A Data Mining Approach to Reading Order Detection

Michelangelo Ceci Margherita Berardi Giuseppe A. Porcelli
Donato Malerba
University of Bari - Dipartimento di Informatica
Via Orabona, 4 - 70125 Bari - Italy
{ceci, berardi, malerba}@di.uniba.it, giu.porcelli@gmail.com

Abstract

Determining the reading order for layout components extracted from a document image can be a crucial problem for several applications. It enables the reconstruction of a single textual element from texts associated to multiple layout components and makes both information extraction and content-based retrieval of documents more effective. A common aspect for all methods reported in the literature is that they strongly depend on the specific domain and are scarcely reusable when the classes of documents or the task at hand changes. In this paper, we investigate the problem of detecting the reading order of layout components by resorting to a data mining approach which acquires the domain specific knowledge from a set of training examples. The input of the learning method is the description of the "chains" of layout components defined by the user. Only spatial information is exploited to describe a chain, thus making the proposed approach also applicable to the cases in which no text can be associated to a layout component. The method induces a probabilistic classifier based on the Bayesian framework which is used for reconstructing either single or multiple chains of layout components. It has been evaluated on a set of document images.

1 Introduction

Documents are characterized by two important structures: the layout structure and the logical structure. The former is based on the presentation of the document content, while the latter is based on the human-perceptible meaning of the content. Document image understanding refers to the process of extracting the logical structure of a document image. Most works on document image understanding aim at associating a "logical label" with some components of the layout structure. However, in its broader sense, document image understanding cannot be considered synonymous of "logical labeling", since relationships among logical components are also possible and their extraction can be equally important for an application domain. Some examples of relations are cross references between captions and figures, as well as cross references between affiliations and authors.

An important class of relations investigated in this paper is represented by the *reading order* of some parts of the document. More specifically, we are interested in determining the reading order of the most abstract layout components on each page of a multi-page document. By following the reading order recognized in a document image, it is possible to cluster together text regions labelled with the same logical label into the same textual component (e.g., section on "material and method" of a scientific paper). Once a single textual component is reconstructed, advanced text processing techniques can be subsequently applied.

Several papers on reading order detection have already been published in the literature. Some are based only on the spatial properties of the layout components [10] [6] [8], while others also exploit the textual content of parts of documents [9] [2] [1]. Moreover, some methods have been devised for properly ordering layout components (independent of their logical meaning), while others consider the recognition of some logical components, such as "title" and "body", as preliminary to reading order detection [1]. A common aspect of all methods is that they strongly depend on the specific domain and are not "reusable" when the classes of documents or the task at hand change. For instance, the classification of blocks as "title" is appropriate for magazine articles, but not for administrative documents. Moreover, Western and Japanese articles have different document encoding rules. To the best of our knowledge, no work investigates the reading order problem by resorting to data mining techniques which can generate the required knowledge from a set of training layout structures whose correct reading order has been provided by the user.

In this paper we exploit the high degree of adaptivity of data mining methods for inducing a predictive model to be used in reading order reconstruction. The predictive model is induced from training examples which are sets of ordered

layout components described by means of both their spatial properties and their possible logical label. Therefore, no textual information is exploited to understand document images. The ordering of the layout components is defined by the user and does not necessarily reflect the traditional Western-style document encoding rule according to which reading sequence proceeds from top to bottom and from left to right. For instance, the user can specify a reading order according to which the affiliation of an author immediately follows the author's name, although the former is reported at the bottom of a page while the latter at the top. In multi-page articles, such as those considered in this paper, ordering is defined at the page level. More precisely, different "chains" of layout components can be defined by the user when independent pieces of information are represented on the same page (e.g., the end of an article and the beginning of a new one). Chains are mutually exclusive, but not necessarily exhaustive, sets of the most abstract layout components in a page, so that their union defines a partial (and not necessarily a total) order on the set of layout objects.

2 Problem Definition

To formalize the problem we intend to solve, some useful definitions are necessary.

