[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 17/21 (2006.01)

[12] 发 明 专 利 说 明 书

专利号 ZL 200410091432.4

[45] 授权公告日 2007 年 6 月 6 日

[11] 授权公告号 CN 1320481C

[22] 申请日 2004.11.22
[21] 申请号 200410091432.4
[73] 专利权人 北京北大方正技术研究院有限公司
地址 100871 北京市海淀区城府路 298 号方正大厦四楼
共同专利权人 北京大学
[72] 发明人 贾 娟 陈晓鸥 陈堃銶
[56] 参考文献
CN1393806A 2003.1.29
WO9502221 A1 1995.1.19
审查员 鲍 薇

[74] 专利代理机构 北京英赛嘉华知识产权代理有限责任公司
代理人 田 明 王达佐

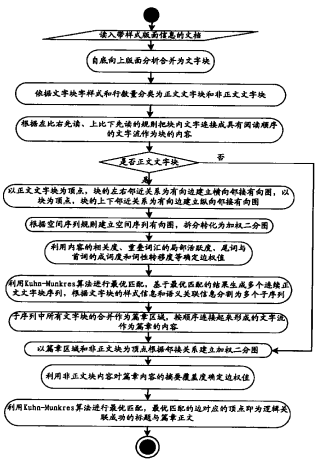
权利要求书 2 页 说明书 6 页 附图 4 页

[54] 发明名称

一种对报纸版面进行标题与正文逻辑关联的方法

[57] 摘要

本发明属于智能文字与图形信息处理技术，具体涉及一种对报纸版面进行标题与正文逻辑关联的方法。 针对现有版面理解技术只依赖样式信息分类版面逻辑对象且对多篇章多标题的报纸版面缺乏语义结构提取的缺陷，本发明首次以图论理论进行数学建模，利用二分图匹配模型描述非正文区域集和正文区域集匹配粒度一对一的特性，根据空间关系建立加权二分图，并首次采用自然语言处理技术计算二分图边权值，最优匹配结果的配对饱和顶点作为逻辑关联成功的标题和正文篇章。 本发明提出用最优匹配的库恩 – 曼克勒(Kuhn – Munkres)算法和人工智能相结合来解决标题正文的逻辑关联问题，使得匹配准确率非常高，可应用于历史数据结构化和元数据提取处理中。



1. 一种对报纸版面进行标题与正文逻辑关联的方法，包括以下步骤:

（1）读入版面分析后的报纸文档，对每个文字块按字体样式和块中行数量分类为正文文字块和非正文文字块，把正文文字块按阅读顺序和块样式分割成多个内容独立的篇章区域;

（2）建立加权二分图，二分图的两个顶点集分别包含所有的非正文文字块和篇章区域，二分图的边与非正文文字块和篇章区域在版面二维空间的相邻关系对应;

（3）二分图边的权值采用自然语言处理技术，由顶点对应的非正文文字块内容和篇章区域内容的语义确定，方法是利用标题是文章内容主题摘要的特点，把正文文字块中的文字进行词法分析后得到词集 a，共有 m 个不同的词，并计算词集 a 中每个词的分散度和共指度，分散度以该词在篇章正文中最后一次出现和第一次出现的句子间距离表示，共指度以该词在篇章中出现的次数表示；同样把非正文文字块中的文字进行词法分析得到词集 b，共有 n 个不同的词，并计算词集 b 中每个词在篇章正文中的相对分散度和相对共指度，相对分散度以该词在篇章正文中最后一次出现和第一次出现的句子间距离表示，相对共指度以该词在篇章中出现的次数表示；词集 a 中前 n 个最大分散度的和作为词集 a 的总分散度，词集 a 中前 n 个最大共指度的和作为词集 a 的总共指度，词集 b 中所有相对分散度的和作为词集 b 的总相对分散度，词集 b 中所有相对共指度的和作为词集 b 的总相对共指度，分散系数的计算通过词集 b 的总相对分散度除以词集 a 的总分散度得到，共指系数的计算通过词集 b 的总相对共指度除以词集 a 的总共指度得到；标题对篇章正文的词覆盖度，以词集 b 的词在篇章正文中出现的个数除以词集 b 所有词的个数表示，分散系数、共指系数和词覆盖度的线性加权即为边的权值;

