Published in ICDAR 2001

Logical Labeling using Bayesien Networks

Souad Souafi-Bensafi[1,2], Marc Parizeau[2], Franck Lebourgeois[1], Hubert Emptoz[1]

[1]Reconnaissance de Formes et Vision,
I.N.S.A. de LYON - Bt 403, 20 Av. A. Einstein
69621 Villeurbanne Cedex FRANCE

[2]Laboratoire de Vision et de Systèmes Numériques,
Université Laval, Département de génie–électrique,
Québec, CANADA, G1K 7P4

Abstract

This paper discusses logical labeling in documents, which is one basic step in logical structure recognition. Logical labels have to be attributed to text blocks composing the layout structure. Our study is based on physical characteristics having a visual aspect: typographic, geometric and/or topologic attributes. Our objective is to map a low level logical structure, which consists of a set of logical labels, on the extracted layout structure components. We have to build a model that allows this mapping. However, the documents we consider have various layout and logical structures, thus, we chose to perform this task by supervised learning on the basis of a set of training documents. This allows us to define a generic method to solve this problem, without imposing any constraint on document structure. We propose a probabilistic model represented by a Bayesian Network (BN), which is a graphical model used in our problem as a classifier. A prototype has been implemented, and applied to tables of contents in periodics.

1 Introduction

The work presented in this paper is linked to the recognition part of a Document Analysis and Recognition system [3]. It consists of recognizing the logical structure of documents using their layout structure. The logical structure is usually composed of a set of logical functions or labels that must be assigned to layout structure components on the one hand, and of the relations between these components on the other hand. The mapping between logical and physical components is called logical labeling. This task has been the subject of much research in the last few decades.

The most common representation of layout and logical structures is the tree representation. An algorithm for tree transformation has been developed [23] to pass from the layout structure to the logical structure. In [4], the tree representation describes a hierarchy on physical blocks, and a perceptual approach was adopted based on tree–grammar inference to build the generic model. This approach was applied to various structured documents. Tree representation has also been used in another context. Automatic generation of decision trees is discussed in [10], and was applied to business letters as a model for logical labeling. The approach uses non supervised learning with GTree (Geometric Tree), a decision tree classifier. Another method [21], combines a decision–tree classification and syntactic analysis using a matricial grammar. It was applied to scientific papers. Finally, a statistical representation called tri-gram tree was used for document logical structure recognition[5].

Logical structure models are often based on the form of the document and the spatial organization of its components. Reference [22] discusses the extraction of Japanese newspaper articles using layout heuristics, while [13] proposes a rule–based segmentation method using a form definition language. Text lines provided by the segmentation phase are labeled by topological matching of the model [28]. In [27], logical structure recognition is based on contructive rules for the layout structure which is represented by a binary tree in which branches correspond to the neighbourhood among the physical objects. This approach was applied to business letters, library cards and visit cards [26]. Reference [7] proposes a system for document classification and understanding, in which the logical object description is based on spatial organization, using general qualitative properties that are independent of the class and a list of facts for each class. In [24], the layout structure is described with an oriented and valued graph expressing neighbourhood using Allen relations [1]. The model is built by learning the graphs corresponding to training set documents, and the inference is made by sub–graphs isomorphisms. This method was also applied to business letters and visit cards [25].

In this study, we are interested in the task of logical labeling in general. However, logical structure can have a multiple level representation according to the document content semantics, because logical labels, according to their function in the document, can be composite, i.e. constituted of a lower level labels. Let's take the example of magazine documents. The table of contents has an hierarchical organization that must be reconstituted. In the lower level, the labels (like *title, author*) are just assigned to text blocks without any relation between them. A succession of these labels can contitute an

article label, and a *section* label is composed of a group of articles, etc. Thus, logical labeling task can be decomposed according to different levels. This decomposition implies different phases, starting by the lower level, and by logical grouping in order to go up to the higher levels. We first focused on the lower level in logical labeling, in which labels must be assigned to text blocks. It is assumed that can be found more than one label on any given line of text. Therefore, text blocks are considered at the word level. This task has to be performed on the basis of physical (geometric, typographical and spatial) features of text blocks. It corresponds to a classification problem and consists in associating to each word of a given document, a logical label according to its physical description. In a previous work [20], a relaxation–based classification method was developed using only word typographical information, and the present work aims at exploring other possible attributes in order improve results. We propose a generic probabilistic approach to build a first logical level classifier for text blocks, using a bayesian network. In parallel, a study is performed on a multi–level relaxation process to recognize the whole logical structure [18].