Let A be a set of blocks in a document page,

Definition Partial order over a set of blocks
A *partial order PO over A* is a relation $PO \in A^2$ such that
• PO is reflexive: $\forall s \in A, (s,s) \in PO$;
• PO is antisymmetric: $\forall s_1, s_2 \in A$, $(s_1,s_2) \in PO \wedge (s_2,s_1) \in PO \Leftrightarrow s_1 = s_2$;
• PO is transitive: $\forall s_1, s_2, s_3 \in A$, $(s_1,s_2) \in PO \wedge (s_2,s_3) \in PO \Rightarrow (s_1,s_3) \in PO$.

When PO satisfies the antisymmetric, the transitive and the irreflexive ($\forall s \in A, (s,s) \notin PO$) properties, it is called a *weak partial order* over A.

Definition Total order over a set of blocks
A partial order T over the set A is a *total order* iff $\forall s_1, s_2 \in A, (s_1,s_2) \in T \vee (s_2,s_1) \in T$.

Definition Complete chain, Chain reduction
Let D be a weak partial order over A, let $B = \{a \in A | (\exists b \in A \; s.t. \; (a,b) \in D \vee (b,a) \in D)\}$ be the set of elements in A related to any element in A itself. If $D \cup \{(a,a) | a \in B\}$ is a total order over B then D is a *complete chain over A*. Furthermore, $C = \{(a,b) \in D | \neg \exists c \in A \; s.t. \; (a,c) \in D \wedge (c,b) \in D\}$ is the *reduction of the chain D over A*.

Example 1 *Let $A = \{a,b,c,d,e\}$. If $D = \{(a,b),(a,c),(a,d),(b,c),(b,d),(c,d)\}$ is a complete chain over A, then $C = \{(a,b),(b,c),(c,d)\}$ is its reduction.*

For our purposes, it is equivalent to deal with complete chains or their reduction. Henceforth, the term *chain* will denote the reduction of a complete chain.

Let Π be a set of pages, i.e., document images. For each page $Pa \in \Pi$, suppose that a set of blocks $\{b_1, b_2, \ldots b_n\}$ are extracted. We write down $Pa = \{b_1, b_2, \ldots b_n\}$ and we assume that each block b_i is associated with a logical label denoted by $label(b_i)$. The term $chains(Pa)$ will denote the set of all possible (reduced) chains over $\{b_1, b_2, \ldots b_n\}$. For a given chain $C \in chains(Pa)$, $b_i \prec_C b_j$ states that b_i precedes b_j in C, i.e. $(b_i, b_j) \in C$. For a set of chains $\{C_u\} \subseteq chains(Pa)$, the extended notation $b_i \prec_{\{C_u\}} b_j$ means that $b_i \prec_C b_j$ for some $C \in \{C_u\}$. The reading order induction problem can be formalized as follows:
Given:
• A set $TP = \{TP_k | TP_k \in \Pi\}$ of *training pages*.
• A set $TC = \{TC_u | \exists TP_k \in TP, TC_u \in chains(TP_k)\}$ of *training chains*.
• A set $Labels$ of logical labels involved in the reading order identification process.
Find: A function $f : \Pi \rightarrow \{C | C \in chains(Pa), Pa \in \Pi\}$ which returns a set of chains over blocks of a page $Pa \in \Pi$ such that the posterior probability $P(b_i \prec_{f(Pa)} b_j |\; b_i, b_j \in Pa, label(b_i), label(b_j) \in Labels, TC)$ is maximized.

This formalization permits us to represent and identify distinct reading orders on the same page and to drop blocks which should not be in the reading order (e.g., figures).

3 Mining Reading Order

To approximate the above mentioned posterior probability, we exploit machine learning studies for ranking examples [3, 7]. In particular, we resort to the naïve Bayesian classification framework [7, 4], according to which an example x is classified by maximizing the posterior probability $P(C_i|x)$ that the observation x is of class C_i: $class(x) = argmax_i P(C_i|x)$. With the Bayes theorem, $P(C_i|x) = P(C_i)P(x|C_i)/P(x)$. Since $P(x)$ does not depend on C_i, then $class(x) = argmax_i P(C_i)P(x|C_i)$. If we assume that x is defined by a vector $(x_1, x_2, \ldots, x_m)$ of m independent features (*naïve Bayes assumption*), the probability $P(x|C_i) = P(x_1, \ldots, x_m|C_i)$ can be factorized as: $P(x|C_i) = P(x_1|C_i) \times \ldots \times P(x_m|C_i)$. Even when the conditional independence assumption is grossly violated, NB classifiers has been been proved accurate [4].

