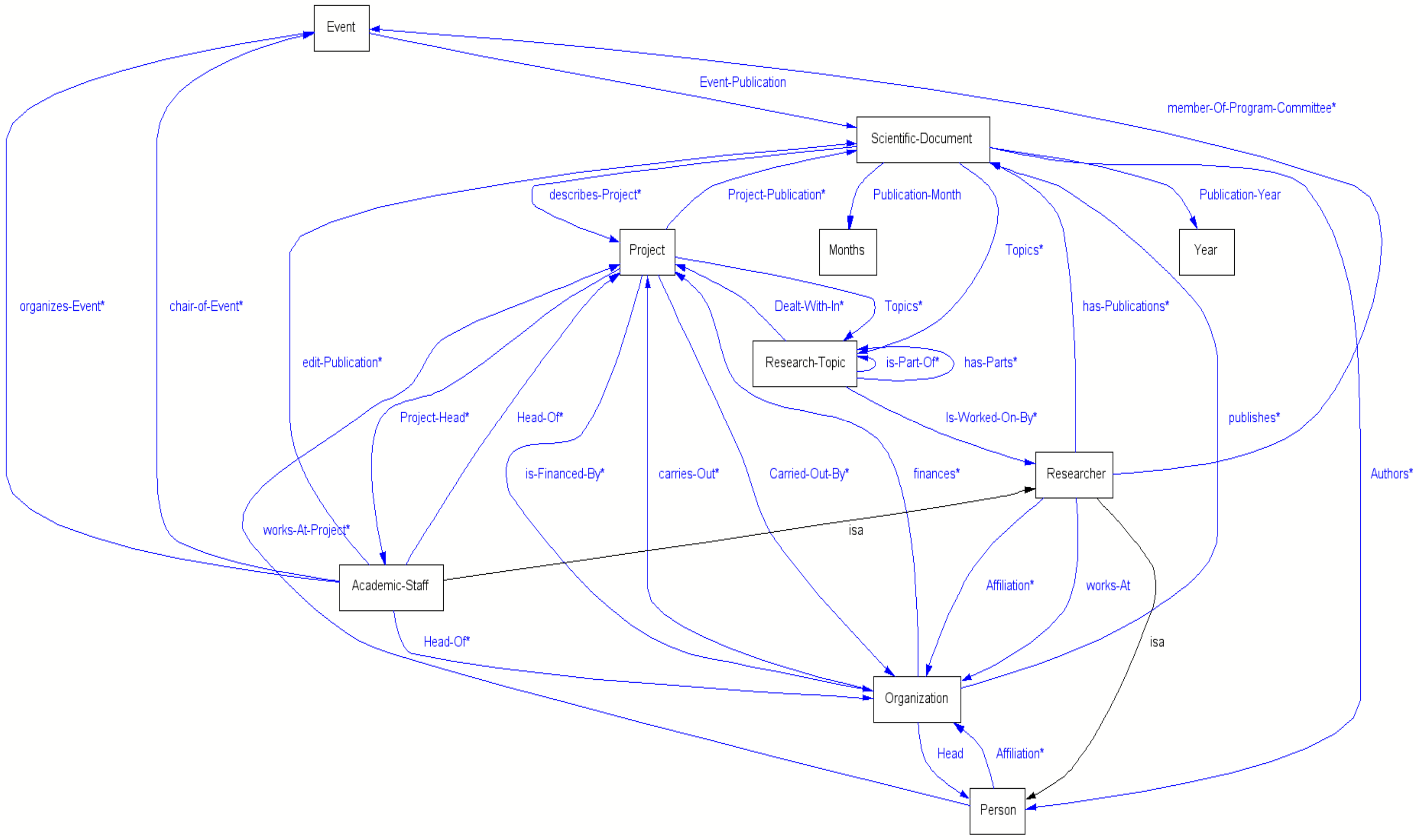
- MASTER-Web
  - Multi-Agent System for Text Extraction and Retrieval and classification over the Web
- CFP agent : scientific events
- Scientific articles' agent
- Slot identification, instead of extraction
- Tests performed two classifications with each page:
  - Identification of the functional category
  - Identification of the concrete subclass of the vision by contents (e.g., CFP of a conference, workshop, journal,...)

Science Ontology



- Available at the Protégé repository
- Reused from the European project (KA)2 [Benjamins et al 98] ontology available at the Ontolingua mirror in Madrid
- Refined in granularity

Relations in the Science Ontology



Concepts, Cases and Recognizers

```
([abstract] of Concept (name "abstract")
     (Synonyms "summary"))
([thesis] of Concept (name "thesis")
     (Keywords "partial fulfillment“))

([ppr_00356] of Case(Description "aff,1st,loc")
     (Absent-Concepts [thesis])
     (Concepts-in-the-Beginning [abstract])
     (Slots-in-the-Beginning [First-Name] [name]
          [Location-Place]))

([Part-Publication] of Class-Recognizer
     (Cases [ppr_00536] [ppr_00356])
     (Class [Part-Publication]))
```

Example

Prev | Index | Next

Lightweight Deductive Databases on the World-Wide Web

S.W. Loke, A. Davison, and L. Sterling

Department of Computer Science
The University of Melbourne
Parkville, Victoria 3052, Australia
Email: {swloke,ad,leon}@cs.mu.oz.au

Abstract:
We investigate a Web information structuring mechanism called *lightweight deductive databases*. Lightweight deductive databases enable more sophisticated automated searching, extraction, and processing, and can facilitate agent-based programming. We also explore how these deductive databases benefit from being distributed on the Web.