Next, after presenting an overview of our framework in Section 2, the Bayesian Networks and their use in our model will described in Section 3, together with the learning and inference tasks. Then, some results on tables of contents in periodics are presented.

2 Framework description

Our study is led in a global system framework depicted in figure 1. This framework insures a chain of document process from their digitization to their content extraction and storage, having two main steps: document analysis and document understanding or recognition. We apply the system principally to tables of contents in periodic magazines, because of their complexity and variety in form, and of their content structure. The information to extract is organized in different text categories, that must be recognized and stored in a re–usable format. The recognition phase must be able to identify the magazine, to locate regions that contain textual information, and then classify this information according to the existing text categories. The present work is focused on the third task, and especially on labeling text blocks, that is assigning a logical function. However, because logical structure can change very much from magazine to magazine, we chose to build through supervised learning, a specialized logical structure model for each magazine. The magazine is thus the document class. Model construction is based on physical description of text blocks and their association with their logical labels. The model has to supply this association, which corresponds to a classification task and the model to build is thus a classifier.

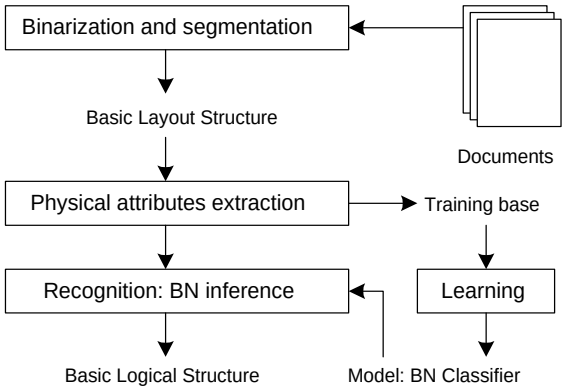

Figure 1: Global system architecture

We use tools developed in our laboratory [16], that allows physical level analysis. A basic layout structure is supplied by segmentation, containing a hierarchy of geometric text blocks: characters, words, lines and paragraphs. Typographical information for word level blocks can also be extracted [17], giving a set of typographical families and for each word the corresponding family. However, this is not sufficient to describe the blocks in order to recognize their logical function. So, we decide to describe each block by an attribute vector, noting that we work at the word level of text blocks. The vector is based on the following features: the typographical family of the block, and its left and right neighbours, the alignment and the horizontal and vertical spacing. A probabilistic model has been used, represented by a bayesian network classifier, for which a general description is done in the next section.

3 Probabilistic model

The proposed probabilistic model uses a bayesian network classifier to represent the relations among the set of attributes and the corresponding class label. Before exposing its application and adaptation to the current problem, it is necessary to describe it in a general manner.

3.1 Bayesian Networks

BNs are Probabilistic Networks that have been used for problems involving reasoning under uncertainty in Artificial Intelligence, in different applications including medical diagnostics, classification systems and software debugging [6]. They are a representation of a set of random variables and the probabilistic relationships among them. For a set of variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, a corresponding BN is represented by Direct Acyclic Graph (DAG) on the one hand, and a set of conditional probabilities accord-

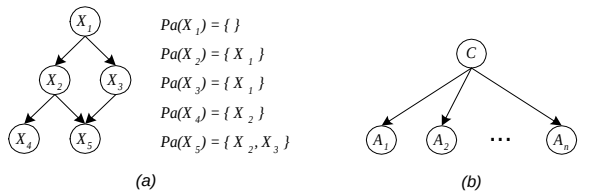

Figure 2: (a): Example of BN, (b): Naïve Bayesian Classifier (NBC)

ing to the graph on the other hand (figure 2a). In the DAG structure, the nodes represent the variables and the edges express a dependence among the variables. To each variable X_i corresponds the set of its parents $Pa(X_i)$ formed the variables it depends upon. The probability among the set $\mathbf{X}$ can thus be decomposed by: $P[X_1, X_2, \ldots, X_n] = \prod_{i=1}^{n} P[X_i|Pa(X_i)]$.

