Page Analysis by 2D Conditional Random Fields

Atsuhiro Takasu
National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo, Japan

Keywords: Page Analysis, Information Extraction, 2D CRF.

Abstract: This paper applies two-dimensional conditional random fields (2D CRF) to page analysis and information extraction. In this paper we discuss features and labels for information extraction by 2D CRF. We evaluated the method by applying it to the problem of extracting bibliographic components from scanned title pages of academic papers. The experimental results show that 2D CRF improves the performance of information extraction compared to chain-model CRF.

1 INTRODUCTION

Meta data is important for document utilization and retrieval. It can help focusing on a specific facet in retrieval. For example, we can retrieve documents written by specific author if the meta data include a field of authors. Meta data have been usually made manually. However, it is hard to enumerate all possible facets used in retrieval in advance. In addition, manual creation of meta data is labor-intensive work.

Information extraction (IE) plays an important role for meta data creation because documents usually contain the information necessary for meta data. Although IE is a traditional research topic, it still attracts researchers and various machine learning techniques have been examined to improve the extraction accuracy. Among them, conditional random fields (CRF) (Lafferty et al., 2001) is a popular one. Council et al. develops a reference string parser called ParsCit based on CRF (Councill et al., 2008). It detects reference strings in academic papers and extracts bibliographic components such as author's name and article title. It is a string parser and the chain-model CRF is used. Zhu et al. proposed two dimensional CRF for IE from Web pages (Zhu et al., 2005). They exploit both layout and textual information for IE. Some researchers applied the two dimensional CRF for page image understanding (Nicolas et al., 2007; Montreuil et al., 2007).

The purpose of information extraction discussed in this paper is to segment tokens in two-dimensional space into logical units and assign labels to them as in (Takasu, 2008). The intended application of the proposed method is extraction of bibliographic information such as titles and authors from academic papers. Academic papers usually contain bibliographic information in the first page as shown in Figure 1 and in reference sections. We first separate each page into portions via a page layout analysis that are shown by red rectangles in Figure 1. We call them *a cell*. Usually the cells do not always correspond to bibliographic component, which is represented with black rectangles in Figure 1. For example, title in the figure is segmented into multiple cells. We call the bounding rectangles corresponding to bibliographic component *a logical component*. The problem discussed in this paper is to reconfigure cells into logical components, and assign a label to each component.

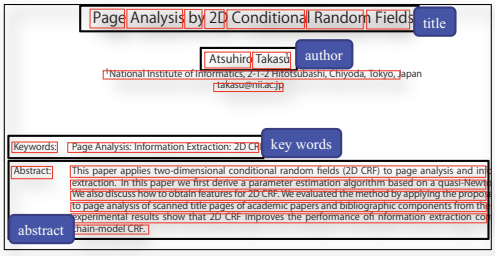

Figure 1: Example of page layout.

In this paper we apply two dimensional CRF to bibliographic component extraction from pages. The task is similar to the study (Councill et al., 2008), but we use two dimensional CRF to exploit both layout and textual information. The rest of this paper is organized as follows. Section 2 describes the two dimensional CRF used in this paper. Section 3 reports an experimental evaluation by bibliographic component extraction from scanned academic papers.

Takasu A..
Page Analysis by 2D Conditional Random Fields.
DOI: 10.5220/0004266505640567
In *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2013), pages 564-567
ISBN: 978-989-8565-41-9

$$\begin{matrix} x_{11}, z_{11} & \cdots & x_{1m}, z_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1}, z_{n1} & \cdots & x_{nm}, z_{nm} \end{matrix}$$

Figure 2: Data Structure for 2D CRF.

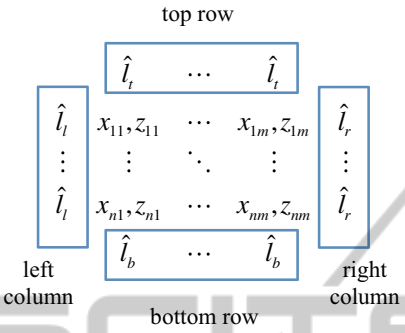

Figure 3: Augmented Data Structure.

2 2D MODEL FOR PAGE ANALYSIS

2.1 Data Structure

We assume that a page is decomposed into rectangles called *cell* through some page segmentation procedure and they forms an $n \times m$ matrix. There usually exist cells spanning over multi-columns and/or multi-rows. However, we assume they are decomposed to form an $n \times m$ matrix as in (Zhu et al., 2005). We denote ijth cell as $x_{i,j}$ that contains two kinds of information:

- layout information such as height, width, font types, and
- text included in the cell.

For a page, $x := \{x_{i,j}\}_{i,j}$ denotes the set of cells in the page.

