(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104517106 A
(43)申请公布日 2015.04.15

(21)申请号 201310455068.4

(22)申请日 2013.09.29

(71)申请人 北大方正集团有限公司
地址 100871 北京市海淀区成府路 298 号方正大厦 9 层
申请人 北京方正阿帕比技术有限公司

(72)发明人 许灿辉 汤帜 徐剑波 陶欣

(74)专利代理机构 北京三聚阳光知识产权代理有限公司 11250
代理人 寇海侠

(51)Int.Cl.
*G06K 9/20*(2006.01)
*G06K 9/62*(2006.01)

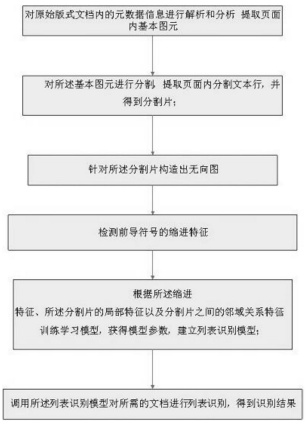
权利要求书2页 说明书8页 附图3页

(54)发明名称

一种列表识别方法与系统

(57)摘要

本发明所述的列表识别方法及系统，对原始版式文档内的元数据信息进行解析和分析，提取页面内基本图元；对所述基本图元进行分割，提取页面内分割文本行，并得到分割片；针对所述分割片构造出无向图；根据所述基本图元的属性，检测前导符号的缩进特征；根据所述缩进特征、所述分割片的局部特征以及分割片之间的邻域关系特征，训练学习模型，获得模型参数，建立列表识别模型；调用所述列表识别模型对所需的文档进行列表识别，得到识别结果。这样以机器学习的方式可以识别列表首行和列表续行的上下文关系，最终实现对版式文档的列表的版面分析及理解，即使列表首行的前导符号变化多样，也能进行识别，提高了版式文档中列表识别的准确性。




CN 104517106 A

1. 一种列表识别方法，其特征在于，包括以下步骤：

对原始版式文档内的元数据信息进行解析和分析，提取页面内基本图元；

对所述基本图元进行分割，提取页面内分割文本行，并得到分割片；

针对所述分割片构造出无向图；

根据所述基本图元的属性，检测前导符号的缩进特征；

根据所述缩进特征、所述分割片的局部特征以及分割片之间的邻域关系特征，训练学习模型，获得模型参数，建立列表识别模型；

调用所述列表识别模型对所需的文档进行列表识别，得到识别结果。

2. 根据权利要求 1 所述的列表识别方法，其特征在于，所述根据所述缩进特征、所述分割片的局部特征以及分割片之间的邻域关系特征，训练学习模型，获得模型参数，建立列表识别模型的过程中，所述学习模型为条件随机场模型，过程包括：

提取所述无向图中每个分割片的局部特征，进行分类，然后将分类得分转化为伪概率，作为条件随机场模型的一元特征函数；

根据无向图邻域关系，提取分割片之间的邻域关系特征作为二元特征函数。

3. 根据权利要求 1 或 2 所述的列表识别方法，其特征在于，所述对所述基本图元进行分割，提取页面内分割文本行，并得到分割片的过程中，将文本行中连续的文本分割到一个分割片中。

4. 根据权利要求 1 或 2 或 3 所述的列表识别方法，其特征在于，所述提取页面内分割文本行时，采用聚类方法。

5. 根据权利要求 1-4 中任一权利要求所述的列表识别方法，其特征在于，在所述针对所述分割片构造出无向图的过程中，利用所述分割片的邻域关系构造无向图。

6. 根据权利要求 1-5 中任一权利要求所述的列表识别方法，其特征在于，在所述构造无向图的过程中，采用最小生成树方法或三角剖分方法构造无向图。

7. 根据权利要求 1-6 中任一权利要求所述的列表识别方法，其特征在于，所述根据所述基本图元的属性，检测前导符号的缩进特征的过程，包括检测所述前导符号缩进级别、缩进量以及与其他前导符号缩进是否一致。

8. 根据权利要求 1-7 中任一权利要求所述的列表识别方法，其特征在于，所述分割片的局部特征包括分割片的长宽比、归一化面积、缩进级别、图像纹理特征。

9. 根据权利要求 2-8 中任一权利要求所述的列表识别方法，其特征在于，所述提取所述无向图中每个分割片的局部特征，进行分类，然后将分类得分转化为伪概率的过程，包括：通过 SVM 分类器进行分类，选择 RBF 径向基核函数，将分类得分转化为伪概率。