In our domain, we intend to exploit this formulation to define the posterior probability that $b_1 \prec b_2$ for any two blocks b_1, b_2 of a document page Pa such that $label(b_1), label(b_2) \in Labels$. Formally:

$$P(b_1 \prec b_2 | Pa) \propto P(b_1 \prec b_2) P(\bigwedge_{r=1}^{m} B_r(b_1, b_2) | b_1 \prec b_2) \quad (1)$$

where $B_r(b_1, b_2),\ r = 1, \ldots, m$ are predicates describing properties of the document blocks b_1 and b_2.

Following the naïve Bayes assumption, $P(\bigwedge_{r=1}^{m} B_r(b_1, b_2)|b_1 \prec b_2)$ is factorized as follows:

$$P(\bigwedge_{r=1}^{m} B_r(b_1, b_2)|b_1 \prec b_2) = \prod_{r=1}^{m} P(B_r(b_1, b_2)|b_1 \prec b_2)$$

Each $P(B_r(b_1, b_2)|b_1 \prec b_2)$ can be estimated using the Laplace estimator which avoids null probabilities:
$P(B_r(b_1, b_2)|b_1 \prec b_2) =$

$$\frac{\#\left\{ (b_{k1}, b_{k2}) \;\middle|\; \begin{array}{c} TP_k \in TP \wedge b_{k1}, b_{k2} \in TP_k \wedge \\ b_{k1} \prec_C b_{k2} \wedge C \in TC \\ \wedge B_r(b_1, b_2) = B_r(b_{k1}, b_{k2}) \end{array} \right\} + 1}{\#\left\{ (b_{k1}, b_{k2}) \;\middle|\; \begin{array}{c} TP_k \in TP \wedge b_{k1}, b_{k2} \in TP_k \wedge \\ b_{k1} \prec_C b_{k2} \wedge C \in TC \end{array} \right\} + F}$$

where F represents the number of values which B_r can assume. Here, F=2 since B_r are boolean properties.

The predicates used to describe the document page involve several different descriptors which can be classified in locational descriptors, such as the coordinates of the centroid of a logical component (*x_pos_centre*, *y_pos_centre*), geometrical descriptors, such as the dimensions of a logical component (*width*, *height*), and topological descriptors, such as relations between two components (*on_top*, *to_right*, *alignment*). We use the aspatial descriptor *type_of* that specifies the content type of a logical component (e.g., image, text, line). The logical descriptor *logic_label*, is an aspatial descriptor used to state labels associated to the logical components (e.g. affiliation, figure of scientific papers).

In our case, predicates using locational descriptors are:
$B_1(b_1, b_2) \Leftrightarrow x\_pos\_centre(b_1) \leq x\_pos\_centre(b_2)$
$B_2(b_1, b_2) \Leftrightarrow y\_pos\_centre(b_1) \leq y\_pos\_centre(b_2)$
$B_3(b_1, b_2) \Leftrightarrow width(b_1) \leq width(b_2)$
$B_4(b_1, b_2) \Leftrightarrow height(b_1) \leq height(b_2)$
Predicates using aspatial descriptors are:
$B_5(b_1, b_2) \Leftrightarrow type\_of(b_1) = type\_of(b_2)$
$B_6(b_1, b_2) \Leftrightarrow logic\_label(b_1) = logic\_label(b_2)$
Finally, predicates involving topological descriptors are:
$B_7(b_1, b_2) \Leftrightarrow on\_top(b_1, b_2)$
$B_8(b_1, b_2) \Leftrightarrow to\_right(b_1, b_2)$
$B_{9+i}(b_1, b_2) \Leftrightarrow alignment(b_1, b_2) = k_i, i = 0, \ldots, 5$
The last six predicates correspond to different alignments, since $k_i \in \{$ *only_left_col*, *only_right_col*, *only_middle_col*, *only_upper_row*, *only_lower_row*, *only_middle_row*$\}$.

Our method differs from that in [7], since we use predicates instead of equality relations among attribute values which require a prediscretization of continuous attributes.