2.2 Feature Vector

Let's consider a set F of feature functions. For a cell $x_{i,j}$, each feature function $f \in F$ calculates a feature value $f(x_{i,j})$. A feature function can be an indicator function that judges whether a cell contains a specific word. Or, it can be distance function that calculates the distance between text of the cell and a specific string. It can be extractor of a specific layout feature such as the width or can be a function to calculate the area of the cell, too.

The feature functions F define a feature vector $x_{i,j} := (f(x_{i,j}))_{f \in F}$. For a page, $x := \{x_{i,j}\}_{i,j}$ denote the set of feature vectors of the page.

2.3 Labels

CRF associates observed feature vector of each cell to a *hidden* label with probability. Let L denote the set of labels. We denote the label of ijth cell as $z_{i,j}$. For a page, $z := \{z_{i,j}\}_{i,j}$ denotes the set of labels of the page as in Figure 2.

Basically the labels correspond to types of logical components such as an author and a title. However, a logical component may be split into multiple cells. To treat the situation, we use additional labels that denote the boundary of cells constituting a logical component.

2.4 Page Boundary

In order to treat boundary conditions of pages, we introduce special labels $\{\hat{l}_l, \hat{l}_r, \hat{l}_t, \hat{l}_b\}$. $\hat{L}$ denotes the augmented label set, i.e., $L \cup \{\hat{l}_l, \hat{l}_r, \hat{l}_t, \hat{l}_b\}$. We add cells of

- top row whose label is $\hat{l}_t$, i.e., $z_{0,j} = \hat{l}_t$ for all columns,
- bottom row whose label is $\hat{l}_b$, i.e., $z_{n+1,j} = \hat{l}_b$ for all columns,
- left column whose label is $\hat{l}_l$, i.e., $z_{i,0} = \hat{l}_l$ for all rows, and
- right column whose label is $\hat{l}_r$ i.e., $z_{i,m+1} = \hat{l}_r$ for all rows,

as in Fig. 3.

2.5 Likelihood of 2D CRF

Two types of parameters are introduced in CRF. One is about observed feature vectors. For each feature function $f \in F$ and a label $l \in L$, let λ_{fl} denote a weight of the combination of f and l. Then, the likelihood that we observe the feature value $f(x_{i,j})$ for the hidden label l is proportional to

$$exp\left(\lambda_{fl} f(x_{i,j})\right) \quad (1)$$

The other is about label relationship between adjacent cells. It is further split into horizontal and vertical relationship. For each pair $(l,g) \in L^2$ of labels, let's introduce horizontal and vertical parameters ϕ_{lg} and θ_{lg}, respectively. Then, the likelihood that two labels l and g appear in horizontally and vertically adjacent cells are proportional to

$$exp(\phi_{lg}) \quad (2)$$

and

$$exp(\theta_{lg}) \ , \tag{3}$$

respectively.

Using these types of likelihood, the joint likelihood of the page feature vectors x and their labels z is given by

$$\begin{aligned}&\Pr(x,z)\\ &\propto \quad exp\left(\sum_{i,j} g(i,j)\right) \cdot \\ &\qquad \underbrace{exp\left(\sum_{i=1}^{n} \phi_{z_{i,m+1}\hat{l}_r}\right)}_{\text{for right column}} \cdot \underbrace{exp\left(\sum_{j=1}^{m} \theta_{z_{n+1,j}\hat{l}_b}\right)}_{\text{for bottom row}} , \end{aligned} \tag{4}$$

where we abbreviate $\sum_{i=1}^{n}\sum_{j=1}^{m}$ to $\sum_{i,j}$, and

$$g(i,j) := \sum_{f\in F} \lambda_{f z_{i,j}} f(x_{i,j}) + \phi_{z_{i-1,j} z_{i,j}} + \theta_{z_{i,j-1} z_{i,j}} \ . \tag{5}$$

Note that n and m differ depending on a page, but we use the same symbol for simplicity.

The conditional likelihood for CRF is given by

$$\Pr(z \mid x) := \frac{\Pr(x,z)}{\Pr(x)} = \frac{\Pr(x,z)}{\sum_{z'} \Pr(x,z')} \ , \tag{6}$$

where the denominator is called partition function and it is denoted as $Z(x)$.

2.6 Parameter Estimation

For parameter estimation, we use training data T consisting of pages represented by a pair (x,z) of feature vectors and labels of cells in the page.

Let us consider a regulalized log likelihood of the training data given by

$$\begin{aligned}&L(\lambda,\phi,\theta)\\ &:= \quad \log\left(\prod_{(x,z)\in T} \Pr(z \mid x)\right)\\ &\qquad -\sum_{f,l}\frac{\lambda_{lf}^2}{\sigma^2} - \sum_{l,l'}\frac{\phi_{ll'}^2}{\sigma^2} - \sum_{l,l'}\frac{\theta_{ll'}^2}{\sigma^2} \ , \end{aligned} \tag{7}$$

where the second, third, and fourth terms are L_2 regularization of the parameters. We estimate the parameters that maximize the likelihood, i.e.,

$$\underset{\lambda,\phi,\theta}{\operatorname{argmax}} L(\lambda,\phi,\theta) \ . \tag{8}$$

We can solve the optimization problem by the quasi-Newton method such as L-BFGS in the same way as (Zhu et al., 2005). We omit the details due to the page limitation.