Bayesian Network Classifiers (BNC): When BNs are used for classification problems, the set of variables is composed of the class C and of attributes (features) $A_1, A_2, \ldots, A_n$, with n being the number of attributes. The Naive Bayesian Classifier (NBC) is a particular case of BNs, in which the attributes are assumed mutually independent(figure 2b). Such strong hypothesis, however, is usually unfounded, and the general BN structure is much powerful. Different structures of BNCs [11], [19], have been defined. The main goal is to impose the minimum of contraints on the structure to be as general as possible, but at the same time to make their manipulation as simple as possible.

Inference: For the general case, the BNs inference process consists in determining various probabilities of interest within the model, which is equivalent to responding to probabilistic requests for any variable. This problem being NP–Hard, several approximate algorithms have been proposed [12], using arc reversing techniques in the graph structure, message–passing schemes, or by transforming the DAG graph to a tree structure. But for the classification, BN inference problem is much simpler. It only requires to compute the class probabilities given the attribute values : $P[C|A_1, A_2, \ldots, A_n] = \alpha \cdot P[C|Pa(C)] \cdot \prod_{i=1}^{n} P[A_i|Pa(A_i)]$, with α being a normalization factor.

Learning: BN learning consists of two parts: learning the DAG–structure and learning the conditional probabilities. For structure learning, the most used approach, consists of a search procedure with a score function, which allows the BN evaluation. Two main types of score functions are used, MDL (Minimum Description Length) functions [11], and bayesian functions [9], [12]. The BN structure learning is also NP–Hard [8]. For this reason, methods for reducing the BN structure search space have been developed: imposing constraints among the random variables or the structure types, or using non–deterministic approaches, as for example genetic algorithms[14], [15].

3.2 Model description

We propose to use the BNC to model the relations between the physical description of a given text block and its label. The attributes represent the physical features of the blocks and the class variable corresponds to the label that has to be assigned to the block.

We are currently developing genetic–programming [2] approach for learning the structure of BNs, but the work is still in progress. Therefore, results presented in the next section will uses a Naïve Bayesian Classifier instead. The NBC however, uses the same step for learning the conditional probabilities of the attributes as the BNC, and the inference process is also the same.

The conditional probabilities are simply computed by extracting the sufficient statistics, corresponding in our discrete case to counting the occurrences of each group composed by a variable and its parents, from the training set. Let $\mathbf{U} = \{u_1, u_2, \ldots, u_N\}$ be the training set containing N vectors, for the NB classifier learning. Each vector is composed of the values of attributes and of the class variables that describe a text block. $Val(A_i) = \{a_{i1}, a_{i2}, \ldots, a_{ir_i}\}$ ($Val(C) = \{c_1, c_2, \ldots, c_{r_0}\}$) is the value domain of A_i (C), with r_i (r_0) being the the number of the possible values of the attribute A_i (the class C). Let q_i (q_0) be the number of the distinct values of $Pa(A_i)$ ($Pa(C)$) according to the training set $\mathbf{U}$: pa_{ij} (pa_{0j}) with $j \in \{1, 2, \ldots, q_i\}$. We note N_{ijk} the number of cases in $\mathbf{U}$ for which $A_i = a_{ik}$ and $Pa(A_i) = pa_{ij}$. The probabilities learning is performed by counting, according to $\mathbf{U}$ and the BN structure, the N_{ijk}, which are the sufficient statistics to estimate these probabilities: $P[A_i|Pa(A_i)] = \frac{N_{ijk}}{N_{ij}}$, with $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$.

The inference process uses directly these probabilities, and for the NBC, by replacing the attribute parent sets, the classification probabilities are computed by : $P[C|A_1, A_2, \ldots, A_n] = \alpha \cdot P[C] \cdot \prod_{i=1}^{n} P[A_i|C]$.

4 Experiments

For the application of this classifier, we used the attributes that have been listed in Section 2. They all have discrete values as follow: *typographical families* take numeric values from 0 to the number of the existing families, with 0 corresponding the reject family; the *alignment* has the values (1) left, (2) right and (3) center; the *previous and next line space* are valued between 0 and 4, according to the vertical distance and the line height, and the -1 value corresponds to the absence of a previous or next

	Learning		Test		Recognition rates(%)		
periodics	#pages	#words	#pages	#words	mean	max	min
Cahiers...	3	489	7	1328	94,7	97,9	88,3
NewsWeek	3	641	11	2445	93,2	98,2	85,6
Nature	4	1718	14	6059	92,8	96,6	84,0
Biofutur	3	462	7	1139	82,6	99,3	65,9

Table 1: Recognition results.

line; the *horizontal distance (left and right)* are valued between 0 et 2 according to the distance with the left or right neighbour and the mean horizontal space in the text line, and the value -1 is used for the case where the neighbour does not exist. The class values are the different labels existing in the considered document class. We principally use the following labels: *section, title, author, page number, summary* and a value associated to the *not labeled* words.

