现代图书情报技术 2007 (2): 18-23 ISSN: 1003-3513 CN: 11-2856/G2

数字图书馆

基于XSLT的PDF论文元数据的优化抽取

陈俊林[1];张文德[1,2]

福州大学图书馆[1]

收稿日期 2006-11-10 修回日期 2006-11-30 网络版发布日期 2007-3-2 接受日期

摘要 简述PDF信息抽取过程中采用的转换工具及抽取语言，简析PDFTOHTML格式转换后的中间文档，分析PDF科技论文首页元数据存在的问题，给出对以上问题的解决方案。

Abstract This paper firstly introduces a format transforming tool and XSLT which is the language used to produce extraction rules, then simply analyses the middle documents generated from PDF to HTML. Thirdly, discusses the problem of metadata existed in the science documents in PDF format, finally gives the methods to solve this problem.

关键词 PDF PDF to HTML XSLT 元数据

Key words PDF;PDF to HTML;XSLT;Metadata

分类号 TP311.13

DOI:

通讯作者:
陈俊林 bluesea_cc@163.com
作者个人主页: 陈俊林 张文德