



[12] 发明专利申请公开说明书

[21] 申请号 200410055697.9

[43] 公开日 2005 年 2 月 9 日

[11] 公开号 CN 1577328A

[22] 申请日 2004.7.28

[21] 申请号 200410055697.9

〔30〕 優先権

[32] 2003. 7.28 [33] US [31] 10/628,766

[71] 申请人 微软公司

地址 美国华盛顿州

[72] 发明人 文继荣 俞诗鹏 蔡 登 马维英

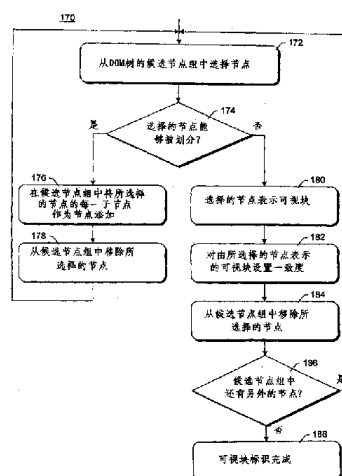
[74] 专利代理机构 上海专利商标事务所
代理人 谢喜堂

权利要求书 12 页 说明书 25 页 附图 11 页

[54] 发明名称 基于视觉的文档分割

[57] 摘要

基于视觉的文档分割标识文档的语义内容的一个或多个部分。所述一个或多个部分通过在所述文档中标识多个可视块，并检测所述多个可视块的可视块之间的一个或多个分隔符来标识。至少部分地基于所述多个可视块和所述一个或多个分隔符对所述文档构造内容结构，并且所述内容结构标识所述文档的语义内容的一个或多个部分。使用基于视觉的文档分割所获得的内容结构能够在文档检索过程中被可任选地使用。



1. 一种标识文档的一个或多个部分的方法，其特征在于，所述方法包括：
在所述文档中标识多个可视块；
- 5 检测所述多个可视块的可视块之间的一个或多个分隔符；以及
 至少部分地基于所述多个可视块以及所述一个或多个分隔符，对所述文档构造内容结构，其中，所述内容结构将所述不同的可视块标识为所述文档的语义内容的不同部分。
2. 如权利要求1所述的方法，其特征在于，所述文档包括 web 页。
- 10 3. 如权利要求1所述的方法，其特征在于，所述文档由具有多个节点的树结构描述，并且其中，在所述文档中标识所述多个可视块包括：
 标识所述多个节点的候选节点组；
 对于所述候选节点组中的每一节点：
 确定所述节点是否能够被划分，以及
- 15 如果所述节点不能被划分，则将所述节点标识为表示一可视块。
4. 如权利要求3所述的方法，其特征在于，如果所述节点不能被划分，则对由所述节点表示的所述可视块设置一致度。
5. 如权利要求3所述的方法，其特征在于，如果所述节点不能被划分，则从所述候选节点组中移除所述节点。
- 20 6. 如权利要求3所述的方法，其特征在于，确定所述节点是否能被划分包括如果所述节点具有带有<HR>超文本标记语言（HTML）标签的子节点，确定所述节点能够被划分。
7. 如权利要求3所述的方法，其特征在于，确定所述节点是否能被划分包括如果所述节点的背景色不同于所述节点的子节点的背景色，确定所述节点能够被划分。
- 25 分。
8. 如权利要求3所述的方法，其特征在于，它还包括检查所述节点是否具有其宽度和高度大于零的子节点，并且如果所述节点没有其宽度和高度大于零的子节点，则从所述候选节点组中移除所述节点。
9. 如权利要求3所述的方法，其特征在于，确定所述节点是否能被划分包括
- 30 如果所述节点的大小比所述节点的子节点的大小的总和大至少一个阈值，确定所述

节点能够被划分。

10. 如权利要求 3 所述的方法，其特征在于，确定所述节点是否能被划分包括如果所述节点具有多个连续子节点，每一子节点具有
超文本标记语言（HTML）标签，确定所述节点能够被划分。

5 11. 如权利要求 1 所述的方法，其特征在于，所述文档由具有多个节点的树结构描述，并且其中，在所述文档中标识所述多个可视块包括至少部分地基于所述多个节点的超文本标记语言（HTML）标签标识不同的可视块。

12. 如权利要求 1 所述的方法，其特征在于，所述文档由具有多个节点的树结构描述，并且其中，在所述文档中标识所述多个可视块包括至少部分地基于所述
10 多个节点的背景色标识不同的可视块。

13. 如权利要求 1 所述的方法，其特征在于，所述文档由具有多个节点的树结构描述，并且其中，在所述文档中标识所述多个可视块包括至少部分地基于所述多个节点是否包括文本以及所述多个节点的大小来标识不同的可视块。

14. 如权利要求 1 所述的方法，其特征在于，检测所述一个或多个分隔符包
15 括：
检测所述可视块之间的一个或多个水平分隔符；以及
检测所述可视块之间的一个或多个垂直分隔符。

15. 如权利要求 1 所述的方法，其特征在于，检测所述一个或多个分隔符包
括：
20 初始化包括所述可视块之间的一个或多个可能分隔符的分隔符列表；
对所述可视块的每一个，分析所述可视块是否与所述分隔符列表的分隔符交迭，并且如果是这样，分析所述可视块如何与所述分隔符交迭；以及
基于所述可视块是否与所述分隔符交迭确定如何处理所述分隔符，并且如果是这样，确定所述可视块如何与所述分隔符交迭。

