(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 101515272 B
(45)授权公告日 2012.10.24

(21)申请号 200810080786.7

(22)申请日 2008.02.18

(73)专利权人 株式会社理光
地址 日本东京都

(72)发明人 杜成

(74)专利代理机构 北京市柳沈律师事务所 11105
代理人 邵亚丽

(51)Int.Cl.
G06F 17/30(2006.01)

(56)对比文件
CN 1567303 A,2005.01.19, 全文.
CN 1577328 A,2005.02.09, 全文.
CN 1763740 A,2006.04.26, 全文.
US 2006149775 A1,2006.07.06, 全文.

审查员 李燕东

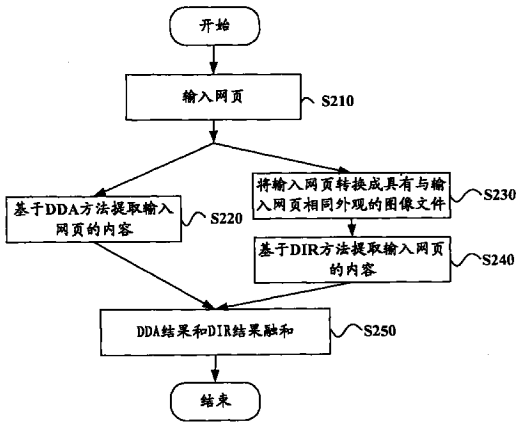
权利要求书 5 页 说明书 8 页 附图 10 页

(54)发明名称

提取网页内容的方法和装置

(57)摘要

本发明提供一种用于提取网页内容的方法和装置。所述方法包括:基于数字文档分析(DDA)方法提取输入网页的网页内容,产生DDA提取结果;基于文档图像识别(DIR)方法提取输入网页的网页内容,产生DIR提取结果;融合所述DDA提取结果和DIR提取结果,产生融合结果。根据本发明,能够得到比传统技术更优的网页提取结果。



CN 101515272 B

1. 一种用于提取网页内容的方法，包括：

基于数字文档分析 (DDA) 方法提取输入网页的网页内容，产生 DDA 提取结果；

基于文档图像识别 (DIR) 方法提取输入网页的网页内容，产生 DIR 提取结果；

融合所述 DDA 提取结果和 DIR 提取结果，产生融合结果，

其中，所述提取结果包括至少一个目标，所述目标代表与网页中的矩形区域相对应的网页内容，所述目标至少包含相应矩形区域的位置信息和类型信息，并且所述类型包括文字、图片和表格，

其中，所述融合 DDA 提取结果和 DIR 提取结果包括：

确定 DDA 提取结果包含的 DDA 目标与 DIR 提取结果包含的 DIR 目标之间的对应关系；

基于所述对应关系以及目标类型执行 DDA 提取结果与 DIR 提取结果的融合。

2. 如权利要求 1 所述的方法，其中，确定 DDA 目标和 DIR 目标之间的对应关系包括计算 DDA 目标与 DIR 目标的重叠尺度。

3. 如权利要求 2 所述的方法，其中，DIR 提取结果表示为 $G = \{G_1, G_2, \ldots, G_M\}$，DDA 提取结果表示为 $D = \{D_1, D_2, \ldots, D_N\}$，则 DDA 目标 D_j 与 DIR 目标 G_i 的重叠尺度通过下式计算：

$$\sigma_{ij} = \frac{Area(G_i \cap D_j)}{Area(G_i)} \text{ 以及 } \tau_{ij} = \frac{Area(G_i \cap D_j)}{Area(D_j)}, \quad i = 1, 2, \cdots, M, \; j = 1, 2, \cdots, N,$$

其中，$Area(D_j)$ 是 D_j 对应的矩形区域的面积，$Area(G_i)$ 是 G_i 对应的矩形区域的面积，$Area(G_i \cap D_j)$ 是 D_j 对应的矩形区域与 G_i 对应的矩形区域之间的重叠面积，并且 N 和 M 分别是 DDA 提取结果和 DIR 提取结果包含的目标个数。

4. 如权利要求 3 所述的方法，其中，基于所述对应关系和目标类型执行 DDA 提取结果与 DIR 提取结果的融合包括：

基于 DDA 目标和 DIR 目标之间的对应关系以及目标类型对 DDA 目标和 DIR 目标进行分类；以及

按照目标的类别来融合 DDA 提取结果和 DIR 提取结果，从而产生融合结果。

5. 如权利要求 4 所述的方法，其中，按照以下规则对 DDA 目标和 DIR 目标进行分类：

对于一个 DIR 目标 G_i，如果存在 DDA 目标 D_j，使得 $\tau_{ij} \approx 1$ 并且 $\sigma_{ij} \approx 1$，而且 G_i 和 D_j 的类型相同，则将该 G_i 和该 D_j 分类到匹配类；

对于一个 DIR 目标 G_i，如果存在 DDA 目标 D_j，使得 $\tau_{ij} \approx 1$ 并且 $\sigma_{ij} \approx 1$，而且 G_i 和 D_j 的类型不同，则将该 G_i 和该 D_j 分类到类型错误类；

对于一个 DIR 目标 G_i，如果对于所有 DDA 目标均有 $\sigma_{ij} \approx 0$，则将该 G_i 分类到漏检类；

对于一个 DDA 目标 D_j，如果对于所有 DIR 目标均有 $\tau_{ij} \approx 0$，则将该 D_j 分类到虚警类；

对于一个 DIR 目标 G_i，如果 $\sum_{j=1}^{N} \sigma_{ij} > T_1$，并且将与之重叠的 DDA 目标合并后得到的合并目标与该 G_i 匹配，则将该 G_i 以及与该 G_i 重叠的 DDA 目标分类到分割类，其中，T_1 是第一预定阈值；

对于一个 DDA 目标 D_j，如果 $\sum_{i=1}^{M} \tau_{ij} > T_2$，并且将与之重叠的 DIR 目标合并后得到的合并目标与该 D_j 匹配，则将该 D_j 以及与之重叠的 DIR 目标分类到合并类，其中，T_2 是第二预定阈值；以及

将不属于以上类别的 DDA 目标和 DIR 目标分类到其他类。

6. 如权利要求 5 所述的方法，其中，按照目标的类别来融合 DDA 提取结果和 DIR 提取结果以产生融合结果包括：

将匹配类中的 DDA 目标添加到融合结果中；

将类型错误类中的 DIR 目标的位置信息和相应的 DDA 目标的类型信息结合产生一个新的目标，并把该新的目标添加到融合结果；

将虚警类中所有的目标都添加到融合结果中；

将分割类中的 DIR 目标添加到融合结果中；

对于合并类，如果与 DDA 目标重叠的 DIR 目标都是图片类型的目标，则将相应的 DIR 目标添加到融合结果中；如果与 DDA 目标重叠的 DIR 目标中既包括图片类型的目标又包括文字类型的目标，则将相应的 DDA 目标添加到融合结果中；如果与 DDA 目标重叠的 DIR 目标都是文字类型的目标，则将与 DDA 目标重叠的 DIR 目标合并而成的合并目标添加到融合结果中；以及

将其他类中的 DDA 目标添加到融合结果中。

7. 如权利要求 1 所述的方法，其中，基于 DDA 方法提取输入网页的网页内容包括：

提取输入网页的文档对象模型 (DOM) 树，并至少保存 DOM 树中每个节点的父节点、子节点、标签名称、内部文字和位置的属性信息；

利用 DOM 树分别提取输入网页中的文字目标、图片目标和表格目标。

8. 如权利要求 7 所述的方法，其中，利用 DOM 树提取文字目标包括：

对于 DOM 树中的每个节点，如果该节点的内部文字属性不为空，而且该节点的子节点中不包含块节点，则确定该节点表示的元素为候选文字目标；

参考候选文字目标的属性信息对所确定的候选文字区域执行合并操作，以得到文字目标，

其中，如果节点的标签名称不是“INPUT”、“！”、“A”、“B”、“U”、“I”、“BIG”、“SMALL”、“FONT”、“HR”、“BR”、“PRE”、“TT”、“S”、“BLOCKQUOTE”、“ADDRESS”、“DFN”、“SAMP″ 、“KBD″ 、“VAR”、“CODE”、“CITE”、“ABBR”、“ACRONYM”、“SUB”、“SUP”、“INS”、“DEL”、“P”、“EM”、“TEXT”、“STRONG”、“/A”之一，则该节点为块节点。

