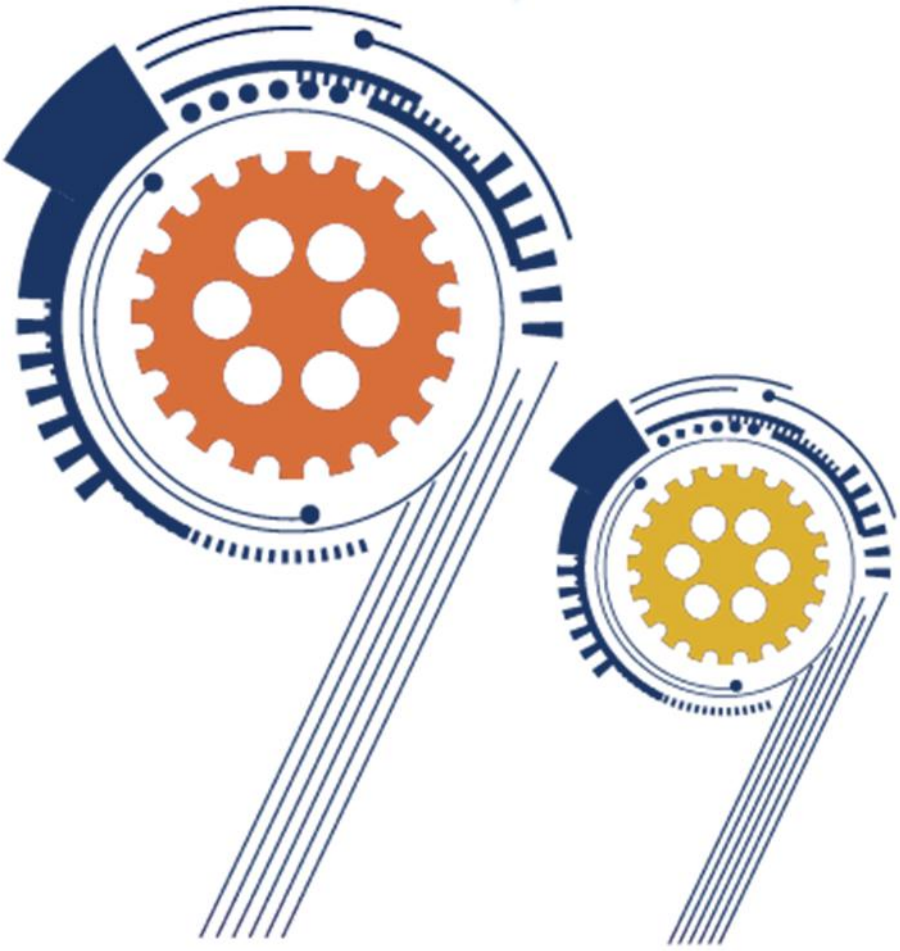
报纸扫描标引加工

Newspaper scanning indexing processing

HUI DU TECHNOLOGY 慧度



慧度99CMS | 北京慧舟普度科技有限公司

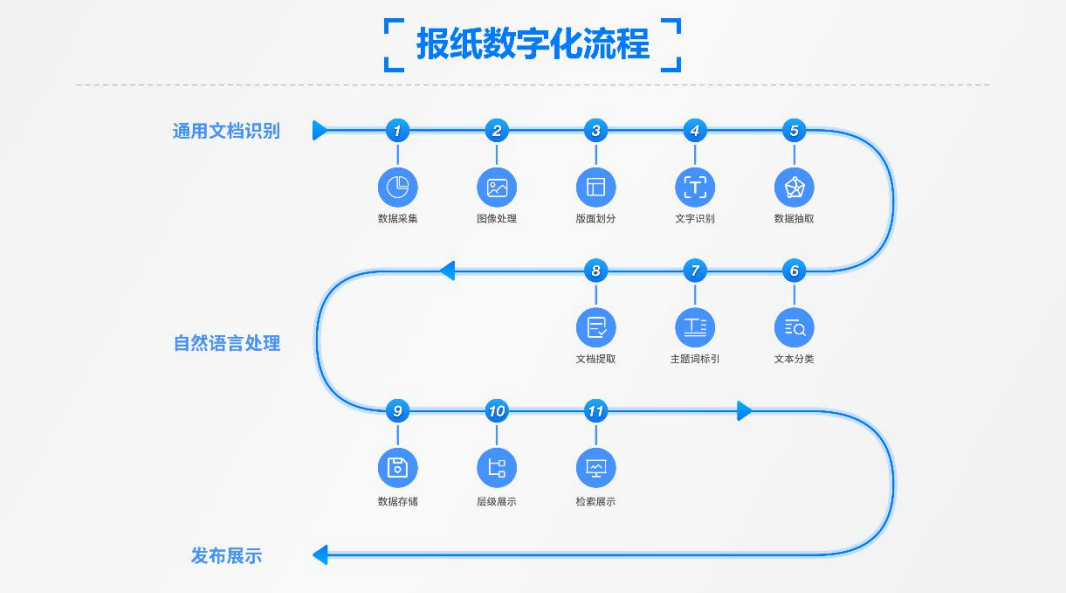
Beijing Hui Zhou Pu Du technology Co.Ltd

一、产品介绍

报纸扫描标引加工平台

报纸扫描标引加工平台：对老旧报纸进行数字化的制作，针对重点又是难点的老旧报纸数字化工作，提供了从报纸扫描、数据清洗、OCR 识别，PDF 反解、质检入库到数字化发布的一整套完整的报纸数字化解决方案。

