# DOMINUS$^{plus}$ - DOcument Management INtelligent Universal System (*plus*)

Stefano Ferilli[1], Floriana Esposito[1], Teresa M.A. Basile[1],
Domenico Redavid[2], and Incoronata Villani[2]

[1] Computer Science Department, University of Bari "Aldo Moro"
{ferilli,esposito,basile}@di.uniba.it
[2] Artificial Brain S.r.l., Bari
{redavid,villani}@abrain.it

**Abstract.** Activities of most organizations, and of universities in particular, involve the need to store, process and manage collections of different kinds of documents. Examples that require advanced solutions to such issues include the management of libraries, scientific conferences, research projects. DOMINUS$^{plus}$ is an open project born with the aim of harmonizing the Artificial Intelligence approaches developed at the LACAM laboratory with the research on Digital Libraries in a general software backbone for document processing and management, extensible with ad-hoc solutions for specific problems and context (such as universities).

## 1 Introduction and Motivation

The DOMINUS$^{plus}$ project springs from the interest of researchers in the LACAM laboratory of the University of Bari in applying Artificial Intelligence (and particularly Machine Learning) methodologies to Document Processing and Digital Library Management, as critical activities in many real-world domains. With specific reference to the context of universities, often practical problems do not facilitate the normal course of research activities, often due to the unavailability of digital material produced in the various activities. In many cases this depends on the lack of intelligent tools that can facilitate on one hand the inclusion of the material and on the other the retrieval. Some practical situations include:

- Libraries. Manual cataloging may generate conceptual errors (e.g., an Operating Systems book might be catalogued as Artificial Intelligence);
- Scientific conferences. Often, the reviewer assignment does not take into account the reviewer expertise (e.g., a paper concerning AI planning might be assigned to a reviewer from a different research field);
- Research project documentation. Generally there is no common repository indexed for a department, faculty, or university that contains the documents produced under the various research projects in which the institution was involved, so finding information for the preparation of consistent and locally agglomerative project proposals is very complex.

M. Agosti et al. (Eds.): IRCDL 2011, CCIS 249, pp. 123–126, 2011.

This resulted in a general framework known as DOMINUS (Document Managemente Intelligen Universal System) (1) developed by the LACAM lab. The need for blending scientific research with technical expertise has led to the involvement of Artificial Brain company whose mission is to enhance local resources in the development of a knowledge-based economy according to the EU policies[1]. The engineered artifact developed by Artificial Brain is DOMINUS$^{plus}$.

## 2 Scientific Challenges

The project's objective was to create a flexible and extensible framework to cover all aspects and functionality involved in the management of digital libraries, from the acquisition of new documents to the retrieval of documents considered as interesting for a given search query, focusing in particular on the semantics of the documents content. The DOMINUS framework processes digital documents through a pipeline consisting of several steps, aimed at acquiring increasingly abstract information from each incoming document, and specifically:

- Acquisition. Documents in various digital formats are acquired and converted to a unified representation expressing both structure and content.
- Layout Analysis. The various components that make up the structure of document pages are extracted and organized into a hierarchical structure.
- Document Image Understanding. The kind/class of the document is identified and each component of the layout structure is associated with a label that expresses its logical role in the document (title, author, summary, etc.).
- Text Analysis. The text in the relevant components is extracted and then processed using NLP techniques.
- Categorization. The document is assigned to a category expressing its domain of interest.
- Information Extraction. Additional relevant information is extracted from the document.

Each of these steps poses specific research problems, most addressed and tackled using Artificial Intelligence techniques, several of which developed at LACAM.

## 3 Key Technologies

The role of Artificial Brain in the project was to engineer and implement the framework developed at LACAM for supporting different tasks according to current state-of-the-art technologies that could provide added value to the functionality of the single components, especially in the usability and scalability perspective. The DOMINUS$^{plus}$ architecture consists of several related components (see Figure 1), each designed to implement a particular functionality in the context of three specific tasks: Layout Analysis, Document Image Understanding and Document Understanding and Indexing. For document acquisition

[1] The EC Community Strategic Guidelines on Cohesion 2007-2013 - `http://ec.europa.eu/regional_policy/information/guidelines/index_en.cfm#1`

and layout analysis tasks, formats that are complaint to the Open Document Architecture and Interchange Format (ODA) standard (PDF, PS, ODT, etc..) have been considered.

The Document Image Understanding task is carried out using the Inductive Logic Programming system INTHELEX, developed at LACAM, and transforming it into a Web service providing the following functionality:

- Knowledge Base Management. It associates, to each user, a workspace organized in projects, in turn consisting of sets of theories, within which the user can create/import theories and make on them the subsequent operations.
- Classification. Given an observation and a theory, it allows to calculate the confidence that the observation belongs to each concept defined in the theory.
- Unification. Given two or more theories, it allows to create a unified theory containing all the concepts contained within these theories.
- Refinement. Given a theory and a new example on which it fails, it allows to refine the theory so that it becomes consistent with all the previously observed examples and the current one.