4 Chain identification

Once $P(b_1 \prec b_2|Pa)$ has been computed for each pair of blocks b_1, b_2 of a new document page Pa, the estimated probabilities can be used to reconstruct chains over blocks of Pa. In our approach, we propose two different solutions:

1. identification of multiple chains of layout components
2. identification of a single chain of layout components.

Both approaches make use of a *labeled* directed graph $G = \langle V, E\rangle$, where $V = \{b \in Pa|label(b) \in Labels\}$ and $E = \{(b_1, b_2, w_{b_1,b_2}) \in V^2 \times [0, 1]|w_{b_1,b_2} = P(b_1 \prec b_2|Pa)\}$ is the set of weighted edges where weights are the probabilities $P(b_1 \prec b_2|Pa)$ computed according to (1).

Identification of multiple chains

This approach aims at identifying a (possibly empty) set of chains over the set of logical components in the same document page. It is two-stepped. The first step aims at identifying the set $Heads$ of heads (first elements) of the possible chains. In order to identify $Heads$, we introduce the graph G' defined as follows: $G' = \langle V, E'\rangle$ where $E' = \{(b_1, b_2)|\exists(b_1, b_2, w_{b_1,b_2}), (b_2, b_1, w_{b_2,b_1}) \in E \wedge w_{b_1,b_2} > (1 + \gamma)w_{b_2,b_1}\}$ where γ is a user-defined parameter used to filter-out connections for which there is scarce evidence that $b_1 \prec b_2$. $Heads$ is the set of nodes having the lowest number of incoming edges in G'. More formally:
$Heads = \{b_1 \in V|\#\{b_2 \in V|(b_2, b_1) \in E'\} = min_{b_3 \in V}\#\{b_2 \in V|(b_2, b_3) \in E'\}\}$

Once the set $Heads$ has been identified, the distinct chains can be reconstructed. Intuitively, each chain is the list of nodes forming a path in G' which begins with a node in $Heads$ and ends with a node without outgoing edges. Formally, an extracted chain $C \subseteq E'$ is defined as follows: $C = \{(b_1, b_2), (b_2, b_3), \ldots, (b_k, b_{k+1})\}$, such that:

- $b_1 \in Heads$,
- $\forall i = 1..k : (b_i, b_{i+1}) \in E'$ and
- $\forall b \in V\ (b_{k+1}, b) \notin E'$.

In order to avoid cyclic paths, we impose that a node cannot appear more than once in the same chain.

Identification of a single chain

The result of the second approach is a single chain. Following the proposal reported in [7], we aim at iteratively evaluating the most promising node to be appended to the resulting chain. More formally, let $SUMPREF_G : V \rightarrow [0, \#V]$ be a preference function defined as follows:

$$SUMPREF_G(b_1) = \sum_{b_2 \in V, b_2 \neq b_1} w_{b_1,b_2} \qquad (2)$$

Algorithm 1 fully specifies the method for the single chain identification. The rationale is that at each step, a node is added to the final chain. Such a node is that for which $SUMPREF_G(\_)$ is the highest. Higher values of $SUMPREF_G(\_)$ are given to nodes which have a high sum of probabilities to precede other nodes. Indeed, the

Algorithm 1 Single chain identification algorithm

1: **findChain** $(G = \langle V, E \rangle)$**: Chain of nodes L**
2: L$\leftarrow \emptyset$;
3: **while** $(\#L <> \#V)$ **do**
4: $L.add\left(\underset{b_i \in V/L}{\arg\max} \; SUMPREF_G(b_i) \right)$;
5: **end while**

algorithm returns an ordered list of nodes which could be straightforwardly transformed into a chain.

5 Experiments

To evaluate the applicability of the proposed approach to reading order identification, we considered a set of multi-page articles published in an international journal. In particular, we considered 24 papers, published as either regular or short articles, in the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) in two issues of 1996. Each paper is a multi-page document, therefore, we processed 211 document images. Initially, document images were pre-processed by WISDOM++[1] in order to segment them, perform layout analysis, identify the membership class and map the layout structure of each page into the logical structure. In all, 206 training chains were manually specified and 1,629 blocks were involved in the training chains (i.e. the average chain length is 7.72). In this task, $Labels$={*abstract*, *affiliation*, *author*, *biography*, *formulae*, *index_term*, *reference*, *section_title*, *paragraph*, *subsection_title*, *title*} and the entire set of logical labels is $Labels \cup$ {*caption*, *figure*, *table*, *page_no*, *running_head*}.