（4）对加权二分图利用库恩-曼克勒(Kuhn-Munkres)算法进行最优匹配，最优匹配结果非正文文字块顶点集中的饱和顶点对应的非正文文字块内容即是标题，而与之有边相连的另一个篇章区域顶点集中的饱和顶点对应的就是这个标题所逻辑关联的正文篇章，二者分别作为 XML 篇章结构中的标题和正文项输出;

上述的逻辑关联指的是把报纸版面二维空间上平铺的各个文字块按其语义功能逻辑分类为标题、正文、报头、引语，然后把表示同一消息的标题和正文作为一个结构的项关联起来。

2. 如权利要求 1 所述的一种对报纸版面进行标题与正文逻辑关联的方法，其特征在于：步骤（1）中报纸文档包括扫描纸介质报纸并经 OCR 识别得到的

文档、PDF、专业排版软件生成的文档，版面分析是自底向上把版面分割为各个块区域，并物理分类为文字块和图像块，对每个文字块按字体样式和块中行数量分类为正文文字块和非正文文字块，把正文文字块的邻接关系表示为有向图，并拆分转化为加权二分图，采用自然语言处理技术计算二分图边权值，通过最优匹配得到多个连续序列，每个序列再根据文字块样式信息分割成多个子序列，合并子序列对应的区域即是独立的篇章区域，其对应内容连接成的文字流作为篇章区域的内容。

3. 如权利要求 1 所述的一种对报纸版面进行标题与正文逻辑关联的方法，其特征在于：步骤（4）中，最优匹配结果的非饱和点对应的文字块既非标题也非正文，只是版面中的报头、引语，既解决了版面对象逻辑分类问题又完成了标题与正文的逻辑关联，计算最优匹配的库恩-曼克勒(Kuhn-Munkres)算法如下：

1）给出初始标号 $l(x_i)=\max_j \omega_{ij}, l(y_j)=0, i,j=1,2...,t$，$t=\max(n,m)$；

2）求出边集 $E_l=\{(x_i,y_j)\,|\,l(x_i)+l(y_j)=\omega_{ij}\}$、$G_l=(X,Y_k,E_l)$及 G_l 中的一个匹配 M；

3）如 M 已饱和 X 的所有结点，则 M 即是 G 的最优匹配，计算结束，否则进行下一步；

4）在 X 中找一 M 非饱和点 x_0，令 $A\leftarrow\{x_0\}, B\leftarrow\phi$，$A,B$ 是两个集合；

5）若 $N_{G_l}(A)=B$，则转第 9）步，否则进行下一步，其中，$N_{G_l}(A)\subseteq Y_k$，是与 A 中结点邻接的结点集合；

6）找一结点 $y\in N_{G_l}(A)-B$；

7）若 y 是 M 饱和点，则找出 y 的配对点 z，令 $A\leftarrow A\cup\{z\}, B\leftarrow B\cup\{y\}$，转第 5）步，否则进行下一步；

8）存在一条从 x_0 到 y 的可增广路 P，令 $M\leftarrow M\oplus E(P)$，转第 3）步；

9）按下式计算 a 值：$a=\min_{\substack{x_i\in A\\ y_j\notin N_{G_l}(A)}}\{l(x_i)+l(y_j)-\omega_{ij}\}$，修改标号：

$$l'(v)=\begin{cases} l(v)-a, 若 v\in A \\ l(v)+a, 若 v\in B \\ l(v), 其它 \end{cases},$$

