Chapter VII
Discovering Spatio–Textual Association Rules in Document Images

Donato Malerba
Università degli Studi di Bari, Italy

Margherita Berardi
Università degli Studi di Bari, Italy

Michelangelo Ceci
Università degli Studi di Bari, Italy

ABSTRACT

This chapter introduces a data mining method for the discovery of association rules from images of scanned paper documents. It argues that a document image is a multi-modal unit of analysis whose semantics is deduced from a combination of both the textual content and the layout structure and the logical structure. Therefore, it proposes a method where both the spatial information derived from a complex document image analysis process (layout analysis), and the information extracted from the logical structure of the document (document image classification and understanding) and the textual information extracted by means of an OCR, are simultaneously considered to generate interesting patterns. The proposed method is based on an inductive logic programming approach, which is argued to be the most appropriate to analyze data available in more than one modality. It contributes to show a possible evolution of the unimodal knowledge discovery scheme, according to which different types of data describing the units of analysis are dealt with through the application of some preprocessing technique that transform them into a single double entry tabular data.

Copyright © 2008, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.
INTRODUCTION

Business processes have always been based on the flow of documents around an organization. The concept of flow is almost synonymous with the concept of paper flow in typical office environments, where the main problem is the amazing number of printed documents that are generated and filed. In fact, much time and effort is wasted in ineffective searches through multiple information sources. Organizations need to extend the scope of Business Intelligence especially to their internal collections of textual data in order to make decisions on the basis of knowledge captured by these collections. Therefore new document management systems with abilities to catalog and automatically organize these documents are necessary. Personal document processing systems that can provide functional capabilities like classifying, storing, retrieving, and reproducing documents, as well as extracting, browsing, retrieving and synthesizing information from a variety of documents are in continual demand (Fan, Sheng, & Ng, 1999). However, they generally operate on electronic documents (e.g., text, word, rtf, pdf, html, and xml files) and not on the more common paper documents, which are made anyway computationally processable through digital scanning.

The pressing need for systems to be used as intelligent interfaces between paper and electronic media has led to the development of a large number of techniques for document image analysis and recognition. The conversion of document images into a symbolic form appropriate for subsequent modification, storage, retrieval, reuse, and transmission is a complex process articulated into several stages. Initially, the document image is preprocessed, for instance, to remove noise. Then it is decomposed into several constituent items, which represent coherent components of the document layout (e.g., text lines or half-tone images). Finally, logically relevant layout components (e.g., title and abstract) are recognized.

Domain-specific knowledge appears essential for document image analysis and understanding: in the literature, there are no examples of attempts to develop document analysis systems that can interpret arbitrary documents (Nagy, 2000). In many applications presented in the literature, a great effort is made to hand-code the necessary knowledge according to some formalism, such as block grammars (Nagy, Seth, & Stoddard, 1992), geometric trees (Dengel, Bleisinger, Hoch, Fein, & Hönes, 1992), and frames (Wenzel & Maus, 2001). However, hand-coding domain knowledge is time-consuming and limits the application of document analysis systems to predefined classes of documents.

To alleviate the burden in developing and customizing document analysis systems, data mining methods can be profitably applied to extract the required domain-specific knowledge. Document image mining denotes the synergy of data mining and document analysis system technology to aid in the analysis and understanding of large collections of document images. It is an interdisciplinary endeavor that draws upon expertise in image processing, data mining, machine learning, and artificial intelligence. The fundamental challenge in document image mining is to determine how low-level, pixel representation contained in a raw image of a scanned document can be efficiently and effectively processed to identify high-level spatial objects and relationships. Since the beginning of the 1990’s, when the first attempts in applying machine learning techniques to document images were reported in the literature (Esposito, Malerba, & Semeraro, 1990), there has been a growing research focus on document image mining (Aiello, Monz, Todoran, & Worring, 2002; Akindele & Belaïd, 1995; Berardi, Ceci, & Malerba, 2003; Cesarini, Francescani, Gori, Marinai, Sheng, & Soda, 1997; Dengel, 1993; Dengel & Dubiel, 1995; Esposito, Malerba & Semeraro, 1994; Kise, Yajima, Babaguchi, Fukunaga, 1993; Walischewski, 1997).