# Cooperation of Multiple Strategies for Automated Learning in Complex Environments

Floriana Esposito, Stefano Ferilli, Nicola Fanizzi,
Teresa Maria Altomare Basile, and Nicola Di Mauro

Dipartimento di Informatica
Università di Bari
via E. Orabona, 4 - 70125 Bari - Italia
{esposito, ferilli, fanizzi, basile, nicodimauro}@di.uniba.it

**Abstract.** This work presents a new version of the incremental learning system INTHELEX, whose multistrategy learning capabilities have been further enhanced. To improve effectiveness and efficiency of the learning process, pure induction and abduction have been augmented with abstraction and deduction. Some results proving the benefits that the addition of each strategy can bring are also reported. INTHELEX will be the learning component in the architecture of the EU project COLLATE, dealing with cultural heritage documents.

## 1 Introduction

Automatic revision of logic theories, that empirical results have shown to yield more accurate definitions from fewer examples than pure induction, is a complex and computationally expensive task. In fact, most systems for theory revision deal with propositional logic and try to modify an existing incorrect theory to fit a set of pre-classified training examples. Other systems revise first-order theories, to overcome the expressive and representational limits showed by the propositional ones. Most of them try to limit the search space by exploiting information and, generally, require a wide, although incomplete, domain theory or a deep knowledge acquired from the user. Some others strongly rely on the interaction with the user, or adopt sophisticated search strategies or more informative search structures. Others do not allow negative information items to be expressed in the theories because of computational complexity considerations. Many of such systems adopt multi-strategy approaches integrating several types of inferential mechanisms. Such considerations, plus the need of testing theoretical results on the Object Identity paradigm [6] in practice, led to the design and implementation of INTHELEX. Its most characterizing features (compared in [5] with similar systems) are in its incremental nature, in the reduced need of a deep background knowledge, in the exploitation of negative information and in the peculiar bias on the generalization model, which reduces the search space and does not limit the expressive power of the adopted representation language.

The following Section presents the inductive core of INTHELEX; Section 3 shows how other reasoning strategies were added and provides some results; Section 4 introduces the EU project COLLATE; Section 5 draws some conclusions.

M.-S. Hacid et al. (Eds.): ISMIS 2002, LNAI 2366, pp. 574–582, 2002.

## 2 INTHELEX: The Inductive Core

INTHELEX (INcremental THEory Learner from EXamples) is a learning system for the induction of *hierarchical* logic theories from examples [5]: it is *fully incremental* (in addition to the possibility of refining a previously generated version of the theory, learning can also start from an empty theory); it is based on the *Object Identity assumption* (terms, even variables, denoted by different names within a formula must refer to different objects)[1]; it learns theories expressed as sets of Datalog$^{OI}$ clauses [12] from positive and negative examples; it can learn simultaneously *multiple concepts*, possibly related to each other according to a given hierarchy (recursion is not allowed); it retains all the processed examples, so to guarantee validity of the learned theories on all of them; it is a *closed loop* learning system (i.e. a system in which feedback on performance is used to activate the theory revision phase [1]).

Incremental learning is necessary when either incomplete information is available at the time of initial theory generation, or the nature of the concepts evolves dynamically. Both cases are very frequent in real-world situations, hence the need for incremental models to complete and support the classical batch ones, that perform learning in one step and thus require the whole set of observations to be available since the beginning. INTHELEX incorporates two refinement operators, one for generalizing hypotheses that reject positive examples, and the other for specializing hypotheses that explain negative examples. It exploits a (possibly empty) previous theory, a graph describing the dependence relationships among concepts, and a historical memory of all the past examples that led to the current theory. Whenever a new example is taken into account, it is stored in such a repository and the current theory is checked against it.

If it is positive and not covered, generalization must be performed. One of the clauses defining the concept the example refers to is chosen by the system for generalization. The lgg$_{OI}$ of this clause and the example is computed [12], by taking into account a number of parameters that restrict the search space according to the degree of generalization to be obtained and the computational budget allowed. If one of the lgg$_{OI}$'s is consistent with all the past negative examples, then it replaces the chosen clause in the theory, or else a new clause is chosen to compute the lgg$_{OI}$. If no clause can be generalized in a consistent way, the system checks if the example itself, with the constants properly turned into variables, is consistent with the past negative examples. If so, such a clause is added to the theory, or else the example itself is added as an exception.