25 16. 如权利要求 15 所述的方法，其特征在于，它还包括，如果所述可视块包含在所述分隔符内，确定将所述分隔符分裂成多个分隔符。

17. 如权利要求 15 所述的方法，其特征在于，它还包括，如果所述可视块跨越所述分隔符，确定修改所述分隔符的参数使所述可视块不再跨越所述分隔符。

18. 如权利要求 17 所述的方法，其特征在于，所述修改包括如果所述分隔符
30 是水平分隔符，则减小所述分隔符的高度。

19. 如权利要求 17 所述的方法, 其特征在于, 所述修改包括如果所述分隔符是垂直分隔符, 则减小所述分隔符的宽度。

20. 如权利要求 15 所述的方法, 其特征在于, 它还包括, 如果所述可视块覆盖所述分隔符, 确定从所述分隔符列表中移除所述分隔符。

5 21. 如权利要求 1 所述的方法, 其特征在于, 它还包括, 对所述一个或多个分隔符的每一个, 基于所述分隔符的两侧的可视块的特征分配权值。

22. 如权利要求 21 所述的方法, 其特征在于, 分配所述权值包括基于所述分隔符的两侧的两个可视块间的距离分配所述权值。

23. 如权利要求 21 所述的方法, 其特征在于, 分配所述权值包括基于所述分
10 隔符是否与<HR> HTML 标签在同一位置分配所述权值。

24. 如权利要求 21 所述的方法, 其特征在于, 分配所述权值包括基于所述分隔符的两侧的两个可视块中使用的字体大小分配所述权值。

25. 如权利要求 21 所述的方法, 其特征在于, 分配所述权值包括基于所述分隔符的两侧的两个可视块中使用的背景色分配所述权值。

15 26. 如权利要求 1 所述的方法, 其特征在于, 它还包括:

检查所述多个可视块的每一个是否满足一致度阈值; 以及

对于不满足所述一致度阈值的所述多个可视块的每一个, 在所述可视块内标识多个新的可视块, 并且使用所述多个新的可视块重复所述检测和构造。

27. 如权利要求 1 所述的方法, 其特征在于, 构造所述内容结构包括:

20 基于所述多个可视块生成一个或多个虚拟块; 以及

在所述内容结构中包括所述一个或多个虚拟块。

28. 如权利要求 27 所述的方法, 其特征在于, 生成所述一个或多个虚拟块包括通过组合所述多个可视块的两个可视块生成所述一个或多个虚拟块。

29. 如权利要求 27 所述的方法, 其特征在于, 它还包括:

25 对所述一个或多个虚拟块的每一个确定一致度。

30. 如权利要求 29 所述的方法, 其特征在于, 对虚拟块确定所述一致度包括至少部分地基于用来生成所述虚拟块的两个可视块之间的分隔符的权值对所述虚拟块确定所述一致度。

31. 在其上储存了多个指令的一个或多个计算机可读介质, 其特征在于, 当
30 由装置的一个或多个处理器执行所述指令时, 使所述一个或多个处理器:

在文档中标识可视块；

检测所述可视块之间的可视分隔符；以及

至少部分地基于所述可视块和所述可视分隔符对所述文档构造标识表示所述文档的语义内容的文档区域的内容结构。

- 5 32. 如权利要求 31 所述的一个或多个计算机可读媒质，其特征在于，所述文档由具有多个节点的树结构描述，并且其中，促使所述一个或多个处理器在所述文档内标识可视块的所述指令包括使所述一个或多个处理器执行以下行动的指令：

标识所述多个节点的候选节点组；

对所述候选节点组中的每一节点：

- 10 确定所述节点是否能够被划分，以及

如果所述节点不能被划分，则将所述节点标识为表示可视块。

33. 如权利要求 31 所述的一个或多个计算机可读媒质，其特征在于，促使所述一个或多个处理器检测所述可视分隔符的所述指令包括促使所述一个或多个处理器执行以下行动的指令：

- 15 检测所述可视块之间的一个或多个水平分隔符；以及

检测所述可视块之间的一个或多个垂直分隔符。

34. 如权利要求 31 所述的一个或多个计算机可读媒质，其特征在于，促使所述一个或多个处理器检测所述可视分隔符的所述指令包括促使所述一个或多个处理器执行以下行动的指令：

- 20 初始化包括所述可视块之间的一个或多个可能可视分隔符的分隔符列表；

对所述可视块的每一个，分析所述可视块是否与所述分隔符列表的分隔符交迭，并且如果是这样，分析所述可视块如何与所述分隔符交迭；以及

基于所示可视块是否与所述分隔符交迭确定如何处理所述分隔符，并且如果是这样，确定所述可视块如何与所述分隔符交迭。

- 25 35. 如权利要求 31 所述的一个或多个计算机可读媒质，其特征在于，所述指令还促使所述一个或多个处理器：

检查所述可视块的每一个是否满足一致度阈值；以及

对于不满足所述一致度阈值的所述可视块的每一个，在所述可视块中标识新可视块，并且使用所述新可视块重复所述检测和构造。

- 30 36. 一种搜索多个文档的方法，其特征在于，所述方法包括：

接收对应于查询的查询标准;

访问对应于所述多个文档的多个块, 其中, 所述多个块的不同块对应于所述多个文档的不同文档, 其中, 所述多个块通过可视地分割所述多个文档的每一个获得;

- 5 至少部分地基于所述块与所述查询标准匹配到何种程度来对所述多个块的一个或多个生成等级;

对所述多个文档的一个或多个生成等级, 其中, 所述多个文档的每一个的等级至少部分地基于对应于所述文档的所述多个块的等级; 以及

返回所述一个或多个已分级文档的至少一个的指示。

- 10 37. 如权利要求 36 所述的方法, 其特征在于, 所述多个文档的每一个包括 web 页。

38. 如权利要求 36 所述的方法, 其特征在于, 对所述多个文档之一生成所述等级包括:

标识对应于所述一个文档的所述多个块的每一个的等级;

- 15 选择所述标识的等级的最高等级作为所述一个文档的等级。

39. 如权利要求 36 所述的方法, 其特征在于, 对所述多个文档之一生成所述等级包括:

标识对应于所述一个文档的所述多个块的每一个的等级;

组合所述标识的等级来组成所述一个文档的等级。

- 20 40. 如权利要求 39 所述的方法, 其特征在于, 所述组合包括对所述标识的等级求平均值。

41. 如权利要求 36 所述的方法, 其特征在于, 所述可视地分割文档包括:

在所述文档中标识多个可视块;

检测所述多个可视块的可视块之间的一个或多个分隔符; 以及

- 25 至少部分地基于所述多个可视块和所述一个或多个分隔符对所述文档构造内容结构, 其中, 所述内容结构将所述不同的可视块标识为所述文档的语义内容的不同部分, 并且其中, 所述不同的可视块是对应于所述文档的所述多个块的块。

42. 如权利要求 41 所述的方法, 其特征在于, 所述文档由具有多个节点的树结构描述, 并且其中, 在所述文档中标识所述多个可视块包括:

- 30 标识所述多个节点的候选节点组;

对所述候选节点组中的每一节点：

确定所述节点是否能够被划分，以及

如果所述节点不能被划分，则将所述节点标识为表示可视块。

43. 在其上储存了多个指令的一个或多个计算机可读媒质，其特征在于，当
5 由装置的一个或多个处理器执行所述指令时，促使所述一个或多个处理器：

接收包括一个或多个搜索条件的查询；

基于多个块与所述一个或多个搜索条件匹配到何种程度对所述多个块进行分
级，其中，所述多个块的每一个是多个文档的一个文档的部分，并且其中，所述多
个块的每一个通过所述多个文档之一的可视分割获得；

10 对所述多个文档的每一个，至少部分地基于对作为所述文档的部分的所述块
的分级对所述文档进行分级；以及

响应于所述查询，返回所述多个文档的一个或多个的所述等级的指示。

44. 如权利要求 43 所述的一个或多个计算机可读媒质，其特征在于，促使所
述处理器对所述文档进行分级的所述指令包括促使所述一个或多个处理器执行以
15 下行动的指令：

对作为所述文档的部分的每一块标识所述等级；

选择所述标识的等级的最高等级作为所述文档的等级。

45. 如权利要求 43 所述的一个或多个计算机可读媒质，其特征在于，促使所
述一个或多个处理器对所述文档进行分级的所述指令包括促使所述一个或多个处
20 理器执行以下行动的指令：

对作为所述文档的部分的每一块标识所述等级；

组合所述每一块的等级来生成所述文档的等级。

46. 如权利要求 43 所述的一个或多个计算机可读媒质，其特征在于，文档的
所述可视分割包括：

25 在所述文档中标识多个可视块；

检测所述多个可视块的可视块之间的一个或多个分隔符；以及

至少部分地基于所述多个可视块以及所述一个或多个分隔符对所述文档构造
内容结构，其中，所述内容结构将所述不同的可视块标识为所述文档的语义内容的
不同部分，并且其中，所述不同的可视块是作为所述文档的部分的所述多个块的块。

30 47. 如权利要求 46 所述的一个或多个计算机可读媒质，其特征在于，所述文

档由具有多个节点的树结构描述,并且其中,在所述文档中标识所述多个可视块包括:

标识所述多个节点的候选节点组;

对所述候选节点组中的每一节点:

5 确定所述节点是否能够被划分,以及

如果所述节点不能被划分,则将所述节点标识为表示可视块。

48. 一种搜索多个 web 页的方法,其特征在于,所述方法包括:

接收搜索所述多个 web 页的请求;

基于所述请求对所述多个 web 页的子集生成第一等级组;

10 通过可视地分割所述 web 页子集中的每一 web 页对所述 web 页子集生成第二等级组;

至少部分地基于所述第二等级组获取所述 web 页子集的最终等级组。

49. 如权利要求 48 所述的方法,其特征在于,获取所述最终等级组包括使用所述第二等级组作为所述最终等级组。

15 50. 如权利要求 48 所述的方法,其特征在于,获取所述最终等级组包括选择所述第一组中的 web 页的等级以及所述第二组中的 web 页的等级中之较高者作为 web 页的所述最终等级。

51. 如权利要求 48 所述的方法,其特征在于,获取所述最终等级足包括对所述第一组中的 web 页的等级和所述第二组中的 web 页的等级求平均值来获得 web
20 页的所述最终等级。

52. 如权利要求 48 所述的方法,其特征在于,所述可视地分割 web 页包括:
在所述 web 页中标识多个可视块;

检测所述多个可视块的可视块之间的一个或多个分隔符;

25 至少部分地基于所述多个可视块和所述一个或多个分隔符对所述 web 页构造内容结构,其中,所述内容结构将所述不同的可视块标识为所述 web 页的语义内容的不同部分。

53. 如权利要求 52 所述的方法,其特征在于,所述 web 页由具有多个节点的树结构描述,并且其中,在所述 web 页中标识所述多个可视块包括:

标识所述多个节点的候选节点组;

30 对所述候选节点组中的每一节点:

确定所述节点是否能够被划分, 以及

如果所述节点不能被划分, 则将所述节点标识为表示可视块。

54. 在其上储存了多个指令的一个或多个计算机媒质, 其特征在于, 当由装置的一个或多个处理器执行所述指令时, 促使所述一个或多个处理器:

- 5 基于多个文档与搜索标准匹配到何种程度对所述多个文档生成第一等级;
 通过可视地分割所述多个文档的每一个对所述多个文档生成第二等级; 以及
 至少部分地基于所述第二等级对所述多个文档生成最终等级。

55. 如权利要求 54 所述的一个或多个计算机可读媒质, 其特征在于, 促使所述一个或多个处理器生成最终等级的所述指令包括促使所述一个或多个处理器使用所述第二等级作为最终等级的指令。

56. 如权利要求 54 所述的一个或多个计算机可读媒质, 其特征在于, 促使所述一个或多个处理器生成最终等级的所述指令包括促使所述一个或多个处理器选择所述文档的所述第一等级以及所述文档的所述第二等级的任一较高者作为所述多个文档的文档最终等级。

15 57. 如权利要求 54 所述的一个或多个计算机可读媒质, 其特征在于, 促使所述一个或多个处理器生成最终等级的所述指令包括促使所述一个或多个处理器通过对所述文档的所述第一等级以及所述文档的所述第二等级求平均值对所述多个文档的一个文档生成最终等级。

58. 如权利要求 54 所述的一个或多个计算机可读媒质, 其特征在于, 促使所述处理器可视地分割文档的所述指令包括促使所述一个或多个处理器执行以下行动的指令:

 在所述文档中标识多个可视块;

 检测所述多个可视块的可视块之间的一个或多个分隔符; 以及

25 至少部分地基于所述多个可视块和所述一个或多个分隔符对所述文档构造内容结构, 其中, 所述内容结构将所述不同的可视块标识为所述文档的语义内容的不同部分。

59. 如权利要求 58 所述的一个或多个计算机可读媒质, 其特征在于, 所述文档由具有多个节点的树结构描述, 并且其中, 促使所述一个或多个处理器在所述文档中标识所述多个可视块的所述指令包括促使所述一个或多个处理器执行以下行动的指令:

标识所述多个节点的候选节点组；

对所述候选节点组中的每一节点：

确定所述节点是否能够被划分，以及

如果所述节点不能被划分，则将所述节点标识为表示可视块。

5 60. 一种搜索多个文档的方法，其特征在于，所述方法包括：

接收搜索所述多个文档的请求，其中，所述请求包括查询标准；

基于所述查询标准标识所述多个文档的子集；

对所述文档的子集的每一个，通过可视地分割所述文档标识多个块；

基于所述多个块的内容扩展所述查询标准；以及

10 基于所述扩展的查询标准标识所述多个文档的第二子集。

61. 如权利要求 60 所述的方法，其特征在于，响应于所述请求返回文档的所述第二子集的标识符。

62. 如权利要求 60 所述的方法，其特征在于，对所述多个文档的所述第二子集的每一文档进行分级；以及

15 响应于所述请求，返回文档的所述第二子集的标识符以及文档的所述第二子集的每一文档的等级的指示。

63. 如权利要求 60 所述的方法，其特征在于，可视地分割所述文档包括：

在所述文档中标识多个可视块；

检测所述多个可视块的可视块之间的一个或多个分隔符；以及

20 至少部分地基于所述多个可视块和所述一个或多个分隔符对所述文档构造内容结构，其中，所述内容结构将所述不同的可视块标识为所述文档的语义内容的不同部分，并且其中，所述不同的可视块是所述文档的多个块。

64. 如权利要求 63 所述的方法，其特征在于，所述文档由具有多个节点的树结构描述，并且其中，在所述文档中标识所述多个可视块包括：

25 标识所述多个节点的候选节点组；

对所述候选节点组的每一节点：

确定所述节点是否能够被划分，以及

如果所述节点不能被划分，则将所述节点标识为表示一可视块。

65. 在其上储存了多个指令的一个或多个计算机可读介质，其特征在于，当
30 由装置的一个或多个处理器执行所述指令时，促使所述一个或多个处理器：

接收一个或多个搜索条件；

标识满足所述一个或多个搜索条件的多个文档；

在所述多个文档的每一个上执行基于视觉的文档分割来标识所述多个文档的每一个的块；

- 5 基于所述块与所述一个或多个搜索条件匹配到何种程度对所述标识的块的每一个生成等级；

从所述标识的块的一个或多个衍生一个或多个扩展条件；以及

标识满足所述一个或多个搜索条件以及所述扩展条件的另一多个文档。

66. 如权利要求 65 所述的一个或多个计算机可读媒质，其特征在于，促使所述一个或多个处理器衍生所述一个或多个扩展条件的所述指令促使所述一个或多个处理器从一组顶级标识的块衍生所述一个或多个扩展条件。

67. 如权利要求 65 所述的一个或多个计算机可读媒质，其特征在于，促使所述一个或多个处理器执行基于视觉的文档分割的所述指令包括促使所述一个或多个处理器执行以下行动的指令：

- 15 在所述文档中标识多个可视块；

检测所述多个可视块的可视块之间的一个或多个分隔符；以及

至少部分地基于所述多个可视块和所述一个或多个分隔符对所述文档构造内容结构，其中，所述内容结构将所述不同的可视块标识为所述文档的语义内容的不同部分，并且其中，所述不同的可视块是所述文档的块。