10. 根据权利要求 1-9 中任一权利要求所述的列表识别方法，其特征在于，所述缩进特征包括前导符号缩进级别、缩进量以及与其他前导符号缩进是否一致。

11. 一种列表识别系统，其特征在于，包括：

提取单元：对原始版式文档内的元数据信息进行解析和分析，提取页面内基本图元；

分割单元：对所述基本图元进行分割，提取页面内分割文本行，并得到分割片；

构造单元：针对所述分割片构造出无向图；

检测单元：根据所述基本图元的属性，检测前导符号的缩进特征；

建模单元：根据所述缩进特征、所述分割片的局部特征以及分割片之间的邻域关系特

征，训练学习模型，获得模型参数，建立列表识别模型；

调用单元：调用所述列表识别模型对所需的文档进行列表识别，得到识别结果。

12. 根据权利要求 11 所述的列表识别系统，其特征在于，

所述学习模型为条件随机场模型，所述建模单元中，包括：

第一特征提取子单元：提取所述无向图中每个分割片的局部特征，进行分类，然后将分类得分转化为伪概率，作为条件随机场模型的一元特征函数；

第二特征提取子单元：根据无向图邻域关系，提取分割片之间的邻域关系特征作为二元特征函数。

13. 根据权利要求 11 或 12 所述的列表识别系统，其特征在于，所述分割单元中，将文本行中连续的文本分割到一个分割片中。

14. 根据权利要求 11 或 12 或 13 所述的列表识别系统，其特征在于，所述提取页面内分割文本行时，采用聚类方法。

15. 根据权利要求 11-14 中任一权利要求所述的列表识别系统，其特征在于，所述构造单元中，根据所述分割片的邻域关系构造无向图。

16. 根据权利要求 11-15 中任一权利要求所述的列表识别系统，其特征在于，所述构造单元中，在所述构造无向图时，采用最小生成树方法或三角剖分方法构造无向图。

17. 根据权利要求 11-16 中任一权利要求所述的列表识别系统，其特征在于，所述检测单元中，检测所述前导符号缩进级别、缩进量以及与其他前导符号缩进是否一致。

18. 根据权利要求 11-17 中任一权利要求所述的列表识别系统，其特征在于，所述分割片的局部特征包括分割片的长宽比、归一化面积、缩进级别、图像纹理特征。

19. 根据权利要求 12-18 中任一权利要求所述的列表识别系统，其特征在于，所述第一特征提取子单元中，通过 SVM 分类器进行分类，选择 RBF 径向基核函数，将分类得分转化为伪概率。

20. 根据权利要求 11-19 中任一权利要求所述的列表识别系统，其特征在于，所述缩进特征包括前导符号缩进级别、缩进量以及与其他前导符号缩进是否一致。

# 一种列表识别方法与系统

## 技术领域

[0001] 本发明涉及电子文档格式转换技术领域，具体地说是一种列表识别方法与系统。

## 背景技术

[0002] 根据版式文档的生成过程，文档是数据和结构的集合，具体包括内容数据、物理结构和逻辑结构。文档分析是对文档物理结构进行抽取，而文档理解则是在物理结构和逻辑结构之间建立映射关系。在实际应用中，移动设备的可读性需求使物理和逻辑结构的恢复尤为重要。页面内列表的检测及识别是文档理解的重点之一。列表具有其独立的逻辑功能，需要对其进行物理划分和逻辑标签标定。但列表从视觉上与正文文本段的特征十分近似，且列表首行的前导符号变化多样，列表续行不具备明显的可区分性特征，根据规则的方法其识别效果不能满足实际需求。

[0003] 列表是文档的重要组成部分，如何准确地识别列表及其列表中的内容，对版式文档的分析尤其重要。现有技术中有一些识别并转换版式文档中列表的方法，如使用一组规则来检测基于矢量图形的文档中的至少一个列表。模式检测逻辑标识可能开始列表的各字符、符号、数字、字母和 / 或图像。另外的模式检测逻辑确定列表是否存在。该系统可以标识和分析标项目符号的列表、标号的或标字母的列表、以及作为两者的任意组合的嵌套列表。该方案的不足在于没有考虑列表的邻域信息，邻域信息包括文本模式、缩进基本、标点、对齐等特征，当文档页面中存在多个列表时，该方案不能识别列表续行和列表首行的上下文关系，文档整体的识别效果不理想。

## 发明内容

[0004] 为此，为此，本发明所要解决的技术问题在于现有技术中的列表识别方法不能识别列表续行和列表首行的上下文关系，从而提出一种可以识别列表首行和续行的基于概率图模型的列表识别方法。

[0005] 为解决上述技术问题，本发明的提供一种列表识别方法与系统。

[0006] 一种列表识别方法，包括以下步骤：

[0007] 对原始版式文档内的元数据信息进行解析和分析，提取页面内基本图元；

[0008] 对所述基本图元进行分割，提取页面内分割文本行，并得到分割片；

[0009] 针对所述分割片构造出无向图；

[0010] 根据所述基本图元的属性，检测前导符号的缩进特征；

[0011] 根据所述缩进特征、所述分割片的局部特征以及分割片之间的邻域关系特征，训练学习模型，获得模型参数，建立列表识别模型；

[0012] 调用所述列表识别模型对所需的文档进行列表识别，得到识别结果。

[0013] 所述的列表识别方法，所述根据所述缩进特征、所述分割片的局部特征以及分割片之间的邻域关系特征，训练学习模型，获得模型参数，建立列表识别模型的过程中，所述学习模型为条件随机场模型，过程包括：