If the example is negative and covered, specialization is needed. Among the theory clauses occurring in the SLD-derivation of the example, INTHELEX tries to specialize one at the lowest possible level in the dependency graph by adding to it one (or more) positive literal(s), which characterize all the past positive examples and can discriminate them from the current negative one. Again, parameters that bound the search for the set of literals to be added are considered.

[1] This often corresponds to human intuition, while allowing the search space to fulfill nice properties affecting efficiency and effectiveness of the learning process [12].

In case of failure on all of the clauses in the derivation, the system tries to add the negation of a literal, that is able to discriminate the negative example from all the past positive ones, to the clause related to the concept the example is an instance of. If this fails too, the negative example is added to the theory as an exception. New incoming observations are always checked against the exceptions before applying the rules that define the concept they refer to.

## 3 Multistrategy Learning

While at the beginning ML research focused on single-strategy methods that apply a primary type of inference and/or computational mechanism, more recently the limitations of these methods led to exploit/combine various, different and complementary learning strategies together. This mimes the typical ability of humans to apply a great variety of learning strategies depending on the particular situation and problem faced. A theoretical framework for integrating different learning strategies is the Inferential Learning Theory [10].

Another peculiarity in INTHELEX is the integration of multistrategy operators that may help in the solution of the theory revision problem by preprocessing the incoming information [6]. Namely, deduction is exploited to fill observations with information that is not explicitly stated, but is implicit in their description, and hence refers to the possibility of better representing the examples and, consequently, the inferred theories. Conversely, abduction aims at completing possibly partial information in the examples (adding more details), whereas abstraction removes superfluous details from the description of both the examples and the theory. Thus, even if with opposite perspectives, both aim at reducing the computational effort required to learn a correct theory with respect to the incoming examples. More details on the theoretical foundations of the cooperation of these strategies in our environment are given in [3], whereas this paper focuses on their implementation and cooperation into a single system.

### 3.1 Deduction

INTHELEX requires the observations to be expressed only in terms of the set of predicates that make up the description language for the given learning problem. To ensure uniformity of the example descriptions, such predicates have no definition. Nevertheless, since the system is able to handle a hierarchy of concepts, combinations of these predicates might identify higher level concepts that is worth adding to the descriptions in order to raise their semantic level. For this reason, INTHELEX implements a saturation operator that exploits deduction to recognize such concepts and explicitly add them to the examples description.

The system can be provided with a Background Knowledge, supposed to be correct and hence not modifiable, containing (complete or partial) definitions in the same format as the theory rules. This way, any time a new example is considered, a preliminary saturation phase can be performed, that adds the higher level concepts whose presence can be deduced from such rules by subsumption and/or resolution. In particular, the generalization model of implication under Object Identity is exploited [4]. Given a set of terms $T$, a substitution $\sigma$ is an *OI-substitution w.r.t.* $T$ iff $\forall t_1, t_2 \in T : t_1 \neq t_2$ implies $t_1\sigma \neq t_2\sigma$. In this setting, an interpretation $I$ is an *OI-model* for the clause $C$ iff for all ground OI-substitutions $\gamma$ it holds that $I \cap C\gamma \neq \emptyset$. A set of clauses $\Sigma$ *OI-implies* a clause $C$ ($\Sigma \models_{OI} C$) iff all OI-models $I$ for $\Sigma$ are also OI-models for $C$. A sound and refutation-complete proof-theory has been built upon this semantics, by defining notions of OI-unifiers, OI-resolution and OI-derivation ($\vdash_{OI}$). It holds that:

**Theorem 1 (Subsumption Theorem).** *Let $\Sigma$ be a finite set of clauses and $C$ be a clause. Then $\Sigma \models_{OI} C$ iff there exists a clause $D$ such that $\Sigma \vdash_{OI} D$ and $D$ $\theta_{OI}$-subsumes $C$.*

Differently from abstraction (see next), all the specific information used by saturation is left in the example description. Hence, it is preserved in the learning process until other evidence reveals it is not significant for the concept definition, which is a more cautious behaviour. This is fundamental if some concept to be learnt are related, since their definition could not be stable yet, and hence one cannot afford to drop the source from which deductions were made in order to be able to recover from deductions made because of wrong rules.

### 3.2 Abduction

Induction and abduction are, both, important strategies to perform hypothetical reasoning (i.e., inferences from incomplete information). Induction means inferring from a certain number of significant observations regularities and laws valid for the whole population. Abduction was defined by Peirce as hypothesizing some facts that, together with a given theory, could explain a given observation.