1. Introduction

Our aim is to enhance the Web with information which is more susceptible to sophisticated automated searching, extraction,

```
FIRE 1 MAIN::i_901_start f-411
FIRE 2 MAIN::v_314_valid f-869, f-867
FIRE 3 MAIN::i_905_filling f-870
FIRE 4 MAIN::i_907_fill-fields f-871
Found country ("Australia")
FIRE 5 MAIN::r_450_slots_hi_funct f-894, f-450,,
SLOT FOUND : Location-Place
FIRE 6 MAIN::r_450_slots_hi_funct f-894, f-68,,
SLOT FOUND : First-Name
FIRE 7 MAIN::r_430_slots_hi_ccpt_bgn f-894, f-593,,
SLOT FOUND : name
FIRE 8 MAIN::r_900_slots_hi_funct f-894, f-957, f-377, f-30
FIRE 9 MAIN::c_600_recognized_default f-960, f-301
CLASS Generic-Part-Publication
FIRE 10 MAIN::e_203_links f-963, f-748, f-964,, f-790, f-53
FIRE 11 MAIN::e_203_links f-963, f-378, f-964,, f-845, f-52
FIRE 12 MAIN::s_002_result f-963, f-971, f-966, f-964
fact : CLASSIFIED
Inserting as recognized...
CLASS Generic-Part-Publication
```

Asking links containing concepts

```
(ask-all :sender cfp
        :receiver ppr
        :language JessTab
        :ontology Science
        :content (object (is-a Anchor) (Link-Text ?l))
   (Result (Page-Status CLASSIFIED) (Class "Conference-Paper"))
   (object (is-a Web-Page) (Contents ?co))
   (test (and  (if-occur ?l (begin-until "abstract" ?co))
               (if-occur (slot-get [Conference] Concepts) ?l))))
```

Receiving a reply

(t*ell*

:sender **PPR-Agent**

:receiver **CFP-Agent**

:in-reply-to **id1**

:reply-with **id2**

:language **JessTab**

:ontology **Science**

:content **(object (is-a Link)**

(URL "http://lcn2002.cs.bonn.edu")

(anchor " IEEE Conference on Local Computer Networks (LCN 2002)")))

Tests

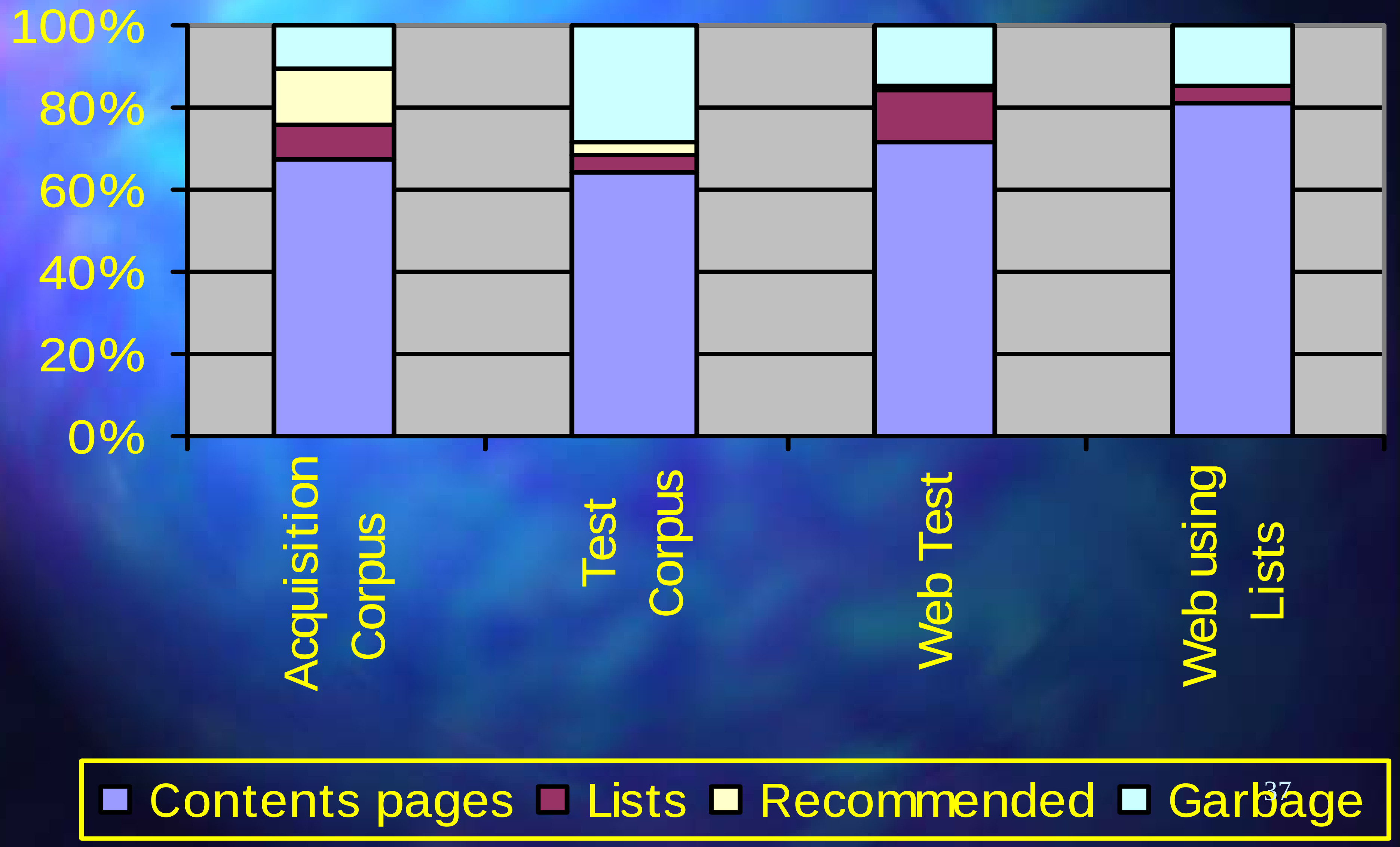
- In each test, an agent processed between 150 and 200 pages
- Tests with each agent:
  - A *corpus* for knowledge acquisition
  - A test *corpus*
  - A Web test
- Cooperation worked, but only 3 links were suggested
  - The CFP agent suggested 30 correct and 7 wrong links to a future researcher's agent

CFP Agent's Results

- Classes : Conference, Workshop, Journal, Magazine, Generic-Live-Sc-Event, Generic-Sc-Publication-Event and Special-Issues for Journal and Magazine
- Templates for 21 slots
- 28 cases for the classifications

