

# 基于XML的HTML自动阅卷算法设计与实现

丁俊辉, 龚沛曾

(同济大学电子与信息学院, 上海 201804)

**摘要:**自动阅卷技术在当今计算机基础教育中起着重要作用。当前主流的基于文档对象模型的HTML网页自动阅卷方法存在准确性低及在大数据量时阅卷效率差的问题。文中介绍了一种基于XML的HTML网页自动阅卷算法,它根据XML与HTML格式上的相似性将HTML文件转换成XML文件,然后利用各种程序语言对XML良好的支持来进行HTML网页自动阅卷过程中的信息处理。该算法不仅可以避免传统人工阅卷的低效率及结果的主观性,而且在准确率及稳定性方面比文档对象模型方法有很大提高,为HTML网页制作考核提供了一种有效可行的方法。

**关键词:**可扩展标记语言;超文本标记语言;自动阅卷

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2012)01-0001-04

## Design and Implementation of Auto-Mark HTML Exam Paper Algorithm Based on XML

DING Jun-hui, GONG Pei-zeng

(School of Electronics and Information, Tongji University, Shanghai 201804, China)

**Abstract:** Automatic marking technology plays an important role in today's computer elementary education. As a principle method to auto-mark HTML exam paper, the method based on document object model is low efficiency when come across with many papers. It introduces an automatically marking algorithm for auto marking the HTML files based on XML. It converts HTML files to XML files based on their formats similarity, and then get information from the XML file easily to data process with the support by several program languages. This algorithm will improve the efficiency and accuracy of the automatic marking process. And make the automatic marking process feasible and effective.

**Key words:** XML; HTML; auto marking

### 0 引言

随着互联网的普及,大家对上网浏览网页已不再陌生。各大院校也把html静态网页的制作当成大学计算机基础教学的一部分。传统情况下,对大批量的学生操作题,阅卷时只能逐个打开试卷进行人工阅卷,分数的主观性大而且阅卷效率低下。随着计算机技术的发展,自动阅卷取代人工阅卷必将成为一种发展趋势。

目前,对自动阅卷方面的研究一直是国内外教学考核的重要研究方向<sup>[1]</sup>。针对客观题的阅卷由于答案比较固定而且容易实现,因此技术比较成熟<sup>[2]</sup>。操作题方面,OFFICE系列操作题的自动阅卷技术也已经比较成熟<sup>[3]</sup>,但针对HTML网页制作的自动阅卷还有

待进一步的提高。对HTML网页的自动阅卷最直接的方法是对静态HTML文件内标签直接分析与匹配的参数配置法,即对于每个所要考查的知识点通过参数配置的形式给出在文档中的位置及属性,分析学生提交的文件,匹配正确则得分。由于HTML的语法规则要求不严格,这种文本分析代码的方法复杂且容错性较差。另一种应用的较多的方法是基于文档对象模型的方法<sup>[4]</sup>,借助Microsoft提供的文档对象模型(即HTML对象组件)来识别提取网页内元素对象属性的方法,该方法能方便、正确地提取页面中的相关信息,但在数据量大的情况下效率比较低且稳定性差。相对而言,文中使用的基于XML的自动阅卷方法具有更好的效率及稳定性。

### 1 HTML与XML

HTML(超文本标记语言)<sup>[5]</sup>是用于创建网页和进行信息发布的通用语言,它以纯文本形式存储,以标签定义文档的组织。HTML文档是由一系列的元素和

收稿日期:2011-05-29;修回日期:2011-09-04

基金项目:教育部“高等理工教育教学改革与实践项目”(110)

作者简介:丁俊辉(1988-),男,河南周口人,硕士研究生,研究方向为模式识别与智能系统;龚沛曾,教授,研究方向为智能技术、计算机辅助教育。

标签组成。其基本的语法结构与 XML 类似,但属于松散的语法结构。XML 是一种可扩展标记语言,它提供描述结构化资料的格式。XML 以其良好的数据存储格式、可扩展性、便于网络传输等优势在许多领域得以广泛应用,具有良好的可靠性与互操作性<sup>[6]</sup>。HTML 与 XML 都是标记语言,相对 HTML 而言,XML 的语法格式更为严谨。在进行信息提取<sup>[7]</sup>的时候,为了高效地进行信息的提取处理,采用将 HTML 文档转换成语法严格规范的 XML 文档,然后利用 XML 开发组件(如微软的 MSXML)来进行信息的提取处理<sup>[8]</sup>。XML 文件必须满足以下基本格式要求<sup>[9]</sup>:

(1) 必须在文件第一行出现小写的 xml 和 version 声明;

(2) 有且只能有一个根节点,并且这个根节点位于所有嵌套标签的最外层;

(3) 标签必须以嵌套式(树状)排列,不同组的标签只能有包含而不能有交错、重叠;

(4) 对于非空的标签必须包含开始标签和结束标签;

(5) 空标签可以单独出现,但结尾必须加上“/”;

(6) 标签名称与属性名称大小写敏感,同时要满足命名规则;

(7) 属性值前后必须被双引号或单引号引起来;一些特殊字符,如“<”,应通过实体表示,如“&lt;”。

## 2 算法概述

本算法思想是:首先根据标准答案来生成答案配置文件。阅卷时先将学生上传 HTML 文档转换成 XML 文档,然后根据答案配置文件从转换后的 XML 文档中提取相应考点的信息,根据所提取信息与答案的匹配程度来判定相关知识点是否得分。HTML 文档自动阅卷流程如图 1 所示,可分为 HTML 文档预处理,配置文件的生成与阅卷三个基本的过程。

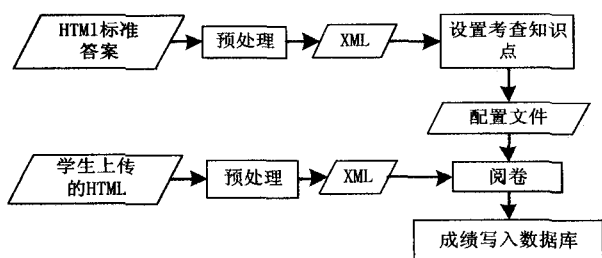


图 1 HTML 文档自动阅卷流程示意图

### 2.1 HTML 文档的预处理

由于 HTML 语法的不严谨,当使用基于文档对象模型的方法提取某一特定信息时可能会出现提取不准确以及耗时间长等问题。为了保证阅卷的准确性,需要在配置答案及阅卷之前均把所给文档进行预处

理,即将 HTML 文档转化为结构严谨的 XML 文档<sup>[10]</sup>。

根据 XML 的语法要求,在预处理的过程中,主要包含对不匹配标签的修复以及对 CSS 部分的转换过程。在转换过程中需满足:不改变原文档的基本结构;不丢失原有数据信息。

预处理主要包括以下过程:

(1) 修复不匹配标签。由于 HTML 语法的不规范,在存在一定错误的情况下,浏览器一般也能显示出页面。而 XML 语法严格要求标签必须成对出现或以空标签形式存在。因此在预处理的过程中必须对 HTML 文档中不匹配的标签进行匹配修复。若不匹配的标签为开始标签,就在相对应的位置添加结束标签;若不匹配的标签为结束标签,就把它直接从文档中删除。

(2) 标签属性的验证。XML 要求标签属性值前后必须被双引号或单引号括起来,而 HTML 文档中存在一些属性值(例如 radioBox 的 checked 属性,在不同编辑器下可能会出现<radio checked />或<radio checked =“true”>)不满足此条件。因此在此部分需要对每个标签的属性进行验证修复,对每个只有属性名而没有属性值的属性设置其值为空。

(3) CSS 部分转化为 XML。由于 CSS 的发展,在当今的网页中起着非常重要的作用。当 CSS 存在于 HTML 文档时,而其形式并不是以标签的形式存在,因此对此部分应按如下例所指定的方式将其转化为 XML。

例: . styleName { font-size: 14px }

=> < styleName ><font-size>14px</font-size></ styleName >

(4) 特殊字符的处理。在 HTML 文档中存在像“&nbsp;”等之类的一些字符串,此类字符串在 XML 中会被解释为实体引用,因此在预处理过程中应将其转化为其它形式或将其直接删除。

### 2.2 生成答案配置文件

为了统一阅卷过程及方便地保存标准答案和考点信息,采用配置文件<sup>[11]</sup>的方式。配置文件在此处为一个由标准答案生成的 XML 文件。配置答案时,首先从标准答案预处理后得到的 XML 文件中提取各项数据显示在界面上,出题老师可以选择是否对某些标签或属性设置知识点。配置文件是答案在系统中的另一种存在形式,它不仅包含标准答案的信息,还包含了 HTML 网页制作题目中所要考查的知识点及分数。一般情况下所要考查知识点均为 HTML 中某一节点的属性或文字。因此生成配置文件方法如下:

(1) 保持原 HTML 的文档结构。对于原 HTML 文档中已有的结构,像 TABLE, FORM 标签等直接在该标签节点上添加“SCORE”属性,其值即为该节点的分

数,如图2中所示。

(2)对于原文档中没有显式存在的知识点,像表格的行数列数等,将行及列的值计算出来之后添加到表格属性部分,当作表格的一个属性,如图2中表格属性部分所示。

(3)需要另外添加的部分。若需要对某一些特定文字格式设置考点,首先在配置文件中添加节点名为“SPECIFIEDTEXT”的节点,然后在标准答案中通过先查找属性后查找样式的方式检索该文字的格式并将其设置为该节点属性,同时设置该节点文字部分为待阅文字。检索时应同时验证文字在文档的唯一性来避免答案的二异性。文本属性查找设置界面如图3中界面所示。

### 2.3 阅 卷

阅卷的过程是对配置文件与学生提交的文件预处理后的XML文件进行信息提取匹配的过程。采用XML组件来进行数据的提取,通过判断查询到的节点是否有SCORE属性来确定其是否为考点。

例:对表单进行阅卷。可使用 `set form = xmldoc. selectSingleNode("//form")` 来提取页面中第一个form节点。判断 `Form. Attributes. getNamedItem("SCORE")` 是否为空来确定是否已将该form节点设

置为考点。

为了方便统计各个知识点的分数,采用各个部分分别阅卷的方式,即对表格、表单、超链接、图片和指定文字等部分分别阅卷。根据在配置文件中设置的考点来进行阅卷。对某一考点,在待阅文档中查询是否存在相关的信息,若存在则根据一定的规则来进行判分。规则分为以下几类:

(1)只需对节点的属性进行判分。在这种情况下该节点除分数属性外,其余属性的分值相等。

例如: `<img src = "title1. jpg" width = "605" height = "62" score = "3" />`

即除 score 属性外,其余三个属性分值均为一分。

(2)需对节点的属性及文字进行判分。有时表单项有一些说明性的文字,希望在阅卷时把这些文字也作为一项考点,而在网页制作时,这些文字并不一定是紧邻它所说明的表单项。在阅卷时采用文字是否包含在此表单的上一级节点中来判定其是否存在。

(3)对试卷中容易出现的一些部分正确的考点,按一定比例给分。主要包括:若要求在页面中只有一个form,若页面有一个以上form标签,则本部分得分为部分的80%;超链接部分,超链接的文字应占到总分数的1/2,其它属性分数相等。

| 表格   | 表单      | 超链接 | 列表                                     | 页面属性old             | 滚动字幕 | 预览                                     | 文本属性 |
|--|---------|-----|--|---------------------|------|--|------|
| <b>表格属性</b>                                |         |     |  |                     |      |  |      |
| 表格总分数为: 0                                  |         |     |  |                     |      |  |      |
| <input checked="" type="checkbox"/> 本知识点分数 | 0       |     |  |                     |      |  |      |
| <input checked="" type="checkbox"/> 宽      | 893     |     |  |                     |      |  |      |
| <input checked="" type="checkbox"/> 高      | 464     |     |  |                     |      |  |      |
| <input checked="" type="checkbox"/> 边框     | 1       |     |  |                     |      |  |      |
| <input checked="" type="checkbox"/> 列数     | 3       |     |  |                     |      |  |      |
| <input checked="" type="checkbox"/> 行数     | 3       |     |  |                     |      |  |      |
| <b>表格内容</b>                                |         |     |  |                     |      |  |      |
| 第1行1列                                      |         |     |  |                     |      |  |      |
| <input checked="" type="checkbox"/> 本知识点分数 | 0       |     | <input checked="" type="checkbox"/> 高  | 107                 |      | <input checked="" type="checkbox"/> 跨列 | 3    |
| <input checked="" type="checkbox"/> 背景色    | #CCCC00 |     | <input checked="" type="checkbox"/> 文字 | 低碳生活网 2011年1月6日 星期六 |      |  |      |