报纸扫描标引加工平台是一套科技含量极高的软件系统，由扫描生产流程子系统和报纸发布子系统组成，包括中文 OCR(文字识别)技术、电子版面恢复技术、数字化生产流程控制等若干核心技术。该系统的基本原理是通过扫描设备，将纸介质的报纸扫描成数字图像，再经过图像处理、版面分析、文字识别、文字校对、版面重构、标引加工、文档精细加工等一系列步骤，最终形成可以方便应用的精美的电子文档，在此基础上，对这些电子文档进行数字化发布，形成可以供全网使用的数字资源。

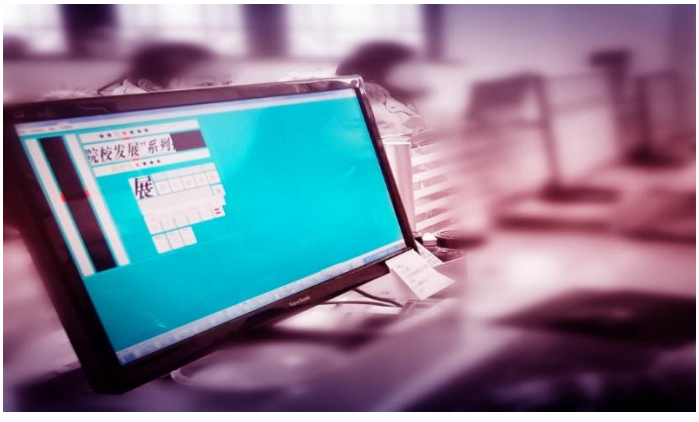


1.扫描和修图

利用大幅面扫描仪扫描纸质报刊，用胶片扫描仪扫描缩微胶片。对扫描得到的TIF 图片进行修正，去除污渍、裂纹等。

2. OCR 文字识别与校对

OCR(光学字符识别)是一种通过计算机自动识别图片上文字的技术，标准印刷汉字的 OCR 识别正确率可达到 99%以上。早期报纸印刷技术简单识别率可能稍低一些，需要进行多次校对以保证最终的质量。校对包括人工校对和智能化自动校对。



3.PDF 还原、输出

输出的 PDF 让其图像层和文字层的文字定位准确，保证反显区域与文字区域相差 1/3 字符以内。软件可以通过对字、行、块来调整。

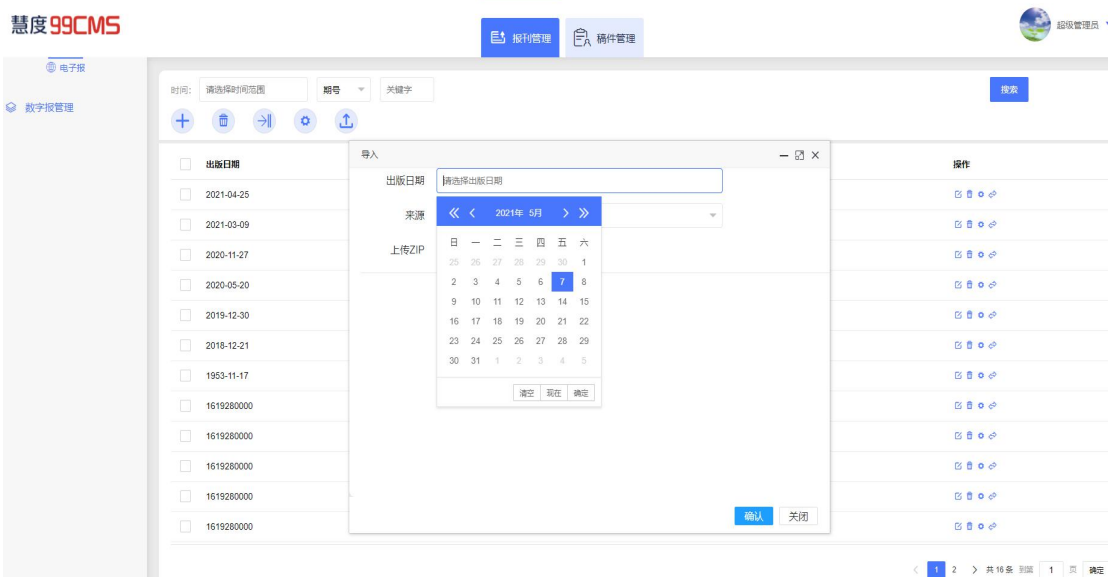

新清华
XIN QINGHUA
1999 年 1 月 22 日 第 1359 期 共 4 版
CIMS 中心完成重大关键技术
支持我国企业快速高效实现自动化
李衍达院士主持研制出极具创新意义的成果
多媒体数据广播系统今日开播
简讯

4.元数据抽取和标引

文章的元数据抽取包括对文章主题、副题、引题、作者、来源、关键词、摘要、引文、外部特征等信息的自动识别和自动抽取。

5.xml 文件解析入库

支持对扫描生成的 xml 文件解析入库。支持自动、批量入库支持后台手动导入



6.自然语义数据智能著录技术

每篇文章在线展示：文章摘要、报道对象、人物、地点、关键词、体裁等文章核心标签。文章入库时通过人工智能数据标注技术，实现 30+的标签智能化标注。

7.数字报多端发布

原版原样看报纸

网站+APP+微信小程序+语音读报

提供好的产品和服务
让生活更美好

Provide good products and services

Make life better

北京慧舟普度科技有限公司

总部地址：北京市昌平区回龙观西大街龙冠商务中心银座3层D310

联系电话：010-57466093　　服务QQ：3172548446

E-mail：service@99cms.cn

扫一扫，关注慧度科技