According to the framework proposed in [8], an *abductive logic theory* is made up by a normal logic program [9], a set of *abducibles* and a set of *integrity constraints* (each corresponding to a combination of literals that is not allowed to occur). Abducibles are the predicates about which assumptions (*abductions*) can be made: They carry all the incompleteness of the domain (if it were possible to complete these predicates then the theory would be correctly described). Integrity constraints provide indirect information about them and, since several explanations may hold for this problem setting, are also exploited to encode preference criteria for selecting the best ones. The proof procedure implemented in INTHELEX starts from a goal and a set of initial assumptions and results in a set of consistent hypotheses (abduced literals) by intertwining *abductive* and *consistency derivations*. Intuitively, an abductive derivation is the standard Logic Programming derivation suitably extended in order to consider abducibles. As soon as an abducible atom $\delta$ is encountered, it is added to the current set of hypotheses, provided that any integrity constraint containing $\delta$ is satisfied. This is checked by starting a consistency derivation. Every integrity constraint containing $\delta$ is considered satisfied if the goal obtained by removing $\delta$ from it fails. In the consistency derivation, when an abducible is encountered, an abductive derivation for its complement is started in order to prove its falsity.

An experiment was run to test if and how much the addition of abduction could improve the learning process. It aimed at learning definitions for a class of paper documents starting from the empty theory. The learning set consisted of 11 positive and 11 negative examples; the test set for performance evaluation was composed of 6 positive and 9 negative examples. Incomplete documents were described by about 30 literals, complete ones by about 100 literals. The description language included predicates concerning the size, type and relative position of the layout blocks. Running the system with abduction not enabled, the resulting theory was made up of 2 clauses (including 10 and 14 literals, respectively), obtained through 6 successive generalizations. The predictive accuracy of such a theory on the test set was 86%. In order to exploit abduction, all the basic predicates in the description language were considered as abducibles, while integrity constraints expressed the mutual exclusion among layout block sizes, types and positions, and the non-reflexivity of the relative positions among blocks. Abduction makes sense in this environment since the absence of a layout block in a document could be due to the writer not fulfilling the style requirements, and not to the insignificance of that block to a correct definition. In other words, a block should not be drop from the definition just because a few examples miss it; conversely, integrity constraints are in charge of avoiding that superfluous blocks that are found in the first few examples introduce unnecessary blocks that can be always abduced in the future. The resulting theory was now made up of just 1 clause of 18 literals, obtained through only 2 generalizations and 7 abductions. This means that in some cases abduction succeeded in covering the examples without firing the refinement operators, and hence the system was able to characterize the target concept by means of less clauses. Again, an 86% accuracy was reached, that grew up to 100% if allowing INTHELEX to exploit abduction also when evaluating the documents in the test set.

### 3.3 Abstraction

Abstraction is a pervasive activity in human perception and reasoning. When we are interested in the role it plays in ML, inductive inference must be taken into account as well. The exploitation of abstraction concerns the shift from the language in which the theory is described to a higher level one.

