关于Word文档到HTML的转化

Contents

[关于Word文档到HTML的转化 1](#_Toc421268208)

[1. 使用到的类库: 1](#_Toc421268209)

[2. Word文档的格式: 1](#_Toc421268210)

[3. 转化过程: 1](#_Toc421268211)

## 使用到的类库:

POI(<http://poi.apache.org/>):

Docx4j(<http://www.docx4java.org/trac/docx4j>):

## Word文档的格式:

Office 97-2003 基本上是微软的二进制格式.

Office 2007 ([OOXML](http://baike.baidu.com/view/1080868.htm)) 是一个xml格式压缩之后的格式.

## 转化过程:

首先确定一个InputStream中的文档格式. 调用

com.cyznj.common.utils.WordToHtmlUtils.getDocInfo(InputStream) 方法

|  |
| --- |
| **public** **static** **enum** Version {  ***OFFICE\_97\_2003***, ***OFFICE\_2007***;  }  **public** **static** **class** DocInfo {  **public** Version version; //版本  **public** String appName; //格式(Office Word, 或者Office Excel, 及其他)  @Override  **public** String toString() {  **return** "DocInfo [version=" + version + ", appName=" + appName + "]";  }  } |

根据文档版本,调用POI的org.apache.poi.hwpf.converter.WordToHtmlConverter类转化***OFFICE\_97\_2003*** 格式.

WordToHtmlConverter采用composite模式将文档分割成多个section, section再分成paragraph. Section用Div标签转化,paragraph用p标签转化. 但是页头页脚没有转化. 目录链接被转化成#链接. 如果需要定制转化过程,可以继承WordToHtmlConverter重写某些转化方法.

调用Docx4j的org.docx4j.Docx4J.toHTML(**HTMLSettings**, OutputStream, int)方法转化***OFFICE\_2007*** 文档. org.docx4j.convert.out.HTMLSettings 类包含了转化html的某些配置 (例如设置值定义的css和javascript,页头,页脚等.