CFP Agent	Acquis. Corpus	Test Corpus	Web with lists	Web w/out lists
Recognition	97.1	93.9	96.1	96.3
Functional categorization	93.8	93.9	93.8	95.7
Contents classification	94.9	93.3	92.9	91.7
Processed pages	244	147	129	188

Functional categories distribution in the CFP agent



Articles' Agent Results

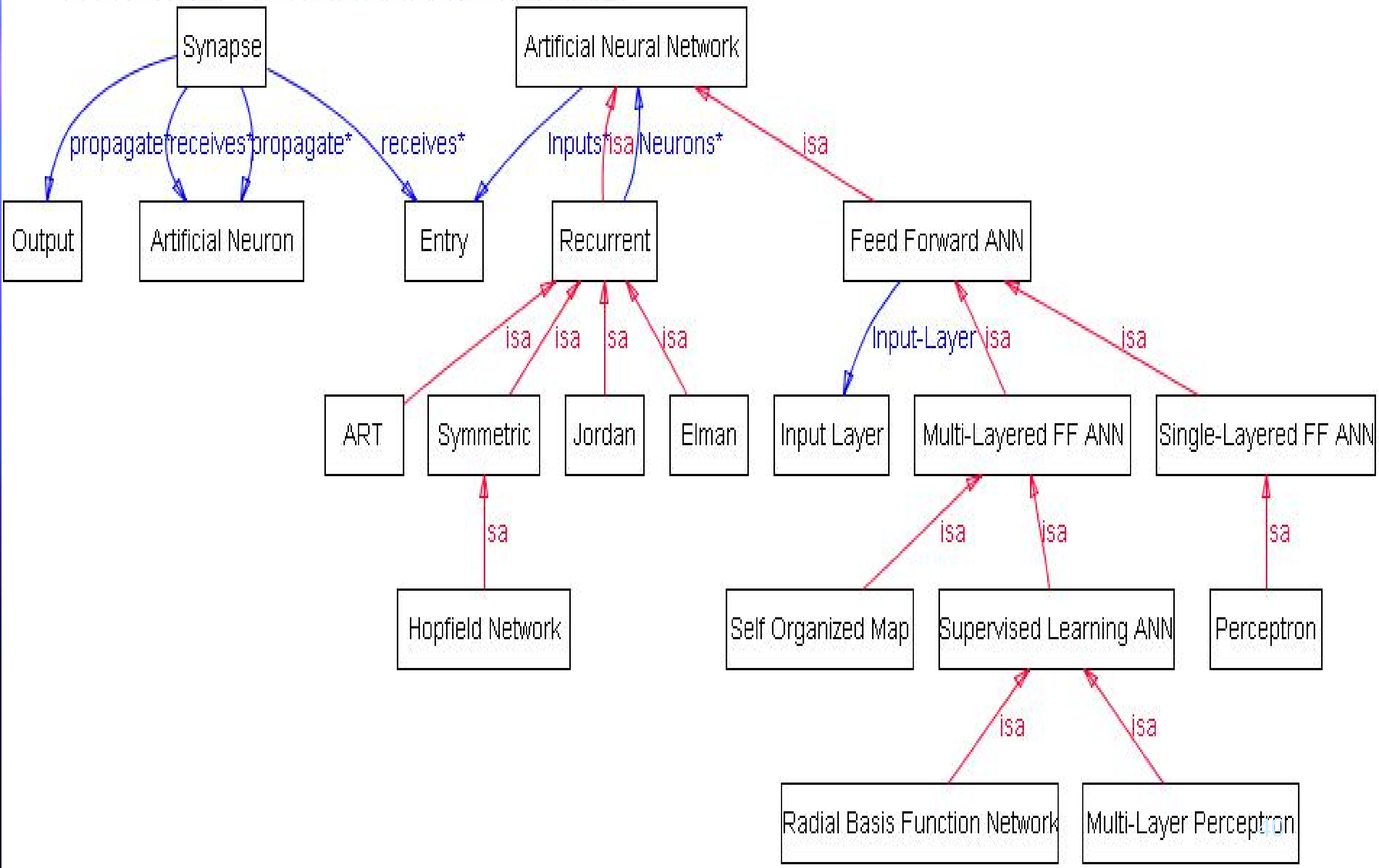
- Classes: Conference, Workshop, Journal and Magazine Articles, Thesis, Dissertation, Technical and Project reports, Book chapter, Generic Publication
- Templates for 8 slots
- 52 cases and templates for the classifications

Scientific articles' agent	Acquis. Corpus	Test corpus	Web w/out lists
Recognition	93.1	82.7	87.8
Functional Categorization	96.8	94	95.1
Classification	97	93	81.4
Processed pages	190	150	184

Case 2 : AI articles’ classification

- Construction of an Artificial Intelligence (AI) ontology
- Classification of scientific articles into multiple sub-areas of AI

Part of the AI ontology: Neural Nets definitions



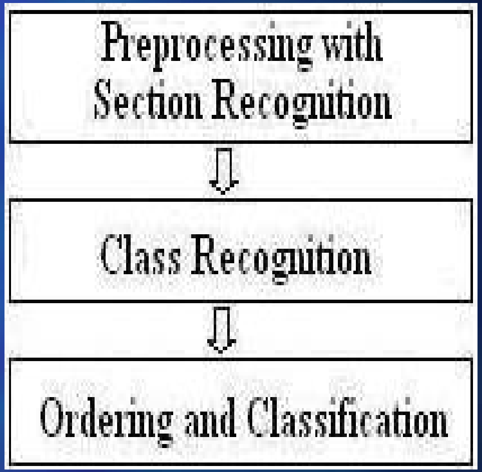
MASTER-Web for Textual Classification

Classification strategies

• **Preprocessing with Section Recognition** - recognizes the relevant sections, identifying and extracting from them terms found in the ontology of the domain.

