

Automatic Extraction of Table Metadata from Digital Documents^{*}

Ying Liu, Prasenjit Mitra, C. Lee Giles, Kun Bai
College of Information Sciences and Technology, Pennsylvania State University
University Park, Pennsylvania, 16802
{yliu,pmitra,giles,kbai}@ist.psu.edu

ABSTRACT

Tables are used to present, list, summarize, and structure important data in documents. In scholarly articles, they are often used to present the relationships among data and highlight a collection of results obtained from experiments and scientific analysis. In digital libraries, extracting this data automatically and understanding the structure and content of tables are very important to many applications. Automatic identification extraction, and search for the contents of tables can be made more precise with the help of metadata. In this paper, we propose a set of medium-independent table metadata to facilitate the table indexing, searching, and exchanging. To extract the contents of tables and their metadata, an automatic table metadata extraction algorithm is designed and tested on PDF documents.

Categories and Subject Descriptors: H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval

General Terms: Algorithms, Experimentation, Documentation, Performance.

Keywords: Metadata extraction, table detection, table structure recognition, searching, exchanging.

1. INTRODUCTION

Digital libraries contain large collections of electronic documents. A table is a popular document element to present structured data and relational information, using text and layouts. Understanding the structure and sharing contents of tables are very important for many applications. For example, in chemistry or biology fields, researchers often use tables to present and share their experimental results. However, converting tables into texts and processing with traditional search methods without the ability to query the contents of the tables cannot give satisfactory results especially for numerical data presented in the tables.

Tables are pervasively used in many media, such as HTML, PDF, images, etc. Researchers in the table-understanding field largely focus on algorithms to detect the table boundary and analyze the table structure in a specific document media with different process methodologies: OCR [3], X-Y cut [1], tag classification, etc. For a defined query, the matched

tables may exist in diverse document formats. Limited attention has been paid to enabling table searching and table exchanging. Although table models and representation schemes are proposed [2], the physical and logical structures are not clearly set apart. This results in a long learning curve and many useless attributes for users' operation/application.

In this paper, we propose a set of medium-independent table metadata and build up the metadata foundation for table representing, searching, and exchanging. The metadata not only includes the information related to the table itself, but also the accessorial information ignored by current researchers, such as table footnote and referenced annotation. With these metadata, we can build up shareable databases to enable more specific searching of tabular data. An automatic table-metadata-extraction algorithm is designed and tested on PDF. It will be easy for users to find the target tables across several mediums and reverse engineering the original table layout, store the extracted data in a database, present it in a medium same/different than their original target, and allow querying on the table contents if required.

2. TABLE METADATA

Extracting sharable table metadata is a challenging problem for several reasons: diverse medium types, no formal table designing rules/standards, different presentation schemes in different mediums, different table layout requirements of different publishers, diverse table cell types, many affiliated elements, etc. In order to characterize tables spanning over a wide range of documents, a rich and flexible set of representation metadata is required. A fundamental extraction tenet is as follows: The metadata should have meaningful names; they should be easy to store in a database to assist future searches; they can be combined into different levels according to the users' purposes; Not all metadata has to be present for all the tables. For different purposes, we may only need a subset of the table metadata. We classify the table metadata into six categories:

Table environment/geography metadata: It records the document metadata and the table position. The metadata includes Document Type, Document Address, Document Name (the name/title of the document.), Document Authors, Document Page Number (the index number of the page where the table is located in), Document Column Number (the number of columns in the page where a table is located in), Table Starting Position (The X and Y-axis position of the starting place of the table). The Table Starting Positionis metadata, together with the Document Page Number, and the Document Column Number, can uniquely

^{*}This work was partially supported by NSF grants 0454052 and 0535656.

locate the table position in a document.

Table frame metadata: Outer Table Ruling (This metadata records the border information of a table, the values can be left, right, top, bottom, all, none, top and bottom, left and right);

Table affiliated metadata: Table Index Number (such as “Table 2”), Table Caption, Caption Position (the caption can be put above/below the table body. If the table does not have a caption, the value should be “none”), Table Footnote, and Referenced Annotation;

Table layout metadata: It helps us capture the visualization of the original table layout. They include Table Width, Table Length, the Number of Rows, Stub Separator (vertical ruling), Boxhead Separator (horizontal ruling), Column Width, Row Length, Column Headers, Row Headers, Alignment (horizontal alignment. The values can be flush left, right, central, left and right);

Table content metadata: Table Content Data (such as the content in cell[i, j], i: the row number, j: the column number); This metadata records the values in each cell and enables searching tables based on cell contents.