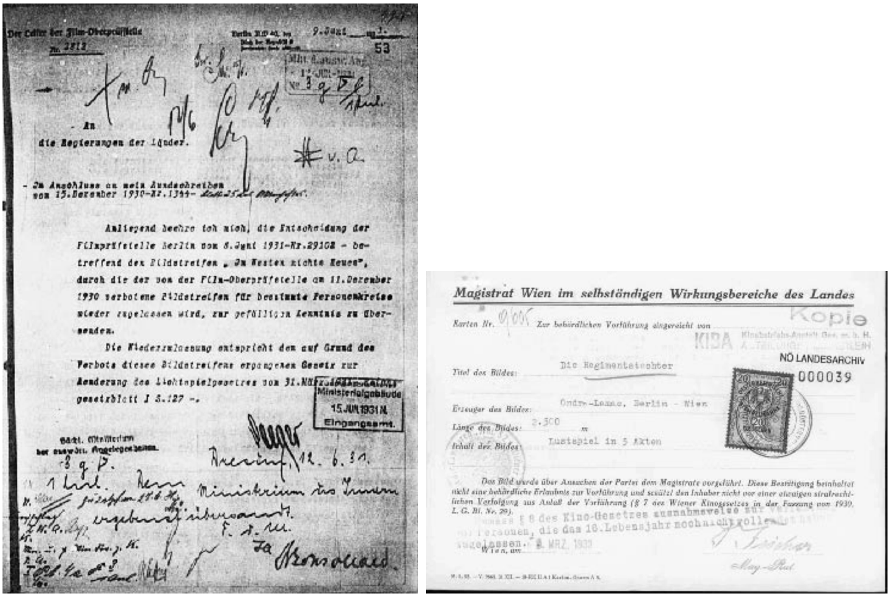
According to the framework proposed in ([14]), concept representation deals with entities belonging to three different levels. Concrete objects reside in the *world* $W$, but any observer's access to it is mediated by his *perception* of it $P(W)$. To be available over time, these stimuli must be memorized in an organized *structure* $S$, i.e. an *extensional* representation of the perceived world. Finally, to reason about the perceived world and communicate with other agents, a *language* $L$ is needed, that describes it *intensionally.* If we assume that $P(W)$ is the source of information, that is recorded into $S$ and then described by $L$, modifications to the structure and language are just a consequence of differences in the perception of the world (due, e.g., to the medium used and the focus-of-attention). Thus, abstraction takes place at the world-perception level by means of a set of operators, and then propagates to higher levels, where it is possible to identify operators corresponding to the previous ones. An abstraction theory contains information for performing the shift specified by the abstraction operators. In INTHELEX, it is assumed that the abstraction theory is already given (i.e. it has not to be learned by the system), and that the system automatically applies it to the learning problem at hand before processing the examples. The implemented abstraction operators allow the system to replace a number of components by a compound object, to decrease the granularity of a set of values, to ignore whole objects or just part of their features, and to neglect the number of occurrences of some kind of object.

The effectiveness of the abstraction operator introduced in INTHELEX was tested on the problem of Text Categorization [11], in order to infer rules for recognizing the subject of a document. Natural language is particularly suitable for abstraction, owing to the presence of many terms that are synonyms or whose meaning differs just slightly, since without binding them onto a common concept it would be impossible for an automatic learning system to grasp the similarities between two lexically different sentences. In order to obtain, from raw text, the structured representations of sentences that can be expressed in the input language required by the symbolic learner, a parser was used as a preprocessor. In the formal representation of texts, we used descriptors expressing the logical/grammatical role and the stem of the words in a sentence.

Experiments were run on the documents used for abduction, concerning foreign commerce. One aimed at learning the concept of "import". Starting from the empty theory, INTHELEX was fed with a total of 67 examples, 39 positive (not all explicitly using verb 'to import') and 28 negative, and yielded a theory composed by 9 clauses. Some of them were only slightly different, considering that: 'enterprise', 'society', 'firm' and 'agency' all can be seen as instances of the concept '*company*'; 'provider' and 'distributor' play the same role (let us call it '*providing_role*'); 'to look for' and 'to be interested in' are almost synonyms (and hence may be grouped in one category, say '*interest_cat*'); 'to buy' and 'to import' bear more or less the same meaning of acquiring something (*'acquisition_cat'*). Exploiting such an ontological information as an abstraction theory results in a theory made up of just 3 rules (67% savings), thus confirming that the use of abstraction improves compactness and readability. Another experiment, aimed at learning the concept of "specialization" (of someone in some field), confirmed the above findings. The system was run on 40 examples, 24 positive and 16 negative. The resulting theory, originally made up of 5 clauses, by exploiting abstraction was reduced to 2 rules (60% savings).

## 4 The COLLATE Project

Many important historic and cultural sources, which constitute a major part of our cultural heritage, are fragile and distributed in various archives, which causes severe problems to full access, knowledge and usage. Moreover, many informal and non-institutional contacts between archives constitute specific professional communities, which today still lack effective and efficient technological



Der Leiter der Film-Oberprüfstelle
Nr. 3812

Berlin, 9. Juni 1931.
53

Rücks. Sachs. Ang.
12. JUN. 1931
Nr. 3 g P.

An
die Regierungen der Länder.

- Im Anschluss an mein Rundschreiben vom 15.Dezember 1930-Nr.1344-

Anliegend beehre ich mich, die Entscheidung der Filmprüfstelle Berlin vom 8.Juni 1931-Nr.29102 - betreffend den Bildstreifen „Im Westen nichts Neues", durch die der von der Film-Oberprüfstelle am 11.Dezember 1930 verbotene Bildstreifen für bestimmte Personenkreise wieder zugelassen wird, zur gefälligen Kenntnis zu übersenden.