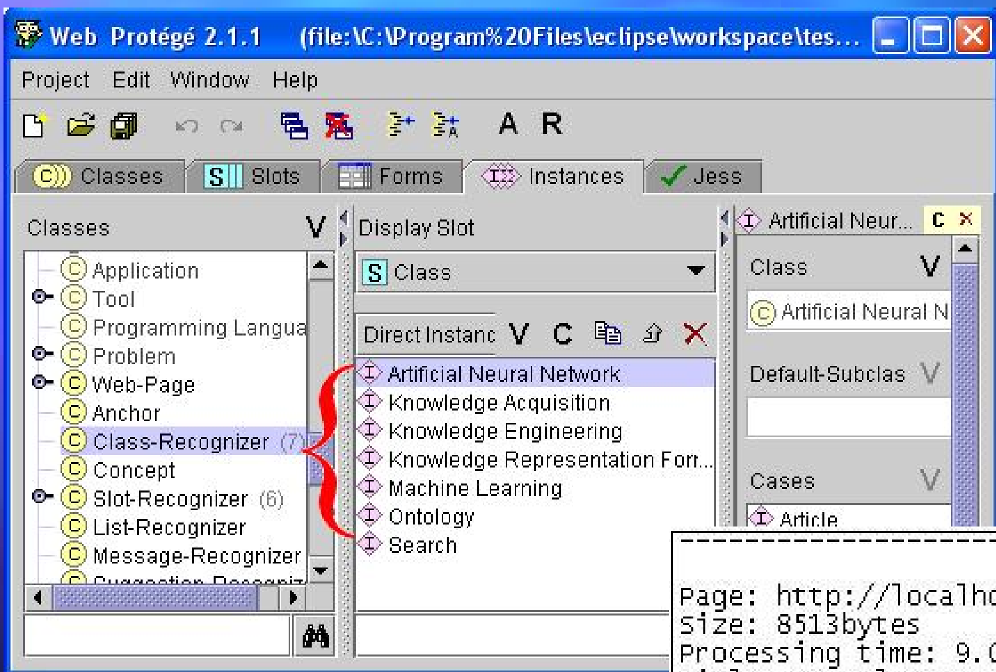
• **Class Recognition** - within the AI domain, recognizes the "main classes" of the upper subareas. 3 methods are being applied:

- **Direct Recognition of Main Classes**
- **Class Recognition Through Attributes**
- **Class Recognition Through an Indirect Relation**



MASTER-Web for Textual Classification

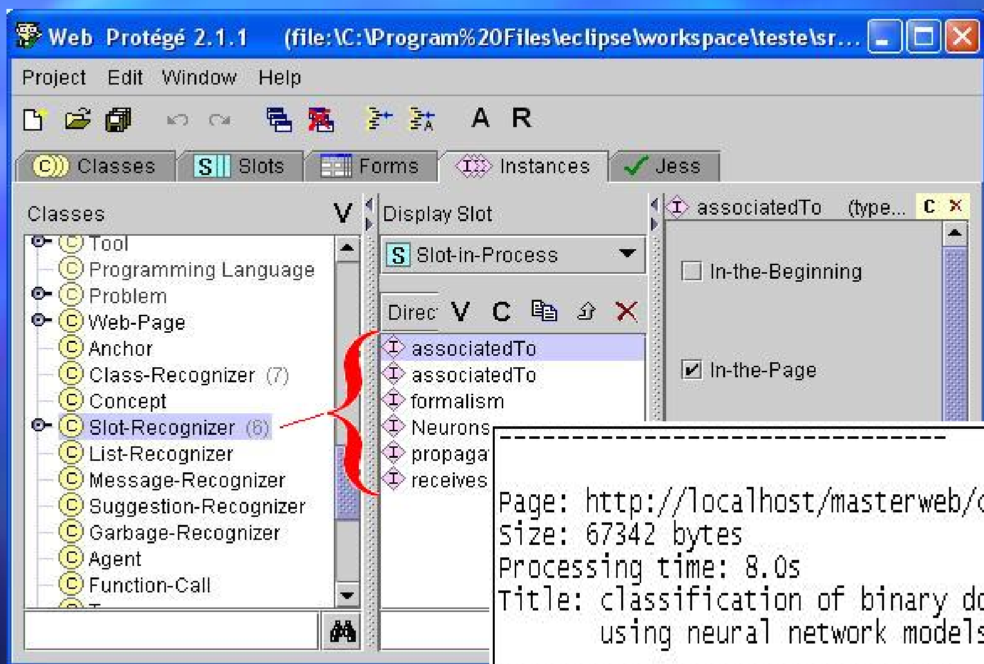
Direct Recognition of Main Classes

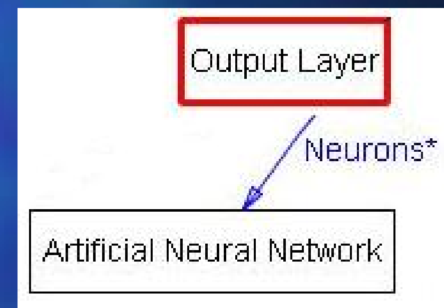



```
--------------------------------

Page: http://localhost/masterweb/cfp/Teste/AI/AI41.htm
Size: 8513bytes
Processing time: 9.0s
Title: neural networks
--------------------------------

Direct Recognition of Main Classes: Artificial Neural network
--------------------------------

Direct Recognition of Main Classes: Knowledge Representation Formalisms
--------------------------------
```

MASTER-Web for Textual Classification

Class Recognition Through Attributes






```
--------------------------------

Page: http://localhost/masterweb/cfp/Teste/AI/AI20.htm
Size: 67342 bytes
Processing time: 8.0s
Title: classification of binary document images into texture or non-textual data blocks
       using neural network models
--------------------------------

Class Recognition Through Attributes:
Output Layer is linked to Artificial Neural Network by slot Neurons
--------------------------------
```

Experiments and Results

- Experimental corpus
  - 406 HTML documents
  - Domain:
    Artificial Intelligence
    Computing
    Medicine,
    Biology
    Economy
    Philosophy
  - Heterogeneous with respect to the sections' division

