Header Metadata Extraction from Semi-structured Documents Using Template Matching*

Zewu Huang, Hai Jin, Pingpeng Yuan, and Zongfen Han

Cluster and Grid Computing Lab
Huazhong University of Science and Technology, Wuhan, 430074, China
hjin@hust.edu.cn

Abstract. With the recent proliferation of documents, automatic metadata extraction from document becomes an important task. In this paper, we propose a novel template matching based method for header metadata extraction form semi-structured documents stored in PDF. In our approach, templates are defined, and the document is considered as strings with format. Templates are used to guide *finite state automaton* (FSA) to extract header metadata of papers. The testing results indicate that our approach can effectively extract metadata, without any training cost and available to some special situation. This approach can effectively assist the automatic index creation in lots of fields such as digital libraries, information retrieval, and data mining.

1 Introduction

Compared with hard copy documents, it is easy to keep and spread digital scientific documents, thus the digital scientific documents are widely accepted as the main document type in libraries and other places. Whether documents are digital or hard-copy ones, it is important to extract metadata of documents since it is impossible to retrieve documents without metadata of documents. Metadata is data about data, used to describe other information based on some rules or policies. Metadata of documents is used in many document processing fields such as search, browsing, and filtering [1].

Providing metadata is the responsibility of each data provider with the quality of the metadata. Many data providers [2] have had significant harvesting problems with XML syntax and encoding issues, even leading to unavailability of service [3]. Therefore, how to automatically extract metadata from documents turns out to be an indispensable research issue. However, because format of digital scientific documents varies according to publication type, e.g., books, journals, conference papers, research reports, and technical reports, it is difficult to automatically extract metadata. Therefore, it is very important to design tools to do that.

In this paper, we take a template matching based approach to extract metadata, including title, author(s), affiliation(s), abstract, and keywords, from semi-structured

* This paper is supported by the National 973 Key Basic Research Program under grant No.2003CB317003, and the Cultivation Fund of the Key Scientific and Technical Innovation Project, Ministry of Education of China under grant No.705034.

R. Meersman, Z. Tari, P. Herrero et al. (Eds.): OTM Workshops 2006, LNCS 4278, pp. 1776 – 1785, 2006.

scientific literature. Experimental results discussed below indicate that our approach is efficient for header metadata extraction from semi-structured documents. Our approach can be effectively used in a variety of situations such as: (1) automatic creation of indices for digital libraries, (2) conversion of documents to semantically richer representation, (3) metadata mining and, (4) the search engines base on metadata.

The rest of the paper is organized as follows. In section 2, we describe our method of header metadata extraction. Section 3 gives our experimental results. In section 4, we introduce related work. We make concluding remarks in section 5.

2 Header Metadata Extraction Method

2.1 Document Outline

In many circumstances, it is fundamental to disaggregate a paper into its basic components [4]. We propose a layout information based approach to document disaggregation. Our approach is based on exploiting the layout information while reading a document. Some layout information is listed as follows:

- The section properties, such as number of columns, section break type, position of page number.
- The paragraph properties, such as flush left, flush right, flush centre, left indent, right indent, and first indent in the first line.
- The font properties, such as size, style, color, bold, underline, Italic, and xy position information.
- The document properties such as page width, page height, left margin, right margin, bottom margin, top margin, starting page number.

In a PDF file, the typical word boundary based structure of a document is broken down into fragments [5]. However, it remains all spatial and font-related data contained in the input PDF file. The styles of documents vary greatly. There are various scientific paper layouts. Fig. 1 shows four typical layouts.

In general, within a document category, a certain visual layout can be identified for all documents within that category. Concretely, the layout information for each metadata is different. The characteristics of header metadata are listed as follow:

Title. Location is always on the upper portion of the first page; Font size is always the largest. Position is always in the middle of the line and flush centre. Font style is always bold. Section break symbol among metadata is always “enter”.

Authors. Location is always immediately under the title; Font size is always smaller than title font. Font is always the same for all authors. Break symbols are always “ ”, “,” or “and”. Authors may or may not be listed in separated lines and flush centre. Author(s) and affiliation(s) may be listed on one column or more. Authors and affiliations may or may not always have superscripts.

Affiliations. Location is always immediately under the authors' list and before abstract. Characteristic words are always “university”, “department”, “@”, “{”, “}”, etc. Font is always the same for all affiliations. Only one affiliation appears when all authors are associated with it, or affiliations mapping to authors are listed separately.