[0014] 提取所述无向图中每个分割片的局部特征，进行分类，然后将分类得分转化为伪概率，作为条件随机场模型的一元特征函数；

[0015] 根据无向图邻域关系，提取分割片之间的邻域关系特征作为二元特征函数。

[0016] 所述的列表识别方法，所述对所述基本图元进行分割，提取页面内分割文本行，并得到分割片的过程中，将文本行中连续的文本分割到一个分割片中。

[0017] 所述的列表识别方法，所述提取页面内分割文本行时，采用聚类方法。

[0018] 所述的列表识别方法，在所述针对所述分割片构造出无向图的过程中，根据所述分割片的邻域关系构造无向图。

[0019] 所述的列表识别方法，在所述构造无向图的过程中，采用最小生成树方法构造无向图。

[0020] 所述的列表识别方法，所述根据所述基本图元的属性，检测前导符号的缩进特征的过程，包括检测所述前导符号缩进级别、缩进量以及与其他前导符号缩进是否一致。

[0021] 所述的列表识别方法，所述分割片的局部特征包括分割片的长宽比、归一化面积、缩进级别、图像纹理特征。

[0022] 所述的列表识别方法，所述提取所述无向图中每个分割片的局部特征，进行分类，然后将分类得分转化为伪概率的过程，包括：通过 SVM 分类器进行分类，选择 RBF 径向基核函数，将分类得分转化为伪概率。

[0023] 所述的列表识别方法，所述缩进特征包括前导符号缩进级别、缩进量以及与其他前导符号缩进是否一致。

[0024] 一种列表识别系统，包括：

[0025] 提取单元：对原始版式文档内的元数据信息进行解析和分析，提取页面内基本图元；

[0026] 分割单元：对所述基本图元进行分割，提取页面内分割文本行，并得到分割片；

[0027] 构造单元：针对所述分割片构造出无向图；

[0028] 检测单元：根据所述基本图元的属性，检测前导符号的缩进特征；

[0029] 建模单元：根据所述缩进特征、所述分割片的局部特征以及分割片之间的邻域关系特征，训练学习模型，获得模型参数，建立列表识别模型；

[0030] 调用单元：调用所述列表识别模型对所需的文档进行列表识别，得到识别结果。

[0031] 所述的列表识别系统，所述学习模型为条件随机场模型，所述建模单元中，包括：

[0032] 第一特征提取子单元：提取所述无向图中每个分割片的局部特征，进行分类，然后将分类得分转化为伪概率，作为条件随机场模型的一元特征函数；

[0033] 第二特征提取子单元：根据无向图邻域关系，提取分割片之间的邻域关系特征作为二元特征函数。

[0034] 所述的列表识别系统，所述分割单元中，将文本行中连续的文本分割到一个分割片中。

[0035] 所述的列表识别系统，所述提取页面内分割文本行时，采用聚类方法。

[0036] 所述的列表识别系统，所述构造单元中，根据所述分割片的邻域关系构造无向图。

[0037] 所述的列表识别系统，所述构造单元中，在所述构造无向图时，采用最小生成树方法构造无向图。

[0038] 所述的列表识别系统，所述检测单元中，检测所述前导符号缩进级别、缩进量以及与其他前导符号缩进是否一致。

[0039] 所述的列表识别系统，所述分割片的局部特征包括分割片的长宽比、归一化面积、缩进级别、图像纹理特征。

[0040] 所述的列表识别系统，所述第一特征提取子单元中，通过 SVM 分类器进行分类，选择 RBF 径向基核函数，将分类得分转化为伪概率。

[0041] 所述的列表识别系统，所述缩进特征包括前导符号缩进级别、缩进量以及与其他前导符号缩进是否一致。

[0042] 本发明的上述技术方案相比现有技术具有以下优点：

[0043] (1)本发明所述的列表识别方法及系统，对原始版式文档内的元数据信息进行解析和分析，提取页面内基本图元；对所述基本图元进行分割，提取页面内分割文本行，并得到分割片；针对所述分割片构造出无向图；根据所述基本图元的属性，检测前导符号的缩进特征；根据所述缩进特征、所述分割片的局部特征以及分割片之间的邻域关系特征，训练学习模型，获得模型参数，建立列表识别模型；调用所述列表识别模型对所需的文档进行列表识别，得到识别结果。这样对列表进行抽取，并根据其逻辑功能进行逻辑标签的标定，以机器学习的方式不仅可以识别列表，还能识别列表首行和列表续行的上下文关系，最终实现对版式文档的列表的版面分析及理解，即使列表首行的前导符号变化多样，也能通过对列表逻辑功能的分析进行识别，提高了版式文档中列表识别的准确性。

[0044] (2)本发明所述的列表识别方法，采用条件随机场模型，根据由分割片局部特征获得一元特征函数、分割片之间的邻域关系特征作为二元特征函数，训练条件随机场模型(CRF)，多特征设计分为一元局部特征和二元邻域特征。一元特征主要来自分割片本身的特征，二元特征主要来自无向图的邻居分割片的关系特征。CRF 模型的目标函数为负对数自然函数。利用多特征以及各种上下文信息可以极大地减少标注分类的不确定性和模糊性对最终标记的负面影响。

[0045] (3)本发明所述的列表识别方法，对文本进行分割时，将文本行中连续的文本分割到一个分割片中，根据文本图元、图像图元一级绘制操作图元来进行分割，获得分割片，将有具有较多相关性的图元分在同一个分割片中，为无向图的构造以及分割片特征的提取奠定基础。

[0046] (4)本发明所述的列表识别方法，所述无向图构造步骤中，根据所述分割片的邻域关系构造无向图，这样在无向图中可以体现出分割片的相对位置关系，通过其邻居的位置关系来生成无向图，采用最小生成树方法或三角剖分构造无向图，由于无向图可以很好的表示邻域关系特征，为提取分割片的局部特征和邻域关系特征创造了方便，保证了提取特征的准确性和高效性。