Experiments and Results

- classification results of the articles by area

Recognition	Correct	False +	False -	Hits (%)
Artificial Neural Network	48	2	1	92,3
Knowledge Acquisition	17	0	1	94,4
Knowledge Engineering	3	0	0	100,0
Knowledge Representation Formalisms	56	9	1	84,8
Machine Learning	51	2	6	86,4
Ontology	19	0	0	100,0
Search	38	1	1	95,0
Other domains	228	7	11	92,7
Total	**460**	**21**	**21**	**91,6**

Experiments and Results

- Results inferred and shown by system

```
------------------ CLASSIFICATION  ARTICLE --------------
http://localhost/masterweb/cfp/Teste/AI/AI41.htm
 Title: neural networks
--- The article is about:
Feed Forward ANN
Supervised Learning ANN
--- Citations:
Knowledge Representation Formalisms
Expert System
---------------------------------------------------------
```

NEURAL NETWORKS

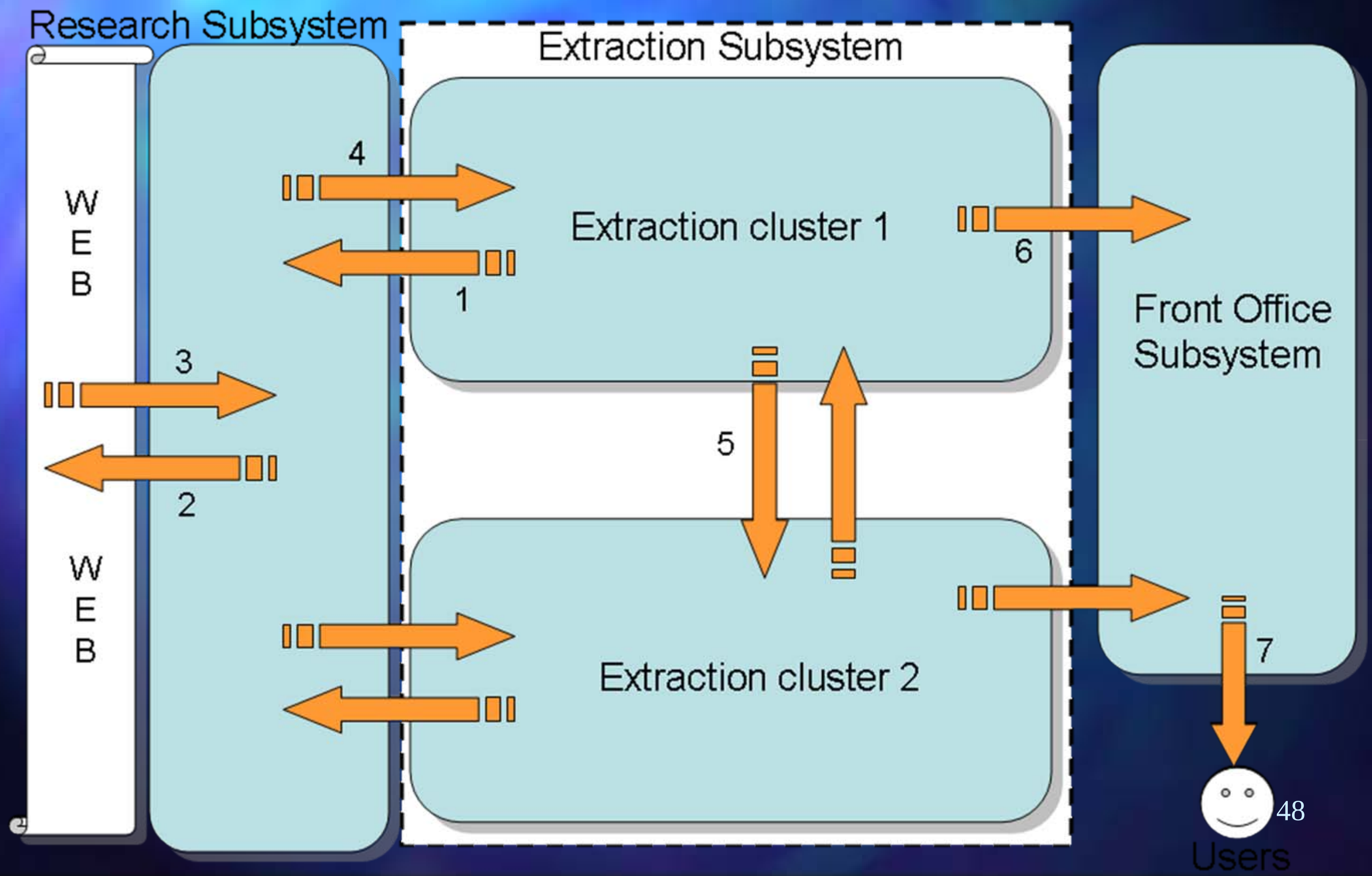
by Christos Stergiou and Dimitrios Siganos