9. 如权利要求 8 所述的方法，其中，对候选文字目标执行合并操作包括：

如果两个文字目标在位置上重叠，则将它们合并为一个文字目标。

10. 如权利要求 8 所述的方法，其中，对候选文字目标执行合并操作包括：如果一个文字目标被另外一个文字目标包含，则删除被包含的文字目标。

11. 如权利要求 8 所述的方法，其中，对候选文字目标执行合并操作包括：

如果两个文字目标所对应的矩形区域在垂直方向上位置相邻，并且它们的字体和文字高度属性相同，它们的左边缘相近，而且宽度相似，则将它们合并为一个文字区域。

12. 如权利要求 7 所述的方法，其中，利用 DOM 树提取图片目标包括：

如果 DOM 树节点的标签名称是“IMG”，并且其尺寸大于第三预定阈值，则确定该节点为图片目标。

13. 如权利要求 7 所述的方法，其中，利用 DOM 树提取表格目标包括：

如果 DOM 树节点的标签名称为“TABLE”，并且该 DOM 树节点包含至少 3 个“TR”子节点，

而且多数的“TR”子节点包含多于一个的“TD”子节点，则确定该节点为表格区域。

14. 一种用于提取网页内容的装置，包括：

数字文档分析 (DDA) 网页内容提取单元，其基于 DDA 方法提取输入网页的网页内容，产生 DDA 提取结果；

文档图像识别 (DIR) 网页内容提取单元，其基于 DIR 方法提取输入网页的网页内容，产生 DIR 提取结果；

融合单元，其融合所述 DDA 提取结果和 DIR 提取结果，产生融合结果，

其中，提取结果表示为至少一个目标的集合，所述目标代表与网页中的矩形区域相对应的网页内容，所述目标至少包含相应矩形区域的位置信息和类型信息，并且所述类型包括文字、图片和表格，

其中，所述融合单元包括：

对应关系确定单元，其确定 DDA 目标与 DIR 目标之间的对应关系；

融合执行单元，其基于 DDA 目标和 DIR 目标之间的对应关系以及 DDA 目标和 DIR 目标的类型执行 DDA 提取结果与 DIR 提取结果的融合，以生成融合结果。

15. 如权利要求 14 所述的装置，其中，所述对应关系确定单元通过计算 DDA 目标与 DIR 目标的重叠尺度来确定 DDA 目标和 DIR 目标之间的对应关系。

16. 如权利要求 15 所述的装置，其中，DIR 提取结果表示为 $G = \{G_1, G_2, \ldots, G_M\}$，DDA 提取结果表示为 $D = \{D_1, D_2, \ldots, D_N\}$，则 DDA 目标 D_i 与 DIR 目标 G_j 的重叠尺度通过下式计算：

$$\sigma_{ij} = \frac{Area(G_i \cap D_j)}{Area(G_i)} \text{以及} \tau_{ij} = \frac{Area(G_i \cap D_j)}{Area(D_j)}, \quad i = 1,2, \cdots, M, \; j = 1,2, \cdots, N,$$

其中，$Area(D_j)$ 是 D_j 对应的矩形区域的面积，$Area(G_i)$ 是 G_i 对应的矩形区域的面积，$Area(G_i \cap D_j)$ 是 D_j 对应的矩形区域与 G_i 对应的矩形区域之间的重叠面积，并且 N 和 M 分别是 DDA 提取结果和 DIR 提取结果中的目标个数。

17. 如权利要求 14 所述的装置，其中，所述融执行合单元包括：

分类单元，其基于 DDA 目标和 DIR 目标之间的对应关系以及目标类型对 DDA 目标和 DIR 目标进行分类；以及

选择单元，其按照 DDA 目标和 DIR 目标的类别来确定所述融合结果中包含的目标。

18. 如权利要求 17 所述的装置，其中，所述分类单元按照以下规则对 DDA 目标和 DIR 目标进行分类：

对于一个 DIR 目标 G_i，如果存在 DDA 目标 D_j，使得 $\tau_{ij} \approx 1$ 并且 $\sigma_{ij} \approx 1$，而且 G_i 和 D_j 的类型相同，则将该 G_i 和该 D_j 分类到匹配类；

对于一个 DIR 目标 G_i，如果存在 DDA 目标 D_j，使得 $\tau_{ij} \approx 1$ 并且 $\sigma_{ij} \approx 1$，而且 G_i 和 D_j 的类型不同，则将该 G_i 和该 D_j 分类到类型错误类；

对于一个 DIR 目标 G_i，如果对于所有 DDA 目标均有 $\sigma_{ij} \approx 0$，则将该 G_i 分类到漏检类；

对于一个 DDA 目标 D_j，如果对于所有 DIR 目标均有 $\tau_{ij} \approx 0$，则将该 D_j 分类到虚警类；

对于一个 DIR 目标 G_i，如果 $\sum_{j=1}^{N} \sigma_{ij} > T_1$，并且将与之重叠的 DDA 目标合并后得到的合并目标与该 G_i 匹配，则将该 G_i 以及与该 G_i 重叠的 DDA 目标分类到分割类，其中，T_1 是第一预定

阈值；

对于一个 DDA 目标 D_j，如果 $\sum_{i=1}^{M}\tau_{ij} > T_2$，并且将与之重叠的 DIR 目标合并后得到的合并目标与该 D_j 匹配，则将该 D_j 以及与之重叠的 DIR 目标分类到合并类，其中，T_2 是第二预定阈值；以及

将不属于以上类别的 DDA 目标和 DIR 目标分类到其他类。

19. 如权利要求 18 所述的装置，其中，所述选择单元如下确定融合结果中包括的目标：

将匹配类中的 DDA 目标添加到融合结果中；

将类型错误类中的 DIR 目标的位置信息和相应的 DDA 目标的类型信息结合产生一个新的目标，并把该新的目标添加到融合结果；

将虚警类中所有的目标都添加到融合结果中；

将分割类中的 DIR 目标添加到融合结果中；

对于合并类，如果与 DDA 目标重叠的 DIR 目标都是图片类型的目标，则将相应的 DIR 目标添加到融合结果中；如果与 DDA 目标重叠的 DIR 目标中既包括图片类型的目标又包括文字类型的目标，则将相应的 DDA 目标添加到融合结果中；如果与 DDA 目标重叠的 DIR 目标都是文字类型的目标，则将与 DDA 目标重叠的 DIR 目标合并而成的合并目标添加到融合结果中；以及

将其他类中的 DDA 目标添加到融合结果中。

20. 如权利要求 14 所述的装置，其中，所述 DDA 网页内容提取单元包括：

文档对象模型 (DOM) 树提取单元，其提取输入网页的 DOM 树，并至少保存 DOM 树中每个节点的父节点、子节点、标签名称、内部文字和位置的属性信息；

文字目标提取单元，其利用 DOM 树提取输入网页中的文字目标；

图片目标提取单元，其利用 DOM 树提取输入网页中的图片目标；以及

表格目标提取单元，其利用 DOM 树提取输入网页中的表格目标。

21. 如权利要求 20 所述的装置，其中，根据所述文字目标提取单元包括：

候选文字目标提取单元，对于 DOM 树中的每个节点，如果该节点的内部文字属性不为空，而且该节点的子节点中不包含块节点，则其确定该节点表示的元素为候选文字目标；

合并单元，其参考候选文字目标的属性信息对所确定的候选文字区域执行合并操作，以得到文字目标，

其中，如果节点的标签名称不是“INPUT”、“！”、“A”、“B”、“U”、“I”、“BIG”、“SMALL”、“FONT”、“HR”、“BR”、“PRE”、“TT”、“S”、“BLOCKQUOTE”、“ADDRESS”、“DFN”、“SAMP”、“KBD”、“VAR”、“CODE”、“CITE”、“ABBR”、“ACRONYM”、“SUB”、“SUP”、“INS”、“DEL”、“P”、“EM”、“TEXT”、“STRONG”、“/A”之一，则该节点为块节点。

22. 如权利要求 21 所述的装置，其中，如果两个文字目标在位置上重叠，则所述合并单元将它们合并为一个文字目标。

23. 如权利要求 21 所述的装置，其中，如果一个文字目标被另外一个文字目标包含，则所述合并单元删除被包含的文字目标。

24. 如权利要求 21 所述的装置，其中，如果两个文字目标所对应的矩形区域在垂直方向上位置相邻，并且它们的字体和文字高度属性相同，它们的左边缘相近，而且宽度相似，

则所述合并单元将它们合并为一个文字区域。

25. 如权利要求 20 所述的装置，其中，如果 DOM 树节点的标签名称是“IMG”，并且其尺寸大于第三预定阈值，则所述图片目标提取单元确定该节点为图片目标。

26. 如权利要求 20 所述的装置，其中，如果 DOM 树节点的标签名称为“TABLE”，并且该 DOM 树节点包含至少 3 个“TR”子节点，而且多数的“TR”子节点包含多于一个的“TD”子节点，则所述表格目标确定单元确定该节点为表格区域。

提取网页内容的方法和装置

技术领域

[0001] 本发明涉及网页处理，更具体地说，本发明涉及提取网页内容的装置和方法。

背景技术

[0002] 如今，因特网已经成为最大的信息来源，人们的日常生活越来越依赖于网络。随着网络的普及，网页内容提取（也称为网页分割）的应用越来越广泛。

[0003] 举例来说，网页内容提取可以使得网页搜索的速度更快，结果更加精确。和传统的文本文档相比，网页的内容更加多样化，同一个网页的不同区域可以包含不同的主题。而且，出于浏览和发布的需要，网页中往往包含很多和主题无关的内容，如广告、导航条、装饰、版权信息以及联系方法等。由于网页的以上特征，相较于把整个网页作为一个信息检索单元，通过对网页进行分割，把每个分割单元作为独立的信息检索单元会使网页搜索结果更加精确。并且，通过网页分割可以排除与网页主题无关的内容，从而使网页搜索的速度更快，结果更加精确。

