Automatic Induction of Rules for Classification and Interpretation of Cultural Heritage Material

S. Ferilli, F. Esposito, T.M.A. Basile, and N. Di Mauro

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

Abstract. This work presents the application of incremental symbolic learning strategies for the automatic induction of classification and interpretation rules in the cultural heritage domain. Specifically, such experience was carried out in the environment of the EU project COLLATE, in whose architecture the incremental learning system INTHELEX is used as a learning component. Results are reported, proving that the system was able to learn highly reliable rules for such a complex task.

1 Introduction

Many important historic and cultural sources, which constitute a major part of our cultural heritage, are fragile and distributed in various archives, which still lack effective and efficient technological 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). The chosen sample domain concerns a large corpus of multi-format documents concerning rare historic film censorship from the 20's and 30's, but includes also newspaper articles, photos, stills, posters and film fragments, provided by three major European film archives. 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 aims at providing suitable task-based interfaces and knowledge management tools to support individual work and collaboration. Continuously integrating valuable knowledge about the cultural, political and social contexts into its digital data and metadata repositories, it will provide improved content-based functionality to better retrieve and interpret the historic material.

Supported by previous successful experience in the application of symbolic learning techniques to classification and understanding of paper documents [4, 6, 9], our aim 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. 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.

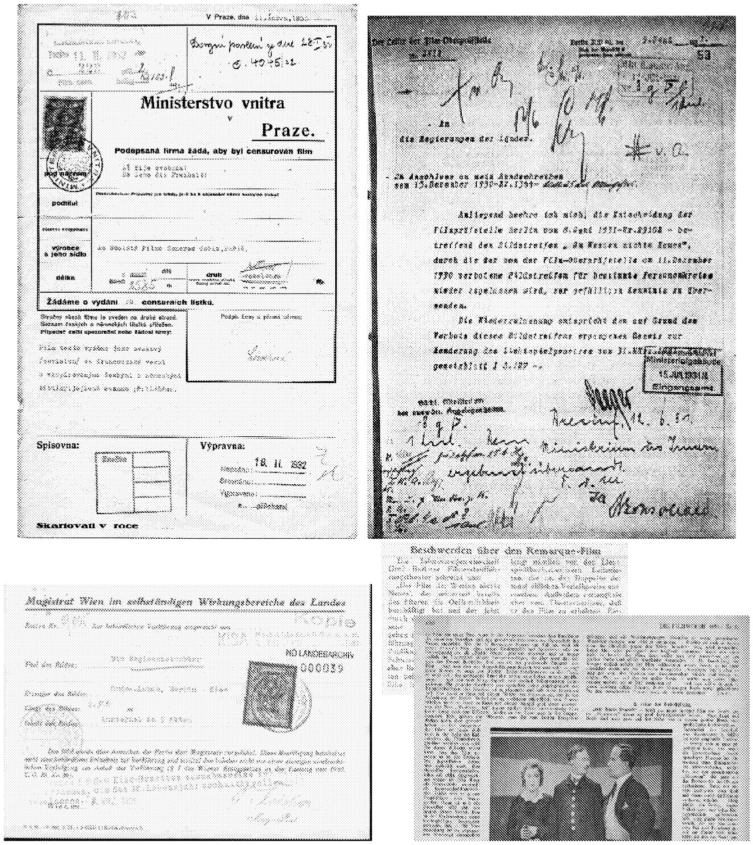
Our proposal, supported by successful experiments reported in this paper, is to exploit symbolic (first-order logic) learning techniques, whose high level representation can better manage the complexity of the task and allows the use of different reasoning strategies than pure induction with the objective of making the learning process more effective and efficient.

The following section introduces and describes the classes of documents that were considered for the experiments. Then, Section 3 presents the system INTHELEX along with its features, and Section 4 shows experimental results. Lastly, Section 5 draws some conclusions and outlines future work directions.