We applied our classifier to four periodic magazines (figure 3), and we obtained the results shown in the Table 1. Every magazine is represented by a document class for which training and testing sets have been constructed. The number of document pages and the number of words are indicated for each set. Recognition is measured by the fraction of the number of recognized words over the total number of words in the test set, for each document. The mean, maximal and minimal rates are presented in the table. Results are very good for certain documents, but somewhat lower for others. The low recognition rates can be explained by the physical features which are not stable and affect the probabilities used for the classification. For example, in the periodic "Biofutur", the minimal recognition rate is $65,9\%$, because the correcponding document had problems in segmentation and physical features extraction. We can deduce that it is necessary to select more discriminant attributes, and this can be performed by the BN structure learning that we did not use in the present work.

5 Conclusion

In this paper, a generic probabilistic model is proposed for a first level logical labeling using Bayesian Networks Classifiers. The goal is to perform this labeling automatically. The model is built with a supervised learning task on the basis of a training set. It is used in the recognition step at the word blocks level, which consists of assigning labels to the blocks according to their physical description. A prototype for the Naïve Bayesian Classifier has been implemented and applied to periodical magazines. Significant results have been obtained, however for some documents, recognition rates were not satifying. It is due to instability at the physical and the logical levels, for example, the quality of documents does not always allow a perfect segmentation and feature extraction, which affect directly the following steps.

We expect improving our results by focusing on the selection of the most representative attributes. Bayesian Networks structure learning can be used for this selection. A combination is also in study, with a relaxation–based classification process. The next step of this study is the recognition of the whole logical structure, by considering all existing hierarchical levels.

References

[1] J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, November 1983.

[2] W. Banzhaf, P. Nordin, R.E. Keller, F.D. Francone. *Genetic Programming : An Introduction On the Automatic Evolution of Computer Programs and Its Applications*. dpunkt.verlag and Morgan Kaufmann Publishers, 1998.

[3] A. Belaïd. *Analyse et Reconnaissance de Documents*, pages 49–92. Editions ADBS, Aix–en–Provence, Octobre 1994.

[4] A. Belaïd. Conception assistée de modèles de page en vue de leur utilisation en reconnaissance de documents. *Lausane–Atelier sur les modèles de pages électroniques*, Lausane, Septembre 1997.

[5] R. Brugger, A. Zramdini, R. Ingold. Modeling documents for structure using generalized N–Grams. 4^{th} *ICDAR*, volume 1, pages 56–60, Ulm, Germany, August 1997.

[6] W. Buntine. A Guide to Literature on Learning Probabilistic Networks from Data. *IEEE Transactions on Knowledge and Data Engineering*, 8(2):195–210, April 1996.

[7] F. Cesarini, E. Francesconi, M. Gori, G. Soda. A Two Level Knowledge Approach for Understading Documents of Multi–Class Domain. 5^{th} *ICDAR*, pages 135–138, Bangalore, India, september 1999.

[8] D. M. Chickering, D. Geiger, D. Hecherman. Learning Bayesian Networks is NP–Hard. Technical report, MSR–TR–96–08, Microsoft Research, March 1996.

[9] G.F. Cooper E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

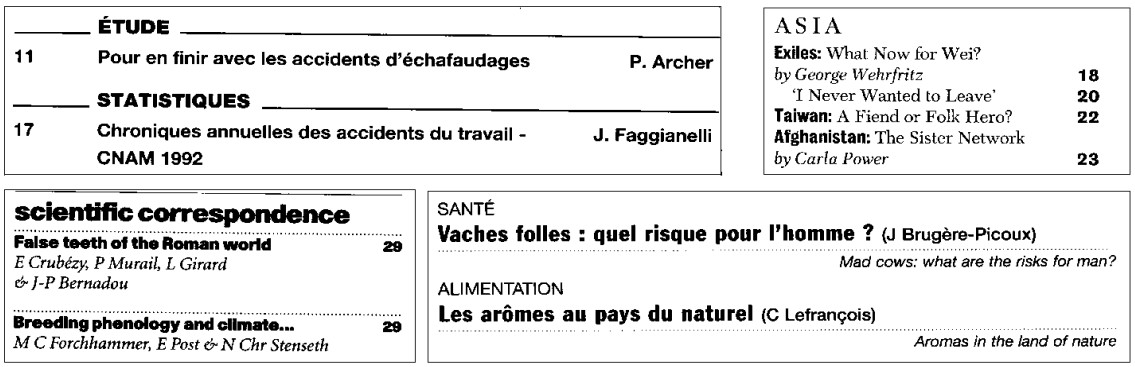

Figure 3: Text blocks examples in four different magazines.