We evaluated the performance of the proposed approach by means of a 6-fold cross-validation, that is, the dataset was divided into six *folds* and then, for every fold, the learner was trained on the remaining folds and tested on it.

For each learning problem, statistics on precision and recall were recorded. Such measures refer to the $\prec$ relation obtained from the chain identification and permitted us to *locally* evaluate the method. To *globally* evaluate the ordering returned by the proposed approach, we resorted to metrics used in information retrieval for the evaluation of the returned rankings [5]. Herein we considered the metrics valid for partial order evaluation. In particular, we considered the *normalized Spearman footrule distance* which, given two complete lists L and L_1 on a set S (L and L_1 are two different permutations without repetition of all the elements in S), is defined as:

$$F(L, L_1) = 2/|S|^2 \sum_{b \in S} abs(pos(L, b) - pos(L_1, b)) \quad (3)$$

[1] http:/www.di.uniba.it/%7Emalerba/wisdom++/

where the function $pos(L, b)$ returns the position of the element b in the ordered list L.

This measure can be straightforwardly generalized in the case of several lists and modified in order to consider partial orders instead of total ones (*induced footrule distance*):

$$F'(L, L_1, \ldots, L_k) = 1/k \sum_{i=1 \ldots k} F(L|_{L_i}, L_i) \quad (4)$$

where $L|_{L_i}$ is the projection of L on L_i.

In this study, $F(L, L_1)$ was used in the evaluation of single chain identification, while $F'(L, L_1, \ldots, L_k)$ was used in the evaluation of multiple chain identification.

Results reported in Figure 1 show that, the higher the γ value, the higher the precision. This behavior was expected since by increasing γ, more block pairs involved in the $\prec$ relation are pruned. What is somewhat surprising is that the recall remains approximately unchanged for $\gamma \leq 0.3$.

Results reported in Table 1 show that in both cases the approach presents high precision and recall rates. Furthermore, we note that in the case of multiple chains, better precision and recall rates than in the case of single chains are obtained. This can be explained by the fact that some ordered pairs of blocks are lost in single chain identification, thus reducing both precision and recall.

Experimental results concerning the reconstruction of single/multiple chains are reported in Table 1 as well. We recall that the lower the distance value, the better the reconstruction of the original chain(s). In this case, although the reconstruction of multiple chains shows better results than the reconstruction of single chains, there is no clear difference between the two approaches. Indeed, the choice of the best reconstruction solution to be adopted does not depend on their performances, but on the task at hand. In figure 2, an example of the application of the two different reconstruction solutions is shown.

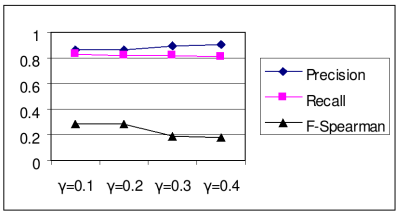

Figure 1. Multiple chains: results varying γ

Measure	Multiple Chains	Single Chain
Precision	0.900 ± 0.055	0.817 ± 0.039
Recall	0.818 ± 0.086	0.727 ± 0.031
F-Spearman	F':0.185 ± 0.102	F:0.240 ± 0.067

Table 1. 6-fold CV results. Average and standard deviation are reported. ($\gamma = 0.3$)

6 Conclusions

In this paper, we present a novel approach for automatically determining the reading order in a document image understanding process. It aims at mining a prediction function from user-defined reading order chains of layout components to be used when processing new documents.

Peculiarities of the proposed approach are: (a) the exploitation of data mining techniques which permit reaching a high degree of adaptivity, and (b) the reconstruction of reading order chains which may not necessarily define a total ordering. This last aspect permits us to consider independent pieces of information represented on the same page (e.g., the end of an article and the beginning of a new one) and to exclude layout components that should not be included in the reading order (e.g. images or page numbers).

