Development of the formal model of the optical recognition mechanism for tabular documents

A V Rabin and A A Petrushevskaya
Saint-Petersburg State University of Aerospace Instrumentation (SUAI), ul. Bolshaya Morskaya, 67, lit. A, St. Petersburg, 190000, Russia

E-mail: alexey.rabin@guap.ru

Abstract. The article describes the tasks of the optical recognition system for tabular documents and the development of a formal model of the optical recognition mechanism for tabular documents. The aim of the study is to use the developed system of automatic analysis of platform-type documentation as expert and training to support decision making, semantic analysis or audit of any content, including in other industries based on the principle of automatic filling of ontologies. The proposed system will allow to overcome this barrier in a cardinal way – by automatically creating a database. In this case, the user's role will be reduced to the choice of sources for filling ontologies (the choice of thesauruses, corpus of texts in machine-readable or natural language, etc.) and the configuration of inference rules, i.e., essentially, to the configuration operations.

1. Introduction
Automated processing of modern business documents - paper input, sorting, routing, etc. - An actively developing field of computer technology. Over the past 20 years, the optical recognition quality of a single character and line of text (OCR) has been brought to a quality sufficient for industrial applications. Text recognition programs have become an important (and for some types of business processes - necessary) part of office software that successfully solves the problem of entering textual information.

In recent years, the ability to “understand” the structure of the source document and create an adequate electronic presentation has become an increasingly relevant property of systems for automated input of documents from a scanned graphic image.

Documentation operations, performed even with the help of existing automation tools, are extremely laborious, especially when it comes to a fairly large enterprise. The aim of the study is to use the developed system of automatic analysis of platform-type documentation as expert and training to support decision making, semantic analysis or audit of any content, including in other industries based on the principle of automatic filling of ontologies.

For tabular documents, the composition, location and structure of information are fixed in a certain way. For systems specialized in the input and processing of tabular documents, a priori knowledge of the structure and filling rules is an important input information, significantly expanding the technological capabilities of document processing and improving the quality of recognition.
2. Formal model of optical recognition mechanism for tabular documents

The creation of optical recognition systems for tabular documents requires the formalization of the concept of a table and the creation of methods for identifying tables in images. A table is defined as an information object displayed on a medium.

As part of a tabular document, the development of a formal model of the table is considered. A comprehensive table model consists of the following consistent models:

1. a model of a tabular information object that describes classes that allow a flat representation of an information object and provides:
   a. a method for describing a data structure of an instance of a table;
   b. a method for describing non-structural dependencies and data constraints;
   c. formalization of valid operations on a table instance.

2. A flat representation model that defines:
   a. media model;
   b. algorithms for constructing table instance representations;
   c. a method for describing the selection of a particular variant of an algorithm and a set of its parameters.

Parts of the table model must be consistent so that:

- any copy of the table contains the facts of the subject area in accordance with the model of the subject area;
- for any instance of the table, you can build a flat view, in accordance with the algorithm for constructing a flat view and data model;
- operations on the table are determined so that at any transition between instances of the table the previous paragraphs are performed;
- the table object presented in accordance with the flat presentation model can be selected, identified, and disassembled in accordance with the recognition model.

Figure 1. Table view diagram.

To conduct a full analysis of the text in natural language, the system should be able to analyze the source text in terms of syntax (sentence structure), semantics (concepts used in the text) and pragmatics.
(the correct use of concepts and the purposes of their use). After that, the system should generate its response in an internal representation suitable for logical inference, and synthesize its response in a natural language.

A common table model is a matrix, as a function of two index arguments. Function values can be represented in concatenated cells. Cell merging and formatting is often used to visualize the subordination and grouping of variable values in the header areas of a flat representation, but the matrix does not reflect the structure of the table object.

The general scheme of an instance of a flat representation of a tabular information object in a document is shown in figure 1 and consists of the following types of areas:

- identification zone;
- corner heading zone;
- horizontal heading zone;
- vertical header zone;
- embedded heading zone;
- matrix core;
- information zone.