根据 l' 求 $E_{l'}$ 及 $G_{l'}$；

10）$l\leftarrow l', G_l\leftarrow G_{l'}$，转第 6）步。

一种对报纸版面进行标题与正文逻辑关联的方法

技术领域

本发明属于智能文字与图形信息处理技术,具体涉及一种对报纸版面进行标题与正文逻辑关联的方法。

背景技术

报纸标题在分类、检索等内容管理系统中起着重要作用，Dublin Core 和 NewsML 都把标题作为一种重要元数据，特别是在跨媒体出版中，标题作为元数据及 XML 消息结构的重要元素，与正文逻辑关联的正确与否直接影响到数字资产管理系统中信息的重用与深加工，如检索、再发布和超级链接等。逻辑关联指的就是把报纸版面二维空间上平铺的各个文字块按其语义功能逻辑分类为标题、正文、报头、引语等，然后把表示同一消息的标题和正文作为一个结构的项关联起来。作为传统的媒体形式，不同于书籍、杂志，报纸的信息传播具有密集性，即在一个版面上进行多个篇章的排版，为了提高易读性，每个篇章都有一个对其内容进行概括的标题，在位置上标题内嵌于篇章区域或与篇章邻接，在表现形式上标题具有通栏、加黑加大字体等醒目特点。但在纸媒介、排版软件、PDF 等各种载体的报纸版面中，篇章正文与标题并没有内在的结构化的关联，只是版面空间上的平铺罗列，且标题位置随意、字体大小不固定、横竖排不固定、一个标题与多个正文块位置邻近，使得判断一个标题与哪个正文匹配存在歧义，另外一些类标题块如报头、引语等在样式上与标题同质，仅仅利用样式信息无法正确对文字块进行逻辑分类。

另外，人们通过视觉思维能力和语义进行正文与标题的逻辑关联，但计算机却无法从直接的信息“理解”这种结构关联。由于报纸历史资产数量巨大，采用人工辅助干预方法既耗时成本又太大，如何在版面理解和结构化重构过程中使计算机智能、自动的进行报纸版面标题与正文的逻辑关联成为迫切需求。

标题与正文逻辑关联和对文字块逻辑分类需要交替进行，即首先粗分类文字块为非正文块和正文块，然后进行逻辑关联，再利用匹配的结果确定哪些非正文文字块是真正的标题，但目前对标题的逻辑分类都是利用样式信息独立进行的，如文献“Document page similarity based on layout visual saliency: Application to query by example and document classification”(Proceedings of the Seventh International Conference on

Document Analysis and Recognition. 2003, 1208～1212)；而文献TOC（Table Of Content）目录提取方法“Automated Detection and Segmentation of Table of Contents Page from Document Images”（作者是S.Mandal, S.P.Chowdhury和A.K.Das. 发表于Proceedings of the Seventh International Conference on Document Analysis and Recognition,2003, 398～402.）只适合书籍版面，对复杂版面的报纸无能为力；文献“复杂中文报纸的版面分析、理解和重构”（作者陈明、丁晓青、梁健.清华大学学报自然科学版 2001 年第 41 卷第 1 期. 页码 29~32,59）的匹配模型规则方法只能处理规则区域的常见类型，当正文区域是不规则形状或标题与正文的位置关系复杂时匹配模型没有描述的情况就无法正确匹配，另外一个标题与多个篇章位置邻接时存在歧义会导致错误匹配。现有技术缺乏统一的数学模型定量整体评估匹配优劣，都没有考虑语义信息，仅仅依据样式和位置信息处理复杂报纸版面是不够的。由于版面重构中标题与正文的逻辑关联处理是版面生成中为正文撰写标题的逆过程，自然语言处理技术中的标题生成的方法“Description of the UAM system for generationg very short summaries at DUC-2004”(Enrique alfonseca,Jose Maria Guirao,Antonio Moreno-Sandoval.Document Understanding Conference 2004)值得借鉴。

发明内容

针对现有技术中对报纸版面标题匹配效果不太理想的缺陷，本发明的目的是提供一种对报纸版面进行标题与正文逻辑关联的方法，该方法能对报纸版面进行篇章结构提取，可以大大提高标题匹配效果。

为达到以上目的，本发明采用的技术方案是：一种对报纸版面进行标题与正文逻辑关联的方法，包括以下步骤：

（1）读入版面分析后的报纸文档，对每个文字块按字体样式和块中行数量分类为正文文字块和非正文文字块，把正文文字块按阅读顺序和块样式分割成多个内容独立的篇章区域；

（2）建立加权二分图，二分图的两个顶点集分别包含所有的非正文文字块和篇章区域，二分图的边与非正文文字块和篇章区域在版面二维空间的相邻关系对应；

（3）二分图边的权值采用自然语言处理技术，由顶点对应的非正文文字块内容和篇章区域内容的语义确定，方法是利用标题是文章内容主题摘要的特点，把正文文字块中的文字进行词法分析后得到词集a，共有m个不同的词，并计算词集 a 中每个词的分散度和共指度，分散度以该词在篇章正文中最后一次出现