[0004] 再例如，网页内容提取也可用于在手持设备上浏览网页。近年来，手持设备，如掌上电脑、个人数字助理 (PDA)、移动电话等发展迅速。但是使用手持设备上网仍然很大程度上受到显示器过小的限制。传统的网页都是针对个人计算机设计的，对于手持设备用户来说，如果需要不停地滚动网页来寻找所需的信息，上网将变得枯燥和费事。通过网页分割，可以将网页的内容一块一块地显示在手持设备上，从而解决了这一问题。

[0005] 此外，如果用户需要使用已有文档来产生新的文档，则对已有的版面进行分割是必不可少的步骤。

[0006] 由于其广泛的应用背景，用户对于网页内容提取的需求很大。研究人员已经提出了一些用于网页内容提取的系统和方法。

[0007] 例如，美国专利申请公开 No.2006/0149775A1 公开了一种基于文档的可视模型分割文档的方法。在该方法中，根据文档中可视的空白或间隙来确定可视模型，利用该可视模型确定文档的层次结构，并利用所确定的层次结构进行文档分割。但是对于那些逻辑结构和物理结构不一致的文档，该方法容易造成错误分割。

[0008] 再例如，美国专利申请公开 No.2006/0106798A1 公开了一种自上而下的、和标签树无关的用于检测网页结构的方法。该方法基于目标的尺寸、位置、颜色以及背景等，通过投影的方法把文档分成若干块，之后，通过比较块之间的视觉相似程度来判断是否继续分为更小的块或与其他块合并。

[0009] 现有的文档内容提取方法主要可以分为两类。第一类方法专注于文档图像处理，通过图像处理的方法来实现文档内容提取，本文中将其称为文档图像处理 (DIR) 方法。第二类方法专注于文档文件格式分析，通过分析输入文件描述的文档结构来提取内容，本文中将这类方法称为数字文档分析 (DDA)。但是无论是 DDA 方法还是 DIR 方法都有其自己的局限性。

[0010] 参考文献

[0011] 专利文献1：美国专利申请公开No.2003/0215136A1，METHOD ANDSYSTEM FOR DOCUMENT SEGMENTATION，Hui Chao等，2003年11月20日；

[0012] 专利文献2：美国专利申请公开No.2006/0149775A1，DOCUMENTSEGMENTATION BASED ON VISUAL GAPS，Daniel Egnor，2006年7月6日；

[0013] 专利文献3：美国专利申请公开No.2006/0106798A1，VISION-BASEDDOCUMENT SEGMENTATION，Ji-Rong Wen等，2006年5月18日；

[0014] 非专利文献1：JL Fisher，SC Hinds and DP D'amato，"A rule-based systemfor document image segmentation"，Proc.10th ICPR，第567-572页，1990年7月；

[0015] 非专利文献2：Deng Cai，Shipeng Yu，Ji-Rong Wen and Wei-Ying Ma，"Extracting Content Structure for Web Pages based on Visual Representation"，The Fifth Asia Pacific Web Conference(APWeb2003)，2003年。

发明内容

[0016] 本发明提出了一种网页内容提取方法，其融和DDA和DIR方法的网页提取结果，从而产生比两种方法都更优的网页提取结果。本发明可用于网页检索，以及涉及网页分割、网页信息重用的文档解决方案。

[0017] 根据本发明的一个方面，一种用于提取网页内容的方法包括：基于数字文档分析(DDA)方法提取输入网页的网页内容，产生DDA提取结果；基于文档图像识别(DIR)方法提取输入网页的网页内容，产生DIR提取结果；融合所述DDA提取结果和DIR提取结果，产生融合结果。其中，提取结果可以表示为至少一个目标的集合，所述目标代表与网页中的矩形区域相对应的网页内容，所述目标至少包含相应矩形区域的位置信息和该目标的类型信息，并且所述类型包括文字、图片和表格。

[0018] 根据本发明的一方面，所述融合DDA提取结果和DIR提取结果包括：确定DDA目标与DIR目标之间的对应关系；基于DDA目标和DIR目标之间的对应关系以及DDA目标和DIR目标的类型执行DDA提取结果与DIR提取结果的融合。其中，确定DDA目标和DIR目标之间的对应关系包括计算DDA目标与DIR目标的重叠尺度。如果DIR提取结果表示为$G=\{G_1,G_2,\ldots,G_M\}$，DDA提取结果表示为$D=\{D_1,D_2,\ldots,D_N\}$，则DDA目标D_i与DIR目标G_j的重叠尺度通过下式计算：

[0019] $$\sigma_{ij}=\frac{Area(G_i\cap D_j)}{Area(G_i)}\text{以及}\tau_{ij}=\frac{Area(G_i\cap D_j)}{Area(D_j)},\ i=1,2,\ldots,M,\ j=1,2,\ldots,N,$$

其中，$Area(D_j)$是D_j对应的矩形区域的面积，$Area(G_i)$是G_i对应的矩形区域的面积，$Area(G_i\cap D_j)$是D_j对应的矩形区域与G_i对应的矩形区域之间的重叠面积，并且N和M分别是DDA提取结果和DIR提取结果中的目标个数。

[0020] 根据本发明的一方面，执行DDA提取结果与DIR提取结果的融合包括：基于DDA目标和DIR目标之间的对应关系以及目标类型对DDA目标和DIR目标进行分类；以及按照目标的类别来融合DDA提取结果和DIR提取结果，从而产生融合结果。

[0021] 根据本发明的以方面，对于一个DIR目标G_i，如果存在DDA目标D_j，使得$\tau_{ij}\approx 1$并且$\sigma_{ij}\approx 1$，而且G_i和D_j的类型相同，则将该G_i和该D_j分类到匹配类；对于一个DIR目标G_i，如果存在DDA目标D_j，使得$\tau_{ij}\approx 1$并且$\sigma_{ij}\approx 1$，而且G_i和D_j的类型不同，则将该

G_i 和该 D_j 分类到类型错误类；对于一个 DIR 目标 G_i，如果对于所有 DDA 目标均有 $\sigma_{ij} \approx 0$，则将该 G_i 分类到漏检类；对于一个 DDA 目标 D_j，如果对于所有 DIR 目标均有 $\tau_{ij} \approx 0$，则将该 D_j 分类到虚警类；对于一个 DIR 目标 G_i，如果 $\sum_{j=1}^{N} \sigma_{ij} > T_1$，并且将与之重叠的 DDA 目标合并后得到的合并目标与该 G_i 匹配，则将该 G_i 以及与该 G_i 重叠的 DDA 目标分类到分割类，其中，T_1 是第一预定阈值；对于一个 DDA 目标 D_j，如果 $\sum_{i=1}^{M} \tau_{ij} > T_2$，并且将与之重叠的 DIR 目标合并后得到的合并目标与该 D_j 匹配，则将该 D_j 以及与之重叠的 DIR 目标分类到合并类，其中，T_2 是第二预定阈值；以及将不属于以上类别的 DDA 目标和 DIR 目标分类到其他类。

[0022] 根据本发明的一方面，将匹配类中的 DDA 目标添加到融合结果中；将类型错误类中的 DIR 目标的位置信息和相应的 DDA 目标的类型信息结合产生一个新的目标，并把该新的目标添加到融合结果；将虚警类中所有的目标都添加到融合结果中；将分割类中的 DIR 目标添加到融合结果中；对于合并类，如果与 DDA 目标重叠的 DIR 目标都是图片类型的目标，则将相应的 DIR 目标添加到融合结果中；如果与 DDA 目标重叠的 DIR 目标中既包括图片类型的目标又包括文字类型的目标，则将相应的 DDA 目标添加到融合结果中；如果与 DDA 目标重叠的 DIR 目标都是文字类型的目标，则将与 DDA 目标重叠的 DIR 目标合并而成的合并目标添加到融合结果中；以及将其他类中的 DDA 目标添加到融合结果中。

[0023] 根据本发明的一方面，基于 DDA 方法提取输入网页的网页内容包括：提取输入网页的文档对象模型 (DOM) 树，并至少保存 DOM 树中每个节点的父节点、字节点、标签名称、内部文字和位置的属性信息；利用 DOM 树分别提取输入网页中的文字目标、图片目标和表格目标。利用 DOM 树提取文字目标包括：对于 DOM 树中的每个节点，如果该节点的内部文字属性不为空，而且该节点的子节点中不包含块节点，则确定该节点表示的元素为候选文字目标；参考候选文字目标的属性信息对所确定的候选文字区域执行合并操作，以得到文字目标，其中，如果节点的标签名称不是“INPUT”、“！”、“A”、“B”、“U”、“I”、“BIG”、“SMALL”、“FONT”、“HR”、“BR”、“PRE”、“TT”、“S”、“BLOCKQUOTE”、“ADDRESS”、“DFN”、“SAMP”、“KBD”、“VAR”、“CODE”、“CITE”、“ABBR”、“ACRONYM”、“SUB”、“SUP”、“INS”、“DEL”、“P”、“EM”、“TEXT”、“STRONG”、“/A”之一，则该节点为块节点。对候选文字目标执行合并操作包括：如果两个文字目标在位置上重叠，则将它们合并为一个文字目标；如果一个文字目标被另外一个文字目标包含，则删除被包含的文字目标；如果两个文字目标所对应的矩形区域在垂直方向上位置相邻，并且它们的字体和文字高度属性相同，它们的左边缘相近，而且宽度相似，则将它们合并为一个文字区域。

