# 版面分割中文本区域最佳结构表示树的生成算法

张 利 朱 颖 吴国威

(清华大学电子工程系,北京 100084)

**摘 要** 版面分割在将印刷品转换成电子版的过程中是必不可少的,而对于分割后的各区域进行理解,达到有效分类的目的显得更为重要。本文给出了一种用最佳树对 Manhattan 文本结构进行描述的算法,利用该算法可以满足那些单靠图象分析而解决不了的高层次版面理解要求。

**关键词** 版面分割,文本结构,最佳表示树

## 1 引 言

随着信息时代的来临,将信息从其它载体转换到电子载体上已成为趋势。书刊报纸的电子化就属此类。版面分割在将报纸、书刊转化成电子版本的过程中是必不可少的。经过底层的区域分割和区域标识后,得到了文本图象的基本元素,如标题、作者、文本段落等等。为了进一步描述和识别这些基本元素之间的空间组织关系,即文本结构,必须赋以一定的表示形式。因为,对文本结构加以描述并与各种类型的文本模式比较、识别,能够实现文本的的自动分类。本文将在区域分割的基础上,对具有一定版面规范信息的 Manhattan 格式的文本图象区域用最佳二叉树直观地描述文本结构的构成和基本元素的相互关系。

## 2 基于自底向上的表示树生成算法

### 2.1 问题的定义

以区域标识得到的基本元素作为叶子节点,经逐层两两合并生成树的各层中间节点直至最后生成根节点。在每一层的合并过程中,会产生多种合并路径。在不加限制的条件下会生成多棵表示树。为了得到对文本结构的正确描述,我们引入文本结构上的特殊规范作为限制条件,在多种合并路径中寻找最符合版面规范的树结构来描述版面,即将问题转化为在一定的求解规则约束下,搜索最佳解的过程。求解最佳表示树的约束条件有 2 类,基本元素在水平垂直方向上的位置关系,以及某些特定基本元素的相互关系。基本元素在水平垂直方向上的位置关系决定着哪些字节点可以相互合并,如图 1 中,基本元素对(2,3)、(4,5)、(2,4)、(3,5)可以合并,而对(1,2)、(1,3)、(1,4)、(1,5)不能合并。特定基本元素的相互位置关系是由排版时的特殊规范条件决定的,它决定哪种合并是最佳合并。如图 1 中(2,3)、(4,5)的合并优于(2,4)、(3,5)的合并方式。

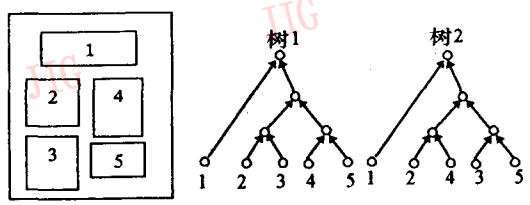


图 1 文本图象的树结构表示

### 2.2 最佳表示树生成算法

本文采用最优解搜索的思想,提出了文本结构的最佳表示树生成算法。算法以表示树生成过程中逐层合并产生的节点代替不同路径形成的表示树,进行最优路径的搜索,避免了对相同节点重复搜索引起的时间和空间复杂度增加。

首先设置节点表 node-list,用于存放叶子节点

收稿日期:1997-09-01;收到修改稿日期:1997-10-30

及合并过程生成的各级中点节点。对每个节点定义如下几个特征：

(1)左、右子节点 lchild、rchild，表示本节点所表示区域是哪两个节点表示区域的合并；

(2)节点所表示的矩形区域的位置(顶点坐标) left、right、top、bottom；

(3)节点所表示区域的标识 label，如标题、作者、文本列、图形、页眉等；

(4)节点所表示区域中包含基本元素的数目 leaf-number；

(5)合并次数 merge-number，表示由叶子节点生成本节点所需合并的次数；

(6)合并失败次数 bad-merge-number，表示在生成本节点过程中，不符合版面规则或版面规则不能解释的合并次数；

(7)本节点已与哪些节点合并过的记录 merge-node。

在后面的阐述中，以“节点→特征”表示节点的特征。

将被处理的图象类型在文本结构上所满足的特殊规范转换成一组求解规则，每条规则对待合并的2个节点做标识和位置关系检查，判决合并是否合法。求解规则形式如下：

rule(node1, relation, node2, label)

其中 node1、node2 是待合并的节点，relation 是 node1 与 node2 之间的关系，求解规则 rule 对节点 node1 和 node2 进行判别，若两者具有 relation 的关系，则判别合并是合法的，并将合并结果作为相应标识 label。在上述定义之后，最佳表示树的生成算法如下：

(1)将所有基本元素放入 node-list 中形成叶子节点，并做相应设置。每个叶子节点的左右节点 lchild、rchild 和合并记录 merge-node 设为空，包含叶子节点数目 leaf-number 设为1，合并次数和合并失败次数设为0。