和第一次出现的句子间距离表示，共指度以该词在篇章中出现的次数表示；同样把非正文文字块中的文字进行词法分析得到词集 b，共有 n 个不同的词，并计算词集 b 中每个词在篇章正文中的相对分散度和相对共指度，相对分散度以该词在篇章正文中最后一次出现和第一次出现的句子间距离表示，相对共指度以该词在篇章中出现的次数表示；词集 a 中前 n 个最大分散度的和作为词集 a 的总分散度，词集 a 中前 n 个最大共指度的和作为词集 a 的总共指度，词集 b 中所有相对分散度的和作为词集 b 的总相对分散度，词集 b 中所有相对共指度的和作为词集 b 的总相对共指度。分散系数的计算通过词集 b 的总相对分散度除以词集 a 的总分散度得到，共指系数的计算通过词集 b 的总相对共指度除以词集 a 的总共指度得到；标题对篇章正文的词覆盖度，以词集 b 的词在篇章正文中出现的个数除以词集 b 所有词的个数表示。分散系数、共指系数和词覆盖度的线性加权即为边的权值；

（4）对加权二分图利用库恩-曼克勒(Kuhn-Munkres)算法进行最优匹配，最优匹配结果非正文文字块顶点集中的饱和顶点对应的非正文文字块内容即是标题，而与之有边相连的另一个篇章区域顶点集中的饱和顶点对应的就是这个标题所逻辑关联的正文篇章，二者分别作为 XML 篇章结构中的标题和正文项输出。

上述的逻辑关联指的是把报纸版面二维空间上平铺的各个文字块按其语义功能逻辑分类为标题、正文、报头、引语等，然后把表示同一消息的标题和正文作为一个结构的项关联起来，在进行标题与正文逻辑关联时，将图论中二分图的理论，算法及结果引入到文字块内容间摘要覆盖性的度量上，具体来说，是将图论中最优匹配的库恩-曼克勒(Kuhn-Munkres)算法用于基于内容的标题与正文逻辑关联。

本发明的效果在于：采用本发明所述的方法，通过信息处理装置能有效地对报纸版面进行篇章结构提取，大大提高了报纸版面中正文与标题的匹配效果。通过对问题的建模和对人类思维的模拟，使得匹配准确率非常高，可广泛地应用于数字资产管理系统的历史数据结构化和元数据提取处理中。

本发明之所以具有这样的效果，是因为本发明针对报纸版面文字区域复杂且文字块之间位置关系多样的特点，提出一种新的对报纸版面中标题逻辑关联正文的方法。本发明利用二分图匹配数学模型精确的描述了标题与正文粒度上一对一的特点，利用样式信息把报纸版面中文字块分类为非正文集和正文集，并根据两集合元素间的空间关系建立初始二分图，特别是首次采用自然语言处理技术，综合考虑抽取型和总结型两种摘要类型，并基于共指词链的长度与分散度计算标题对正文的语义摘要覆盖度作为非正文块与正文块之间逻辑关联的

评判因子即加权二分图的边权值，经过最优匹配后的连接饱和点的边即是标题与正文的关联关系。

附图说明

图 1 是本发明的流程图;

图 2 是版面分析并分类后的报纸示意图;

图 3 是恢复阅读顺序后的具有篇章区域的报纸示意图;

图 4 是非正文文字块与篇章区域根据邻接关系生成的二分图示意图;

图 5 是库恩-曼克勒(Kuhn-Munkres)最优匹配算法结果示意图。

具体实施方式

下面结合附图对本发明作进一步地描述，本发明的流程图如图 1 所示:

（1）读入版面分析后的报纸文档，报纸文档包括扫描纸介质报纸并经 OCR 识别得到的文档、PDF、专业排版软件如方正飞腾生成的文档等，版面分析是自底向上把版面分割为各个块区域，并物理分类为文字块和图像块。对每个文字块按字体样式和块中行数量分类为正文文字块和非正文文字块，如图 2 所示，实线矩形表示正文文字块，虚线矩形表示非正文文字块，把正文文字块的邻接关系表示为有向图，并拆分转化为加权二分图，采用自然语言处理技术计算二分图边权值，通过最优匹配得到多个连续序列，每个序列再根据文字块样式信息分割成多个子序列，合并子序列对应的区域即是独立的篇章区域，其对应内容连接成的文字流作为篇章区域的内容，如图 3 所示，箭头表示阅读顺序的先后，每个连续的箭头序列把正文文字块组成了篇章区域，带圈数字表示篇章区域的编号，普通数字表示非正文文字块的编号;

（2）建立加权二分图，二分图的两个顶点集分别包含所有的非正文文字块和篇章区域，二分图的边与非正文文字块和篇章区域在版面二维空间的相邻关系对应，如图 4 所示，左边顶点集表示非正文文字块，右边顶点集表示篇章区域;