[0024] 根据本发明的一方面，如果 DOM 树节点的标签名称是“IMG”，并且其尺寸大于第三预定阈值，则确定该节点为图片目标。如果 DOM 树节点的标签名称为“TABLE”，并且该 DOM 树节点包含至少 3 个“TR”子节点，而且多数的“TR”子节点包含多于一个的“TD”子节点，则确定该节点为表格区域。

[0025] 根据本发明的另一方面，提供一种用于提取网页内容的装置，包括：数字文档分析 (DDA) 网页内容提取单元，其基于 DDA 方法提取输入网页的网页内容，产生 DDA 提取结果；文档图像识别 (DIR) 网页内容提取单元，其基于 DIR 方法提取输入网页的网页内容，产生 DIR 提取结果；融合单元，其融合所述 DDA 提取结果和 DIR 提取结果，产生融合结果。

附图说明

[0026] 图 1 是示出根据本发明实施例的网页内容提取装置的示例结构的框图；

[0027] 图 2 是示出根据本发明实施例的网页内容提取方法的流程图。

[0028] 图 3 是示出图 1 中的 DDA 网页内容提取单元的示例结构的框图；

[0029] 图 4 示出了网页文件源码及其对应的 DOM 树的例子；

[0030] 图 5 是示出根据本发明示例实施例的 DDA 网页内容提取方法的流程图；

[0031] 图 6 是示出图 1 中的结果融合单元的示例结构的框图；

[0032] 图 7 是示出根据本发明实施例的融合 DDA 和 DIR 提取结果的方法流程图；

[0033] 图 8A-8C 分别示出了 DDA 网页内容提取结果、DIR 网页内容提取结果和融合结果的示例；

具体实施方式

[0034] 下面将参照附图详细描述本发明的示例实施例。附图中，相似的附图标记始终指代相似的元素。

[0035] 图 1 是示出根据本发明实施例的网页内容提取装置 100 的示例性结构的框图。根据本发明的示例实施例，网页内容提取装置 100 包括输入单元 110、DDA 网页内容提取单元 120、网页到图像转换单元 130、DIR 网页内容提取单元 140 以及 DDA 和 DIR 提取结果融合单元 150。输入单元 110 用于输入网页。在本发明的示例实施例中，输入的网页例如可以是超文本标记语言 (HTML) 格式的网页文件。DDA 网页内容提取单元 120 对输入网页进行基于 DDA 方法的网页内容提取处理，产生并输出 DDA 网页内容提取结果。下文中将参照图 3 对 DDA 网页内容提取单元进行更具体地描述。网页到图像转换单元 130 接收输入网页，将其转换成具有与输入网页相同外观的图像文件并输出。DIR 网页内容提取单元 140 对该图像文件进行处理，产生并输出 DIR 网页内容提取结果。这里，DIR 网页内容提取单元可以使用任意的基于图像处理的文档内容提取方法来进行提取。鉴于基于图像处理的文档内容提取方法为公知技术，在此省略对 DIR 网页内容提取单元的详细描述。结果融合单元 150 接收 DDA 和 DIR 网页内容提取结果，对两个结果进行比较，产生并输出融合后的网页内容提取结果。下文中将参照图 6 对结果融合单元 150 进行更详细地描述。在本发明的示例实施例中，网页内容提取结果可以表示为目标集合，该集合中的每一个目标代表网页中一个矩形区域内的网页内容，并且该目标可以包含该相应的网页内矩形区域的位置信息以及类型信息。在本发明的示例实施例中，所述类型可以包括文字、表格和图片。

[0036] 图 2 是示出根据本发明示例实施例的网页内容提取方法的流程图。参照图 2，在步骤 S210 输入网页文件，在步骤 S220 基于 DDA 方法提取输入网页的内容，产生并输出包括至少一个目标（称为 DDA 目标）的 DDA 网页内容提取结果。在步骤 S230 将输入网页转换成具有与输入网页相同外观的图像文件，并在步骤 S240 基于 DIR 方法提取该图像文件的内容，产生并输出包括至少一个目标（称为 DIR 目标）的 DIR 网页内容提取结果。最后，在步骤 S250 将 DDA 提取结果和 DIR 提取结果进行比较，基于 DDA 目标与 DIR 目标的对应关系以及目标类型来融合 DDA 提取结果和 DIR 提取结果，产生新的目标集合作为最终的网页内容提取结果。应当注意，步骤 S220 与步骤 S230-S240 可以以任意次序顺序执行，也可以并行

执行。

[0037] 下面，参照图 3 对 DDA 网页内容提取单元 120 进行具体描述。图 3 是示出根据本发明实施例的 DDA 网页内容提取单元 120 的示例结构的框图。DDA 网页内容提取单元 120 对网页文件结构进行处理，以分别提取文字、表格和图片类型的网页内容（下文中称为文字区域、表格区域和图片区域），并输出 DDA 网页内容提取结果。参照图 3，DDA 网页内容提取单元 120 包括文档目标模型 (DOM) 树提取单元 310、文字区域提取单元 320、图片区域提取单元 340、表格区域提取单元 350 和输出单元 360。

[0038] DOM 树提取单元 310 接收输入网页，提取输入网页的 DOM 树。如上所述，在本发明的示例实施例中，输入的网页可以是超文本标记语言 (HTML) 格式的网页文件。DOM 树是对应于输入网页的树形结构。网页中的每个元素都被表示为该树形结构中的一个节点，并通过不同的路径连接到根节点。图 4 示出了网页文件源码及其对应的 DOM 树的例子。DOM 树提取单元 310 在提取 DOM 树之后，保存网页中每个元素的父节点、子节点、标签名称、内部文字以及位置信息等属性，并使所述属性可以被后续单元访问。注意，在网页的源码中，元素的位置信息并没有被记录，DOM 树提取单元 310 可以考虑特定的网页浏览器，例如微软公司的 Internet Explorer，计算出元素的位置信息。在本发明的示例实施例中，DOM 树提取单元 310 可以借助于微软公司提供的 COM 接口 MSHTML 来计算元素的位置信息。

[0039] 文字区域提取单元 320 利用 DOM 树提取单元 310 提取的 DOM 树来提取文字区域，并将所提取的文字区域输出给输出单元 360。具体来说，文手区域提取单元 320 包括候选文字区域提取单元 321 和候选文字区域合并单元 322。候选文字区域提取单元 321 通过 DOM 树访问每个网页元素，如果该元素的内部文字属性不为空，而且该元素的子节点中不包含块节点，则候选文字区域提取单元 321 确定（提取）该元素为候选文字区域，并将其添加到候选文字区域序列中。这里，如果一个节点的标签名称不是“INPUT”、“！”、“A”、“B”、“U”、“I”、“BIG”、“SMALL”、“FONT”、“HR”、“BR”、“PRE”、“TT”、“S”、“BLOCKQUOTE”、“ADDRESS”、“DFN”、“SAMP”、“KBD”、“VAR”、“CODE”、“CITE”、“ABBR”、“ACRONYM”、“SUB”、“SUP”、“INS”、“DEL”、“P”、“EM”、“TEXT”、“STRONG”、“/A”之一，则该节点被定义为块节点。在访问了每一个网页元素之后，候选文字区域提取单元 321 将生成的候选文字区域序列输出到候选文字区域合并单元 322。候选文字区域合并单元 322 参考候选文字区域的属性信息，对候选文字区域执行合并操作。例如，如果两个文字区域在位置上重叠，则可以将它们合并为一个更大的文字区域。或者，如果一个文字区域被另外一个文字区域包含，则可以删除较小的文字区域。再例如，如果两个文字区域在垂直方向上位置相邻，字体以及文字高度等属性相同，左边缘相近，并且宽度相似，则可以将它们合并为一个更大的文字区域。以上给出了合并单元执行合并操作所遵循的规则的例子，然而本发明不限于此，也可以使用其它规则。文字区域合并单元 322 将合并后的文字区域输出到输出单元 360 以作为文字类型的 DDA 目标。

[0040] 图片区域提取单元 340 利用 DOM 树提取单元 310 提取的 DOM 树来提取图片区域。在本发明的示例实施例中，图片区域提取单元 340 也可以对提取文字区域之后 DOM 树中剩余的元素进行处理来提取图片区域。如果一个元素的标签名称是“IMG”，并且它的尺寸大于预定阈值，则确定该元素为图片区域，并将所确定的图片区域输出到输出单元 360 以作为图片类型的 DDA 目标。