[0047] (5)本发明所述的列表识别方法，在所述检测步骤中，检测所述前导符号缩进级别、缩进量以及与其他前导符号缩进是否一致，这样获得了所述前导符号的特征，可以更好的训练和识别前导符号，便于更好的识别和提取列表。

## 附图说明

[0048] 为了使本发明的内容更容易被清楚的理解，下面根据本发明的具体实施例并结合

附图，对本发明作进一步详细的说明，其中

[0049] 图 1 是本发明的列表识别方法的一个实施例的流程图；

[0050] 图 2 是本发明的列表识别方法的另一个实施例的流程图；

[0051] 图 3 是本发明的列表识别方法的另一个实施例的 MST 最小生成树示意图；

[0052] 图 4 是本发明所述的列表识别方法的一个实施例中列表单元和表注的逻辑标签示意图。

## 具体实施方式

[0053] 实施例 1

[0054] 本实施例提供一种列表识别方法，如图 1 所示，包括以下步骤：

[0055] (1)对原始版式文档内的元数据信息进行解析和分析，提取页面内基本图元。此处采用现有技术中的分析工具可以提取并获得页面内的基本图元。所述基本图元中包括了文本图元、图像图元以及绘制操作信息等。

[0056] (2)对所述基本图元进行分割，提取页面内分割文本行，并得到分割片。此步骤中，将文本行中连续的文本分割到一个分割片中。根据各个基本图元的属性基于周围图元的关系进行合理的分割，得到分割片。提取页面内分割文本行时，采用聚类方法通过聚类分析的手段获得页面内分割问本行。

[0057] (3)针对所述分割片构造出无向图。此时，利用所述分割片的邻域关系，采用最小生成树方法构造无向图。邻域关系也就是与其周围的分割片的邻居关系，位置关系信息等邻域关系信息。

[0058] (4)根据所述基本图元的属性，检测前导符号的缩进特征，即检测所述前导符号缩进级别、缩进量以及与其他前导符号缩进是否一致，得到的缩进特征包括前导符号缩进级别、缩进量以及与其他前导符号缩进是否一致。

[0059] (5)根据所述缩进特征、所述分割片的局部特征以及分割片之间的邻域关系特征，训练学习模型，获得模型参数，建立列表识别模型。此处的训练模型可以选择条件随机场模型，也可以选择结构化的支持向量机模型(structural SVM)，或者其他可以学习的模型，通过上述特征进行训练，机器通过自学习的方式，建立列表识别模型。该方法采用一种可学习的模型继续训练，提高了模型的可训练程度，从而可以提高建模的效率和精度，保证了列表识别的准确性。

[0060] (6)调用所述列表识别模型对所需的文档进行列表识别，得到识别结果。

[0061] 本发明所述的识别方法，以机器学习的方式不仅可以识别列表，还能识别列表首行和列表续行的上下文关系，最终实现对版式文档的列表的版面分析及理解，即使列表首行的前导符号变化多样，也能通过对列表逻辑功能的分析进行识别，提高了版式文档中列表识别的准确性。。

[0062] 作为其他可以替换的实时方式，在所述步骤(5)建立列表识别模型的过程中，所述学习模型可以选择条件随机场模型，此处建模的过程为：

[0063] 提取所述无向图中每个分割片的局部特征，进行分类，然后将分类得分转化为伪概率，作为条件随机场模型的一元特征函数。本实施例中，所述分割片的局部特征包括分割片的长宽比、归一化面积、缩进级别、图像纹理特征，将这些特征通过 SVM 分类器进行分类，

选择 RBF 径向基核函数，将分类得分转化为伪概率，从而获得一元特征函数。

[0064] 并根据无向图邻域关系，提取分割片之间的邻域关系特征作为二元特征函数。然后将所述缩进特征、所述分割片的局部特征以及分割片之间的邻域关系特征输入所述条件随机场模型中，然后获得模型参数，并建立了列表识别模型。

[0065] 实施例 2 ：

[0066] 本实施例提供一种列表识别系统，包括 ：

[0067] 提取单元 ：对原始版式文档内的元数据信息进行解析和分析，提取页面内基本图元。

[0068] 分割单元 ：对所述基本图元进行分割，提取页面内分割文本行，并得到分割片。所述提取页面内分割文本行时，采用聚类方法。将文本行中连续的文本分割到一个分割片中。

[0069] 构造单元 ：针对所述分割片构造出无向图。根据所述分割片的邻域关系，采用最小生成树方法构造无向图。

[0070] 检测单元 ：根据所述基本图元的属性，检测前导符号的缩进特征，即检测所述前导符号缩进级别、缩进量以及与其他前导符号缩进是否一致，得到的缩进特征包括前导符号缩进级别、缩进量以及与其他前导符号缩进是否一致。。