- 20 68. 一种系统，其特征在于，它包括：

从文档中提取可视块的可视块提取器；

可视分隔符检测器，耦合来接收所述提取的可视块并基于所述提取的可视块检测所述提取的可视块之间的一个或多个可视分隔符；以及

- 25 内容结构构造器，耦合来接收所述提取的可视块和所述检测的可视分隔符并使用所述提取的可视块和所述检测的可视分隔符来对所述文档构造内容结构。

69. 如权利要求 68 所述的系统，其特征在于，它还包括：

至少部分地基于对多个文档的一个或多个构造的所述内容结构从所述多个文档检索文档的文档检索模块。

70. 如权利要求 68 所述的系统，其特征在于，所述文档由具有多个节点的树结构描述，并且其中，所述可视块提取器通过以下行动从所述文档提取可视块：

标识所述多个节点的候选节点组；

对所述候选节点组中的每一节点：

确定所述节点是否能够被划分，以及

如果所述节点不能被划分，则将所述节点标识为表示可视块。

- 5 71. 如权利要求 68 所述的系统，其特征在于，所述可视分隔符检测器检测所述可视块之间的一个或多个水平分隔符以及所述可视块之间的一个或多个垂直分隔符。

72. 如权利要求 68 所述的系统，其特征在于，所述可视分隔符检测器通过以下行动检测所述一个或多个分隔符：

- 10 初始化包括所述可视块之间的一个或多个可能分隔符的分隔符列表；

对所述多个可视块的每一个，分析所述可视块是否与所述分隔符列表的分隔符交迭，并且如果是这样，分析所述可视块如何与所述分隔符交迭；以及

基于所述可视块是否与所述分隔符交迭确定如何处理所述分隔符，并且如果是这样，确定所述可视块如何与所述分隔符交迭。

- 15 73. 如权利要求 68 所述的系统，其特征在于，所述内容结构构造器还执行：

检查所述多个可视块的每一个是否满足一致度阈值；以及

对不满足所述一致度阈值的所述多个可视块的每一个，将所述可视块返回到所述可视块提取器以从所述可视块提取多个新的可视块，并且还使用所述多个新的可视块令所述可视分隔符检测器检测一个或多个可视分隔符。

- 20 74. 一种系统，其特征在于，它包括：

用于在文档中标识多个可视块的装置；

用于检测所述多个可视块的所述可视块之间的一个或多个分隔符的装置；以及

- 25 用于至少部分地基于所述多个可视块和所述一个或多个分隔符对所述文档构造内容结构的装置，其中，所述内容结构将所述不同的可视块标识为所述文档的语义内容的不同部分。

75. 如权利要求 74 所述的系统，其特征在于，所述文档由具有多个节点的树结构描述，并且其中，用于在所述文档中标识所述多个可视块的装置包括：

用于标识所述多个节点的候选节点组的装置；

- 30 对所述候选节点组中的每一节点：

用于确定所述节点是否能够被划分的装置，以及
如果所述节点不能被划分，用于将所述节点标识为表示可视块的装置。

基于视觉的文档分割

5 技术领域

本发明涉及划分文档，尤其涉及基于视觉的文档分割。

背景技术

人们能够访问大量的信息。然而，在任一给定的情况下找到他们所需的特定
10 信息是相当困难的。例如，通过因特网，大量的信息以 web 页的形式可由人们访问。这类 web 页的数量可以约为 100 万或更多。另外，可用的 web 页经常变化，一些页被添加，其它被删除，其它被修改。

由此，当有人期望找出某一信息，如对问题的答案时，从这一大信息源提取特定的信息的能力变得十分重要。开发了过程和技术来允许用户通过因特网搜索信
15 息，并通常以搜索引擎的方式对用户可用。然而，大部分由于所搜索的 web 页上的内容的极度宽广的范围，这类搜索引擎缺乏准确度。例如，一些 web 页包括版权和其它商业相关注意事项，一些 web 页包括广告。这类商业相关和广告数据不总是与该 web 页的基础内容相关，并且如果考虑这一因素，由此会降低搜索过程的准确度。作为另一示例，不同的 web 页可以在长度上大不相同，一些可包括多
20 个主题，而其它包含单个主题。

web 页的这些特性会降低搜索过程的准确性。因此，具有一种提高搜索文档的准确性的方法是有益的。

发明内容

25 本发明描述了基于视觉的文档分割。

依照一个方面，标识了文档的语义内容的一个或多个部分。该一个或多个部分通过在文档中标识多个可视块，并检测多个可视块的可视块之间的一个或多个分隔符来标识。至少部分地基于该多个可视块和该一个或多个分隔符构造该文档的内容结构，并且该内容结构将不同的可视块标识为该文档的语义内容的不同部分。

30 依照其它方面，在文档检索过程中使用了采用基于视觉的文档分割所获的内

容结构。

附图说明

文档中，使用相同的标号标识相同的组件和/或特征。

5 图 1 所示是使用基于视觉的文档分割的示例系统的结构图。

图 2 所示是用于执行基于视觉的文档分割的示例过程的流程图。

图 3 所示是用于执行可视块标识的示例过程的流程图。

图 4 所示是用于执行可视分隔符检测的示例过程的流程图。

图 5a、5b、5c、5d、5e 和 5f 所示是检测水平分隔符的示例。

10 图 6 所示是用于执行内容结构构造的示例过程的流程图。

图 7 所示是生成可视块和内容结构树的示例。

图 8 所示是用于使用基于视觉的文档分割执行文档检索的示例过程的流程图。

图 9 所示是用于使用基于视觉的文档分割执行文档检索的另一示例过程的流程图。

15 图 10 所示是用于使用基于视觉的文档分割的文档检索的另一示例过程的流程图。

图 11 所示是可用来实现所描述的技术的一般计算环境的示例。

具体实施方式

20 本发明描述了基于视觉的文档分割。基于文档的可视外观，基于视觉的文档分割标识包括了该文档的语义内容的文档的多个部分。基于视觉的文档分割可以以多种不同的方式来使用。例如，当搜索文档时可以使用分割，以使搜索结果基于该文档的语义内容部分。