[0041] 表格区域提取单元 350 利用 DOM 树提取单元 310 提取的 DOM 树来提取表格区域。

在本发明的示例实施例中，表格区域提取单元 350 也可以对提取文字区域和图片区域之后 DOM 树中剩余的元素进行处理来提取表格区域。表格区域提取单元 350 可以将标签名称是“TABLE”的元素确定为表格区域。或者，考虑到网页文件中“TABLE”元素经常被用来规范版面，而不是表示真正的表格区域，因此，表格区域提取单元 350 也可以对标签名称为“TABLE”的元素进行进一步地判断以确定表格区域。例如，如果一个元素的标签名称是“TABLE”，并且包含至少 3 个“TR”子节点，而且多数的“TR”子节点包含多于一个的“TD”子节点，则确定该元素为表格区域。表格区域提取单元 350 将所提取的表格区域输出到输出单元 360，以作为表格类型的 DDA 目标。

[0042] 输出单元 360 集合文字区域提取单元 320、图片区域提取单元 340 和表格区域提取单元 350 提取的 DDA 目标，以作为 DDA 网页内容提取结果输出。

[0043] 以上，参照图 3 对 DDA 网页内容提取单元进行了详细描述。然而应当理解，以上描述仅仅是示例性的，而非限制性的。本发明的 DDA 网页内容提取单元也可以具有其它结构，或者采用其它的基于 DDA 的方法来提取网页内容。

[0044] 图 5 是示出根据本发明示例实施例的 DDA 网页内容提取方法的流程图。参照图 5，该 DDA 网页内容提取方法首先在步骤 S510 提取输入网页的 DOM 树。在步骤 S520，利用 DOM 树提取候选文字区域，并在步骤 S530 参考候选文字区域的属性信息对候选文字区域执行合并操作，产生类型为文字的 DDA 目标。在步骤 S540，利用 DOM 树提取图片区域作为图片类型的 DDA 目标。在步骤 S550，利用 DOM 树提取表格区域作为表格类型的 DDA 目标。在步骤 S560，集合文字、图片和表格类型的 DDA 目标以作为 DDA 网页内容提取结果输出。注意，上述步骤 S520-S530、步骤 S540 和步骤 S550 被示为并行执行。然而本发明不限于此，以上步骤也可以以任意次序顺序执行。

[0045] 下面，参考图 6 对结果融合单元 150 进行详细描述。图 6 是示出根据本发明示例实施例的结果融合单元 150 的示例结构的框图。结果融合单元 150 接收 DDA 网页内容提取单元 120 输出的 DDA 网页内容提取结果（以下称为 DDA 提取结果）和 DIR 网页内容提取单元 140 输出的 DIR 网页内容提取结果（以下称为 DIR 提取结果），确定 DDA 目标和 DIR 目标之间的对应关系，基于该对应关系和目标类型融合 DDA 提取结果和 DIR 提取结果，由此产生更优的融合的网页内容提取结果（以下称为融合结果）。如图 6 所示，根据本发明示例实施例的结果融合单元 150 可以包括对应关系确定单元 610 和融合执行单元 620。对应关系确定单元 610 接收 DDA 提取结果和 DIR 提取结果并确定 DDA 目标和 DIR 目标之间的对应关系。在一种实现方式中，DDA 目标和 DIR 目标之间的对应关系可以表示为 DDA 目标和 DIR 目标所对应的矩形区域的重叠尺度。如上所述，网页内容提取结果可以表示为代表网页内容的目标的集合。在这里，将 DIR 提取结果表示为目标集合 $G = \{G_1, G_2, \ldots, G_M\}$，将 DDA 提取结果表示为目标集合 $D = \{D_1, D_2, \ldots, D_N\}$，其中，每个目标 G_i 和每个目标 D_j 均对应于网页中的矩形区域，并且至少包含相应矩形区域的位置信息和类型信息，M 和 N 分别为 DIR 网页内容提取单元 120 和 DDA 网页内容提取单元 140 提取的目标个数。则重叠尺度可以定义如下：

[0046] $$\sigma_{ij} = \frac{Area(G_i \cap D_j)}{Area(G_i)}, \quad i = 1,2,\cdots,M, j = 1,2,\cdots,N \qquad (1)$$

[0047] $$\tau_{ij}=\frac{Area(G_i\cap D_j)}{Area(D_j)},\quad i=1,2,\cdots,M,j=1,2,\cdots,N \qquad (2)$$

[0048] 这里 Area(D_j) 是第 j 个 DDA 目标对应的矩形区域的面积，Area(G_i) 是第 i 个 DIR 目标对应的矩形区域的面积，Area($G_i \cap D_j$) 是第 i 个 DIR 目标和第 j 个 DDA 目标所对应的矩形区域之间的重叠面积。也就是说，对应关系确定单元 610 计算任意 D_j 与 G_i 之间的重叠尺度。

[0049] 融合执行单元 620 根据 DDA 目标和 DIR 目标之间的对应关系以及目标类型来融合 DDA 提取结果和 DIR 提取结果。在一种实现方式中，融合执行单元 621 可以包括分类单元 621 和选择单元 622。分类单元 621 根据重叠尺度和目标类型对 DDA 目标和 DIR 目标进行分类。如上所述，目标类型包括文字、图片以及表格。在本发明的示例实施例中，分类单元 620 可以将 DDA 目标和 DIR 目标分成如下 7 类：

[0050] 1)、对于一个 DIR 目标 G_i，如果存在 DDA 目标 D_j，使得 $\tau_{ij} \approx 1$ 并且 $\sigma_{ij} \approx 1$，而且 G_i 和 D_j 的类型相同（同为文字，图片或表格），则 G_i 和 D_j 被分类到匹配类。

[0051] 2)、对于一个 DIR 目标 G_i，如果存在一个 DDA 目标 D_j，使得 $\tau_{ij} \approx 1$ 并且 $\sigma_{ij} \approx 1$，而且 G_i 和 D_j 的类型不同，则将 G_i 和 D_j 分类到类型错误类。

[0052] 3)、对于一个 DIR 目标 G_i，如果对于所有 DDA 目标，均有 $\sigma_{ij} \approx 0$，也就是说不存在与之重叠的 DDA 目标，则该 G_i 被分类到漏检类。

[0053] 4)、对于一个 DDA 目标 D_j，如果对于所有 DIR 目标，均有 $\tau_{ij} \approx 0$，也就是说不存在与之重叠的 DIR 目标，则该 D_j 被分类到虚警类。

[0054] 5)、对于一个 DIR 目标 G_i，如果$\sum_{j=1}^{N}\sigma_{ij} > T_1$，并且将与之重叠的 DDA 目标合并后得到的合并目标与该 G_i 匹配，则将该 G_i 以及与该 G_i 重叠的 DDA 目标分类到分割类，并将与该 G_i 重叠的 DDA 目标定义为该 G_i 的分割。其中，T_1 是可以由用户根据输入网页的特性以及用户的需求来预先确定的预定阈值，T_1 越小，分类到分割类的目标越多。

[0055] 6)、对于一个 DDA 目标 D_j，如果$\sum_{i=1}^{M}\tau_{ij} > T_2$，并且将与之重叠的 DIR 目标合并后得到的合并目标与该 D_j 匹配，则将该 D_j 以及与之重叠的 DIR 目标分类到合并类，并将 D_j 称为与 D_j 重叠的 DIR 目标的合并。其中，T_2 是可以由用户根据输入网页的特性以及用户的需求来预先确定的预定阈值，T_2 越小，分类到合并类的目标越多。

[0056] 7)、将剩余的不属于以上 6 类的 DDA 目标和 DIR 目标分类为其他类。

[0057] 选择单元 622 根据分类单元 621 的分类结果选择目标以构成融合结果 R = {R_1, R_2, ..., R_L} 并输出，其中，L 为融合结果中目标的个数，从而实现对 DDA 和 DIR 提取结果的融合。在本发明的示例实施例中，选择单元 622 可以对不同的类别采取不同的融合策略。举例来说，对于匹配类中的每对目标，选择单元 622 可以选择对应的 DDA 目标并将其添加到融合结果中；对于类型错误类中的每对目标，选择单元 622 可以将 DIR 目标的位置信息和 DDA 目标的类型信息结合产生一个新的目标，并把该新的目标添加到融合结果中。再例如，选择单元 622 可以简单地忽略漏检类中所有的目标；并且将虚警类中所有的目标都添加到融合结果中。此外，对于分割类，选择单元 622 可以将其中的 DIR 目标添加到融合结果中。对于合并类中的每组目标，可以根据目标类型来选择添加到融合结果中的目标。例如，如果

与 DDA 目标（例如 Dj）重叠的至少一个 DIR 目标都是图片类型的目标，则将相应的至少一个 DIR 目标添加到融合结果中；如果与 DDA 目标重叠的至少一个 DIR 目标中既包括图片类型的目标又包括文字类型的目标，则将相应的 DDA 目标（例如 D_j）添加到融合结果中；如果与 DDA 目标（例如 D_j）重叠的至少一个 DIR 目标都是文字类型的目标，则将相应的至少一个 DIR 目标合并为新的目标，并将该新的目标添加到融合结果中。对于其他类中的目标，选择单元 622 可以将其中的 DDA 目标添加到融合结果中。