[0071] 建模单元 ：根据所述缩进特征、所述分割片的局部特征以及分割片之间的邻域关系特征，训练学习模型，获得模型参数，建立列表识别模型。

[0072] 调用单元 ：调用所述列表识别模型对所需的文档进行列表识别，得到识别结果。

[0073] 作为优选的实施方式，所述建模单元中，所述学习模型为条件随机场模型，所述建模单元还包括 ：

[0074] 第一特征提取子单元 ：提取所述无向图中每个分割片的局部特征，进行分类，然后将分类得分转化为伪概率，作为条件随机场模型的一元特征函数。所述分割片的局部特征包括分割片的长宽比、归一化面积、缩进级别、图像纹理特征。所述分割片的局部特征通过 SVM 分类器进行分类，选择 RBF 径向基核函数，将分类得分转化为伪概率

[0075] 第二特征提取子单元 ：根据无向图邻域关系，提取分割片

[0076] 之间的邻域关系特征作为二元特征函数。

[0077] 实施例 3 ：

[0078] 本实施例所述的列表识别系统对应的列表识别方法流程图如图 2 所示，包括以下步骤 ：

[0079] （1）提取步骤 ：通过解析引擎对原始版式文档内的元数据信息进行解析，提取页面内的基本图元，包括文本图元、图像图元以及绘制操作。所述文本图元包括文本编码、字体类型、字体颜色、字体大小等 ；所述图像图元包括自然图像和合成图像 ；所述绘制操作图元信息包括绘制线、绘制图形操作信息。

[0080] （2）分割步骤 ：对所述文本图元、图像图元以及绘制操作图元进行聚类，分割页面内容，并得到分割片。此处采用聚类分析的方法提取页面内分割文本行，如采用 XY-cut 方法。分割片根据其文本图元、图像图元、绘制操作图元的区域类型获得。

[0081] （3）无向图构造步骤 ：针对所述分割片构造出无向图。根据所述分割片的邻域关系构造，所述邻域关系是指分割片与其周围的分割片的邻居关系，在此采用最小生成树的方法构造无向图。

[0082] 最小生成树(Minimum Spanning Tree，MST)方法及原理具体为：一个有 n 个结点的连通图的生成树是原图的极小连通子图，且包含原图中的所有 n 个结点，并且有保持图连通的最少的边。在一给定的无向图 G=(V,E) 中，(u,v) 代表连接顶点 u 与顶点 v 的边(即)，而 w(u,v) 代表此边的权重，若存在 T 为 E 的子集(即)且为无循环图，使得的 w(T) 最小，则此 T 为 G 的最小生成树。

[0083] $$\omega(t)=\sum_{(u,v)\in t}\omega(u,v)$$

[0084] 最小生成树其实是最小权重生成树的简称。

[0085] 因此采用最小生成树的方法将分割片构造出无向图，图 3 给出了一个页面内分割片的 MST 最小生成树示意图。

[0086] 此外，作为其他可以替换的实施方式，还可以采用 Delaunay 三角剖分方法来构造无向图。Delaunay 三角剖分方法，由于其独特性，关于点集的很多种几何图都和 Delaunay 三角剖分相关，如 Voronoi 图，EMST 树，Gabriel 图等。Delaunay 三角剖分有最大化最小角，“最接近于规则化的“的三角网和唯一性(任意四点不能共圆)两个特点。因此，采用现有技术中的 Delaunay 三角剖分方法可以构造无向图。

[0087] (4)单元格检测步骤：根据所述基本图元的属性，检测前导符号的缩进特征，即检测所述前导符号缩进级别、缩进量以及与其他前导符号缩进是否一致，得到的缩进特征包括前导符号缩进级别、缩进量以及与其他前导符号缩进是否一致。

[0088] (5)分类步骤：提取所述无向图中每个分割片的局部特征，通过 SVM 分类器，选择 RBF 径向基函数，采用 Platt 方法将基于局部特征的分类得分转化为伪概率，伪概率作为条件随机场模型(CRF)的一元特征函数。根据无向图邻域关系，提取分割片之间的邻域关系特征作为二元特征函数。。

[0089] 支持向量机SVM(Support Vector Machine)是一种可训练的机器学习方法，SVM的主要思想可以概括为两点：⑴它是针对线性可分情况进行分析，对于线性不可分的情况，通过使用非线性映射算法将低维输入空间线性不可分的样本转化为高维特征空间使其线性可分，从而使得高维特征空间采用线性算法对样本的非线性特征进行线性分析成为可能。在本步骤中，通过 SVM 进行分类。

[0090] 所谓径向基函数 (Radial Basis Function 简称 RBF)，就是某种沿径向对称的标量函数。通常定义为空间中任一点 x 到某一中心 xc 之间欧氏距离的单调函数，可记作 k(||x-xc||)，其作用往往是局部的，即当 x 远离 xc 时函数取值很小。最常用的径向基函数是高斯核函数，形式为 k(||x-xc||)=exp{-||x-xc||^2/2*σ^2)} 其中 xc 为核函数中心，σ 为函数的宽度参数，控制了函数的径向作用范围。通过选择 RBF 径向基函数。采用 Platt 方法将分类得分转化为伪概率。

[0091] (6)训练识别步骤：根据所述缩进特征、所述分割片的局部特征以及分割片之间的邻域关系特征，训练学习模型，获得模型参数，建立列表识别模型。

[0092] 概率图模型是一类用图形模式表达基于概率相关关系的模型的总称，它能够以统一概率框架融合利用多特征和上下文信息，本实施例中将页面内分割片的邻域关系表示为无向图结构，将逻辑标注的问题转换为基于无向概率图模型的分割片标记问题。