2 Documents

This Section aims at briefly describing the documents that were taken into account for the research described in this paper. The COLLATE repository, set up by the film archives DIF (Deutsches Filminstitut, Frankfurt am Main), FAA (Film Archive Austria, Vienna) and NFA (Nrodni Filmov Archiv, Prague), includes a large collection of several thousands comprehensive documents concerning film culture, and focuses on documents related to censorship processes (see Figure 1). The importance of censorship for film production distribution lies mainly in the fact that it is often impossible to identify a unique film. Often, there are lots of different film versions with cuts, changed endings and new intertitles, depending on the place and date of release. Exactly these differences are documented in censorship documents and allow statements about the original film. They define and identify the object of interest. Often they provide the only source available today for the reconstruction of the large number of films that have been lost or destroyed. Censorship documents support this restoration process by identifying and structuring the film fragments. They allow to put together film fragments from various copies in order to obtain a correct reconstruction. Each Country developed its own censorship history embedded in the political history. The collection is complemented by further documents like press articles, correspondence, photos, etc.

The sample document reported in the upper-left part of Figure 1 is an **Application Form** belonging to NFA. This kind of documents was required for applying to get the permission to show a film from a production or distribution company. This was common practice mainly in Czechoslovakia. The consequence of this application was the examination by the censorship office. The “application



Ministerstvo vnitra
v
Praze.
Podepsaná firma žádá, aby byl censurován film
Žádáme o vydání censurních lístků.
Spisovna:
Výpravna:
18. II. 1932
Skartovati v roce

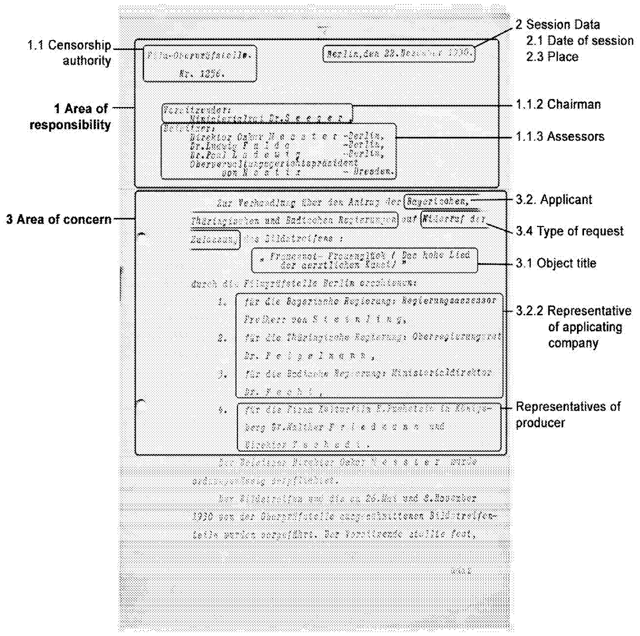
Magistrat Wien im selbständigen Wirkungsbereiche des Landes
NÖ LANDESARCHIV
Kopie

Beschwerden über den Remarque-Film

Fig. 1. Sample COLLATE documents

form" could be a source for information like: *Name of applicant* (production or distribution company), *title of the film*, *year of production*, *length* (before censorship), *brief content*, *information about earlier examinations*, etc. It was usually accompanied by a list of intertitles or dialogue list. Indeed, the applicant was obliged to enclose a list of intertitles or, in case of sound films, a list with the beginnings of the dialogues. These lists served to check whether a film shown in the cinema was the same as the one examined by the censorship office.

As regards the document shown in the upper-right side of Figure 1, it is an instance of **Censorship Decision**. This kind of documents are about the decision whether a film could or could not - and in which version - be distributed and shown throughout a Country. The Censorship Decision is often a protocol of the examination meeting and is issued by the censorship office. It provides information such as: *film title*, *participants in the examination*, *notices on content*, *juridical legitimization for the decision*, *legal motivation*, *conditions for permission* (for example cuts, change of title, etc.), *reference to previous de-*


1.1 Censorship authority
1 Area of responsibility
3 Area of concern
2 Session Data
2.1 Date of session
2.3 Place
1.1.2 Chairman
1.1.3 Assessors
3.2. Applicant
3.4 Type of request
3.1 Object title
3.2.2 Representative of applicating company
Representatives of producer

Film-Oberprüfstelle.

Nr. 1396.

Berlin, den 29. Dezember 1930.

Vorsitzender:
Ministerialrat Dr. Seeger.
Beisitzer:
Direktor Oskar Messter - Berlin,
Dr. Ludwig Fulda - Berlin,
Dr. Paul Ludewig - Berlin,
Oberverwaltungsgerichtspräsident
von Nostitz - Dresden.