图2 配置答案界面一:表格知识点设置

| 表格  | 表单        | 超链接 | 列表                          | 页面属性 | 滚动字幕 | 图片                            | 文本属性 |
|---|-----------|-----|-----------------------------|------|------|-------------------------------|------|
| <b>文本1</b>  |           |     |                             |      |      |                               |      |
| 请先搜索再修改   |           |     |                             |      |      |                               |      |
| 文字 低碳生活 <input type="button" value="搜索"/> <input type="button" value="重置"/> |           |     |                             |      |      |                               |      |
| <input checked="" type="checkbox"/> 字体名称                                    | 楷体_GB2312 |     | <input type="checkbox"/> 斜体 |      |      | <input type="checkbox"/> 对齐方式 |      |
| <input checked="" type="checkbox"/> 字体大小                                    | 36PX      |     | <input type="checkbox"/> 粗体 |      |      |                               |      |
| <input checked="" type="checkbox"/> 字体颜色                                    | #FFFFFF00 |     |                             |      |      |                               |      |
| <input type="checkbox"/> 本知识点分数   | 1         |     |                             |      |      |                               |      |
| 已选属性: 字体大小: 36PX 字体名称: 楷体_GB2312 字体颜色: #FFFFFF00                            |           |     |                             |      |      |                               |      |
| <input type="button" value="添加"/>   |           |     |                             |      |      |                               |      |

图3 配置答案界面二:文本属性知识点设置

### 3 实验与分析

在 HTML 文档阅卷方法中,由于文本分析方法较复杂且因其不够灵活往往只被用于局部分析。因此实践中常用的 HTML 自动阅卷方法为文档对象模型分析方法和基于 XML 的分析方法。文中共有两个实验。其中“可阅卷率”计算公式如下:

$$\text{可阅卷率} = \frac{\text{可阅卷份数}}{\text{页面总数}} = \frac{\text{可阅卷份数}}{\text{学生人数} \times \text{网页文件数}}$$

其中,“可阅卷份数”是指使用 HTML 文档自动阅卷程序能完成阅卷的网页页面数。

“人工阅卷份数”是指从待阅页面中随机抽取一定数量的页面进行人工阅卷,“平均误差”为自动阅卷程序阅卷所得分数和人工阅卷所得分数之差的平均值,其计算公式如下:

$$\text{avgErr} = \frac{\sum_{i=1}^n (|\text{auto}_i - \text{manul}_i|)}{n}$$

其中,  $n$  指人工阅卷份数,  $\text{auto}_i$  是指对第  $i$  个页面采用自动阅卷程序阅卷所得分数,  $\text{manul}_i$  是对第  $i$  个页面采用人工阅卷所得分数。

实验一:针对文档对象模型分析方法以及基于 XML 的方法在阅卷准确率和运行时间方面进行了对比实验。实验数据是从在 2010-2011 学年第一学期期末同济大学计算机基础课程考试中中学生上机考试上传的 HTML 文档中随机选取 html 文档,分别用文档对象模型分析方法和基于 XML 的信息抽取方法对这些文档进行自动阅卷。两种方法的实验结果见表 1。

表 1 文档对象模型分析方法与基于 XML 的分析方法比较

| 文档数量<br>(个) | 文档对象模型分析方法 |         | 基于 XML 的方法 |         |
|-------------|------------|---------|------------|---------|
|             | 可阅卷率       | 运行时间(S) | 可阅卷率       | 运行时间(S) |
| 50          | 100%       | 78      | 100%       | 9       |
| 100         | 99%        | 196     | 99%        | 19      |
| 150         | 97.3%      | 281     | 99.3%      | 27      |

实验二:在 2010 - 2011 学年第一学期期末同济大学举行的大学计算机基础课程考试中,采用本系统对网页设计部分进行阅卷,实验结果如表 2 所示。

表 2 自动阅卷实验结果

| 场次 | 学生<br>人数 | 网页文<br>件数 | 可阅卷情况     |          | 准确情况       |          |
|----|----------|-----------|-----------|----------|------------|----------|
|    |          |           | 可阅卷<br>份数 | 可阅<br>卷率 | 人工阅卷<br>份数 | 平均<br>误差 |
| 1  | 189      | 2         | 376       | 99.47%   | 50         | 0.55     |
| 2  | 218      | 1         | 216       | 99.08%   | 50         | 0.40     |
| 3  | 220      | 1         | 220       | 100%     | 50         | 0.05     |
| 4  | 198      | 2         | 396       | 100%     | 50         | 0.08     |
| 5  | 180      | 2         | 359       | 99.72%   | 50         | 0.10     |
| 平均 | 195      | 1.6       | 440       | 99.65%   | 50         | 0.23     |