（3）二分图边的权值采用自然语言处理技术，由顶点对应的非正文文字块内容和篇章区域内容的语义确定，方法是利用标题是文章内容主题摘要的特点，把正文文字块中的文字进行词法分析后得到词集 a，共有 m 个不同的词，并计算词集 a 中每个词的分散度和共指度，分散度以该词在篇章正文中最后一次出现和第一次出现的句子间距离表示，共指度以该词在篇章中出现的次数表示; 同样把非正文文字块中的文字进行词法分析后得到词集 b，共有 n 个不同的词，并计算词集 b 中每个词在篇章正文中的相对分散度和相对共指度，相对分散度以该词在篇章正文中最后一次出现和第一次出现的句子间距离表示，相对共指度

以该词在篇章中出现的次数表示；词集 a 中前 n 个最大分散度的和作为词集 a 的总分散度，词集 a 中前 n 个最大共指度的和作为词集 a 的总共指度，词集 b 中所有相对分散度的和作为词集 b 的总相对分散度，词集 b 中所有相对共指度的和作为词集 b 的总相对共指度。分散系数的计算通过词集 b 的总相对分散度除以词集 a 的总分散度得到，共指系数的计算通过词集 b 的总相对共指度除以词集 a 的总共指度得到；标题对篇章正文的词覆盖度，以词集 b 的词在篇章正文中出现的个数除以词集 b 所有词的个数表示。分散系数、共指系数和词覆盖度的线性加权即为边的权值；

（4）对加权二分图利用库恩-曼克勒(Kuhn-Munkres)算法进行最优匹配，最优匹配结果非正文文字块顶点集中的饱和顶点对应的非正文文字块内容即是标题，而与之有边相连的另一个篇章区域顶点集中的饱和顶点对应的就是这个标题所逻辑关联的正文篇章，如图 5 所示，被边相连的左边顶点表示标题，右边顶点表示与之逻辑关联的篇章正文，如标题 6 与正文 7 是同一个消息的组成部分，二者分别作为 XML 篇章结构中的标题和正文项输出。最优匹配结果的非饱和点对应的文字块既非标题也非正文，只是版面中如报头、引语等其他类型的内容，既解决了版面对象逻辑分类问题又完成了标题与正文的逻辑关联。计算最优匹配的 Kuhn_Munkres 算法如下：

1）给出初始标号 $l(x_i)=\max_j \omega_{ij}, l(y_j)=0, i,j=1,2...,t$，$t=\max(n,m)$；

2）求出边集 $E_l=\{(x_i,y_j)\,|\,l(x_i)+l(y_j)=\omega_{ij}\}$、$G_l=(X,Y_k,E_l)$ 及 G_l 中的一个匹配 M；

3）如 M 已饱和 X 的所有结点，则 M 即是 G 的最优匹配，计算结束，否则进行下一步；

4）在 X 中找一 M 非饱和点 x_0，令 $A\leftarrow\{x_0\}, B\leftarrow\phi$，$A,B$ 是两个集合；

5）若 $N_{G_l}(A)=B$，则转第 9）步，否则进行下一步，其中，$N_{G_l}(A)\subseteq Y_k$，是与 A 中结点邻接的结点集合；

6）找一结点 $y\in N_{G_l}(A)-B$；

7）若 y 是 M 饱和点，则找出 y 的配对点 z，令 $A\leftarrow A\cup\{z\}, B\leftarrow B\cup\{y\}$，转第 5）步，否则进行下一步；

8）存在一条从 x_0 到 y 的可增广路 P，令 $M\leftarrow M\oplus E(P)$，转第 3）步；

9）按下式计算 a 值：$a=\min_{\substack{x_i\in A\\ y_j\notin N_{G_l}(A)}}\{l(x_i)+l(y_j)-\omega_{ij}\}$，修改标号：

$$l'(v)=\begin{cases} l(v)-a, 若v\in A \\ l(v)+a, 若v\in B \\ l(v), 其它 \end{cases},$$

根据l'求$E_{l'}$及$G_{l'}$;