Four main types of non-empty heading zones for vertical headings are also distinguished (figure 2), and the three types clearly define the hierarchical structure of key information.

![Figure 2](image2.png)

**Figure 2.** Layout of a vertical header zone of type a) matrix, b) T-hierarchy, c) M-hierarchy, and d) D-hierarchy.

Due to the peculiarities of the perception of textual information for horizontal header zones, three types are defined (figure 3), while T-hierarchical layouts often contain only part of the lines of separation into graphs.

![Figure 3](image3.png)

**Figure 3.** Layout of the horizontal header zone of type a) matrix, b) T-hierarchy, c) G-hierarchy.

The header zone is intended to identify the elements of the table, therefore it contains one or more unique elements - keys. Logically, the header zone is represented by a forest of trees in which the paths from the root to the peaks (including the inner ones) define a key unique to the zone, consisting of the
values of the cells in the header zone. The body of the table is generally represented by a matrix table, which is geometrically aligned with the header areas. An inset is a way to compactly place one or more senior levels of vertical zone headers. Identification and information zones contain general unstructured tabular information (name, footnotes).

Thus, identification and information attributes are mapped to the corresponding zones. A tabular information object is represented in the form of a matrix core and, possibly, framing its header zones.

The following are a lot of disjoint rectangles: \( R = \{ R_i \} \), \( \forall R_i, R_i \rightarrow R_i \cap R_i = \emptyset \). Intersections and associations are considered loosely, only in the case of a nonzero area.

Minimum enclosing rectangle (1):

\[
R^* = \text{RECT}(\min_{R_i \in R} R_i x_L, \min_{R_i \in R} R_i y_T, \max_{R_i \in R} R_i x_R, \max_{R_i \in R} R_i y_B).
\]

Let the rectangles \( R_0 \) and \( R_1 \) be in the enterprise-client relationship, which is denoted as \( R_0 \supseteq R_1 \) or \( r^{(F\rightarrow C)}(R_0, R_1) \), if (2)

\[
\begin{align*}
R_0 x_B &= R_1 x_L \\
R_0 y_T &\leq R_1 y_T \\
R_0 y_B &\geq R_1 y_B
\end{align*}
\]

An additional rectangle will be designated as \( R^\text{root} \): \( R^\text{root} \supseteq R^* \) subject to the following conditions (3):

\[
\begin{align*}
R^\text{root} y_T &= R^* y_T \\
R^\text{root} y_B &\leq R^* y_B \\
R^\text{root} &\supseteq R^*
\end{align*}
\]

Let the set \( \mathcal{R}^+ = \mathcal{R} \cup \{ R^\text{root} \} \) be an extension of the original set.

The extension \( \mathcal{R}^+ \) and the set of relations \( R^{(F\rightarrow C)} \) on it define the inheritance graph \( \Gamma = (\mathcal{R}^+, R^{(F\rightarrow C)}) \), which is a tree.

Rectangles \( R_1 \) and \( R_2 \) are geometrically senior manager – junior manager, then the record will be of the form \( R_1 \supseteq R_2 \) or \( r^{(B\rightarrow B)}(R_0, R_1) \), if relation (4) is true

\[
\begin{align*}
(\exists R_0: R_0 &\supseteq R_1, R_0 \supseteq \ R_2) \\
&\quad \ R_1 y_B \leq R_2 y_T
\end{align*}
\]

There are also many \( \mathcal{R}^+_\text{CH}(R_0) \) links for the \( R_0 \) rectangle.

Rectangles of the set \( R_1 \in \mathcal{R} \) form the T– hierarchical covering if they tightly cover the enclosing rectangle \( R^* = \bigcup_{R_i \in \mathcal{R}} R_i \), and for each of them the parent is uniquely defined \( \forall R_i \in \mathcal{R}! R_j \in \mathcal{R}^+: R_j \supseteq \ R_i \).