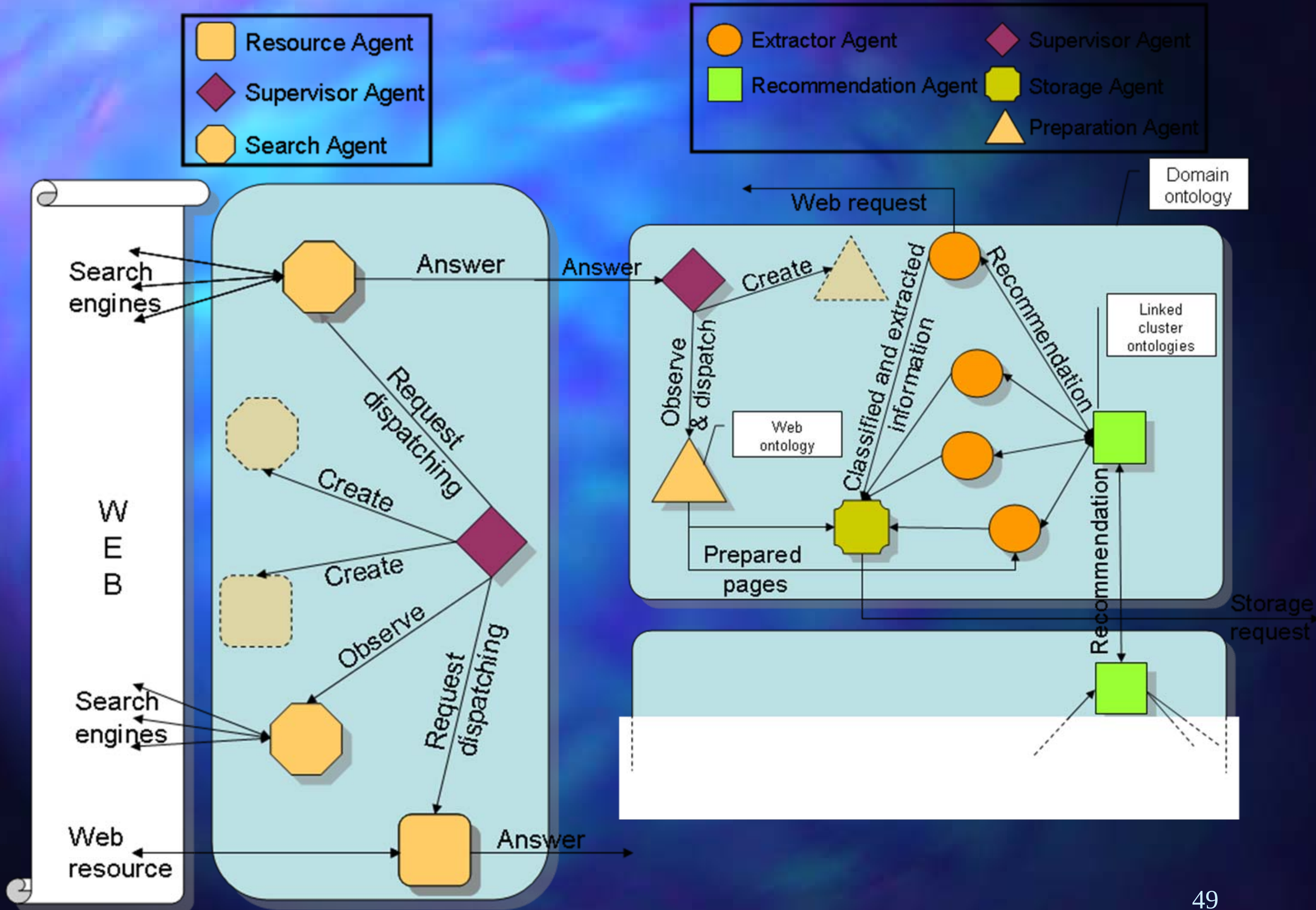
Abstract

This report is an introduction to Artificial Neural Networks. The various types of neural networks are explained and demonstrated, applications of neural networks like ANNs in medicine are described, and a detailed historical background is provided. The connection between the artificial and the real thing is also investigated and explained. Finally, the mathematical models involved are presented and demonstrated.

Evolution

AGATHE: a better agentization





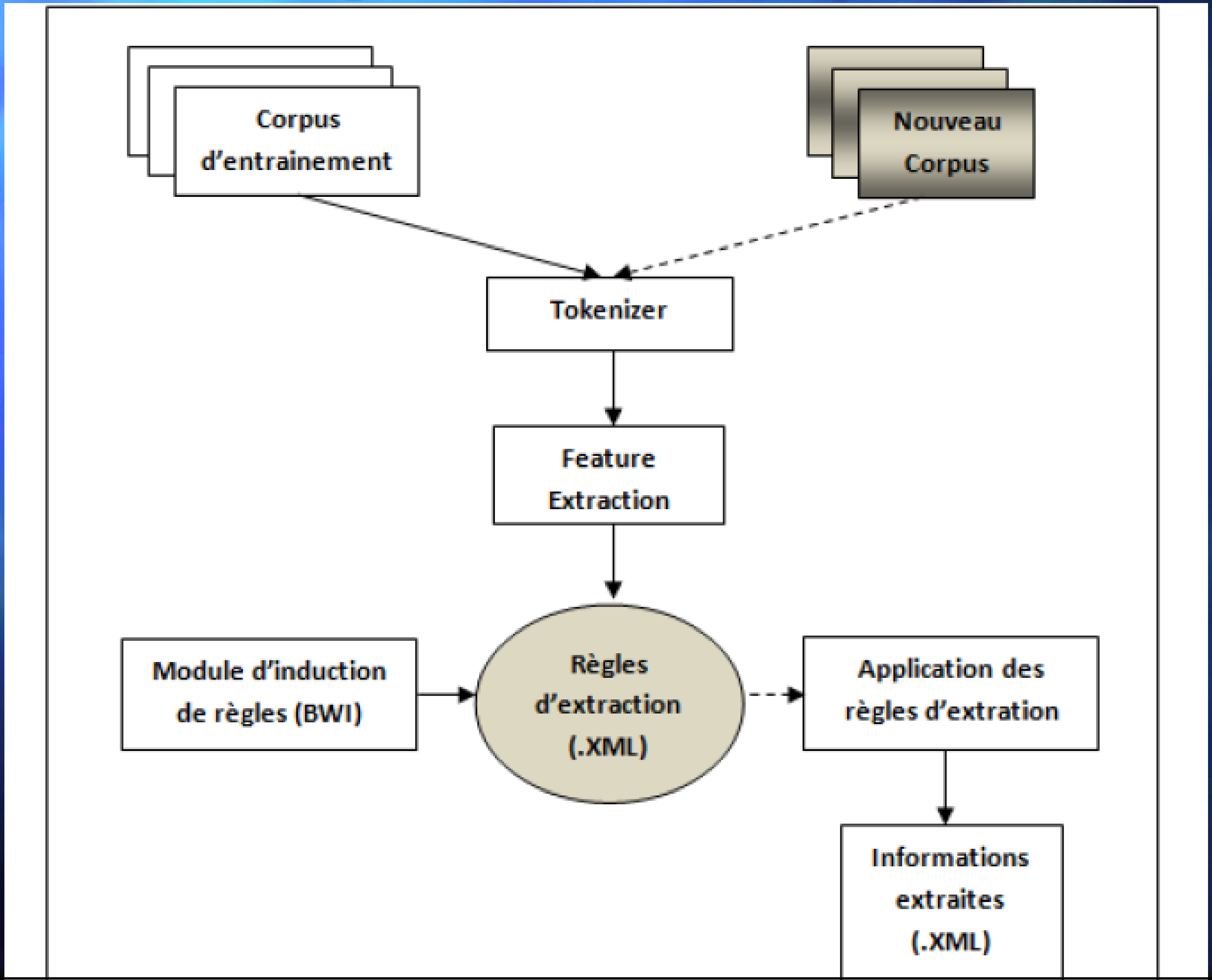
Resource Agent
Supervisor Agent
Search Agent
Extractor Agent
Recommendation Agent
Supervisor Agent
Storage Agent
Preparation Agent
Domain ontology
Web request
Search engines
Answer
Answer
Create
Observe & dispatch
Classified and extracted information
Recommendation
Linked cluster ontologies
Request dispatching
Web ontology
Create
Create
W E B
Prepared pages
Storage request
Recommendation
Observe
Request dispatching
Search engines
Web resource
Answer