Zur Verhandlung über den Antrag der Bayerischen, Thüringischen und Badischen Regierungen auf Widerruf der Zulassung des Bildstreifens :

„Frauennot- Frauenglück / Das hohe Lied der aerztlichen Kunst/ "

durch die Filmprüfstelle Berlin erschienen:

1. für die Bayerische Regierung: Regierungsassessor Freiherr von Steinling,
2. für die Thüringische Regierung: Oberregierungsrat Dr. Feigelmann,
3. für die Badische Regierung: Ministerialdirektor Dr. Fecht,
4. für die Firma Kulturfilm E.Fuchslein in Königsberg Dr.Walther Friedmann und Direktor Tschechi.

Der Beisitzer Direktor Oskar Messter wurde ordnungsmässig verpflichtet.

Der Bildstreifen und die am 26.Mai und 8.November 1930 von der Oberprüfstelle ausgeschnittenen Bildstreifenteile wurden vorgeführt. Der Vorsitzende stellte fest,

Nach

Fig. 2. Sample COLLATE Document Text Structure

cisions, *costs for the procedure*. For instance, Figure 2 shows the first page of one such document, with the interesting items already annotated by the experts. Indeed, it is noteworthy that almost all documents in the COLLATE collection are multipage, but generally just the first page contains information of interest.

The lower-left part of Figure 1 reports an example of **Registration Card**, a certification that the film had been approved for exhibition in the present version by the censoring authority. The registration cards were given to the distribution company which had to pay for this, and had to enclose the cards to the film copies. When the police checked the cinemas from time to time, the owner or projectionist had to show the registration card. Such cards constitute a large portion of the COLLATE collection, mainly provided by FAA and DIF, and are an important source for film reconstruction. They are a source for the following information: *Film title*, *production company*, *date and number of examination*, *length* (after censoring), *number of acts*, *brief content*, *forbidden parts*, *staff*.

The last documents in Figure 1 represent **Articles** from the contemporary film press, newspapers or magazines. They are necessary to reconstruct the context of a film, since they enlighten the reception background. One may also find debates and details on censorship there, because the press of every political direction watched closely the results of the examination.

3 The Learning Component

The need of automatically labelling the huge amount of documents in the COLLATE repository, along with their significant components, suggested the use of a learning system to learn rules for such tasks from a small number of selected and annotated sample documents. In particular, the complexity of the domain and the need for the rules to be understandable by film experts, led to the choice of symbolic first-order learning.

INTHELEX (INcremental THEory Learner from EXamples) is a learning system for the induction of *hierarchical* logic theories from examples [5]: it learns theories expressed in a first-order logic representation from positive and negative examples; it can learn simultaneously *multiple concepts*, possibly related to each other (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]); 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).

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 version of the theory, a graph describing the dependence relationships among concepts, and an 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 definitions of the concept the example refers to is chosen by the system for generalization. If a generalization can be found that is consistent with all the past negative examples, then it replaces the chosen definition in the theory, or else another definition is chosen to be generalized. If no definition can be generalized in a consistent way, the system checks if the exact shape of the example itself can be regarded as a definition that is consistent with the past negative examples. If so, it 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 definitions involved in the example coverage, INTHELEX tries to specialize one at the lowest possible level in the dependency graph by adding to it positive information, which characterize all the past positive examples and can discriminate them from the current negative one. In case of failure on all of the considered definitions, the system tries to add negative information, that is able to discriminate the negative example from all the past positive ones, to the definition 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.

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], according to the theoretical framework for integrating different learning strategies known as Inferential Learning Theory [8]. 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.

To ensure uniformity of the example descriptions, INTHELEX requires the observations to be expressed only in terms of basic predicates that have no definition. Nevertheless, 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 exploits deduction to recognize such concepts and explicitly add them to the examples description. For doing this, it 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.

Abduction was defined by Peirce as hypothesizing some facts that, together with a given theory, could explain a given observation. 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 (each corresponding to a combination of literals that is not allowed to occur) provide indirect information about them. The proof procedure implemented in INTHELEX starts from a goal and a set of initial assumptions and results in a set of consistent hypotheses by intertwining *abductive* and *consistency derivations*.

The exploitation of abstraction concerns the shift from the language in which the theory is described to a higher level one. Abstraction takes place at the world-perception level, and then propagates to higher levels, by means of a set of operators. An abstraction theory contains information for performing the shift specified by such operators, that 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. In INTHELEX the abstraction theory must be given, and the system automatically applies it to the learning problem at hand before processing the examples.

