TechOcr程序进展情况汇报

近两周来，通过图像处理来提高识别率的尝试，进展缓慢；目前转变思路，从模式识别入手，尝试提升OCR过程的识别率。

**一．Tesseract OCR的模式识别工具**

比较幸运的是，Tesseract提供了为OCR模式识别所需要的全部工具。如针对样本识别的修改，通过设定字频、词频的调整等。但需要提供大量的原始素材，通过工具进行样本训练。

这样，对模式识别的改进，就只需要进行样本训练，也就是原始素材的提供，降低了技术难度。

目前已经按照训练流程，完成了测试，针对特定文字的小范围训练样本，可以达到完全正确。

**二．需要提供的图像样本**

汉字OCR最大的困难，就是字符数量巨大，在GBK编码中，使用了23940个码位中的21003个。Tesseract提供的工具，可以通过人工界面的干预，调整程序对边框、轮廓的判定。但手工干预如此多的汉字，不太现实。

通过对.box样本矫正文件的结构分析，发现针对打印体文字的样本，可以通过自行编程，输出同样格式的.box文件，而无需人工界面干预。原来的工具，可作为一种校验工具来使用。

这样，实际的工作就成了编写一个提取汉字点阵、标定汉字轮廓、标定识别结果的程序。这一程序难度不大，但工作量巨大。

完成样本矫正后，余下的工作，就是批处理运行脚本。因需要对不同点阵数、不同字体、不同属性（加粗、斜体）的矫正文件进行处理。还需要编写一个脚本生成程序。

**三．需要提供的语义样本**

是否提供语义样本是可选的，但明显，提供语义样本能够显著提高识别率。

对OCR程序而言，语义样本主要为字频、词频素材。比较现实的想法，是从输入法程序中，抽取词库文件。再编写一个格式转换程序来完成。

此项抽取工作，目前尚未开展。