[0058] 图 7 是示出根据本发明实施例的融合 DDA 提取结果和 DIR 提取结果的方法流程图。在本发明实施例中，基于 DDA 目标和 DIR 目标的对应关系和类型来对 DDA 提取结果和 DIR 提取结果进行融合。参照图 7，首先，在步骤 S710 接收 DDA 提取结果和 DIR 提取结果。然后在步骤 S720 确定 DDA 目标和 DIR 目标之间的对应关系，该对应关系可以通过计算 DDA 目标和 DIR 目标之间的重叠尺度来确定。在步骤 S730，基于重叠尺度和目标类型来对 DDA 目标和 DIR 目标进行分类。在步骤 S740，基于类别和目标类型来确定包括在最终的融合网页内容提取结果中的目标。

[0059] 图 8A-8C 分别示出了 DDA 网页内容提取结果、DIR 网页内容提取结果和融合结果的示例。例如，从图中可以看出，DDA 提取结果中的 D2 到 D5 过于精细，对应的 DIR 提取结果 G2 更好；而 DIR 提取结果漏掉了右下角的页码信息，该信息被 DDA 方法检测到，为 D13。本发明的网页内容提取装置将 DDA 提取结果和 DIR 提取结果进行融合，从而能够得到更好的网页内容提取结果，如图 8C 所示。

[0060] 以上参照附图对本发明进行了描述。应当理解，以上内容仅仅是示例性的，而非限制性的。本领域技术人员可以在不偏离权利要求书所限定的本发明的精神和范围的前提下，对这里公开的装置和方法进行形式和细节上的各种变更。

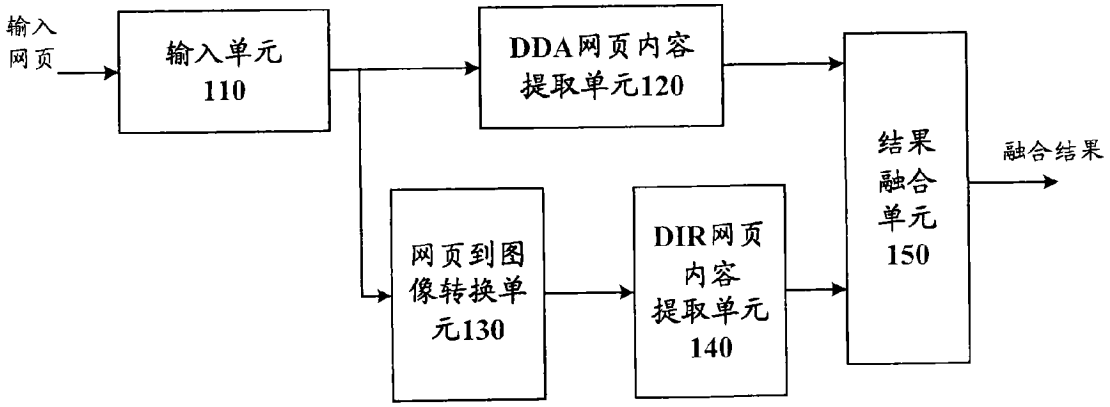

输入
网页
输入单元
110
DDA网页内容
提取单元120
网页到图
像转换单
元130
DIR网页
内容
提取单元
140
结果
融合
单元
150
融合结果

图 1

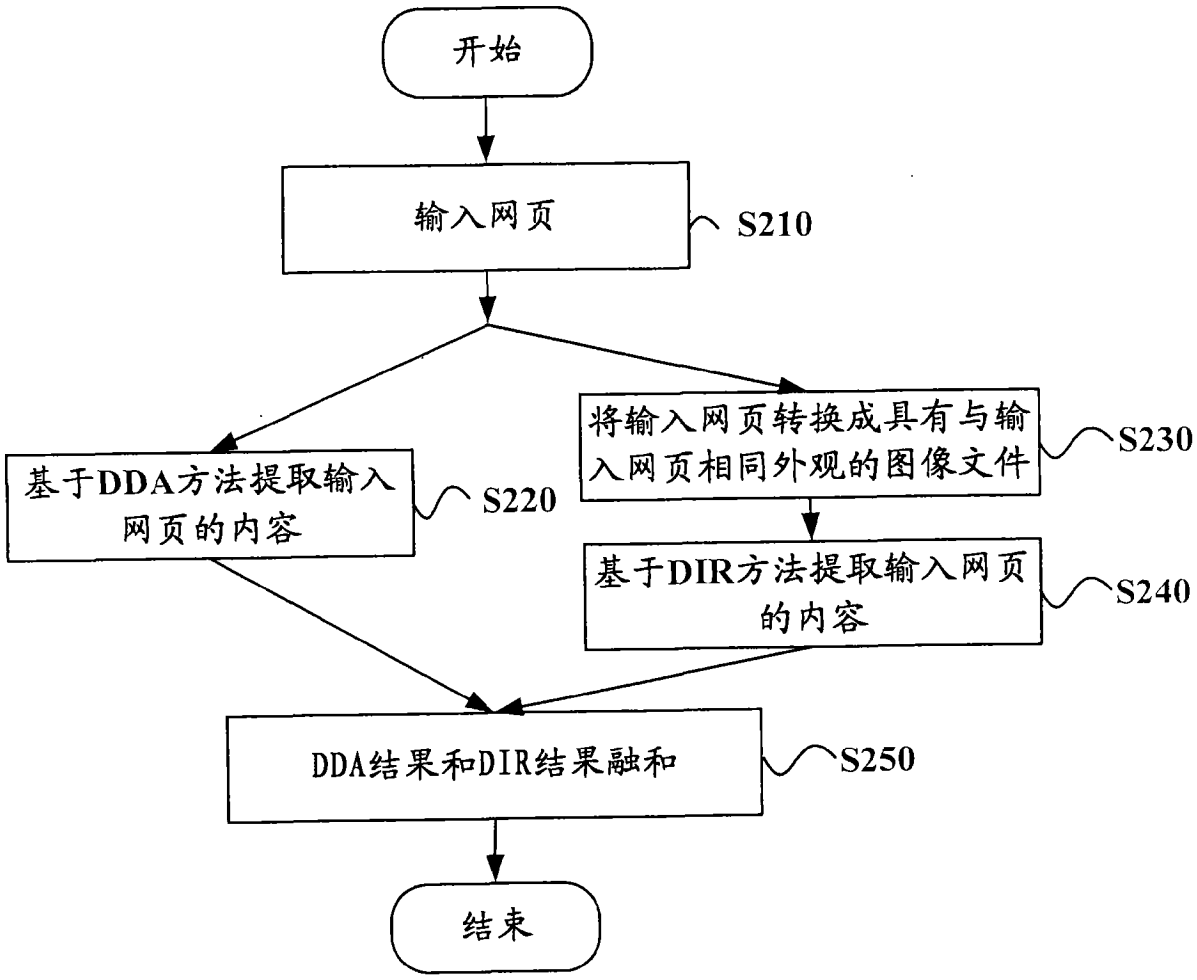

开始
输入网页
S210
基于DDA方法提取输入网页的内容
S220
将输入网页转换成具有与输入网页相同外观的图像文件
S230
基于DIR方法提取输入网页的内容
S240
DDA结果和DIR结果融和
S250
结束

图 2

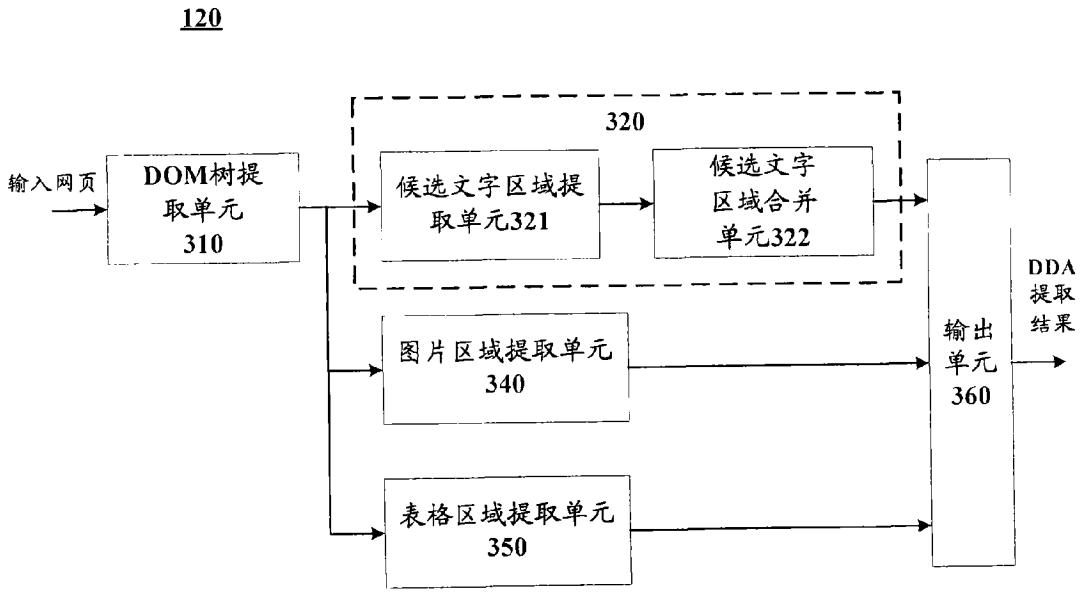
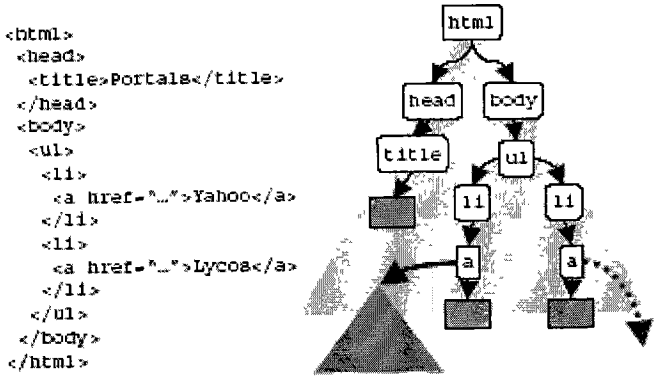