下文的讨论借助于文档以及用于描述文档结构的模型。文档可以是多种格式
25 中的任一种，诸如依照标准通用标记语言（SGML）如可扩充标记语言（XML）格式或超文本标记语言（HTML）格式。在几个实施例中，这些文档是 HTML 格式的 web 页。这里讨论的模型可以是描述文档的结构的多模型中的任一种。在几个实施例中，使用的模型是文档对象模型（DOM）。“文档对象模型”是文档的树结构表示，也称为 DOM 树。在本发明的许多讨论中，描述文档为 HTML 格式
30 （如，web 页），描述模型为 DOM 树，并且该文档的每一 HTML 标签由 DOM 树

中的一个节点表示（DOM 树也可包括另外的节点，如#text 或#comment 节点，可能不表示 HTML 标签）。然而，可以理解，基于视觉的文档分割也可以用于这些其它文档和/模型。

图 1 所示是使用基于视觉的文档分割的示例系统 100 的结构图。系统 100 包括具有可视块提取器 104、可视分隔符检测器 106 以及内容结构构造器 108 的基于视觉的文档分割模块 102。文档描述对基于视觉的文档分割模块 102 是可访问的。这一文档描述是描述文档的结构的模型，如 DOM 树。文档描述可由另一组件（未示出）生成，并令其对模块 102 可用，或者可选地，文档自身可对模块 102 可访问，并且模块 102 可生成文档描述。

10 可视块提取器 104 基于文档描述标识该文档的多个可视块。这一可视块的标识（这里也称为提取）基于文档内的可视提示，如字体大小和/或类型、字体颜色和/或背景信息、HTML 标签类型等等。标识的块是该文档的区域。标识的块由提取器 104 输出并令其对可视分隔符检测器 106 可用。

15 可视分隔符检测器 106 检测不同的标识块之间的分隔符。这些分隔符可以采用多种不同的形式，如文档中的线、文档中的空白区、对不同块的不同背景颜色等等。在特定的实施例中，检测文档中水平和/或垂直方向上的分隔符。令这些检测的分隔符对内容结构构造器 108 可用。

内容结构构造器 108 基于提取器 104 标识的块以及检测器 106 检测的分隔符对文档生成内容结构。内容结构是表示文档的语义内容的文档的一个或多个部分。
20 内容结构由模块 102 输出。

然后可以以多种方式使用文档的内容结构。在几个实施例中，系统 100 包括可任选文档检索模块 110。文档检索模块 110 使用来自基于视觉的分割模块 102 的内容结构来确定返回哪一文档作为对查询的响应。例如，当用户在搜索 web 页时输入一个或多个搜索条件时，可以使用内容结构来确定哪些 web 页满足标准。

25 文档检索过程中使用内容结构的方式可以变化。在一些实施例中，使用内容结构来为响应于查询而检索的文档分级。另外，在一些实施例中，使用内容结构作为扩展搜索标准的基准。这些使用在下文更详细地讨论。

应当注意，这里描述的基于视觉的文档分割使用当显示文档时它所显现的方式。基于视觉的文档分割不需要实际显示文档，并且基于视觉的文档分割也不需要
30 用户实际看见所显示的文档。

图 2 所示是用于执行基于视觉的文档分割的示例过程 140 的流程图。过程 140 由图 1 的基于视觉的文档分割模块 102 实现，并可以以软件、固件、硬件或其组合执行。

最初，访问文档描述（行动 142）。这一文档描述是例如 HTML web 页的 DOM 5 树。采用了文档描述，标识文档中一个或多个可视块（行动 144）。检测标识的可视块之间的分隔符（行动 146），并基于标识的可视块和检测的分隔符为文档构造内容结构（行动 148）。然后检查内容结构是否满足粒度需求（行动 150）。粒度需求是指内容结构的可视块的一致度（DoC），为每一可视块一致到何种程度的量度。如果不满足粒度需求，则过程 140 返回到行动 144，其中在不满足粒度需求的 10 文档内容中标识新可视块。一旦满足了粒度需求，输出该文档的内容结构（行动 152）。对行动 144、146 和 148 的組合的每次循环（即，标识了可视块、检测了标识的块之间的分隔符并基于标识的可视块和检测的分隔符构造了内容结构）这里称为“轮”。

由此，可以看到，基于视觉的文档分割采用自顶向下的方法。一般而言，标识 15 一组初始的可视块，并如下所述令其对可视分隔符检测和内容结构构造可用。在这些可视块内，在随后的轮中适当地标识另外的可视块（即，对不满足粒度需求的每一可视块在随后的轮中标识另外的可视块）。

如这里所讨论的，假定使用的文档在显示时在形状上为矩形。由此，参考矩形或笛卡尔坐标系（如，X，Y 坐标系）描述一些大小、维数、面积等等。可以 20 理解，这些仅作参考，并且矩形形状和坐标系可以不同。在一些实例中，如果使用了非矩形形状，则在执行基于视觉的文档分割之前将其转化为矩形。在其它实例中，在此描述的过程和技术适应于使用这些不同的形状和/或坐标系。

可视块标识

25 可视块标识，这里也称为可视块提取，基于可视提示标识了文档的不同区域。然后输出这些不同区域的指示并在可视分隔符检测中使用，如后文所详细描述。可视块标识（如，在图 2 的行动 144 中或由图 1 的可视块提取器 104 实现）能够如下执行。