(2)对 node-list 中的节点排序，排序规则在算法之后给出。

(3)取出 node-list 中的第1个节点 $n$，若 $n$ 所包含的叶子节点数目 leaf-number 等于基本元素的个数，则搜索结束，$n$ 为最佳表示树的根节点，遍历其子孙节点得到的树结构即为最佳表示树。否则，在 node-list 中找出能与 $n$ 产生合并的所有节点 $n_1, n_2, \cdots, n_m$，将 $n, n_1, n_2, \cdots, n_m$ 逐一合并生成新节点，并放入 node-list 中，若能与 $n$ 合并的节点数为零，则从 node-list 中取出第2个节点做同样处理；若 node-list 中所有节点都不产生合并，则结束搜索过程，搜索失败。能产生合并的判别条件及节点合并过程在算法之后给出。

(4)跳转第(2)步重复搜索。

上述算法中排序规则如下：

• 若节点 $n_1$ 的合并失败次数小于节点 $n_2$ 的合并失败次数，即 $n_1$→bad-merge-number＜$n_2$→bad-merge-number，则 $n_1$ 排在前；

• 若节点 $n_1$ 的合并失败次数与节点 $n_2$ 的合并失败次数相等，则若 $n_1$ 的合并次数大于节点 $n_2$ 的合并次数，即 $n_1$→bad-merge-number＞$n_2$→bad-merge-number，则 $n_1$ 排在前；

• 若节点 $n_1$ 与节点 $n_2$ 的合并失败次数和合并次数均相等，则若 $n_1$ 表示的区域面积大于节点 $n_2$ 表示的区域面积，则 $n_1$ 排在前；否则，$n_1$、$n_2$ 的排列顺序可以任意。

2节点与产生合并的条件是：

• 节点 $n_1$ 与 $n_2$ 节点表示的区域不重叠；

• 节点 $n_1$ 与 $n_2$ 合并后的区域内除 $n_1$ 和 $n_2$ 所包含的叶子节点外，不包含新的叶子节点；

• 节点 $n_1$ 与节点 $n_2$ 未合并过；

节点 $n_1$ 与节点 $n_2$ 合并产生新节点 $n$ 的过程如下：

• 设置节点 $n$ 的左右字节点为 $n$→lchild＝$n_1$，$n$→rchild＝$n_2$；

• 设置节点 $n$ 表示区域位置为 $n$→left＝min ($n_1$→left, $n_2$→left)，$n$→right＝max($n_1$→right, $n_2$→right)，$n$→top＝min($n_1$→top, $n_2$→top)，$n$→bottom＝max($n_1$→bottom, $n_2$→bottom)；

• 设置节点 $n$ 包含的叶子节点数目 $n$→leaf-number＝$n_1$→leaf-number＋$n_2$→leaf-number；

• 设置节点 $n$ 的合并次数 $n$→merge-number＝$n_1$→merge-number＋$n_2$→merge-number＋1；

• 用规则组判别 $n_1$ 与 $n_2$ 合并的合法性，若合法，则设置节点 $n$ 的合并失败次数为 $n$→bad-merge-number＝$n_1$→bad-merge-number＋$n_2$→bad-merge-number，并设置相应节点 $n$ 的标识；否则，设置节点 $n$ 的合并失败次数为 $n$→bad-merge-number＝$n_1$→bad-merge-number＋$n_2$→bad-merge-number＋1；

• 将 $n_2$ 计入节点 $n_1$ 的合并记录 $n_1$→merge-node 中，将 $n_1$ 计入节点 $n_2$ 的合并记录 $n_2$→merge-

node 中;

· 在 node-list 中找出与新节点 $n$ 表示区域相同且区域标识相同的节点设为 $m$,并将 $n$ 与 $m$ 比较,若 $n\rightarrow$bad-merge-number$>m\rightarrow$bad-merge-number,则将新合并的节点 $n$ 从 node-list 中消除;否则,将节点 $m$ 从 node-list 中消除。

### 2.3 算法性能

文本结构表示树的生成算法实现了自底向上的最佳表示树的形成。在搜索过程中,把节点 $n$ 的合并失败次数看作从起始叶子节点到达节点 $n$ 的费用 $g(n)=n\rightarrow$bad-merge-number,上述算法即为搜索最小费用的最优路径过程。搜索过程是文本结构的形成过程,最优路径的搜索是最佳表示树的形成过程。因此,算法获得的树结构一定是在给定规则下对原文本结构的最佳描述。

在搜索过程中,会生成多个表示相同区域的节点,算法在每次新节点生成时,总是将当前状态下非最佳路径形成的节点舍弃掉,避免了后续过程中非最佳的搜索路径及由此产生的非最佳节点。

树生成过程中,时间和空间复杂度取决于基本元素即叶子节点的数目和相互位置关系,以及文本结构的规范性限制条件。基本元素相互合并的数目越多,树生成的复杂度越高;对文本结构的限制条件越多即可利用的知识越多,树生成的复杂度越低。

对图 2 所示的文本图象,运用最佳树生成算法获得文本的结构表示,结果如图 3,运用了 14 条限制规则,搜索过程中共生成 165 个节点,其中基本元素个数为 18 个,在 CPU 为 Pentium120M,存储体为 32M 的实验条件下,搜索过程花费时间为 10 秒。