120
320
输入网页
DOM树提取单元310
候选文字区域提取单元321
候选文字区域合并单元322
图片区域提取单元340
表格区域提取单元350
输出单元360
DDA提取结果

图 3


<html>
<head>
<title>Portals</title>
</head>
<body>

Yahoo

Lycos

</body>
</html>
html
head
body
title
ul
li
li
a
a

图 4

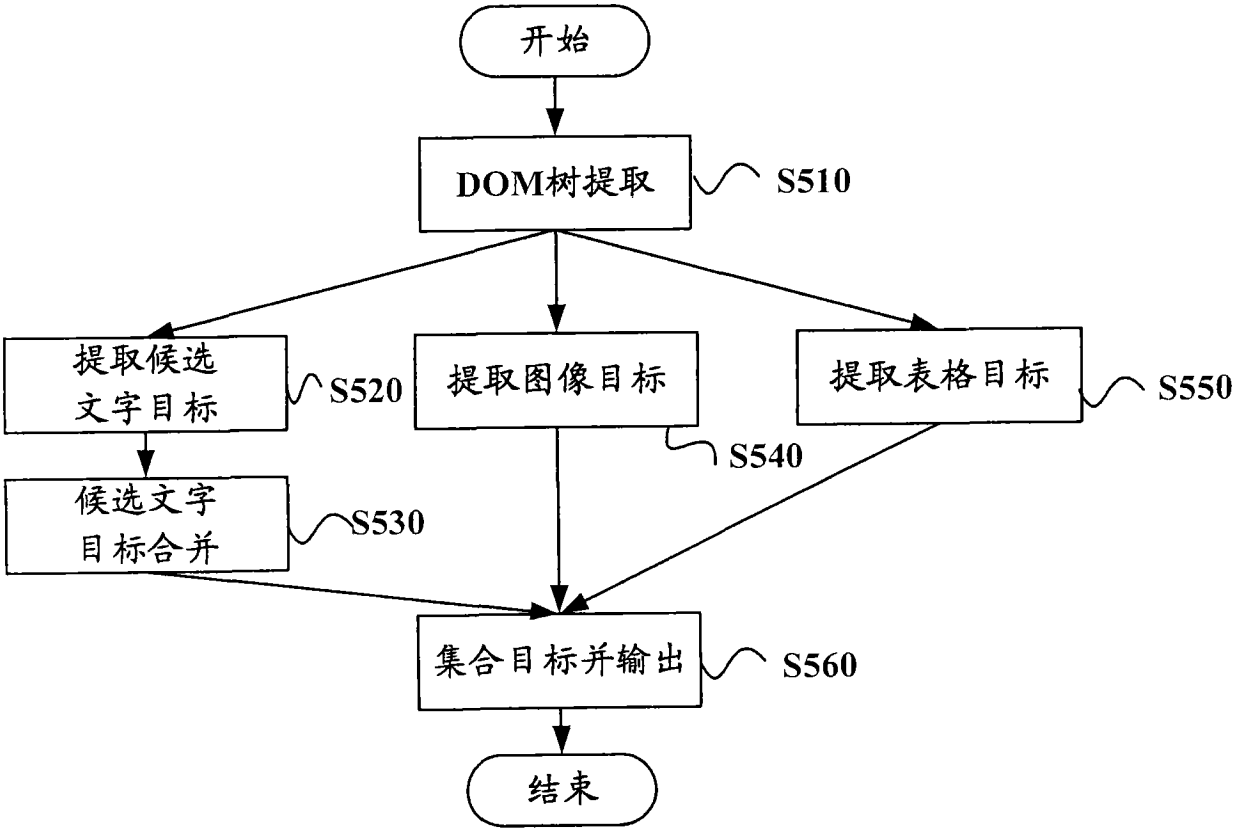

开始
DOM树提取
S510
提取候选文字目标
S520
提取图像目标
S540
提取表格目标
S550
候选文字目标合并
S530
集合目标并输出
S560
结束