一般而言，文档的 DOM 树中的每一节点能够表示一个可视块。一些大节点 30 （如，具有<TABLE>或<P>标签）通常仅用于组织目的，并且不适合表示单个可视

块。因此，这类大节点被进一步划分并由其子节点替代。另一方面，由于叶节点潜在的巨大数量，不应当最初将可视块标识为 DOM 树中的每一叶节点（尽管由于基于视觉的可视文档分割的自上而下特性，一些这样的叶节点最终可被标识为可视块）。

5 图 3 所示是用于执行可视块标识的示例过程 170 的流程图。过程 170 以额外的细节说明了图 2 的行动 144。过程 170 可以软件、固件、硬件或其组合执行。

10 最初，从 DOM 树的候选节点组中选择一个节点（行动 172）。这一候选节点组指在当前轮中为潜在可视块的 DOM 树的节点。当执行过程 170 时可以改变候选节点，如后文更详细讨论的。在特定的实施例中，候选节点组最初为单个节点，单个节点在第一轮中为 DOM 树的根节点或者在随后的轮中为 DOM 树的子树的顶部节点（如，对应于不满足粒度需求的可视块的 DOM 树的一部分的顶部节点）。可选地，最初的候选节点组可以不同，如 DOM 树中根节点以下一层的节点或 DOM 树的子树的顶部节点以下一层的节点。

15 然后检查是否能划分选择的节点（行动 174）。如果能划分选择的节点，则向候选节点组添加选择的节点的每一子节点（行动 176），并且将选择的节点从候选节点组中移除（行动 178）。过程 170 然后返回到行动 172 并从候选节点组中选择另一节点。

20 返回到行动 174，如果选择的节点不能被划分，则选择的节点表示一个可视块（行动 180），并将其添加到由过程 170 标识的可视块池（pool）中。对该可视块设置一致度（DoC）（行动 182），如后文更详细讨论的，并将选择的节点从候选节点组中移除（行动 184）。还检查候选节点组中是否有另外的节点（行动 186）。如果候选节点组中有另外的节点，则过程返回到行动 172，从候选节点组中选择一个节点。如果候选节点组中没有另外的节点，则完成可视块标识过程 170（行动 188），行动 180 中选择的可视块为过程 170 标识的可视块。然后可以在可视分隔符检测中使用过程 170 标识的可视块池，如后文更详细讨论的。

25 基于一个或多个规则确定一个特定节点是否能被划分（行动 174 中）。这些规则利用关于该节点以及可能 DOM 树中一个或多个其它节点的一个或多个提示或信息。这类提示或信息的示例包括标签提示、颜色提示、文本提示和大小提示。可以基于这些提示来指定的规则的示例如下。在特定的实施例中，也可以包括缺省规则（如，该节点可以被划分，或者作为替换该节点不能被划分）。如果没有其它

规则可用，则使用该缺省规则。

顶块规则：顶块规则基于评估的树中节点的位置。顶块规则规定如果该节点是评估的树中的顶部节点（如，第一轮中 DOM 树的根节点），则该节点能够被划分。

- 5 标签规则：标签规则基于该节点的 HTML 标签。一些 HTML 标签如<HR>标签，经常用来从可视的观点分隔不同的主题。标签规则规定，如果该节点的子节点之一的 HTML 标签不等于一组特定的标签之一（如，该 HTML 标签为<HR>标签），则该节点能够被划分。

- 10 颜色规则：颜色规则基于该节点的背景色以及该节点的子节点的背景色。颜色规则规定，如果该节点的背景色不同于其至少一个子节点的背景色，则该节点能够被划分。

- 15 文本和大小规则：可以使用若干文本和大小规则。在这些规则的一些中，对节点的高度、宽度和/或大小作出参考。节点的高度指当显示一个块时由该节点表示的块的可视高度（如，使用 X, Y 坐标系统以像素表示），节点的宽度指当显示一个块时由节点表示的块的可视宽度（如，使用 X, Y 坐标系统以像素表示），节点的面积表示当显示一个块时由节点表示的块所占用的显示面积（如，使用 X, Y 坐标系统以像素表示）（如，在矩形块的情况下为该节点的宽度和高度的乘积）。可以基于包括在 DOM 树的节点中的位置定位信息容易地确定这些大小。

- 20 这些规则的其中一些（以及本描述中其它地方的讨论）参考有效节点、块节点、文本节点、虚拟文本节点和/或导航节点。有效节点指当显示 HTML 文档时可见的节点（如，高度和宽度都不等于零）。块节点指具有不同于<A>、、、<HR>、<I>、<P>、、或<TEXT>的某一标签的节点。文本节点指仅包含自由文本的节点。虚拟文本节点指非块节点且仅具有文本节点作为（多个）子节点的节点。导航节点指其高度大于宽度的两倍，且宽度小于 200 的节点。在特定的实现中，关于导航节点的规则在第一轮中使用，但是在随后的轮中不使用。

25 可以对具有一般标签（即，对于除<P>标签、<TABLE>标签、<TD>标签和标签以外的所有 HTML 标签）的节点使用以下基于文本和大小提示的规则：

- 如果该节点没有有效子节点，则该节点不能被划分，并且将该节点从候选节点组中移除；
- 30 • 如果该节点只有一个有效子节点，并且该子节点不是文本节点，则跟踪到