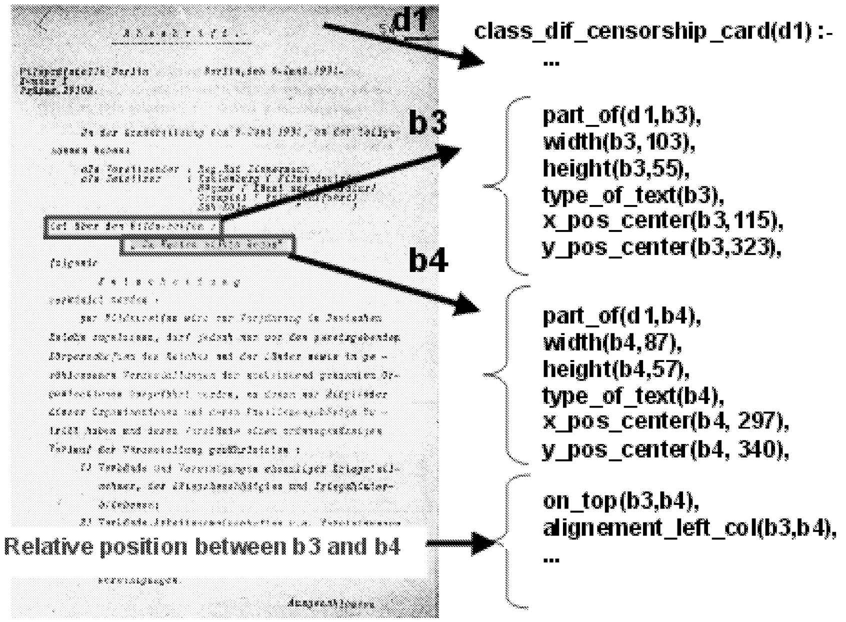


Fig. 3. An example of processed document

4 Experimental results

INTHELEX was considered a suitable learning component for the COLLATE architecture based on its previous successful application to different kinds of documents, indicating a good generality of the approach. Moreover, many of its features met the requirements imposed by the complexity of the documents to be handled. In addition to being a symbolic (first-order logic) incremental system, its multistrategy capabilities seemed very useful. For instance, abduction could make the system more flexible in the absence of particular layout components due to the typist's style, while abstraction could help in focusing on layout patterns that are meaningful to the identification of the interesting details, neglecting less interesting ones. Experimental results, reported in the following, confirm the above expectations.

The COLLATE dataset for INTHELEX consisted of 29 documents for the class `faa_registration_card`, 36 ones for the class `dif_censorship_decision`, 24 for the class `nfa_cen_dec_model_a` and 13 for the class `nfa_cen_dec_model_b`. Other 17 reject documents were obtained from newspaper articles. Note that the symbolic method adopted allows the experiment supervisor to specifically select prototypical examples to be included in the learning set. This explains why theories with good predictiveness can be obtained even from few observations.

The first-order descriptions of such documents, needed to run INTHELEX, were automatically generated by the system WISDOM++ [3]. Starting from

```
pos_left(X):-
    x_pos_centre(X,Y), Y >= 0, Y =< 213.
pos_center(X):-
    x_pos_centre(X,Y), Y >= 214, Y =< 426.
pos_right(X):-
    x_pos_centre(X,Y), Y >= 427.
```

Fig. 4. Abstraction rules for horizontal block position

scanned images, it is able to identify in few seconds the layout blocks that make up a paper document and to describe them in terms of their size (height and width, in pixels), position (horizontal and vertical, in pixels from the top-left corner), type (text, line, picture and mixed) and relative position (horizontal/vertical alignment between two blocks, adjacency). It is not a domain-specific system, since it has already been used to process several other kinds of documents, such as commercial letters and scientific papers. Figure 3 shows an example of a document and its description.

Since the inductive procedure embedded in INTHELEX is not able to handle numeric values (such as the number of pixels in the document descriptions provided by WISDOM++), a change of representation in the description language was necessary, such that final observations were made up of symbolic attributes only. The abstraction operator was used for breaking numeric values into intervals represented by symbolic constants (*discretization*), by providing the system with an Abstraction Theory containing rules that encode such a language shift. Figure 4 shows the rules of the Abstraction Theory that are in charge of discretizing the horizontal position of a layout block in a document.