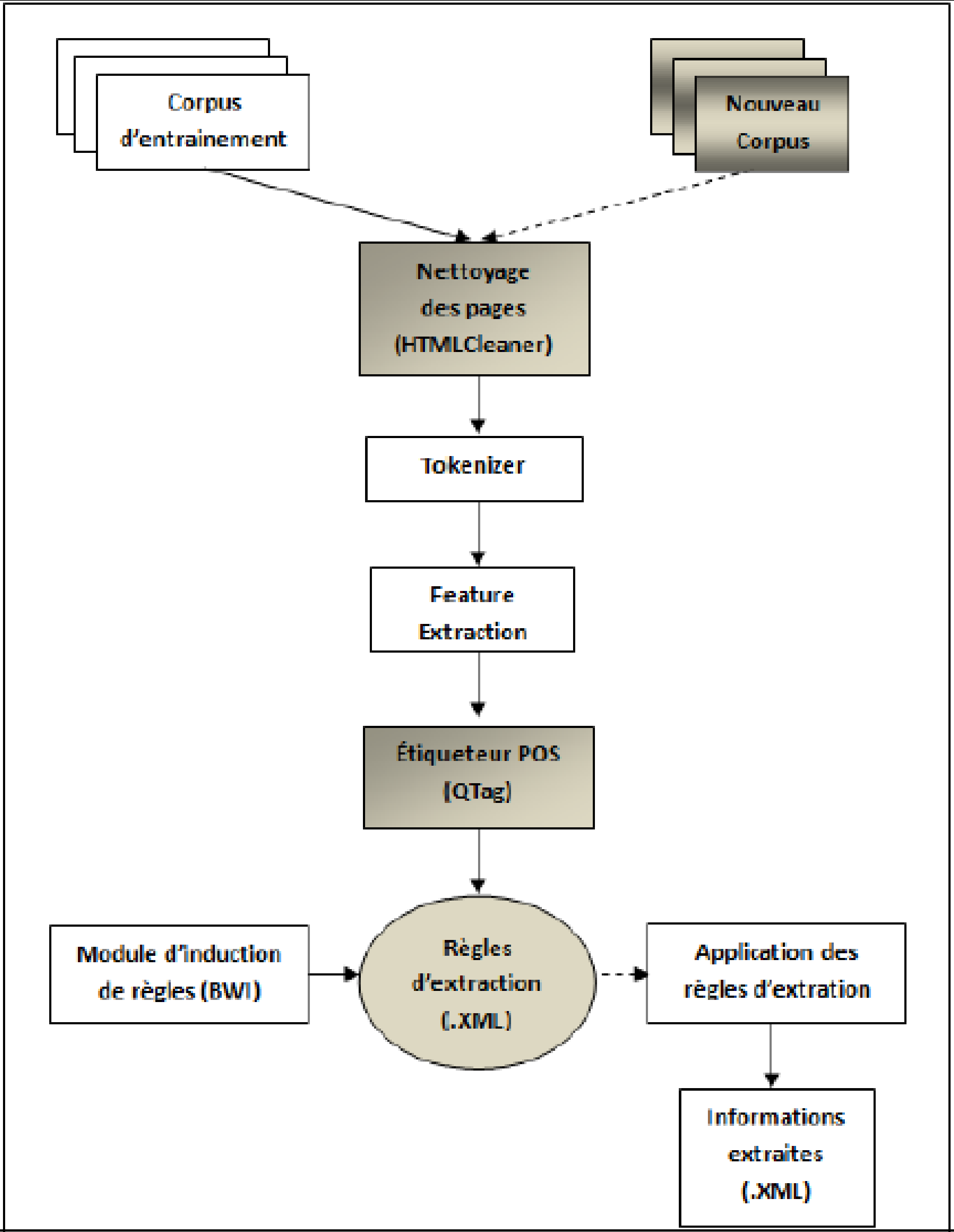
Envisaged advantages

- Better flexibility, extensibility, scalability and reusability
- Cooperation between different domains
  - Ex: information related to accommodation and transport possibilities, touristic information (monuments, galleries and cultural events occurring in the same time period of a scientific event (cluster of Science) should be recommended to the Tourism cluster

Better Extraction with TIES

[Giuliano et al 2004]





WEEPAIES: TIES with some NLP
[Lima et al 2010]

Tab. 11. Perfomances par slot de 5 systèmes sur le corpus *Seminars*.