[0093] 条件随机域(也称作条件随机场)(conditional random fields，简称 CRF，或

CRFs)，是一种判别式概率模型，是随机场的一种，常用于标注或分析序列资料，如自然语言文字或是生物序列。而条件随机场则使用一种概率图模型，具有表达长距离依赖性和交叠性特征的能力，能够较好地解决标注(分类)偏置等问题的优点，而且所有特征可以进行全局归一化，能够求得全局的最优解。条件随机场是一个典型的判别式模型，其联合概率可以写成若干势函数联乘的形式，其中最常用的是线性链条件随机场。CRF 的算法实现目前已经有多个知名的开源项目，并且已经被广泛应用在学术界研究以及工业界应用当中。具体来说，条件随机场(Conditional Random Field，CRF)模型的优势在于可以更好地利用分割片本身的观察信息(observation)和自适应上下文信息(contextual information)。

[0094] 本实施例所述的列表识别方法利用多特征以及各种上下文信息可以极大地减少标注分类的不确定性和模糊性对最终标记的负面影响。在本实施例中，多特征设计分为一元局部特征和二元邻域特征。一元特征主要来自分割片本身的特征(即分割片之间的邻域关系特征)，二元特征主要来自无向图的邻居分割片的关系特征(即分割片之间的邻域关系特征)。CRF 模型的目标函数为负对数自然函数。

[0095] 本步骤具体的过程如下：根据无向图邻域关系，提取文本行之间二元关系特征，主要包括二个分割片是否左对齐、右对齐或中间对齐；是否具有同样是字体和字体尺寸；是否出现重叠；二个分割片宽度比、高度比、面积比等。构造一元和二元的特征函数，训练条件随机场模型得到模型参数，最终得到列表类别的识别结果。

[0096] (7)调用所述列表识别模型对所需的文档进行列表识别，得到识别结果。这样对列表进行抽取，并根据其逻辑功能进行逻辑标签的标定，如图 4 所示，以机器学习的方式不仅可以识别列表，还能识别列表首行和列表续行的上下文关系，最终实现对版式文档的列表的版面分析及理解，即使列表首行的前导符号变化多样，也能通过对列表逻辑功能的分析进行识别，提高了版式文档中列表识别的准确性。

[0097] 显然，上述实施例仅仅是为清楚地说明所作的举例，而并非对实施方式的限定。对于所属领域的普通技术人员来说，在上述说明的基础上还可以做出其它不同形式的变化或变动。这里无需也无法对所有的实施方式予以穷举。而由此所引伸出的显而易见的变化或变动仍处于本发明创造的保护范围之中。

[0098] 本领域内的技术人员应明白，本发明的实施例可提供为方法、系统、或计算机程序产品。因此，本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0099] 本发明是参照根据本发明实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0100] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特

定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能。

[0101] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能的步骤。

[0102] 尽管已描述了本发明的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例作出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本发明范围的所有变更和修改。