The complexity of the domain is confirmed by the description length of the documents, that ranges between 40 and 379 literals (144 on average) for class `faa_registration_card`, between 54 and 263 (215 on average) for class `dif_censorship_decision`; between 105 and 585 (269 on average) for class `nfa_cen_dec_model_a` and between 191 and 384 literals (260 on average) for class `nfa_cen_dec_model_b`. It is worth noting that the description length after the abstraction process on numeric features doesn't change (increase/decrease) with respect to the original one, since each numeric value is now represented by a corresponding symbolic value.

Each document was considered as a positive example for the class it belongs, and as a negative example for the other classes to be learned; reject documents were considered as negative examples for all classes. Definitions for each class were learned, starting from the empty theory, and their predictive accuracy was tested according to a 10-fold cross validation methodology, ensuring that each fold contained the same proportion of positive and negative examples. Table 1 reports the experimental results, averaged on the 10 folds, of the classification process in this environment as regards number of clauses that define the concept (Cl), number of performed generalizations (lgg), Accuracy on the test set (expressed in percentage) and Runtime (in seconds).

```
class_dif_cen_decision(A) :-
    image_lenght_long(A), image_width_short(A),
    part_of(A, B), type_of_text(B),
    width_medium_large(B), height_very_very_small(B),
    pos_left(B), pos_upper(B),
    part_of(A, C), type_of_text(C),
    height_very_very_small(C),
    pos_left(C), pos_upper(C),
    on_top(C, D), type_of_text(D),
    width_medium_large(D), height_very_very_small(D),
    pos_left(D), pos_upper(D).
```

Fig. 5. Example of learned definition

As regards the rules learned by the system, Figure 5 shows a definition for the classification of documents belonging to `dif_censorship_decision` class. An explanation of the concept according to this rule is "a document belongs to this class if it has long length and short width, it contains three components in the upper-left part, all of type text and having very short height, two of which are medium large and one of these two is on top of the third". Two remarks are worth for this class: first, the features in this description are common to all the learned definitions in the 10 folds, which suggests that the system was able to catch the significant components and explains why its performance on this class is the best of all; second, starting with descriptions whose average length was 215, the average number of literals in the learned rules is just 22.

Each document is composed by blocks whose labels regard the role they play in it. For instance, in one of the three blocks appearing in the rule in Figure 5 the experts recognized a "session_data" item. The curiosity of checking the correctness of such a guess was one of the motivations to run additional experiments aimed at learning definitions for the semantic labels of interest for each document class. Indeed, different document classes have different labels. As regard the first class of documents, `faa_registration_card`, the domain experts provided the following labels characterizing the objects belonging to it (in square brackets the number of items in the document dataset): *registration_au* [28+], *date_place* [26+], *department* [17+], *applicant* [11+], *reg_number* [28+] , *film_genre* [20+], *film_length* [19+], *film_producer* [18+], *film_title* [20+]. Like in the classification step, each example is positive for the label(s) it belongs to, and

Table 1. Statistics for Document Classification

	Cl	Lgg	Accuracy	Runtime
DIF	1.00	7.50	99.17	17.13
FAA	3.50	9.70	94.17	334.05
NFA_A	2.80	7.30	93.92	87.71
NFA_B	1.70	5.40	97.56	92.05

negative for all the others. Again, a 10-fold cross-validation was applied, and the results were averaged (see Table 2).

Table 2. Statistics for Understanding FAA

	Cl	Lgg	Accuracy	Runtime
registration_au	5.6	12.5	91.43	3739.366
date_place	6.9	13.5	86.69	7239.625
department	1.9	6.6	98.95	118.625
applicant	2	4.5	97.89	93.993
reg_number	5.1	14.4	91.95	4578.208
film_genre	4	8.4	93.02	2344.899
film_length	5.5	9.9	90.87	3855.391
film_producer	4.9	10.4	94.05	4717.17
film_title	5.4	11.1	89.85	4863.084

The labels specified for class `dif_censorship_decision` were: *cens_signature* [35+], *cert_signature* [35+], *object_title* [36+], *cens_authority* [36+], *chairman* [36+], *assessors* [36+], *session_data* [36+], *representative* [49+]. Table 3 shows the results of a 10-fold cross-validation run on this dataset.