Reading order chains are reconstructed according to two different modalities: single vs. multiple chains identification. Results prove that the reconstruction phase significantly depends on the application at hand. In particular, if the user is interested in reconstructing the actual chain (e.g. text reconstruction for rendering purposes), the best solution is in the identification of single chains. On the contrary, when the user is interested in recomposing a text such that sequential components are correctly linked (e.g. in information extraction applications), the most promising solution is the identification of multiple chains.

For future work we intend to consider the entire document (and not the single page) as the analysis unit in order to reconstruct multiple crossing-pages chains typically found in collections of documents (e.g., conference proceedings).

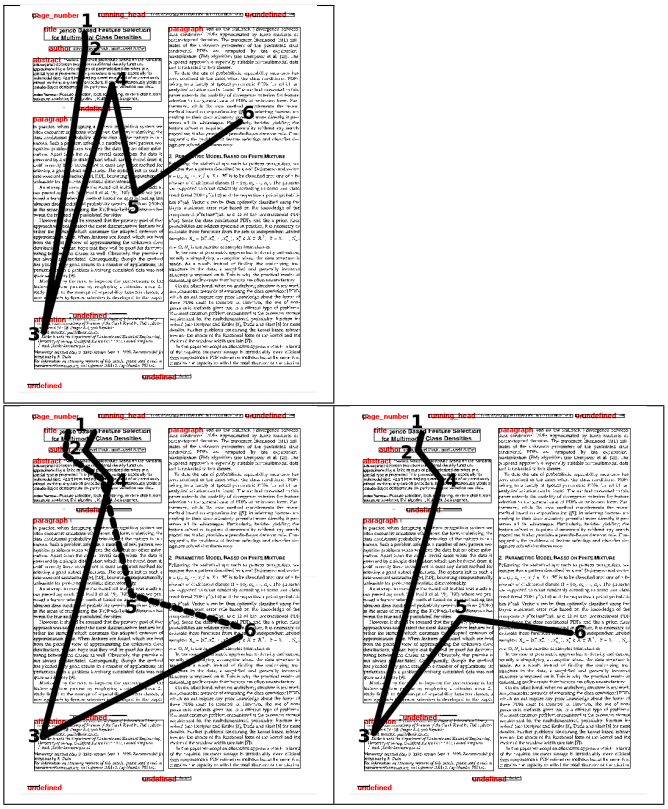

Figure 2. Extracted reading order. Actual reading order(top-left); Identified multiple (bottom-left) and single (bottom-right) chains

7 Acknowledgments

The work presented in this paper is partial fulfillment of the research objective set by the ATENEO-2007 project on "Metodi di scoperta della conoscenza nelle basi di dati: evoluzioni rispetto allo schema unimodale".

References

[1] M. Aiello, C. Monz, L. Todoran, and M. Worring. Document understanding for a broad class of documents. *International Journal on Document Analysis and Recognition-IJDAR*, 5(1):1–16, 2002.

[2] T. M. Breuel. High performance document layout analysis. In *Proceedings of the 2003 Symposium on Document Image Understanding (SDIUT '03)*, 2003.

[3] W. W. Cohen, R. E. Schapire, and Y. Singer. Learning to order things. *J. Artif. Intell. Res. (JAIR)*, 10:243–270, 1999.

[4] P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130, 1997.

[5] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 613–622, New York, NY, USA, 2001. ACM Press.

[6] Y. Ishitani. Document transformation system from papers to xml data based on pivot XML document method. In *ICDAR '03*, page 250. IEEE Computer Society, 2003.

[7] T. Kamishima and S. Akaho. Learning from order examples. In *ICDM*, pages 645–648, 2002.

[8] G. Nagy and S. Seth. Hierarchical representation of optically scanned documents. In *Seventh Int'l Conf. Pattern Recognition*, pages 347–349. IEEE CS Press, 1984.

[9] S. L. Taylor, D. A. Dahl, M. Lipshutz, C. Weir, L. M. Norton, R. Nilson, and M. Linebarger. Integrated text and image understanding for document understanding. In *HLT '94: Proceedings of the workshop on Human Language Technology*, pages 421–426, Morristown, NJ, USA, 1994.

[10] S. Tsujimoto and H. Asada. Understanding multi-articled documents. In *in Proceedings of the 10th International Conference on Pattern Recognition*, pages 551–556, 1990.