The last condition means the uniqueness of the path from any rectangle through the parents to the root of the tree corresponding to \( R^\text{root} \). This is how the cells are located in the vertical header zone.

For a set of rectangles, the following theorem on dense packing in a T-hierarchical header zone is formulated and proved (5):

For a set of rectangles, the following theorem on dense packing in a T-hierarchical header zone is formulated and proved (5):

\[
\begin{align*}
R^* &= \bigcup_{R_i \in \mathcal{R}} R_i \\
\forall R_i \in \mathcal{R}: R_j &\supseteq R_i \iff \forall R_i \in \mathcal{R}^+ \exists R \in \mathcal{R}: R_i \supseteq R \rightarrow R_i \supseteq \mathcal{R}^+_\text{CH}(R_i)
\end{align*}
\]

Further, for all types of header zones, the key cell stacking is considered and the definition of the «correct» flat representation is formulated: a flat representation having header zones is geometrically...
consistent if any cell core border of the matrix core is confirmed by the cell-key boundary of the corresponding header zone. A non-empty corner zone can contain values of only non-essential variables and must be consistent with the vertical and / or horizontal header zone. The corner zone plays the same role for the vertical or horizontal heading zone as the heading zone for the body, so coordination is carried out according to the same rules. The main types of non-empty corner zones are the matrix (figure 4.a) and the T-hierarchy (figure 4.b). Simultaneous coordination of the corner zone with both header zones is possible for the oblique (figure 4.c) and matrix (figure 4.d) type.

![Figure 4](image-url) Layout of the corner header zone of type a) matrix, b) T-hierarchy, c) oblique symmetry, d) matrix symmetry.

By the number of header zones, the flat representation belongs to one of the following classes:

a. the flat view that does not contain header zones;
b. the flat view that contains one header zone;
c. the flat view that contains two header zones.

In total, three basic types of tabular information objects are distinguished, for which basic operations that preserve the structure of the object are determined.

Class I is a multidimensional tabular information object $T_N$, $N>1$ is represented by the tuple $T_N=(\psi_\nu = f_\nu(\Lambda_\nu), I, \Delta, \Phi, \Psi)$, where the measurement vector $\bar{V}$ defines the subspace in which the vector $\Lambda_\nu$ of the full address space $\Delta$ according to the law from the set of functions $\Phi = \{f_\nu(\Lambda_\nu)|\bar{V} \in I\}$ the scalar dependent variable $\psi_\nu = f_\nu(\Lambda_\nu)$ from the set of dependent variables $\Psi = \{\psi_\nu|\bar{V} \in I\}$.

An important particular case is the number of measurements $N=2$:

$$T=\langle \psi_{p,q} = f_{p,q}(\delta_{R_p}, \bar{\delta}_{C_q}), I_R, I_C, \Delta_R, \Delta_C, \Phi_{RC}, \Psi_{RC} \rangle,$$

for which the measurements $I_R, I_C$ and address subspaces $\Delta_R, \Delta_C$ are named row and column, respectively. Law $\aleph$ defines the way of transition from general to flat $\langle \psi_\nu = f_\nu(\Lambda_\nu), I, \Delta, \Phi, \Psi \rangle$ to $\langle \psi_{p,q} = f_{p,q}(\delta_{R_p}, \bar{\delta}_{C_q}), I_R, I_C, \Delta_R, \Delta_C, \Phi_{RC}, \Psi_{RC} \rangle$ (7).

Class II tabular information objects are objects of the following type:

$$T=\langle \Phi \rangle,$$

where the functions of the set of functions (9):

$$\Phi = \{F_0(\xi), F_i(\xi)|i, \xi = V(\eta_{i,1}, ..., \eta_{i,N}), \xi = V(\xi_1, ..., \xi_N) \} \cup \{F_0(\xi)\}$$

are consistent in the number of arguments $N$.