Table type metadata: It records the ontology of a table. The value can be Numerical, Text-based, Symbol-based, Image-based, or Combination.

3. TABLE METADATA EXTRACTION ALGORITHM FOR PDF FILES

Current researches in the field of table understanding mainly recognize tables in HTML or image files. Although PDF is getting more popular, no tools or research extracts tables from it. We designed an automatic table metadata extractor for PDF files to make up the vacancy. The algorithm is composed of three steps: converting a PDF document into a formatted text file, detecting the table candidates based on location analysis and keyword matching (table environment/geography metadata is extracted in this phase), confirming/denying table candidates, and recognizing the table structure. Figure 1 shows the detailed pseudo-codes for the step 2 and step 3. α_1 is a user defined parameter based on observation and experience. We set it as 2 because documents usually leave at least two times of the average space between content lines before the beginning of a table. The predefined keyword list: we combine the keyword matching with position, space analysis in the algorithm. For the table detection, we create a predefined keyword list that records all the possible starting keywords of tables, such as “Table”, “TABLE”, “table”, “Form”, “form”, “FORM”, etc.

4. EXPERIMENT AND DISCUSSION

We tested the algorithms on 120 randomly selected PDF documents from digital libraries. Each extracted metadata is evaluated for precision, recall, and accuracy. We define A as the number of true positive metadata extracted by our algorithms and labeled with the correct metadata labels, B as the number of true positive metadata but overlooked by our algorithm, C as the number of true negative metadata that is misidentified by our algorithm as another metadata, and D as the number of true negative metadata predicted as negative. Experimental results show that the algorithm has good performance on table metadata extraction (an overall precision, recall and accuracy of over 95%). Precision, recall and accuracy are defined as follows: $precision = \frac{A}{A+C}$,

$$recall = \frac{A}{A+B}, \text{ and accuracy} = \frac{A+D}{A+B+C+D}.$$

Our current algorithm has two main limitations: first, it can not extract the table-frame metadata. Second, if there are images in table cells, we also will treat them as empty cells. The reason is that when we convert the PDF documents into text format, all the line/image information is lost. Getting the line/image information, and creating a table database based on these metadata to facilitate table exchanging and searching are left for the future.

Input: the converted PDF file in text format f2

Output: the starting position of a table candidate and Table Environment/Geography metadata

```

Begin
   $\Delta_y \leftarrow$  compute the average space between two lines in f2;
  for each page do
    for each line do
      linespace = the line space between this line and its previous line
      d = the first word of this line
      if ( (linespace >  $\Delta_y * \sigma_1$ ) && (d  $\in$  the predefined keyword list) )
        etm  $\leftarrow$  collect the Table Environment/Geography metadata
        return (etm);
      end if
    end for
  end for
End

```

Input: the starting position of a table candidate detected in Step 2

Output: Other table metadata except for the environment/geography metadata

```

Begin
  nline  $\leftarrow$  read the first text piece in the table candidate
  row=1; column=1;
  x0  $\leftarrow$  the start X-axis position in nline; x1  $\leftarrow$  the end X-axis position in nline;
  y0  $\leftarrow$  the start Y-axis position in nline; y1  $\leftarrow$  the end Y-axis position in nline;
  startX[column] $\leftarrow$ x0; endX[column] $\leftarrow$ x1; startY[row] $\leftarrow$ y0; endY[row] $\leftarrow$ y1;
  while (NOT at the end of a table)
    nline  $\leftarrow$  read the next text piece in the table candidate
    if (nline = new column in the same row)
      column++; adjust startX[column] and endX[column];
    end if
    if (nline = next line in the same cell)
      combining with previous lines in the same cell;
      adjust startX[column] and endX[column];
    end if
    if (nline = text pieces in the next row)
      row++; adjust startY[row] and endY[row];
    end if
    if (nline is special scenarios, e.g. a superscript)
      combine with the previous text line;
      adjust startX[column] and endX[column];
    end if
    tab.add(cells[row][column]);
  end while
  tab  $\leftarrow$  collect other table metadata;
  return (tab);
End

```

Figure 1: Table Metadata Extraction Algorithm

5. REFERENCES

- [1] J. Ha, R. M. Haralick, and I. T. Philips. Recursive x-y cut using bounding boxes of connected components. In *Third Int'l Conference on Document Analysis and Recognition*, pages 952–955, 1955.
- [2] OASIS. The organization for the advancement of structured information standards. www.oasis-open.org.
- [3] D. Pinto, A. McCallum, X. Wei, and W. Bruce. Table extraction using conditional random fields. In *Proceeding of the 26th ACM SIGIR*, July 28-August 1 Toronto, Canada, 2003.