3 EXPERIMENTAL RESULT

We applied the proposed model to extract bibliographic components in title pages of academic journals.

3.1 Data Sets

The task of the experiment is to assign labels to cells from scanned and OCRed title pages. We evaluated proposed model using the following three kinds of academic papers.

- Papers issued by the Information Processing Society of Japan (IPSJ): In this experiment, we used the papers issued in 2003.
- Papers issued by the (IEICE-E) : In this experiment, we used the papers issued in 2003.
- Papers issued by the (IEICE-J) : In this experiment, we used the papers issued in 2003 and 2004.

For each dataset, we applied 5-fold cross validation.

3.2 Evaluation Metric

As for the evaluation metric, we used the accuracy that the model assigns a label to each cell in the test matrices. We regarded the model succeeds in labeling only when it assigns correct labels to all cells in the matrix. The accuracy is defined by

$$\frac{\text{the number of successfully labeled matrices}}{\text{total number of test matrices}} \ .$$

3.3 Experiment Procedure

To make ground truth data, we manually extracted bounding rectangles that correspond to a logical component in the first page of the papers. Extracted logical components are *article title, author, abstract* and *keyword*. Since a logical component usually consists of multiple cells, the label of each cell is assigned with the label of the logical component that contains the cell.

We used the following features of each cell (Ohta et al., 2010):

- abscissa,
- ordinate,
- width,
- height,
- gap between adjacent cells,
- averaged characters' width in the cell,
- averaged characters' height in the cell,

Table 1: Extraction accuracy.

Method	IPSJ	IEICE-E	IEICE-J
Chain	0.938	0.949	0.798
2D	0.962	0.964	0.855

- number of characters in the cell,
- proportion of alphanumerics,
- proportion of hiragana and katakana,
- proportion of symbols, and
- presence of predefined keywords.

3.4 Experimental Result

For comparison, we applied a chain-model CRF examined in (Ohta et al., 2010). The OCR we used in this experiment made a character sequence from scanned document image according to the result of its layout analysis. In this experiment, we converted each character sequences into a word sequence and applied chain-model CRF.

Table 1 shows the extraction accuracy. "Chain" stands for the result when we used the chain model, whereas "2D" stands for the result of the proposed method. As shown in the table, two dimensional CRF achieved better performance than the chain model. We obtained more improvement for the data set "IEICE-J". This is because the OCR often analyzed the layout of "IEICE-J" pages incorrectly. It resulted in generating incorrectly ordered sequences and degraded the accuracy of the chain-model CRF. In contrast, two dimensional CRF is not affected by the order of cells by OCR. Therefore, it can improve the extraction accuracy.

4 CONCLUSIONS

This paper examines a two dimensional CRF for extracting bibliographic components from scanned page images of academic papers. We experimentally showed that the proposed method is effective especially for the pages whose layout is incorrectly analyzed.

Currently we use two dimensional CRF that treats matrices. With this model, we can assign a label to each cell but we need a post-processing that extracts logical components by merging cell. We plan to extend the model to treat tree structured data such as XY-tree. It enables us to extract logical components as well as labeling simultaneously. In this paper we manually determined the augmented labels for merging cells into logical component. We are interested in designing the augmented labels systematically.

REFERENCES

Councill, I. G., Giles, C. L., and Kan, M.-Y. (2008). Parscit: An open-source crf reference string parsing package. In *Intl. Conf. on Language Resources and Evaluation (LREC 2008)*, pages 661 – 667.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML 2001)*, pages 282 – 289.

Montreuil, F., Grosicki, E., Heutte, L., and Nicolas, S. (2007). Unconstrained handwritten document layout extraction using 2d conditional random fields. In *International Conference on Document Analysis and Recognition (ICDAR 2009)*, pages 407 – 411.

Nicolas, S., Dardenne, J., Paquet, T., and Heutte, L. (2007). Document image segmentation using a 2d conditional random field model. In *International Conference on Document Analysis and Recognition (ICDAR 2007)*, pages 407 – 411.

Ohta, M., Inoue, R., and Takasu, A. (2010). "Empirical Evaluation of Active Sampling for CRF-based Analysis of Pages". In *International Conference on Information Reuse and Integration (IEEE IRI2010)*, pages 13–18.

Takasu, A. (2008). "Information Extraction by Two Dimensional Parser". In *Proc. IEEE Intl. Conf. on Tools with Artificial Intelligence*, pages 333–340.

Zhu, J., Nie, Z., Wen, J.-R., Zhang, B., and Ma, W.-Y. (2005). 2d conditional random fields for web information extraction. In *International Conference on Machine Learning (ICML 2005)*.