Table Recognition and Evaluation in PDF Documents

Jing Fang Advisor: Zhi Tang
Institute of Computer Science & Technology, Peking University

Background

The most straightforward motivation of my thesis stems from the requirement of mobile reading of fixed-layout documents (e.g. PDF). Due to the relatively small screens of handheld devices, those documents usually need to be recomposed to avoid readers moving the screen back and forth.

When coming to tables, the ideal way of reading is: table contents can be rendered as a integrated object with original grid structure. For large tables, it is better to be able to hide/show certain columns or rows according to people's reading preference. This proposes requirement of segmenting table cells and finding out the indication relations correctly.

A good number of research efforts have been made on table recognition, which also brings up the evaluation issue. A large quantity and representative ground-truthed dataset and performance metrics are necessary to evaluate which algorithm is better in different applications.

Challenges

● Table detection

Existing works on table detection from PDF format made use of just cell layout information. This works well for regular tables, but become ineffective when dealing with irregular tables or tables in complex-layouted pages.

● Table structure extraction

The challenges are mainly caused by varied ways tables can show up in real-world documents, e.g. nested tables, sparse tables, multi-level headers, tables with spanning or multi-line cells etc. It is not easy to segment cells correctly and extract indication relations.

● Automatic evaluation

Most of existing table recognition algorithms were evaluated on their in-house data set. No general table ground-truth dataset for PDF documents is publicly available.

Another issue of is about performance metrics. Most of existing published papers analyze experiment results using P&R metrics, which are widely used in information retrieve field. But error types of table recognition is more complicated than just true and false. Besides, human judge is subjective and hard to reproduce.

Progress to Date

● Table detection

A table detection method via visual separators and geometric content layout information has been proposed and accepted by ICDAR. The visual separators refer to not only the graphic ruling lines but also the white spaces [1] to handle tables with or without ruling lines. Furthermore, page columns are detected in order to assist table region delimitation in complex layout pages. (Fig 1)

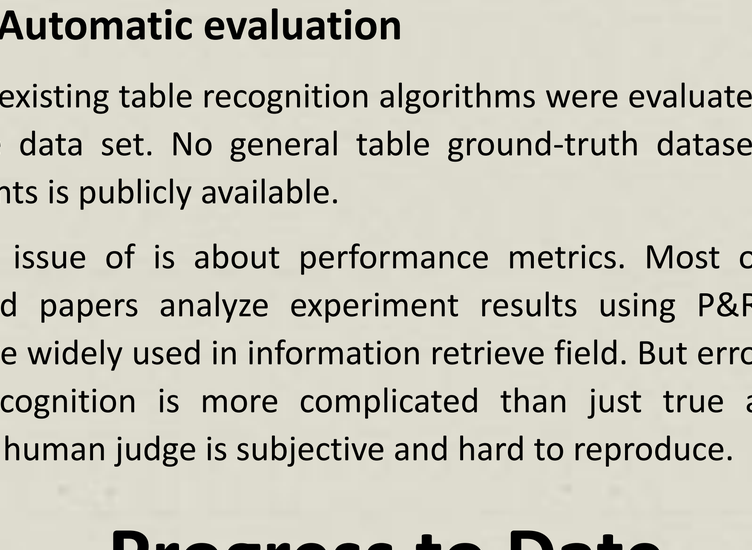

Fig 1. Workflow of our table detection method

Table I compares our results with Liu's sparse line based algorithm [2]. Fig2 provides some examples to illustrate the effectiveness of our methods in different cases.

Table I. Experimental data

Methods	D1		D2	
	Precision	Recall	Precision	Recall
Method in this paper	96.13%	92.07%	94.42%	93.71%
Method in Liu [2]	93.56%	83.20%	96.28%	92.50%

Fig 2. Experimental example illustrations

● Table detection evaluation

1. Both English and Chinese pages dataset with ground-truths have been constructed and will be made publicly accessible soon.
2. The ground-truths are XML-based, containing not only table content data, but also low-level objects parsed from PDF documents. Page images are also provided to compatible with image-based algorithm evaluation.
3. A set of metrics are defined and implemented to automatically evaluate table detection algorithms.

Future Direction

● Table structure extraction

1. Give a physical description of the table, i.e., identify its cells and their relative positions, as well as its rows and columns from the detected table regions.

2. Extract logical structure of table cells, i.e., determine the heading rows and columns, and relationship between the indication cells and body data cells.

3. Identify the affiliated table attributes, such as table caption, table footnotes and descriptions from text paragraphs. These optional components provide basic semantic information of tables.

● Table structure evaluation

For structure analysis, evaluation metrics should work for not only physical segmentation, but also logical relationship and semantic interpretation. Proposing proper evaluation metrics will be covered in my future research.

Bibliography

1. T.M. Breuel, "Two geometric algorithms for layout analysis," Proc. Of Document Analysis Systems (DAS'02), 2002, pp. 188-199.
2. Y. Liu, K. Bai, P. Mitra, and C.L. Giles, "TableSeer: automatic table metadata extraction and searching in digital libraries," Proc. of Joint Conference on Digital Libraries, (JCDL'07), 2007.

北京大学 Institute of Computer Science & Technology, Peking University