java认证辅导:java中对this的理解Java认证考试 PDF转换可能丢失图片或格式，建议阅读原文 https://www.100test.com/kao_ti2020/644/2021_2022_java_E8_AE_A4_E8_AF_81_c104_644616.htm Web-Harvest是一个Java开源Web数据抽取工具。它能够收集指定的Web页面并从这些页面中提取有用的数据。其实现原理是，根据预先定义的配置文件用httpclient获取页面的全部内容（关于httpclient的内容，本博有些文章已介绍），然后运用XPath、XQuery、正则表达式等这些技术来实现对text/xml的内容筛选操作，选取精确的数据。前两年比较火的垂直搜索（比如：酷讯等）也是采用类似的原理实现的。Web-Harvest应用，关键就是理解和定义配置文件，其他的就是考虑怎么处理数据的Java代码。当然在爬虫开始前，也可以把Java变量填充到配置文件中，实现动态的配置。 （友情提示：本博文章欢迎转载，但请注明出处：陈新汉，http://www.blogjava.net/hankchen） 现在以爬取天涯论坛的所有版面信息为例，介绍Web-Harvest的用法，特别是其配置文件。 天涯的版块地图页面时：http://www.tianya.cn/bbs/index.shtml [天涯的部分版面列表] 我们的目标就是要抓取全部的版块信息，包括版块之间的父子关系。 先查看版块地图的页面源代码，寻求规律： 社会民生 天涯杂谈 国际观察 天涯时空 传媒江湖 …… //省略 文学读书 莲蓬鬼话 煮酒论史 舞文弄墨 ……. //省略 ……. //省略 通过页面源码分析，发现每个大板块都是在的包括之下，而大板块下面的小版块都是下面的形式包含的。 xxx，这些规律就是webharvest爬数据的规则。 下面先给出全部的配置：(tianya.xml) ]] 100Test 下载频道开通，各类考试题目直接下