通过对以上数据的分析得出以下几点:

(1)由表 1 可以看出:在准确率方面,在数据量小的情况下,文档对象模型分析方法与基于 XML 方法都有很高的准确率,但当数据量大的时候基于文档对象模型分析方法的准确率会有所下降;在运行时间方面文档对象模型分析方法较长。为了准确获取 HTML 对象,文档对象模型方法加载文档时需要通过死循环来等待 DocumentComplete 事件的发生,降低了阅卷的效率。而基于 XML 的方法处理的是规范的 XML 文档,可以很好地避免这类问题。

(2)由表 2 可以看出平均可阅卷率达到 99.65%,个别文件不能正常阅出分数是因为:

①页面后缀错误。程序只对后缀名为 .html 或 .htm 的网页文件进行评阅,对于其他后缀名的网页文件默认跳过。

②页面文件被损。个别网页文件因未知原因被损坏,导致无法阅卷。

③上传的文件内容非 HTML 文件内容导致预处理失败,导致无法阅卷。

(3)表 2 中可以看出各场次平均误差均比较低,总的平均误差也只有 0.23。误差产生的原因一是由于学生的页面中未严格按照标准来做而由人工阅卷时的主观性而产生。二是由于一部分页面未能成功阅卷导致。

总的来说,基于 XML 的方法在准确性及时间上都优于文档模型方法,并且与人工阅卷误差在 0.23 左右,对不能阅的文件都会标记出来,因此在消耗时间上及准确性上完全可以满足 HTML 阅卷需要。

### 4 结束语

本文首先介绍了现今 HTML 自动阅卷的几种方法,并对几种方法进行了分析比较。之后对基于 XML 的方法进行了设计与实现。最后利用数据进行了实验比较,证明了基于 XML 方法的准确性及可行性。对于其中存在的个别文件不能阅卷的情况可通过增加验证的方式进行后续的改进。

基于 XML 的方法具有很强的扩展性,文中主要介绍了应用程序的方式实现。由于是利用 XML 来实现,可以利用组件技术将其扩展成网络版<sup>[12]</sup>,在提交作业的同时即可完成阅卷的工作,节省大量的时间。

#### 参考文献:

[1] Zhang Liang,Zhuang Yueting,Yuan Zhenming,et al. A Web-Based Examination and Evaluation System for Computer Education [ C ]//Sixth IEEE International Conference on Advanced Learning Technologies. [ s. l. ]:[ s. n. ],2006:120 - 124.

(下转第 8 页)

JSON 字符串解析成 JSON 对象,这样就可以方便地使用 JSON 对象了。

### 3.1.2 JSON 与 XML 对比

①可读性。二者都具有比较强的可读性,在可读性上面,二者相差无几。

②可扩展性。XML<sup>[13]</sup> 的可扩展性远远强于 JSON,因为 XML 天生就具有可扩展性。

③编码难度。XML 基于 DOM 这类去编码,而 JSON 说到底也是 JavaScript 代码,当然 JSON 的编码更加的简洁容易。

④解码难度。JSON 就是一个 JavaScript 的对象,对于它的解码,就是使用对象,解码难度几乎为零,而 XML 需要像解析 DOM 一样去解析,难度较大。

## 3.2 JavaScript 数据格式与 JSON、XML 的比较

### 3.2.1 Stripes 中的 JavaScript 数据格式

①如何解析。在 Stripes 中,内置了一个类叫 JavaScriptResolution,可以通过调用其中的 JavaScriptResolution 的构造函数,构造一 JavaScriptResolution 对象,并通过 return 语句,将一串字符串以 JavaScript 的格式返回,例如: return new JavaScriptResolution(“alert(‘hello word!’)”);

②如何转换。在使用 JQuery 的 Ajax 库中,只需要将 dataType 设置为“script“,就可以以 JavaScript 的格式转换响应的数据。

### 3.2.2 JavaScript 数据格式的优势

①使用 JavaScript 最主要的原因是它在反应数据之间的那种循环关系的时候所表现出来的那种简洁。通过 JavaScript 来表示这种关系,将非常具有可读性和可操作性。

