International Journal of Information Management 40 (2018) 67–75

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

Review

Future paradigms of automated processing of business documents

Matteo Cristani[a,*], Andrea Bertolaso[b], Simone Scannapieco[c], Claudio Tomazzoli[a]

[a] Department of Computer Science, University of Verona, Strada Le Grazie 15, Verona, Italy
[b] EVS s.r.l., Strada Le Grazie 15, Verona, Italy
[c] Research Department, Real T s.r.l., Viale Venezia 7, Verona, Italy

ARTICLE INFO

Keywords:
Automated document processing
Business documents
OCR
Statistical natural processing

ABSTRACT

In this paper we summarize the results obtained so far in the communities interested in the development of automated processing techniques as applied to business documents, and devise a few evolutions that are demanded by the current stage of either those techniques by themselves or by collateral sector advancements. It emerges a clear picture of a field that has put an enormous effort in solving problems that changed a lot during the last 30 years, and is now rapidly evolving to incorporate document processing into workflow management systems on one side and to include features derived by the introduction of cloud computing technologies on the other side. We propose an architectural schema for business document processing that comes from the two above evolution lines.

1. Introduction

Business documents are a large class of documents not specifically determined in the current literature. For a reference, we consider the official document classifications introduced in different countries, or proposed standards, typically for the purpose of electronic processing of these documents as advanced by Kabak and Dogac (2010). There are several aspects of business documents that distinguish them from the generality of the reference class. In particular, three features are unique of business documents:

1. Business documents mandatorily contain predefined *sets of data*. For instance, invoices contain name of the issuing company, amount, taxes.
2. Business documents are schematically organised in *predefined layouts*, in particular certain areas of the page are reserved for logos, in general, and other sections are reserved for title lines.
3. The structure of business documents associates sections to *specific keywords*. For instance, the word carrier is always present in the part of a waybill where the name of the transport company is provided.

Generally speaking, and as referred to also in the field of document processing on the web, we often find the term *semi-structured* documents for those satisfying either Point 1 or 2. We can also find the commonly used term *structured* as referred to documents satisfying both Points 1 and 2. Therefore, we can summarize the above observations by stating that business documents might be either semi-structured or structured, but they can be always classified also by means of keywords. For the generality of keywords to be used, as we shall see in the rest of this paper, there are several concrete proposals in the literature, and these lie on distinctive aspects:

- Language used in issuing the document;
- Issuing country, for legal reasons;
- Product category, as in many conditions, specific categories require specific actions.

Unstructured documents have been dealt with in a number of investigations. However, though many business-related documents such as contracts, notarial deeds or commercial letters/emails that are not structured, are relevant to the process of de-materialization and electronic archiving, not very often they have indeed been processed with the explicit purpose of feeding a database table. This aspect is practically distinctive, and it is also the reason for which it makes sense to develop technologies for business documents processing that are not generic of business-related documents.

For the sake of limitedness of the horizon of the investigation, in this paper we refer explicitly to business documents, namely those documents that result from the interaction of two main aspects:

- They have a legal value;
- They need to be used to fill in a database table.

* Corresponding author.
E-mail addresses: matteo.cristani@univr.it (M. Cristani), andrea.bertolaso@embeddedvisionsystems.it (A. Bertolaso), scannapieco@realt.it (S. Scannapieco), tomazzoli@univr.it (C. Tomazzoli).

https://doi.org/10.1016/j.ijinfomgt.2018.01.010
Received 3 December 2017; Accepted 24 January 2018

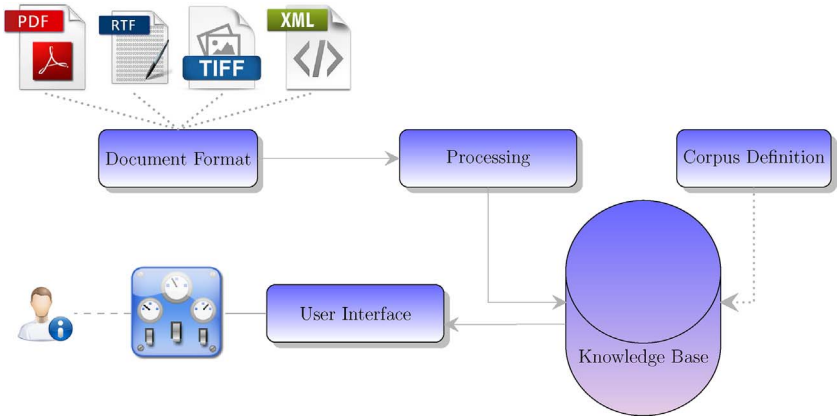

Fig. 1. Processing documents: the common schematic approach of many real systems.