[10] A. Dengel. Initial Learning of Document Structure. 2^{th} *ICDAR*, pages 86–90, Tsukuba Science City, Japan, October 1993.

[11] N. Friedman, D. Geiger, M. Goldszmidt. Bayesian Network Classifiers. *Machine Learning*, (29):131–163, 1997.

[12] D. Hecherman. A Tutorial on Learning with Bayesian Networks. Technical report, MSR–TR–95–06, Microsoft Research, March 1995.

[13] J. Higashino et al. A knowledge–based segmentation method for document understading. 8^{th} *ICPR*, volume 1, pages 745–748, Paris, France, October 1986.

[14] P. Larrañaga, C.M.H. Kuijpers, R. Murga, Y. Yurramendi. Learning Bayesian Networks Structures by Searching for the Best Ordering with Genetic Algorithms. *IEEE Transactions on Systems, Man and Cybernetics*, 26(4):487–493, July 1996.

[15] P. Larrañaga, M. Poza, Y. Yurramendi, R. Murga, C.M.H. Kuijpers. Structural Learning of Bayesian Networks by Genetic Algorithms: A performance Analysis of Control Parameters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(9):912–926, September 1996.

[16] F. Lebourgeois. Localisation de textes dans une image à niveaux de gris. 4^{eme} *Colloque National sur l'Écrit et le Document, CNED'96*, pages 207–214, Nantes, France, Juillet 1996.

[17] F. LeBourgeois H. Emptoz. Document Analysis in Gray Level and Typography Extraction Using Character Pattern Redundancies. 5^{th}*ICDAR*, pages 177–180, Bangalore, India, September 1999.

[18] F. LeBourgeois, H. Emptoz, S. Souafi-Bensafi. Document Understanding using Probabilistic Relaxation: Application on tables of contents of periodicals, *submitted to icdar 2001.*

[19] S. Monti G.F. Cooper. A bayesian Network Classifier that Combines a Finite Mixture Model and a Naive Bayes Model. *Proceeding of* 15^{th} *International Conference on Uncertainty in Artificial Intelligence*, 1999.

[20] S. Souafi-Bensafi, F. LeBourgeois, H. Emptoz. Modélisation et Reconnaissance des Structures de Documents : Application aux Sommaires de Revues. *CIFED 2000*, pages 71–80, Lyon, France, Juillet 2000.

[21] A. Takasu, S. Satoh, E. Katsura. A Document Undestading Method for Database Construction of Electronics Library. 11^{th} *ICPR*, volume 2, pages 463–466, Jerusalem, Israel, October 1994.

[22] J. Toyada et al. Study of extracting Japanese newspaper article. 6^{th} *ICPR*, volume 2, pages 1113–1115, October 1982.

[23] S. Tsujimoto H. Asada. Understanding multi–articled documents. 10^{th} *ICPR*, pages 551–556, Atlanta City, New Jersey USA, June 1990.

[24] H. Walischewski. Automatic Acquisition for Spatial Document Interpretation. 4^{th} *ICDAR*, volume 1, pages 243–247, Ulm, Germany, August 1997.

[25] H. Walischewski. Automatic Acquisition for Spatial Document Interpretation. 5^{th} *ICDAR*, pages –, Bangalore, India, september 1999.

[26] T. Watanabe X. Huang. Automatic Acquisition of Layout Knowledge for Understading Business Cards. 4^{th} *ICDAR*, volume 1, pages 216–220, Ulm, Germany, 1997.

[27] T. Watanabe, Q. Luo, N. Sugie. A Cooperative Document Understanding Method among Mutiple Recognition Procedures. 11^{th} *ICPR*, pages 689–692, August 1992.

[28] A. Yamashita, A. Tomio, H. Takahashi, K. Toyokawa. A Model Based Layout Understading Method for the Document Recogntion System. 1^{st} *ICDAR*, pages 130–138, St Malo, France, september 1991.