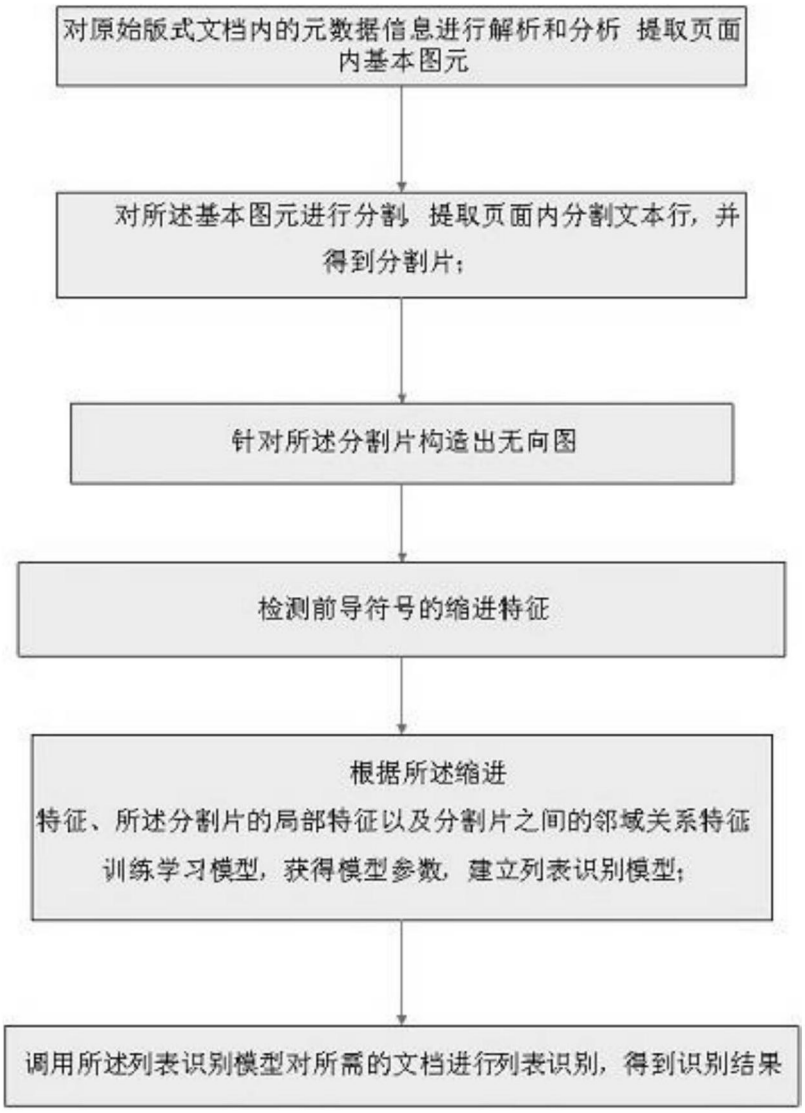

对原始版式文档内的元数据信息进行解析和分析 提取页面内基本图元
对所述基本图元进行分割，提取页面内分割文本行，并得到分割片；
针对所述分割片构造出无向图
检测前导符号的缩进特征
根据所述缩进特征、所述分割片的局部特征以及分割片之间的邻域关系特征训练学习模型，获得模型参数，建立列表识别模型；
调用所述列表识别模型对所需的文档进行列表识别，得到识别结果

图 1

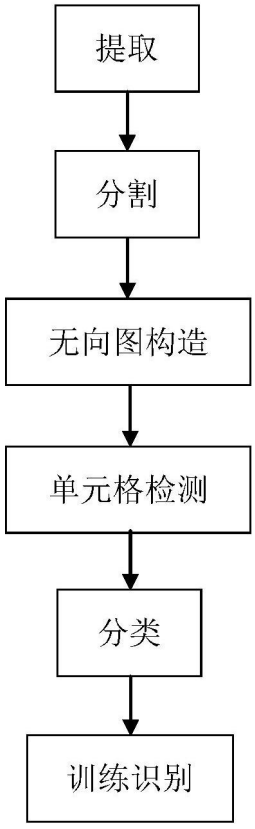

提取
分割
无向图构造
单元格检测
分类
训练识别

图 2

| Symbol | Description |
|---|---|
| $R_i$ | $i$-th device profile |
| $D_j(k)$ | $k$-th version of data item $D_j$ |
| $N_{User}$ | Number of users in the cell |
| $\lambda_{Ctrl.}$ | Aggregate request rate in the cell |
| $\mu_{Ctrl.}$ | Service rate of the control channel |
| $\mu_{Sche.}$ | Service rate of the cache |
| $\mu_{BCast.}$ | Service rate of the broadcast channel |
| $\rho_{Sche.}$ | Standard deviation of the service time of the cache |
| $B_{Ctrl.}$ | Bandwidth of the control channel |
| $B_{BCast}$ | Bandwidth of the broadcast channel |

TABLE I
DESCRIPTION OF SYMBOLS

The rate that a service handoff is forced to terminate is called the *service dropping rate* (abbreviated as SDR).

## III. Analytical and Transcoding Models

### A. Analytical Model

In this subsection, we derive the worst case of the average access time of QTP, and use the derived results to guide the version decision policy and the server admission control scheme proposed in Section IV. To facilitate the following discussion, we make the following assumptions.

1) The employed scheduling scheme of the scheduler is FCFS (standing for first come, first serve).
2) No request merge occurs in the data request queue of the scheduler.
3) One transmission of a data object in the broadcast channel is received by exactly one client.
4) The messages of registration, de-registration and handoff are negligible.

Assumptions 2 and 3 occur when the users' interests are highly diverse, and hence the effect of on-demand broadcast is eliminated. We make these two assumptions since we focus on the worst case of the transcoding proxy. Assumption 4 is made since we focus on the the situation that the number of data requests is much higher than the number of control messages (i.e., registration, de-registration, handoff and service termination). These assumptions will be relaxed in our simulation model described in Section V. For better readability, a list of used symbols is shown in Table I.

We model QTP as a queueing network as shown in Figure 3. Queue 2 is a physical queue which is located in the scheduler. On the contrary, Queue 1 and Queue 3 are logical queues which are only used to model the control and broadcast channels in order to derive the average waiting time of a data request on the control and broadcast channels, respectively.

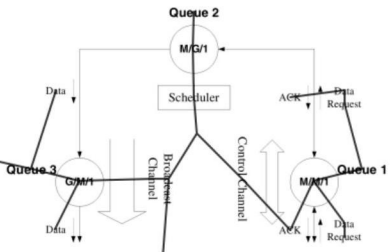



Fig. 3. The analytical model of the proposed transcoding proxy

We also let $B_{Ctrl.}$ be the bandwidth of the control channel, and let the waiting time of the control channel for a data request (denoted as $W_{Ctrl.}$) be the time interval between the user sending a data request and the user receiving the acknowledge. Then, we have the following lemma.

**Lemma 1:** The average waiting time of the control channel can be formulated as below.

$$W_{Ctrl.} = \frac{1}{\frac{B_{Ctrl.}}{s_{Ctrl.}+s_{Ack.}} - \lambda_{Ctrl.}}$$

Let the waiting time of the scheduler for a data request (denoted as $W_{Sche.}$) be the time interval from the arrival of the data request from the scheduler's perspective to the time that the requested data object has been obtained by the scheduler. Note that the service time of a cache manager is affected by several factors such as cached status of required data objects, the employed replacement scheme, the characteristic of the input jobs, and so on. The service time of the cache manager cannot be modeled by a particular mathematical distribution. Therefore, we model the average service time of the cache manager as an arbitrary distribution with mean $\frac{1}{\mu_{Sche.}}$ and variance $\sigma^2_{Sche.}$. Suppose that $\rho_{Sche.} = \frac{\lambda_{Ctrl.}}{\mu_{Sche.}}$ is the load of the scheduler. We then have the following lemma.