The majority of document processing technologies that are focussing on these issues, have been mainly dealing with invoices.

In the history of document processing, we find three aspects that concern the processing itself that have been dealt with:

- The *recognition of characters* (OCR, Optical Character Recognition) or the detection of specific areas containing characters (OMR, Optical Mark Recognition), specifically designed to perform the analysis of marks given to students in multiple-choice automated test correction, an issue that became quite relevant when mass university phenomenon took place.
- The *identification of layout*, that undergoes the process of *document classification* and potentially the process of *information extraction*.
- The *keyword detection* and the subsequent process of *bag of words* definition with the again further process of document classification and information extraction as in the above mentioned phase.

The general schema of document interpretation, for instance the one used in some real systems, such as GATE, devised and discussed by Bontcheva, Tablan, Maynard, and Cunningham (2004), is presented in Fig. 1.

In this paper we survey the researches that investigated the problems arising when processing business documents, and try to devise a development of the future technologies that are going to be developed henceforth.

Specifically, Section 2 discusses general approaches to document processing and related work. In particular, Section 2.1 is devoted to the presentation of some systems and the approaches used for their engineering, while Section 2.2 discusses in details some general approaches. Section 3 introduces the techniques and technologies built for processing documents in form of images, the original goal of document processing, and Section 4 goes to the core problem of business document processing, related to the paradigmatic case of invoices. Section 5 summarises the historical approaches and how they evolved in the past, and Section 6 discusses the evolution we envision henceforth. Section 7 ends the paper by drawing some conclusions.

2. A survey of methods and technologies for business document processing

First of all, in order to better understand the focus of our investigation, let us provide a definition of *document processing technology*.

Document processing is a process that consists in a sequence of steps, not necessarily to be performed completely:

- *Transforms the document from a material one into a digital version*;
- *Establishes the classes to which the document belongs and therefore associates the document to its features*;
- *Finds the structure of the document and splits it into pieces where every piece is associated to a specific content*;
- *Identifies the document contents as related to a specific set of tables in a database*;
- *Transduces the specific contents found in the* **FEATURE SCAN** *step into the database tables designed to be the target.*

Analysis of documents is, in fact, an ubiquitous theme in several investigations related to the web and other domains of interest of business documents. In this section we survey a few relevant works related to development of the field of document processing. Some of these investigations, for instance those of Cesarini, Francesconi, Gori, and Soda (2003), describe technologies that operate specifically on invoices or other similar business documents. Other works, including the effort by Medvet, Bartoli, and Davanzo (2011) or the development proposed by Tuganbaev, Pakhchanian, and Deryagin (2005) can be considered generic, in the sense that they are thought for a wider range of document classes but have been tested, or can be applied, also for automatic invoice/business document processing. Some works, such as that of Altamura, Esposito, and Malerba (2001), have been explicitely thought for other document classes but apply even for the specific purpose of business document processings.

A field that has often been considered contiguous to document processing is *document classification* that has developed a lot in the recent past, especially close to the specific problem of multimodality, namely documents that include images and texts, and are classified based upon these aspects by Cristani and Tomazzoli (2014, 2016).

Dengel (2003) gathers some preview approaches to document analysis and cluster them based on the number of classes (single or multiple) and the type of documents (structured, semi-structured, unstructured) they process.

Some further studies, especially related to techniques folded in the analysis of very specific classes of documents, investigated specific information related to taxes and postal addresses by Cristani and Gabrielli (2010), and by Cristani and Gugole (2008) where no structure actually exists, but content constraints are expressed explicitly.

The rest of this section is organised as devised below. Section 2.1 discusses some real systems as they had been described in the current literature about document processing, and analyses the approaches employed in these applications. This section is the basis of the methodological analysis provided in Section 5, where a schema is provided that is in turn referred to the real systems that have taken market in the past.

Section 2.2 generalizes the study scope and provides a picture of general approaches to document processing, in terms of the view that has been proposed by the scholars of this field. Again, this inspires the novelties proposed in Section 6.