10）$l\leftarrow l', G_l\leftarrow G_{l'}$，转第6）步。

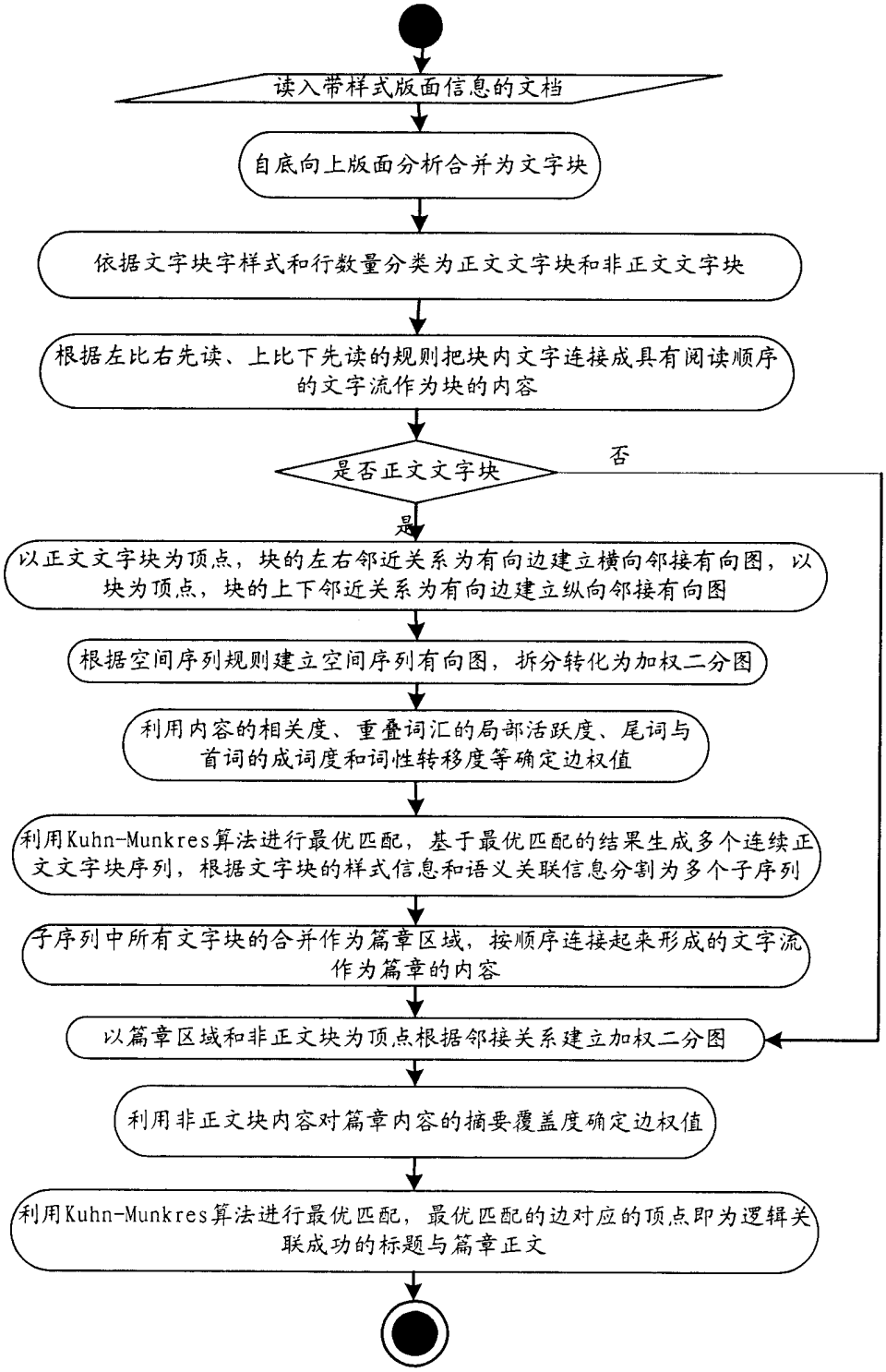

读入带样式版面信息的文档
自底向上版面分析合并为文字块
依据文字块字样式和行数量分类为正文文字块和非正文文字块
根据左比右先读、上比下先读的规则把块内文字连接成具有阅读顺序的文字流作为块的内容
是否正文文字块
否
是
以正文文字块为顶点，块的左右邻近关系为有向边建立横向邻接有向图，以块为顶点，块的上下邻近关系为有向边建立纵向邻接有向图
根据空间序列规则建立空间序列有向图，拆分转化为加权二分图
利用内容的相关度、重叠词汇的局部活跃度、尾词与首词的成词度和词性转移度等确定边权值
利用Kuhn-Munkres算法进行最优匹配，基于最优匹配的结果生成多个连续正文文字块序列，根据文字块的样式信息和语义关联信息分割为多个子序列
子序列中所有文字块的合并作为篇章区域，按顺序连接起来形成的文字流作为篇章的内容
以篇章区域和非正文块为顶点根据邻接关系建立加权二分图
利用非正文块内容对篇章内容的摘要覆盖度确定边权值
利用Kuhn-Munkres算法进行最优匹配，最优匹配的边对应的顶点即为逻辑关联成功的标题与篇章正文