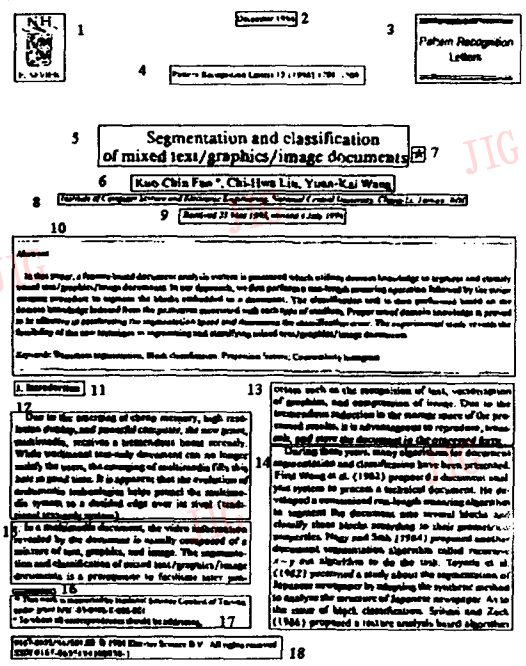


图 2 一个版面分割的例子

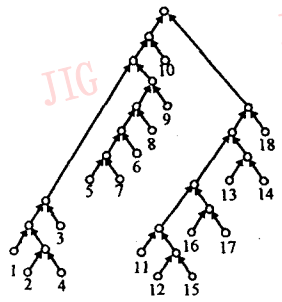


图 3　图 2 的最佳结构表示树

## 3 结束语

描述和理解文本结构,对有效识别、编辑、存储及生成文本起着重要的作用。对于版面分割后的孤立区域,如果只是利用图象分析技术是无法满足一些较高层次要求的,例如图 2 中的区域 18 是脚注,倘若只是利用图象区域的特征分类,很容易把它和区域 17 视为同类,而人们却希望他们能划分为 2 个不同类型的区域。如果把对版面结构的这种特殊要求转化为对文本结构的一种约束规则,即对文本图象结构构成的已知知识,那么通过本文所讨论的算法就能很容易地对 2 种类型区域的关系加以识别。随着版面知识可利用性的增加,算法的结果也将越来越接近人们对版面理解的最终要求。

### 参 考 文 献

1　傅京孙,蔡自兴,徐光佑. 人工智能及其应用. 北京:清华大学出版社,1987.

2　Esposito F, Malerba D, Semeraro G. An experimental page layout recognition system for office document automatic classification: An integrated approach for inductive generalization. Proc. 10th Int. conf. Pattern Recognition, 1990: 557～562.

3　Stephen W L, Srihari S N. Reading newspaper text. Proc. 10th Int. Conf. Pattern recognition, 1990:703～705.

**张　利**　1987 年毕业于清华大学无线电电子学系,1992 年获清华大学电子工程系硕士学位后留校任教。现主要从事图形图象教学和科研任务,研究方向是图文版面自动分割、图象监控与传输等。

**朱 颖** 1994年毕业于清华大学电子工程系，1996年获清华大学电子工程系硕士学位后赴美留学，现为美国普林斯顿大学博士研究生。

**吴国威** 教授，1958年毕业于清华大学电子工程系，从事的科研工作曾多次获得部委级奖励，并获国家发明专利，已发表论文数十篇。研究方向为信号与图象处理、计算机视觉、图象识别和人工智能。

# An Algorithm to Establish Optimal Trees for the Description of Document Structures in Document Segmentation

Zhang Li, Zhu Ying, Wu Guowei

(*Department of Electronic Engineering, Tsinghua University, Beijing* 100084)

**Abstract** In order to save space of newspapers and magazines in an electric way, it is important to classify their contents automatically after they are scanned into computer. An algorithm which can establish optimal trees for the description of document structures in document segmentation is given in this paper. We can get better understanding of the document structures by using this method.

**Keywords** Document understanding, Document structure, Optimal tree

# 比尔·盖茨谈未来PC

在WinHEC 98座谈会上Microsoft公司总裁就PC未来发表演说在谈及当前PC的销售数量节节上升，价格却在不断下降、他颇为振奋地说："计算机发展的速度比以往任何时候都快，所有这一切使PC成了更加得心应手的工具。"

他认为PC未来的趋势是采用速度更快的CPU和更大的内存，64MB成为规范，然后是128MB。

盖茨还预见到一些新型设备。最令他感到兴奋的产品之一是大约5年后推出的高清晰度图形输入板。该设备可让用户在PC上完成大部分阅读。

盖茨担心的一个主要问题就是供消费者使用的高速访问设备的状况。尽管电缆调制解调器和数字租用线路的实验取得了一些成就，但绝大多数家用连接设备的速度依然只有28.8kbps，简直太微不足道了。盖茨相信，Universal DSL标准将推动消费者的需求增长，并形成庞大的市场，但政府的管制将推迟电信公司的推广工作。