As to the Document Understanding and Indexing tasks, several established techniques and libraries were used throughout. For instance, in Text Recognition (e.g., GhostScript for PS/PDF documents and Tesseract for graphic components), Information Retrieval based on both classical term-based and advanced concept-based indexing techniques (including a Vector Space Model based on TF-IDF, Latent Semantic Indexing based on the log-entropy weighting scheme and Concept Indexing), Information Extraction (including several Keyword Extraction techniques based on different perspectives and Formal Concept Analysis for identifying interesting concepts). A full set of NLP techniques underlies all these tasks, including both standard (tokenization, language recognition, stopword removal, PoS tagging, stemming) and advanced ones (syntactic and semantic analysis, e.g., the Stanford parser, WordNet and WordNet Domains).

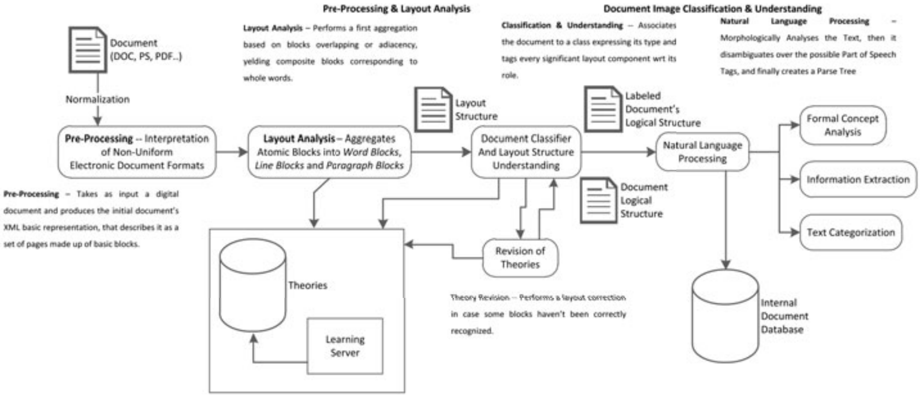
Pre-Processing & Layout Analysis
Document Image Classification & Understanding
Layout Analysis – Performs a first aggregation based on blocks overlapping or adiacency, yelding composite blocks corresponding to whole words.
Classification & Understanding – Associates the document to a class expressing its type and tags every significant layout component wrt its role.
Natural Language Processing – Morphologically Analyses the Text, then it disambiguates over the possible Part of Speech Tags, and finally creates a Parse Tree
Document (DOC, PS, PDF..)
Normalization
Layout Structure
Labeled Document's Logical Structure
Pre-Processing -- Interpretation of Non-Uniform Electronic Document Formats
Layout Analysis – Aggregates Atomic Blocks into Word Blocks, Line Blocks and Paragraph Blocks
Document Classifier And Layout Structure Understanding
Natural Language Processing
Formal Concept Analysis
Information Extraction
Text Categorization
Document Logical Structure
Pre-Processing – Takes as input a digital document and produces the initial document's XML basic representation, that describes it as a set of pages made up of basic blocks.
Revision of Theories
Theories
Learning Server
Theory Revision – Performs a layout correction in case some blocks haven't been correctly recognized.
Internal Document Database

**Fig. 1.** DOMINUS$^{plus}$ Architecture

Overall, the architecture makes extensive use of the object-relational mapping system called Hibernate, which enhances performances and allows the decoupling from a particular DBMS implementation.

## 4 Contribution by Research Group

Among the many techniques and approaches implemented in DOMINUS$^{plus}$, the main contribution from the LACAM, representing the core of the framework and deserving more attention, is the logical learning tool. INTHELEX (Incremental Theory Learner from Examples) is an Inductive Logic Programming (ILP) system capable of learning hierarchical theories from positive and negative examples that adopts Datalog$^{OI}$ as a representation language. Its peculiarity consists in a fully incremental behavior (in addition to refining previously generated concept definitions, learning can also start from an empty theory). INTHELEX is able to learn simultaneously several (possibly inter-related) concepts/classes, ensuring the validity of theories learned in every moment. Specifically, it incorporates two inductive refinement operators for the revision of theories: one for generalizing definitions that reject positive examples, and the other for the specialization of definitions that explain negative examples. If a positive example is not covered, the system first attempts to generalize one of the available concept definitions referred to by the example, so that the resulting revised theory, covers the new example and is consistent with all previous negative examples. If a generalization of this type is found, it replaces the previous definition in the theory, or else a new clause is chosen for the calculation of the generalization. If the system cannot generalize any definition in a consistent manner, it attempts to add the negation of a condition in order to discriminate the negative example from all the previous positive. If this does not lead to results, the negative example is added to the theory as an exception, and each new observation will be compared with the exception before making inferences about theories. In addition to the inductive operators, INTHELEX is endowed with multistrategy reasoning capabilities based on Deduction (to identify information that is implicit in the observations), Abduction (to handle partial information) and Abstraction (to switch to more expressive description languages).

**Acknowledgements.** DOMINUS$^{plus}$ is an open project promoted by LACAM (Laboratorio per l'Acquisizione della Conoscenza e l'Apprendimento nelle Macchine - `http://lacam.di.uniba.it:8000/`) and Artificial Brain S.r.l. (`http://www.abrain.it/`).

## Reference

1. Ferilli, S.: Automatic Digital Document Processing and Management: Problems, Algorithms and Techniques. Springer Publishing Company, Incorporated (2011)