图1


确保建筑幕墙工程质量的
几点措施
为你的
环境
"添丽"
建筑节能、保温材料当先行
天然石材卫士——养护剂
水源热泵中央空调
业界风采
新一代纯润息来水箱
三津管
广西研制多功能窗
青岛流亭装饰
建材批发城

图 2

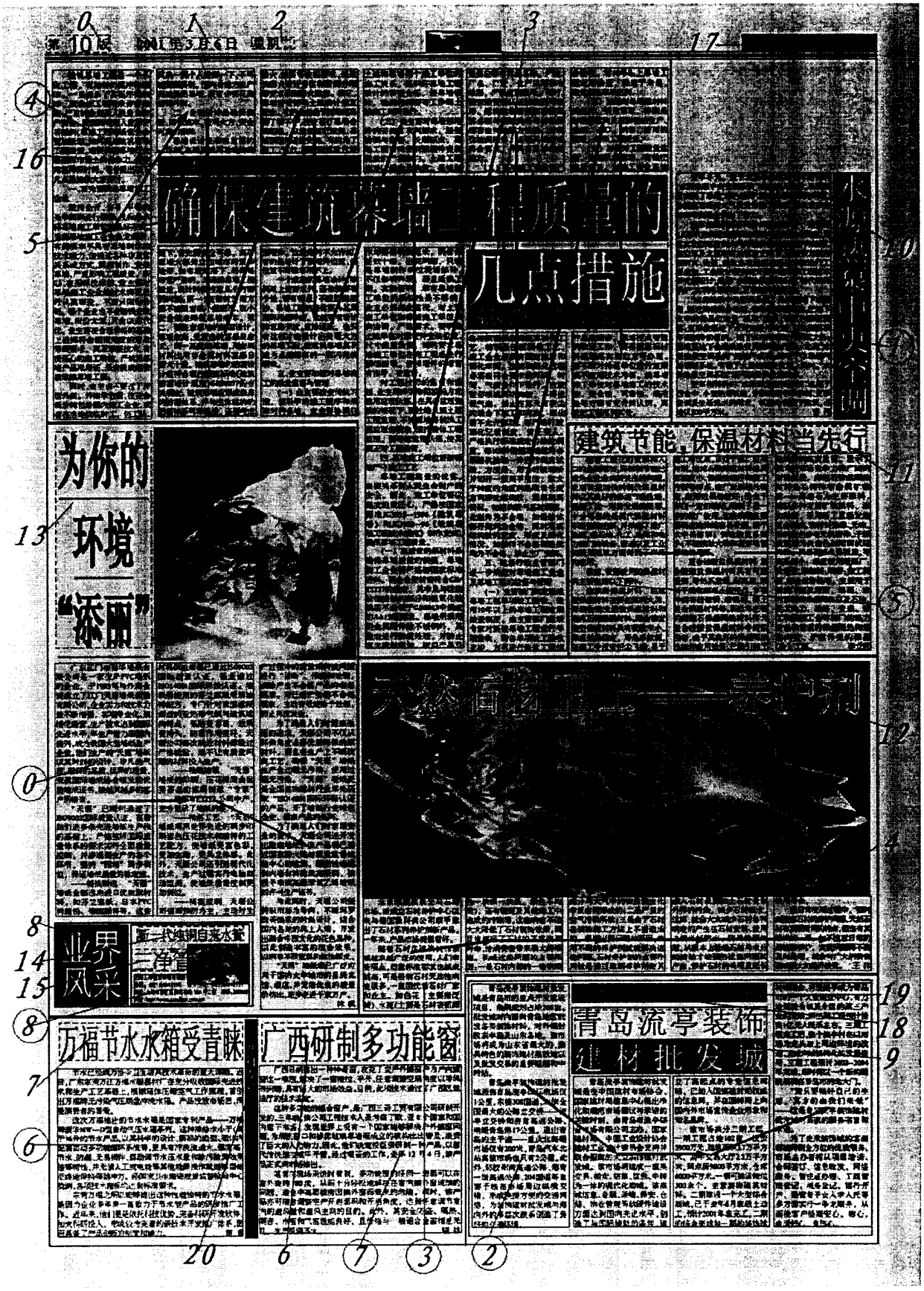

确保建筑幕墙工程质量的
几点措施
为你的
环境
“添丽”
建筑节能，保温材料当先行
青岛流亭装饰
建材批发城
广西研制多功能窗

图 3

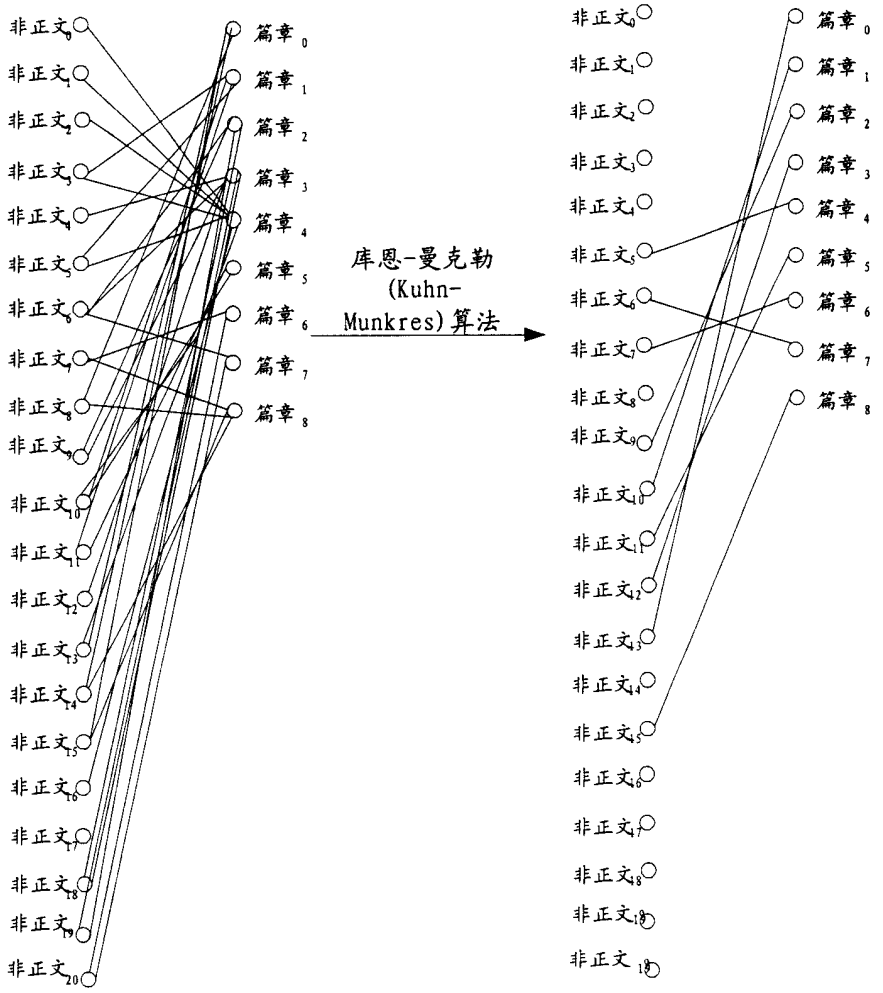

非正文0
非正文1
非正文2
非正文3
非正文4
非正文5
非正文6
非正文7
非正文8
非正文9
非正文10
非正文11
非正文12
非正文13
非正文14
非正文15
非正文16
非正文17
非正文18
非正文19
非正文20
篇章0
篇章1
篇章2
篇章3
篇章4
篇章5
篇章6
篇章7
篇章8
库恩-曼克勒(Kuhn-Munkres)算法
非正文0
非正文1
非正文2
非正文3
非正文4
非正文5
非正文6
非正文7
非正文8
非正文9
非正文10
非正文11
非正文12
非正文13
非正文14
非正文15
非正文16
非正文17
非正文18
非正文19
非正文20
篇章0
篇章1
篇章2
篇章3
篇章4
篇章5
篇章6
篇章7
篇章8

图4 图5