②关于标准化的支持。在对于 JavaScript 格式来说,它就是 JavaScript 代码,是 ECMAScript<sup>[14]</sup> 的标准。

## 4 结束语

文中针对当前流行的 Ajax 技术,利用 Stripes 这个

轻量级的框架分析使用 Stripes+Ajax 架构。同时分析了在 Stripes+Ajax 架构中的数据传输格式,对比这些数据传输格式的利弊,使得读者可以从中得到多种 Stripes+Ajax 的解决方案,可以根据自己的需求,选择最佳的实现方案来提供更高效、便捷、交互性更好的 Web 开放模式。由于 Stripes 简单及其强大的功能,相信将来人们在做 Ajax 时将会选用 Stripes+Ajax 方案。

### 参考文献:

- [1] 王 沛,冯曼菲. 征服 AJAX: Web2.0 开发技术详解[M]. 北京:人民邮电出版社,2005.
- [2] 李万龙,吴雪莉,王艳霞,等. 基于 Struts 框架的 Web 应用程序的实现[J]. 计算机技术与发展,2006,16(4):102-104.
- [3] 李 刚. 基于 J2EE 的 AJAX 开发宝典[M]. 北京:电子工业出版社,2007.
- [4] 余名高,吴海林. AJAX 在 Struts 中的应用[J]. 计算机技术与发展,2007,17(10):69-72.
- [5] 柯自聪. AJAX 开发精要—概念、案例与框架[M]. 北京:电子工业出版社,2006.
- [6] 杨 勇,韩莉英. 基于 Struts2 框架的 AJAX 开发研究[J]. 计算机工程与设计,2009,30(16):3910-3913.
- [7] Beginning Stripes[EB/OL]. 2007. <http://www.stripesframework.org/display/stripes/Home>.
- [8] 朱兴亮. 使用 Stripes 开发 Web 应用[J]. 电脑知识与技术,2007(8):403-404.
- [9] Bibeault B, Katz Y. JQuery 实战[M]. 陈 宁,译. 北京:人民邮电出版社,2009.
- [10] 徐 驰,徐燕凌. Ajax 模式在异步交互 Web 环境中的应用[J]. 计算机技术与发展,2006,16(10):228-230.
- [11] 余名高,王程根. 基于 Web2.0 的 Ajax 技术的开发[J]. 计算机技术与发展,2007,17(5):203-205.
- [12] Introducing JSON[EB/OL]. 2002. <http://www.JSON.org/>.
- [13] XML[EB/OL]. 2000. <http://en.wikipedia.org/wiki/>.
- [14] Standard ECMA-262 ECMAScript Language Specification[EB/OL]. 2001. <http://www.ecma-international.org/publications/standards/Ecma-262.htm>.

(上接第 4 页)

- [2] 赵国英,黄心渊,陈世红. 网上考试系统的设计及实现[J]. 计算机工程,2002,28(1):275-277.
- [3] 吴宏良. OFFICE 文档对象分析与自动阅卷研究[D]. 上海:华东师范大学,2009.
- [4] 马永进,金炳尧. 网页制作自动阅卷的实现方法[J]. 浙江师范大学学报,2007,30(4):424-427.
- [5] W3C. Document Object Model (DOM) Level 1 Specification, Version 1.0[EB/OL]. 1998. <http://www.w3.org/>.
- [6] 孟庆祥. 基于 XML 元素处理的 Web 信息抽取研究与实现[D]. 北京:北京交通大学,2009.
- [7] Shaker M, Ibrahim H. Information Extraction from Hypertext

Mark-Up Language Web Pages[J]. Journal of Computer Science,2009,5(8):596-607.

- [8] Laender A H F, Ribeiro-Neto B A. A brief survey of web data extraction tools[J]. ACM SIGMOD,2002,31(2):84-93.
- [9] Moro M M, Braganholo V. XML: some papers in a haystack[J]. ACM SIGMOD Record Archive,2009,38(2):29-34.
- [10] 戴怡钧. HTML 转换到 XML 格式以及不同 XML 标准格式之间的转换[D]. 上海:上海交通大学,2003.
- [11] 汤克明,陈 岐. Word 自动阅卷系统的设计与实现[J]. 计算机工程与应用,2008,44(35):69-72.
- [12] 董英斌,竹 翠. 基于网络的新型计算机考试系统[J]. 计算机工程,2001,27(8):150-152.