Table 3. Statistics for Understanding DIF

	Cl	Lgg	Accuracy	Runtime
cens_signature	2.2	11.6	98.32	1459.883
cert_signature	2.2	7.6	98.31	176.592
object_title	5	15.2	94.66	3960.829
cens_authority	2.9	12.1	97.64	2519.45
chairman	4.6	13.8	93.10	9332.845
assessors	4.6	15	94.48	12170.93
session_data	2.5	8.6	97.68	1037.96
representative	5.6	20.7	92.98	13761.958

Finally, class `nfa_cen_dec_model_a` was characterized by these labels, almost all different from the others: *dispatch_office* [33+], *applic_notes* [18+], *no_censor_card* [21+], *film_producer* [20+], *no_prec_doc* [20+], *applicant* [22+], *film_genre* [17+], *registration_au* [25+], *cens_process* [30+], *cens_card* [26+], *delivery_date* [16+]. Again, a 10-fold cross-validation was applied, whose averaged results are reported in Table 4.

The reported outcomes reveal that INTHELEX was actually able to learn significant definitions for both the document classes and the layout blocks of interest for each of them. Indeed, the predictive accuracy is always very high, reaching even 99.17% in one case and only in 2 cases out of 32 falling below 90% (specifically, 86.69% and 89.85%). It is very interesting to note that the

Table 4. Statistics for understanding (Model A)

	Cl	Lgg	Accuracy	Runtime
dispatch_office	6.8	13.9	94.28	13149.31
applic_notes	2.5	5.7	98.81	231.05
no_censor_card	5.3	11.2	95.47	8136.796
film_producer	4.9	9.6	93.98	5303.78
no_prec_doc	4.6	11	93.97	5561.14
applicant	6.7	11.5	93.66	15588.15
film_genre	2.8	6.9	98.53	684.35
registration_au	4.1	12.5	94.64	5159.74
cens_process	4.8	10.8	98.51	4027.90
cens_card	5.6	11.8	94.62	3363.86
delivery_date	4	9.1	95.515	3827.34

best accuracy is obtained by a theory made up of only one clause (that we may state has perfectly grasped the target concept), and coincides with the best runtime (classification for class DIF). The learned rules show a high degree of understandability for human experts, which was one of the requirements for the experiment. As expected, the classification problem turned out to be easier than the interpretation one (that is concerned with the semantics of the layout blocks inside documents). This is suggested by the tendential increase in number of clauses, performed generalizations and runtime from Table 1 to Tables 3, 2 and 4. Such an increase is particularly evident for the runtime, even if it should be considered that the high predictive accuracy should ensure that few theory revisions can be expected when processing further documents. Scalability should be ensured, since we expect that very few documents will generate theory revision. Moreover, symbolic representations allow the expert to properly choose the training examples so that few of them are sufficient to reach a correct definition.

5 Conclusions and Future Work

This paper proposed the application of symbolic (first-order logic) multistrategy learning techniques to induce rules for automatic classification and interpretation of cultural heritage material. Experimental results prove the benefits that such an approach can bring. Specifically, the chosen domain comes from the EU project COLLATE, concerned with film censorship documents dating back to the 20s and 30s. The learning component is the incremental system INTHELEX, whose performance on such a task proved very interesting.

Future work will concern finding better and more tailored ways of exploiting the features provided by INTHELEX in order to tackle in a still more efficient and effective way the problems raised by the low layout quality typical of cultural heritage documents. In particular, being able to handle numeric/probabilistic features could provide important support for this aim.

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 P. Constantopoulos and I.T. Sølvberg, editors, *Research and Advanced Technology for Digital Libraries*, number 2163 in LNCS, pages 37–48. Springer, 2001.

[3] F. Esposito, D. Malerba, and F.A. Lisi. Machine learning for intelligent processing of printed documents. *Journal of Intelligent Information Systems*, 14(2/3):175–198, 2000.

[4] F. Esposito, D. Malerba, G. Semeraro, N. Fanizzi, and S. Ferilli. Adding machine learning and knowledge intensive techniques to a digital library service. *International Journal on Digital Libraries*, 2(1):3–19, 1998.

[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):40–46, 1996.

[8] 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, U.S.A., 1994.

[9] G. Semeraro, S. Ferilli, N. Fanizzi, 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 LNCS, pages 59–70. Springer, 2001.