基于图像识别的文献数据挖掘，主要用于获取文献中的文字、插图、表格、及语料库等。技术路线阐述如下：

1. 文献预处理：文献预处理是指将不同文献格式（如pdf、caj等）统一转换成png图像格式，再对图像进行一系列矫正处理。首先把文献（数字文档或扫描文档）中的每一页以图像格式提取，对于不同的文献格式有不同的提取方式，如pdf格式可以使用python的开源包pdf2image提取；caj格式可以使用python的开源包caj2pdf转换成pdf格式再用pdf2image提取。接下来对于提取出的每页图像，使用python的开源包OpenCV进行图像平滑与去噪处理，还可以选择继续进行灰度化、二值化处理（既将图像上的每一个像素转化为灰度等级或0/1黑白状态，对于一些扫描文档来说灰度化、二值化可以提升后续识别效果，但大多数字文档不用执行这步操作）。
2. 布局检测：布局检测是指对图像中的每个标题区域、文字区域、插图区域或表格区域进行定位，以边界框的形式将这些区域分别标记出来。Layout-parser是一个效果出色且开源的文献布局检测包，它基于detectron2这一知名的计算机视觉框架，用深度学习的方法对文献布局进行精准定位，下图为一个基于layout-parser布局检测的案例展示（该案例中仅定位了文字区域）。

图片包含 文本

描述已自动生成

1. 文本识别：文本识别是指把布局识别中提取的每个标题区域与文字区域中的文本识别出来。目前有不少开源的、基于Python的文本识别检测包或模块。其中惠普公司与Google公司共同开发并开源的tesseract OCR为其中最为知名的之一，它支持包括简体中文、繁体中文、英文等在内的众多语言，且拥有14种页面切分模式方便处理不同的文本排列模式（如横排竖排等，但本项目仅需使用它的默认模式，即自动模式）。需说明的一点是，虽然tesseract OCR支持的语言种类繁多，但它相比起中文识别更加擅长处理英文识别，故对于中文文献更推荐使用百度公司开发并开源的PaddleOCR，该模型有着轻量级（仅8.1M）且侧重中文识别的优势与特点。
2. 表格识别：表格识别是指把布局识别中提取的每个表格区域中的内容提取出来。可以使用python的开源包img2table来识别表格并保存为xlsx格式或json格式，需要注意的是，img2table是一个基于OCR的工具，且需要自定义使用的OCR，故依旧建议使用tesseract OCR识别英文表格，PaddleOCR识别中文表格。
3. 语料库构建（可选）：语料库构建是指对文本识别中提取出的文字信息构建向量数据库，把文本映射至高维空间，以便日后进行相似性搜索、聚类、自然语言处理等操作。Facebook AI研究院开发的Faiss是进行语料库构建的通用方案，它基于深度学习框架，构建向量库后可以用L2欧式距离、向量内积等方式轻松高效搜索相似文本，且支持GPU加速。