图 5

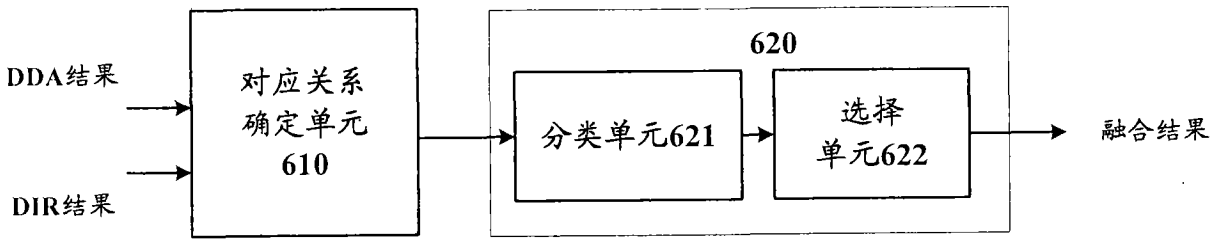


图 6

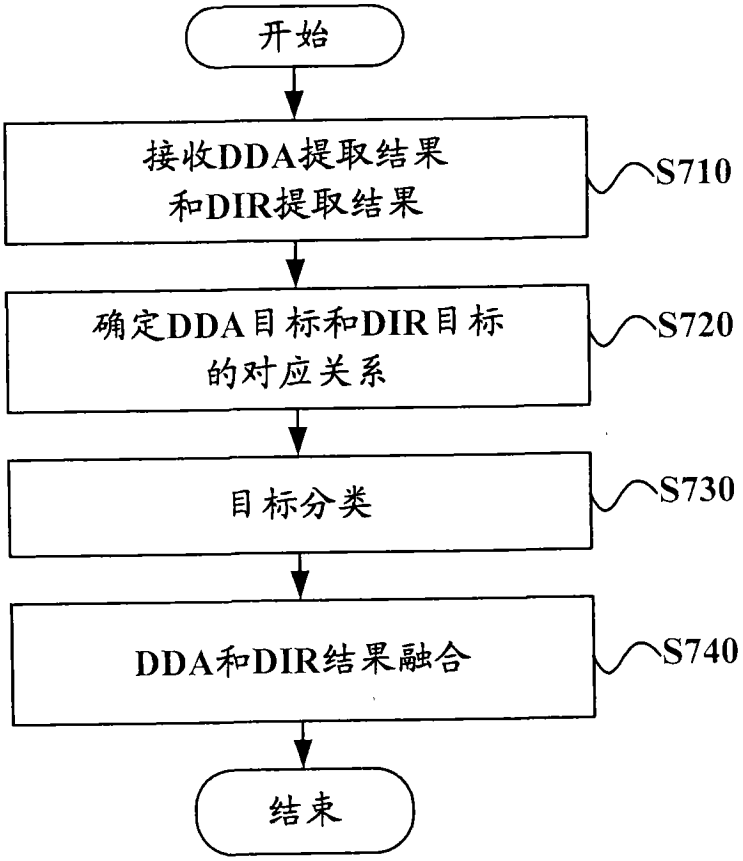

开始
接收DDA提取结果和DIR提取结果
S710
确定DDA目标和DIR目标的对应关系
S720
目标分类
S730
DDA和DIR结果融合
S740
结束

图 7

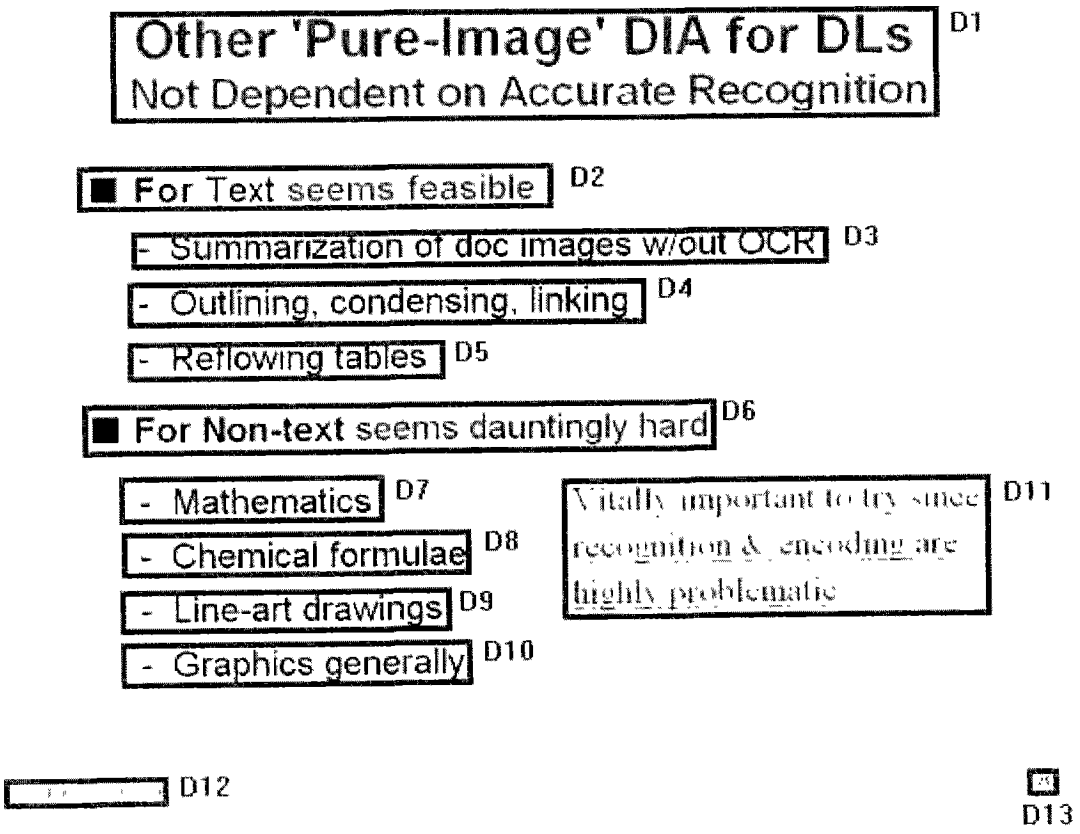

Other 'Pure-Image' DIA for DLs
Not Dependent on Accurate Recognition
D1
For Text seems feasible
D2
- Summarization of doc images w/out OCR
D3
- Outlining, condensing, linking
D4
- Reflowing tables
D5
For Non-text seems dauntingly hard
D6
- Mathematics
D7
- Chemical formulae
D8
- Line-art drawings
D9
- Graphics generally
D10
Vitally important to try since recognition & encoding are highly problematic
D11
D12
D13

图 8A

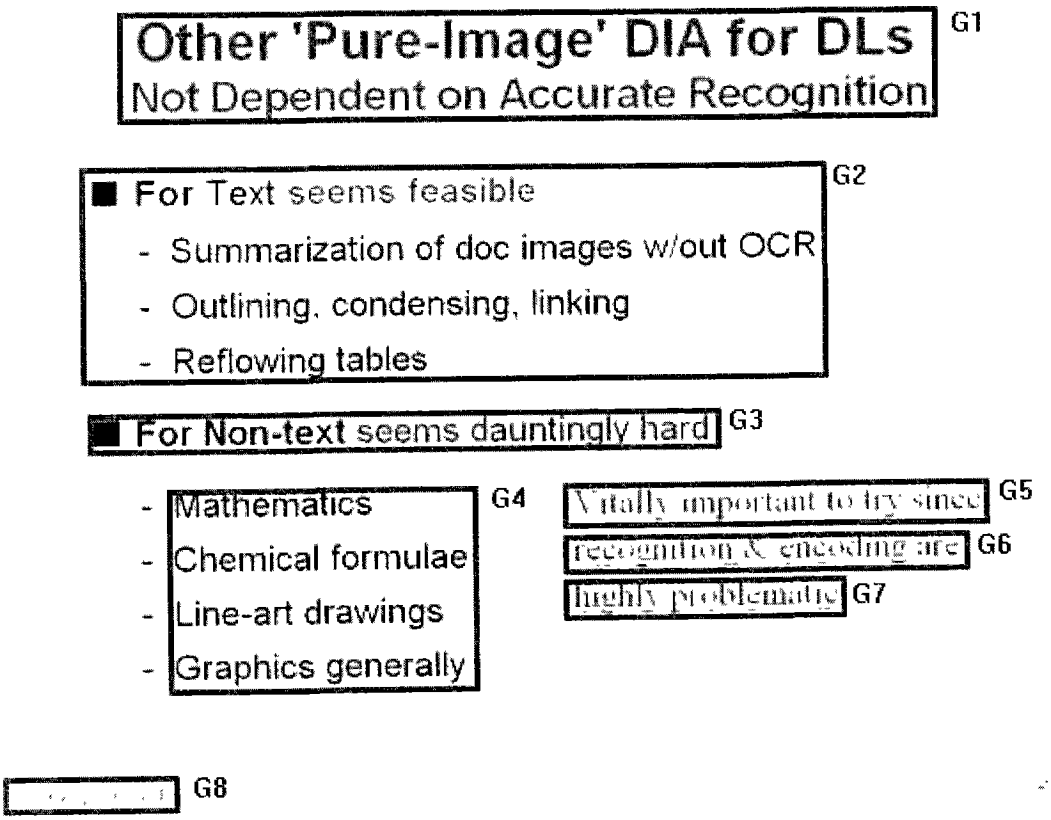

Other 'Pure-Image' DIA for DLs
Not Dependent on Accurate Recognition
G1
For Text seems feasible
- Summarization of doc images w/out OCR
- Outlining, condensing, linking
- Reflowing tables
G2
For Non-text seems dauntingly hard
G3
- Mathematics
- Chemical formulae
- Line-art drawings
- Graphics generally
G4
Vitally important to try since
G5
recognition & encoding are
G6
highly problematic
G7
G8

图 8B

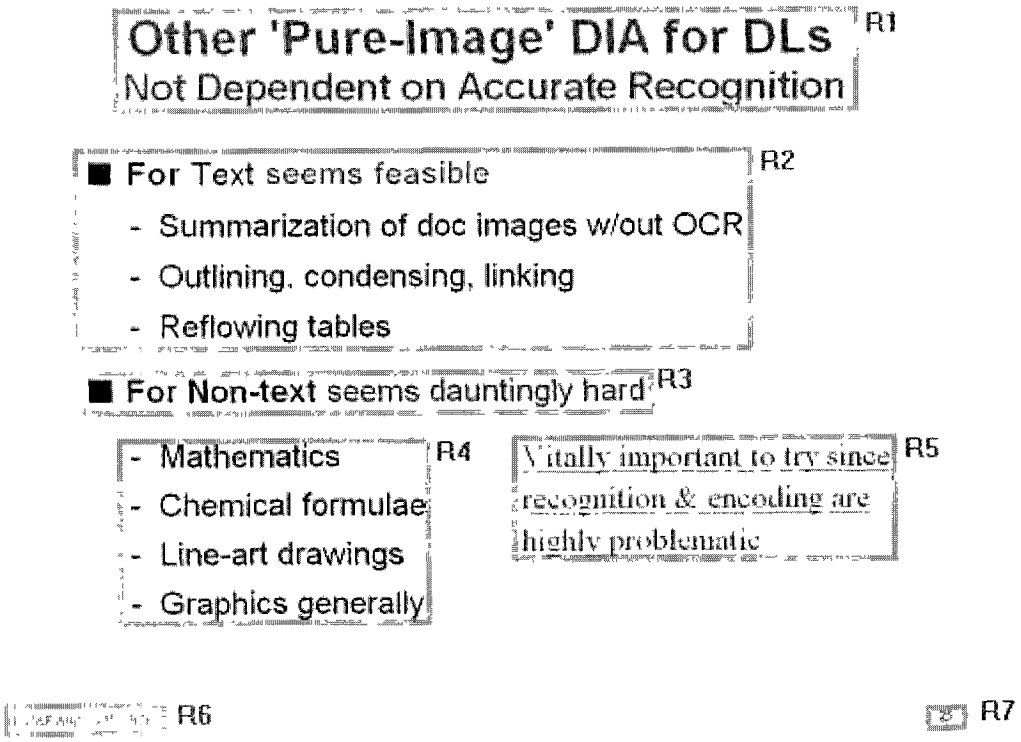

Other 'Pure-Image' DIA for DLs
Not Dependent on Accurate Recognition
R1
For Text seems feasible
- Summarization of doc images w/out OCR
- Outlining, condensing, linking
- Reflowing tables
R2
For Non-text seems dauntingly hard
R3
- Mathematics
- Chemical formulae
- Line-art drawings
- Graphics generally
R4
Vitally important to try since recognition & encoding are highly problematic
R5
R6
R7

图 8C