A METHOD FOR TABLE DETECTION IN METAFILES

Authors: I.V. Bychkov, A.E. Hmelnov, G.M. Ruzhnikov, A.O. Shigarov
Location: Institute for System Dynamics and Control Theory, Siberian Branch of RAS
664033, 134, Lermontov st., Irkutsk, Russia, tel. +7-3952-45-31-02, e-mail: shigarov@icc.ru

Abstract

The poster presentation demonstrates a heuristic method for detection of statistical tables in documents. The method uses Enhanced Metafiles as input data, which allows one to apply it to documents of various formats.

Introduction

For the purpose of solving many research and applied problems it is necessary to extract data from the tables contained in various documents. Methods and systems of table extraction from documents allow one to automate this process. The first stage of table extraction is table detection, i.e. finding on the pages of documents the areas, which may be identified as tables.

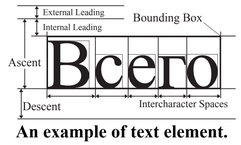
The existing methods of table detection use as a rule such input data as raster images, ASCII-text. At the same time, it would be interesting to consider one of the exchange formats, e.g. PDF, PostScript or Enhanced Metafiles (EMF), as a representation of input data for table extraction algorithms. These formats are more informative in comparison with raster images and ASCII-text.

Our heuristic method of table detection is oriented on metafiles. In contrast to the other exchange formats, the EMF may be interpreted using the functions of GDI API. This fact makes EMF processing rather simple and accessible. Furthermore, documents of various formats, e.g. DOC, XLS, PDF, ASCII-text or HTML may be printed to metafiles (EMF). It is assumed that tables in the documents, which are printed into metafiles, are not represented by raster inclusions.

Extraction Data from Metafiles

Each document page is represented by a separate metafile. Interpretation of data from these records is performed using the information from the device context, which plays metafile.

As a result of processing metafile this way, it is possible to obtain for each record, corresponding to a text output instruction, one or more text elements. Each text element contains the following three sets: 1) characters; 2) intercharacter spaces; 3) font metrics, and also a bounding box.



An example of text element.

Furthermore, as a result of processing the metafile, rules (ruling lines) are extracted from the records corresponding to the instructions of output of graphics.

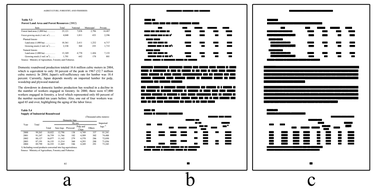
After extraction of text elements they undergo preprocessing to detect text ruling, which is removed from the text and joined with the rest of ruling. Aside from that, the restoration of words (i.e. sequences of non-blank characters) is performed, which takes into account inter character spaces and relative position of text elements. The restoration of words is required because different parts of one word may correspond to different text elements, and vice versa, several words may occur in one text element. After the preprocessing, the majority of text elements shall correspond to separate complete words.

Table Detection

The method is focused on the statistical tables.

An example of statistical table.

The closely located and not separated by any rules text elements are grouped into text blocks.



a b c

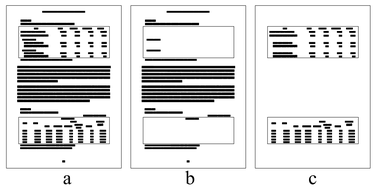
An example of building text blocks on the page: the input page (a); bounding boxes of text elements (b); bounding boxes of text blocks (c).

The text blocks are further grouped into lines. Then, the white space into the bounding box of each of the lines is segmented. Vertical gaps between text blocks are identified among the segments.

Year	Total	By use					Imported logs [1]
		Total	Saw-logs	Plywood	Pulp and Chips	Others	
2000	99,263	18,022	12,798	138	4,749	337	81,241
2001	91,247	16,759	11,766	182	4,509	302	74,488

An example of a fragment of a page with obtained lines: the lines are marked by rectangular frames; vertical gaps in the lines are hatched.

Then table regions are formed by the sequences of consecutive lines, which satisfy of the following heuristic conditions: **(a)** the line should contain at least two text blocks; **(b)** the width of the line's white space with respect to its total width should not exceed a predefined threshold; **(c)** the lower boundary of any vertical gap of the line should coincide with the lower boundary of its bounding box; **(d)** any two lines of the table region should both satisfy the condition that each vertical gap from the upper line should correspond to at least one vertical gap from the lower line, furthermore, the latter gap should be such that its upper boundary coincides with the upper boundary of the lower line, and the intersection of x-axis projections of these two vertical gaps exceeds a predefined threshold.



a b c

An example line division: the input page (a); lines of text (b); lines of table regions (c).

The table regions obtained shall not have common lines. Next, for each table region found, a set of its proper vertical gaps is formed by vertical gaps of its lines.

COUNTRY	January - July				July			
	Quantity		Value		Quantity		Value	
	2004	2005	2004	2005	2004	2005	2004	2005
European Union								
Germany	11,662	10,684	86,690	81,784	1,118	415	6,523	1,685
Belgium-Luxembourg	9,505	5,284	67,820	37,930	27	123	146	648
Netherlands	2,775	4,875	21,429	39,694				
France	5,612	3,030	15,889	12,923				
Other Markets								
Japan	13,352	9,117	90,901	52,604	107	76	550	313
Russian Federation	6,406	8,801	29,026	47,781	1,173		5,123	
Switzerland	1,902	2,899	13,713	21,090	144	58	878	447

An example of table regions: the table regions are marked by rectangular frames; vertical gaps in the table regions are hatched.

The table regions obtained may represent either tables or parts of tables, but they could also represent the text parts having table-like arrangement. The table regions, which form one table, correlate with each other according to the arrangement of the x-axis projections of their vertical gaps. This feature is used in the proposed method for merging the table regions into tables and determining the table boundaries.

Experimental Evaluation

Experimental data

Government statistical reports of different countries, e.g., Russia (www.gks.ru), US (www.fedstats.gov), EU (Eurostat yearbook 2006-07), Japan (Statistical Handbook of Japan 2007), and also in financial reports of open joint stock companies, e.g., Boeing, Aeroflot.

Total 345 pages of statistical reports, which have been represented in such formats as PDF, DOC, XLS and HTML. These pages contained 440 tables. Furthermore, these documents contained text fragments having a table-like form.

Measures

A table is considered to be correctly detected if at least its body has been detected correctly, i.e. every line in the body of the table is identi-fied as being part of the table, and any line, which does not belong to the table, has not been identified as a line of its body.

precision 86.4% (i.e. the percentage of the number of correctly detected tables with respect to the total number of detected tables).

recall 92.6% (i.e. the percentage of the number of correctly detected tables with respect to the total number of existing tables).

The experimental evaluation of the method has demonstrated its efficiency for a wide range of statistical tables.