**Lemma 2:** The average waiting time of the scheduler is

$$W_{Sche.} = \frac{1}{\mu_{Sche.}} + \frac{\frac{\rho_{Sche.}}{\mu_{Sche.}} + \lambda_{Ctrl.}\sigma^2_{Sche.}}{2(1-\rho_{Sche.})}.$$

Let the waiting time of the broadcast channel for a data

图 3

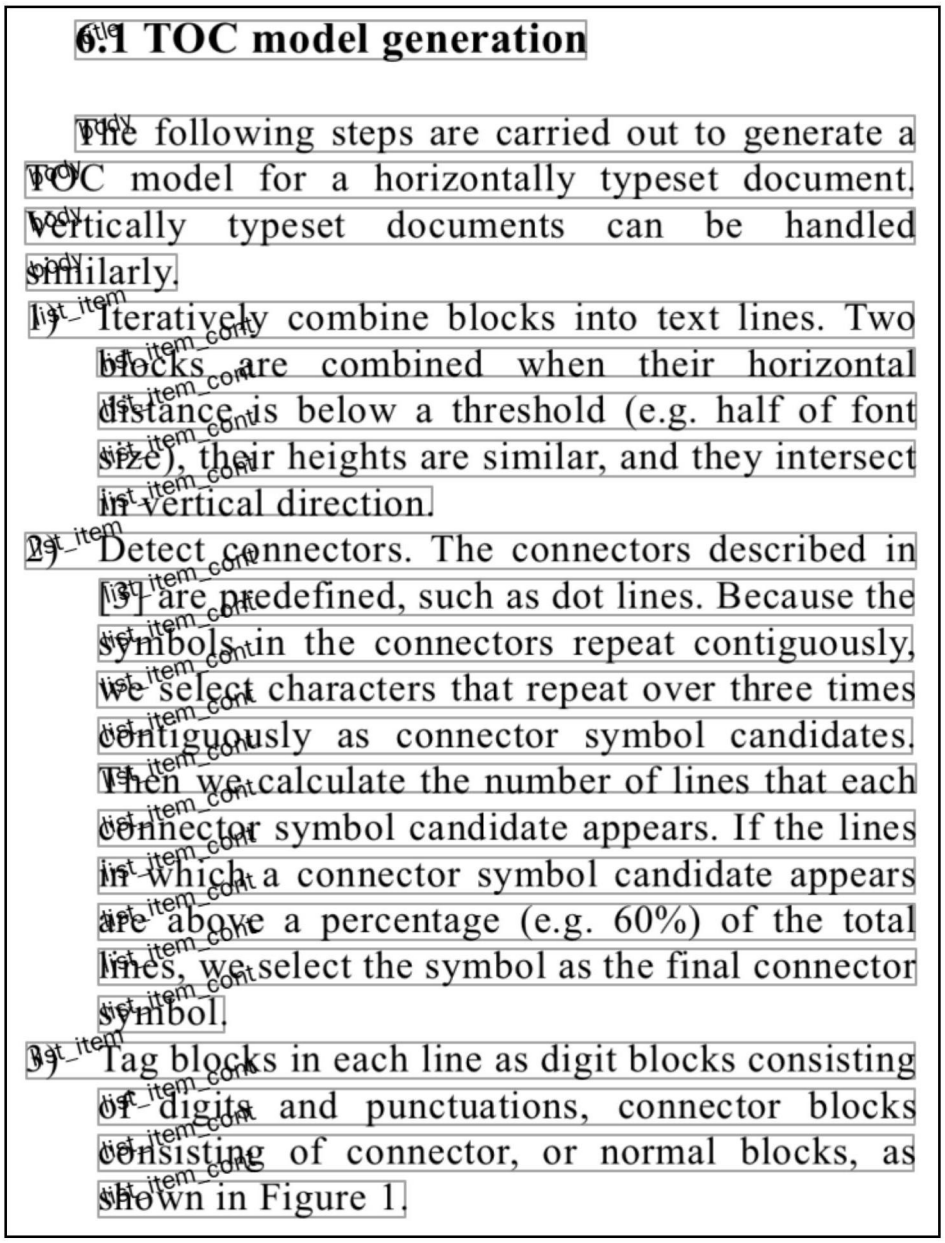

6.1 TOC model generation
The following steps are carried out to generate a TOC model for a horizontally typeset document. Vertically typeset documents can be handled similarly.
1) Iteratively combine blocks into text lines. Two blocks are combined when their horizontal distance is below a threshold (e.g. half of font size), their heights are similar, and they intersect in vertical direction.
2) Detect connectors. The connectors described in [3] are predefined, such as dot lines. Because the symbols in the connectors repeat contiguously, we select characters that repeat over three times contiguously as connector symbol candidates. Then we calculate the number of lines that each connector symbol candidate appears. If the lines in which a connector symbol candidate appears are above a percentage (e.g. 60%) of the total lines, we select the symbol as the final connector symbol.
3) Tag blocks in each line as digit blocks consisting of digits and punctuations, connector blocks consisting of connector, or normal blocks, as shown in Figure 1.

图 4