Die Wiederzulassung entspricht dem auf Grund des Verbots dieses Bildstreifens ergangenen Gesetz zur Aenderung des Lichtspielgesetzes vom 31.März 1931 - Reichsgesetzblatt I S.127 -.

Ministerialgebäude
15.JUN.1931 N.
Eingangsamt.

Sächs. Ministerium
der auswärt. Angelegenheiten.

3 g P.
Dresden, 12. 6. 31.
1 Abl. Dem Ministerium des Innern ergebenst übersandt.

Magistrat Wien im selbständigen Wirkungsbereiche des Landes

Kopie

Karten Nr. 9/005 Zur behördlichen Vorführung eingereicht von
KIBA Kinobetriebs-Anstalt Ges. m. b. H.

NÖ LANDESARCHIV
000039

Titel des Bildes: Die Regimentstochter

Erzeuger des Bildes: Ondra-Lamac, Berlin - Wien

Länge des Bildes: 2.300 m

Inhalt des Bildes: Lustspiel in 5 Akten

Das Bild wurde über Ansuchen der Partei dem Magistrate vorgeführt. Diese Bestätigung beinhaltet nicht eine behördliche Erlaubnis zur Vorführung und schützt den Inhaber nicht vor einer etwaigen strafrechtlichen Verfolgung aus Anlaß der Vorführung (§ 7 des Wiener Kinogesetzes in der Fassung von 1939, L. G. Bl. Nr. 29).

Gemäss § 8 des Kino-Gesetzes ausnahmsweise zur Vorführung vor Personen, die das 16. Lebensjahr noch nicht vollendet haben, zugelassen.

Wien, am 2. MÄRZ 1939

Mag.-Rat

**Fig. 1.** Sample COLLATE documents

support for cooperative and collaborative knowledge working. The IST-1999-20882 project COLLATE (Collaboratory for Annotation, Indexing and Retrieval of Digitized Historical Archive Material) aims at developing a WWW-based *collaboratory* [7] for archives, researchers and end-users working with digitized historic/cultural material (URL: `http://www.collate.de`).

Though the developed tools and interfaces are generic, the chosen sample domain concerns historic film documentation. Multi-format documents on European early 20th century films, provided by three major European film archives, include a large corpus of rare historic film censorship documents from the 20ies and 30ies, but also newspaper articles, photos, stills, posters and film fragments. In-depth analysis and comparison of such documents can give evidence about different film versions and cuts, and allow to restore lost/damaged films or identify actors and film fragments of unknown origin. All material is analyzed, indexed, annotated and interlinked by film experts, to which the COLLATE system will provide suitable task-based interfaces and knowledge management tools to support individual work and collaboration. Continuously integrating the hereby-derived user knowledge into its digital data and metadata repositories, it can offer improved content-based retrieval functionality. Thus, enabling users to create and share valuable knowledge about the cultural, political and social contexts in turn allows other end-users to better retrieve and interpret the historic material.

Supported by previous successful experience in the application of symbolic learning techniques to paper documents [6,13], our aim is applying INTHELEX to these documents. The objective is learning to automatically identify and label document classes and significant components, to be used for indexing/retrieval purposes and to be submitted to the COLLATE users for annotation. Combining results from the manual and automatic indexing procedures, elaborate content-based retrieval mechanisms can be applied [2]. The challenge comes from the low layout quality and standard of such a material, which introduces a considerable amount of noise in its description (see Fig.1). As regards the layout quality, it is often affected by manual annotations, stamps that overlap to sensible components, ink specks, etc.. As to the layout standard, many documents are typewritten sheets, that consist of all equally spaced lines in Gothic type. Such a situation should account for a profitable use of the abduction and abstraction features of INTHELEX: While the former could make the system more flexible in the absence of particular layout components due to the typist's style, the latter could help in ignoring layout details that are meaningless or superfluous to the identification of the interesting ones.

Preliminary experiments showed that INTHELEX is able to distinguish at least 3 classes of COLLATE censorship documents, and to single out a number of logical components inside them. For instance, it learns rules that can separate the censorship authority, applicant and decision in documents like the one on the right in Fig. 1.