Class I tabular information objects can be equivalent to a variety of relational relationships. Class II tabular information objects in the general case does not have a direct relational representation.

The main representations of the tables are paper documents, electronic documents and databases. In input processes using automatic or semi-automatic procedures, information from the source view is converted to a tabular information object.

Isolation and analysis of the structure of image objects or a text document is the first step in solving the complex task of automating the input of tabular documents, where the processing of tables requires specific actions. In accordance with the visualization model considered, a flat representation of a tabular
informational object consists of cells that have explicit or implicit boundaries, therefore it is a geometric analysis of image objects in order to identify compact and structured graphic primitives in accordance with visualization models that is the core of optical table input processes.

Isolation and analysis of the structure of image objects or a text document is the first step in solving the complex task of automating the input of tabular documents, where the processing of tables requires specific actions. In accordance with the visualization model considered, a flat representation of a tabular informational object consists of cells that have explicit or implicit boundaries, therefore it is a geometric analysis of image objects in order to identify compact and structured graphic primitives in accordance with visualization models that is the core of optical table input processes.

3. Conclusions

Another important task of implementing an industrial system for entering tabular documents is the verification of recognition results.

The above model describes the relationship between the values of the variables of a tabular information object in terms of structural elements, which allows you to escape from a specific flat representation. Thus, when entering well-known documents, it is possible to formulate and implement sets of rules to which the data must comply. Automatic verification and presentation to the operator of the results allows you to control the recognition results and actions of the operator, which, in turn, significantly improves the overall quality of the input system.

Acknowledgments

The paper was prepared with the financial support of the Ministry of science and higher education of the Russian Federation in the course of the applied research «The comprehensive project to create high-tech production of software tools for automatic analysis of documentation on paper and digital media using semantic-cognitive technologies for cataloging poorly structured information» (unique identifier of the project 02.G25.31.0305, Decree of the Government of the Russian Federation N 218, 09.04.2010).

References

[1] Rabin A V, Petrushevskaya A A, and Sinitsin O V 2020 Methods and formal models of intelligent analysis of weakly structured data (IOP Conference Series: Materials Science and Engineering) 450 734 012159
[2] Field D V 2007 Development of models, methods and means of processing tabular documents in information systems: abstract of thesis. for a job (Moscow Institute of Physics and Technology MIPT) 7–14
[3] Field D V 2008 About the tabular presentation of information (Proceedings of the 51st scientific conference of MIPT) 78–98
[4] Zarubin A A, Koval A R, Moshkin V S, and Filippov A A 2017 Construction of the problem area ontology based on the syntagmatic analysis of external wiki-resources (URL: http://ceur-ws.org/Vol-1903/paper26.pdf)
[5] Couellan, N, Jan S, Jorquera T, and Georgé J 2015 Self-adaptive support vector machine: a multi-agent optimization perspective (Expert Syst. Appl.) 42(9) 4284–98
[6] Maldonado S, Weber R, and Famili F 2014 Feature selection for high-dimensional class-imbalanced data sets using support vector machines (Inf. Sci.) 286
[7] Yan H, Ye Q, Dong-Jun Yu, Yuan X, Yiqing X, and Liyong F 2018 Least squares twin bounded support vector machines based on L1-norm distance metric for classification (Pattern Recognit) 74
[8] Bakos J D, Zhang F, and Zhang Y 2013 Accelerating frequent itemset mining on graphics processing unit (J. Supercomput) 66(1) 94–117
[9] Huang Y S, Yu K M, Zhou L W, Hsu C H, and Liu S H 2013 Accelerating parallel frequent itemset mining on graphics processors with sorting (LNCS IFIP International Federation for Information, Springer, Berlin) 8147 225–45
[10] Ltifi H, Kolski C, Ayed M B, and Alimi A M 2013 A human-centred design approach for developing dynamic decision support system based on knowledge discovery in databases (Journal of Decision Systems) 22 69–96