

PROSAR-AIDA

PROSAR-AIDA: **AI-based document analysis**

PROSAR-AIDA (**A**rtificial **I**ntelligence for **D**ocument **A**nalysis) is a software module based on PROSAR which is available for content-based document processing.

Technical Description

The module PROSAR-AIDA allows content-based processing of formatted and freeform documents. Based on an initial full-text OCR, PROSAR-AIDA provides two main functions:

1. Recognising the type of document through content-based analysis
2. Finding relevant index data without having to know the position or the layout of the data on the page.

These functions are described in detail below.

Document type detection

Document type detection is a hierarchical process, based on rules which assess the presence or absence of defined key texts. On the highest hierarchical level, PROSAR-AIDA decides between rough categories. Should the attributed category itself contain further definitions, the decision-making process is repeated recursively until the document can clearly be attributed to a single type. The number of sub-categories is unlimited.

Starting from rough distinguishing criteria, it is thus possible to refine the decision step by step. In the terminology of object-oriented programming, one could call the higher hierarchical levels "base classes", from which the lower hierarchical levels "inherit" their distinguishing criteria.

The document type detection is administrated using a graphical user interface which presents the hierarchical structure to the user in an intuitive way. The defined hierarchy of the categories is presented as a tree for this purpose, so that the user can open the folder requiring administration to gain access to the information present at the current hierarchical level (in a similar way in which directory trees are presented in a file system).

At every hierarchical level, the user can delete or modify the definitions or create new branches. The set of rules describing a branch of a tree is created by defining and connecting key terms in the graphical user interface. In addition to this, search areas can be marked on an example document in order to distinguish document types containing the same key text, but in different positions. The set of rules and example documents are saved permanently in the programs data base, so that they can be accessed at any time and modified if necessary.

Besides manually defining rules, PROSAR-AIDA can automatically learn relevant key words and generate appropriate rules in its auto-learning mode. All that is required is a batch of pre-classified example documents. The definitions created by PROSAR-AIDA can be understood by humans and be analysed, extended and changed if necessary by trained personnel. Should additional document classes require subsequent administration, individual nodes or branches of a rule tree can be trained without negatively influencing the other definitions.

Data extraction

In order to find the relevant index data automatically, regardless of its position on the document, it must be possible to give a sufficiently precise definition for the desired data. PROSAR-AIDA provides various mechanisms for this purpose, including the search for data items with a known format (e. g. amounts), geometric and logical relationships of typical key words (e. g. key word "VAT" and "Total" in the same line), or the search for content from existing database tables (e. g. checking whether a client's address exists on the document). All search methods can be combined with each other. This, for instance, makes it possible to filter out the total amount from amongst the many other amounts which may be on the document.

By using such definitions, PROSAR-AIDA checks all possible text areas with the help of the defined extraction rule and delivers all text areas found in the document which fulfil the checking criteria. In this way, data such as contract numbers, bank sort codes and account numbers, invoice dates, sums etc. can be filtered out and used for automatic indexing or for workflow routing purposes.

Should the document type detector recognize an actual form, all the features of a standard forms processing system are also available. Thus, processing all kinds of documents (freeform, formatted documents and forms) is possible in a single system without first having to sort the documents.

Data extraction administration is entirely carried out in the graphical user interface. No programming skills are required.

Add-on module "Table Analysis"

PROSAR-AIDA can be expanded using the table-analysis module. This module enables table-oriented structures to be read in a document. Tables are recognised on the basis of column headings or the structure of the columns' content. The exact layout of the table can vary. Different column widths and column orders are automatically taken into account by the system. The module even reads tables with entries covering several lines, as well as tables in which different entries are contained in the same column.

Areas of use for this module are in the complete processing of table-oriented documents (such as invoices and delivery notes) or tabular lists such as tables in hospital or medical invoices.

Add-on module "Image Processing"

PROSAR-AIDA can also be expanded using the "intelligent image processing" module. With this module, graphic qualities on the page can be used as document type detection criteria. In particular, this module enables the recognition of "shine through" back pages, which can then be classified as blank pages by PROSAR-AIDA, thereby acting as an intelligent back page filter.

Performance

PROSAR-AIDA is specially designed for high performance in environments with complex data requirements and high document volumes. Even with several hundred different document types and extensive data extraction (including tables), the processing times are typically below 2 seconds per

document using standard PC hardware. This allows volumes between 2,000 and 3,000 images per hour to be processed on a single computer. In addition, PROSAR-AIDA supports operation on SMP architectures and multi-server operation, so that using a cluster of recognition Pcs will result in appropriately higher volumes.

Installations with volumes of between 1,000 and 200,000 images per day are in running operation and are available for reference visits. Some installations achieve throughputs of up to 25,000 images per hour.

Availability

PROSAR-AIDA is available for the operating systems OS/2 and Windows NT / Windows 2000 (other operating systems on demand). Under Windows, PROSAR-AIDA is also available as specially adapted module (custom module) for integration into the AscentCapture (Kofax) and InputAccel (ActionPoint) capturing systems.

©1999, 2002 Parodatec GmbH. Information in this leaflet is subject to change without notice. Technical data, prices and delivery times stated in this report are non-binding information and are not grounds for a claim. Parodatec is only contractually bound by a direct written offer. All trademarks acknowledged. "Parodatec" and "PROSAR" are registered trademarks belonging to Parodatec GmbH.