## 5 Conclusions and Future Work

Incremental approaches to machine learning can help in obtaining more efficiency, and are necessary in a number of real-world situations. The incremental system INTHELEX works on first-order logic representations. Its multistrategy learning capabilities have been further enhanced in order to improve effectiveness and efficiency of the learning process, by augmenting pure induction and abduction with abstraction and deduction. This paper presents some sample results proving the benefits that the addition of each strategy can bring. INTHELEX is included in the architecture of the EU project COLLATE, in order to learn rules for automatic classification and interpretation of cultural heritage documents dating back to the 20s and 30s. Future work will concern a more extensive experimentation, aimed at finding tighter ways of cooperation among the learning strategies, and an analysis of the complexity of the presented techniques. Moreover, the addition of numeric capabilities can be considered fundamental for effective learning in some contexts, and hence deserves further study

**Acknowledgements.** This work was partially funded by the EU project IST-1999-20882 COLLATE "Collaboratory for Annotation, Indexing and Retrieval of Digitized Historical Archive Material".

## References

[1] J. M. Becker. Inductive learning of decision rules with exceptions: Methodology and experimentation. B.s. diss., Dept. of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA, 1985. UIUCDCS-F-85-945.

[2] H. Brocks, U. Thiel, A. Stein, and A. Dirsch-Weigand. Customizable retrieval functions based on user tasks in the cultural heritage domain. In *Research and Advanced Technology for Digital Libraries. Proceedings of ECDL2001*, Lecture Notes in Computer Science. Springer, 2001.

[3] F. Esposito, N. Fanizzi, S. Ferilli, and G. Semeraro. Abduction and abstraction in inductive learning. In *Proceedings of the 5th International Workshop on Multistrategy Learning*, Guimarães, Portugal, 2000.

[4] F. Esposito, N. Fanizzi, S. Ferilli, and G. Semeraro. Oi-implication: Soundness and refutation completeness. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 847–852, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers.

[5] F. Esposito, G. Semeraro, N. Fanizzi, and S. Ferilli. Multistrategy Theory Revision: Induction and abduction in INTHELEX. *Machine Learning Journal*, 38(1/2):133–156, 2000.

[6] S. Ferilli. *A Framework for Incremental Synthesis of Logic Theories: An Application to Document Processing.* Ph.D. thesis, Dipartimento di Informatica, Università di Bari, Bari, Italy, November 2000.

[7] R.T. Kouzes, J.D. Myers, and W.A. Wulf. Collaboratories: Doing science on the internet. *IEEE Computer*, 29(8), 1996.

[8] E. Lamma, P. Mello, F. Riguzzi, F. Esposito, S. Ferilli, and G. Semeraro. Cooperation of abduction and induction in logic programming. In A. C. Kakas and P. Flach, editors, *Abductive and Inductive Reasoning: Essays on their Relation and Integration.* Kluwer, 2000.

[9] J. W. Lloyd. *Foundations of Logic Programming.* Springer-Verlag, Berlin, second edition, 1987.

[10] R.S. Michalski. Inferential theory of learning. developing foundations for multistrategy learning. In R.S. Michalski and G. Tecuci, editors, *Machine Learning. A Multistrategy Approach*, volume IV, pages 3–61. Morgan Kaufmann, San Mateo, CA, 1994.

[11] F. Sebastiani. Machine learning in automated text categorization. Technical Report Technical Report IEI:B4-31-12-99, CNR - IEI, Pisa, Italy, December 1999. Rev. 2001.

[12] G. Semeraro, F. Esposito, D. Malerba, N. Fanizzi, and S. Ferilli. A logic framework for the incremental inductive synthesis of Datalog theories. In N.E. Fuchs, editor, *Proceedings of 7th International Workshop on Logic Program Synthesis and Transformation - LOPSTR97*, volume 1463 of *LNCS*, pages 300–321. Springer, 1998.

[13] G. Semeraro, N. Fanizzi, S. Ferilli, and F. Esposito. Document classification and interpretation through the inference of logic-based models. In P. Constantopoulos and I.T. Sølvberg, editors, *Research and Advanced Technology for Digital Libraries*, number 2163 in Lecture Notes in Computer Science, pages 59–70. Springer-Verlag, 2001.

[14] J.-D. Zucker. Semantic abstraction for concept representation and learning. In R. S. Michalski and L. Saitta, editors, *Proceedings of the 4th International Workshop on Multistrategy Learning*, Desenzano del Garda, Italy, 1998.