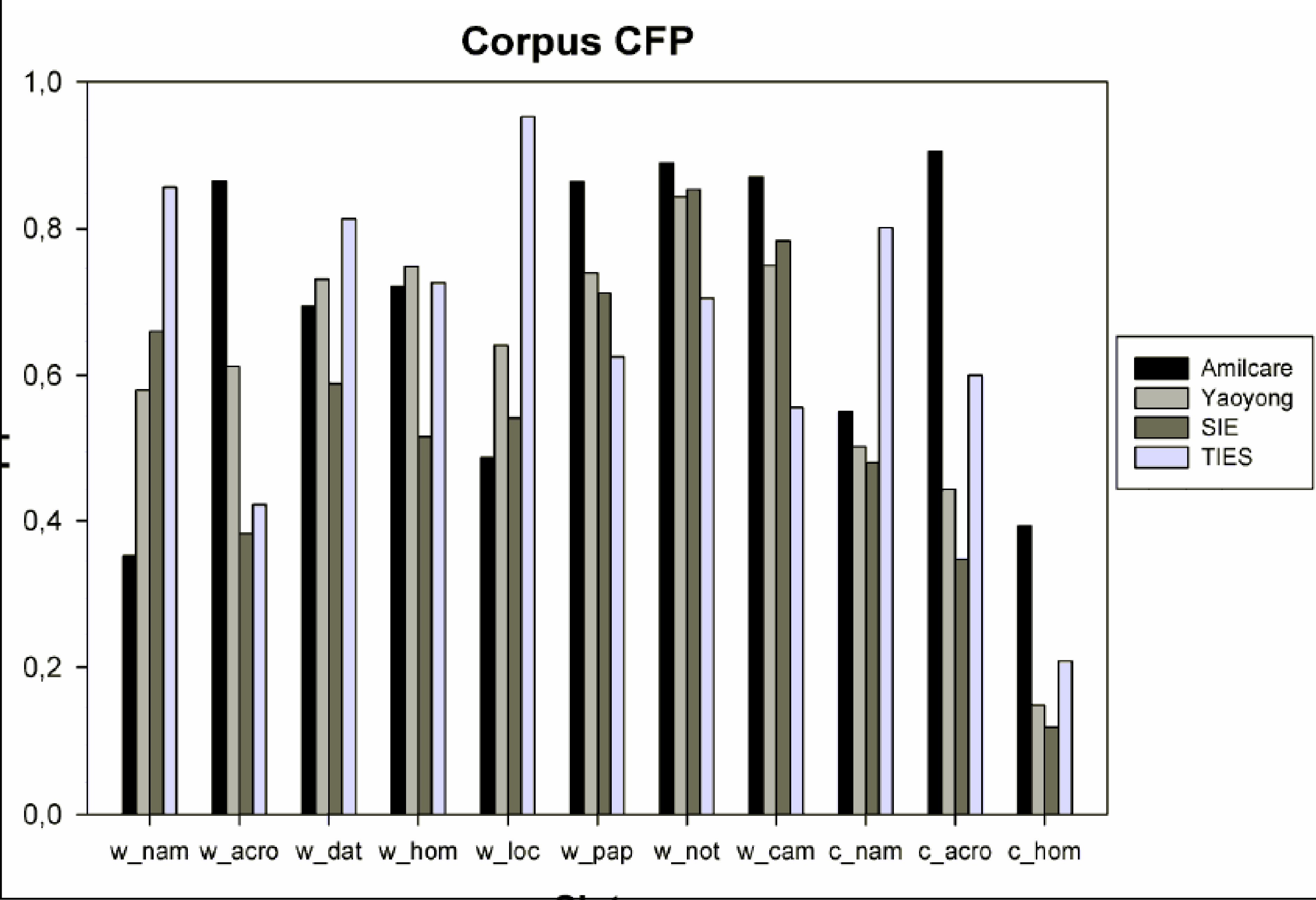
	speaker	*location*	*stime*	*etime*	*All Slots*
SIE	-	-	-	-	86,6
GATE-SVM	69,0	81,3	94,8	92,7	86,2
$(LP)^2$	77,6	75,0	**99,0**	95,5	86,0
Rapier	53,0	72,7	93,4	96,2	77,3
TIES	**86,2**	**88,8**	93,9	**96,7**	**91,4**

Tab. 12. Perfomances par slot de 4 systèmes sur le corpus *Jobs* en utilisant un ensemble d'attributs composé d'information de capitalisation et POS.

Slot	*$(LP)^2$*	*GATE_SVM*	*Rapier*	*TIES*
id	**100,0**	97,7	97,5	98,1
title	43,9	49,6	40,5	**67,4**
company	71,9	77,2	69,5	**78,9**
salary	62,8	86,5	67,4	**89,2**
recruiter	80,6	78,4	68,4	**86,1**
state	86,7	92,8	90,2	**96,9**
city	93,0	95,5	90,4	**96,5**
country	81,0	96,2	93,2	**98,8**
language	**91,0**	86,9	80,6	88,5
plataform	80,5	80,1	72,5	**86,9**
application	**78,4**	70,2	69,3	73,1
area	**66,9**	46,8	42,4	51,6
req_y_exp	68,8	80,8	67,1	**86,4**
des_y_exp	60,4	81,9	87,5	**89,9**
req_degree	84,7	**87,5**	81,5	78,6
des_degree	**65,1**	59,2	72,2	47,6
post date	99,5	99,2	99,5	**100,0**
All slots	**84,1**	**80,8**	**75,1**	**83,8**

Good results in standard corpora

Integration with AGATHE already implemented


Corpus CFP
1,0
0,8
0,6
0,4
0,2
0,0
w_nam
w_acro
w_dat
w_hom
w_loc
w_pap
w_not
w_cam
c_nam
c_acro
c_hom
Amilcare
Yaoyong
SIE
TIES

Possible continuations

- Other agents and domains (researchers, hotels,…)
- Tests with AGATHE and WEEPAIES Integrated
- Duplicity checking
- Benefit from URLs directory structure prefixes information
- Extraction and information cooperation

Conclusions

- CIG systems for specific domains seem to be feasible
- Cooperation among agents can facilitate retrieval in a common domain
- Functional categorisation and a detailed domain ontology seem to be requirements for success
- Current keyword-based search engines can be a basis for more accurate ontology-based domain-restricted cooperative information agents