该子节点中（从候选节点组中移除该节点并且用其子节点来替换它）；

- 如果该节点的大小比其子节点大小的总和大至少三倍，则该节点能够被划分；

- 如果该节点的所有子节点是文本节点或虚拟文本节点，则该节点不能被划分；

- 如果该节点具有至少一个文本节点子节点或者至少一个虚拟文本节点子节点，并且该节点的宽度或高度小于一阈值（如，150），则该节点不能被划分；

- 如果该节点为导航节点，则该节点不能被划分；

- 如果该节点具有至少一个文本节点子节点或者至少一个虚拟文本节点子节点，并且还不具有块节点子节点，则该节点不能被划分；

- 如果该节点具有多于两个连续
子节点，则该节点不能被划分；以及

- 否则，该节点能够被划分。

可以对具有<P> HTML 标签的节点使用以下基于文本和大小提示的规则：

- 如果该节点没有有效子节点，则该节点不能被划分，并且将该节点从候选节点组中移除；

- 如果该节点只有一个有效子节点并且该子节点不是文本节点，则跟踪到该子节点（从候选节点组中移除该节点，并且用其子节点替换它）；

- 如果该节点的所有子节点是文本节点或虚拟文本节点，则该节点不能被划分；

- 如果该节点的所有子节点的大小的总和大于该节点的大小，并且至少一个子节点为块节点，则该节点能够被划分；

- 如果该节点具有至少一个文本节点子节点或者至少一个虚拟文本节点子节点，并且该节点的宽度或高度小于一阈值（如，150），则该节点不能被划分；

- 如果该节点没有块节点子节点，则该节点不能被划分。

- 如果该节点的所有子节点小于一门限（如，宽度<150 且高度<100），则该节点不能被划分；

- 否则，该节点能够被划分。

可以对具有<TABLE> HTML 标签的节点使用以下基于文本和大小提示的规则：

- 如果该节点没有有效子节点，则该节点不能被划分，并且从候选节点组中

移除该节点；

- 如果该节点只有一个有效子节点且该子节点不是文本节点，则跟踪到该子节点（从候选节点组中移除该节点并用其子节点替换它）；

- 如果该节点的最大子节点的大小小于一阈值（如，250*200），则该节点不能被划分；

- 如果该节点的所有子节点小于一阈值（如，宽度<150 或高度<100），则该节点不能被划分；

- 如果该节点的所有子节点大于一阈值（如，宽度>150 且高度>100），则该节点不能被划分；

10 • 如果该节点的一些（大于 0）有效子节点大于一阈值（如，宽度>150 且高度>100）并且该节点的一些（大于 0）有效子节点小于一阈值（如，宽度<150 且高度<100），则该节点不能被划分；

- 否则，该节点能够被划分。

可以对具有<TD> HTML 标签的节点使用以下基于文本和大小提示的规则：

15 • 如果该节点没有有效子节点，则该节点不能被划分，并且从候选节点组中移除该节点；

- 如果该节点只有一个有效子节点并且该子节点不是文本节点，则跟踪到该子节点（从候选节点组中移除该节点并用其子节点替换它）；

- 否则，该节点不能被划分；

20 可以对具有 HTML 标签的节点使用以下基于文本和大小提示的规则：

- 如果该节点没有有效子节点，则该节点不能被划分，并且从候选子节点中移除该节点；

- 如果该节点只有一个有效子节点，并且该子节点不是文本节点，则跟踪到该子节点（从候选节点组中移除该节点并用其子节点替换它）；

25 • 如果该节点没有块节点子节点，则该节点不能被划分；

- 如果该节点的所有子节点具有 HTML 标签，则该节点不能被划分；

- 如果该节点的所有子节点小于一阈值（如，宽度<150 且高度<100），则该节点不能被划分。

- 否则，该节点能够被划分。

30 在可视块标识中，一致度 (DoC) 值是分配给从 DOM 树提取的可视块的量度。

也可以在内容结构构造中对可视块生成不同的一致度 (DoC) 值, 如后文更详细讨论的。可视块 (或虚拟块) 的 DoC 值是该可视块 (或虚拟块) 一致到何种程度的测量。

在可视块标识中分配给块的 DoC 值可以以多种不同的方式分配。在几个实施
5 例中, 基于 DOM 树中其相应的节点 (表示该可视块的节点) 分配可视块的 DoC 值, 并且可以基于该节点的 HTML 标签而变化。以下是在可视块标识中分配 DoC 时可以遵循的一组示例规则。在以下示例中, DoC 值为从 1 到 10 范围内的整数, 尽管可以可选地使用不同的范围 (如, 由整数、实数组成等等)。会出现以下规则的多个应用到特定的块, 并且不能导致该块的多个 DoC 值的情况。这一情况可以
10 以多种不同的方式来解决, 如选择由规则之一分配的最大 DoC 值、选择由规则之一分配的最小 DoC 值、平均由规则分配的 DoC 值、计算由规则分配的 DoC 值的加权平均值等等。

以下示例规则用来向具有一般 HTML 标签 (即, 对除<P>标签、<TABLE>标签、<TD>标签和标签之外的所有 HTML 标签) 的可视块分配 DoC 值:

- 15 • 如果该节点的所有子节点是文本节点或虚拟文本节点, 则设置该块的 DoC 值为 10;
- 如果该节点具有至少一个文本节点子节点或至少一个虚拟文本节点子节点, 并且该节点的宽度或高度小于一阈值 (如, 150), 则设置该块的 DoC 值为 8;
- 如果该节点是导航节点 (如, 仅在第一轮中, 其高度大于其宽度两倍, 并且其宽度小于 200 的节点; 在随后的轮中, 不应用该规则), 则设置该块的 DoC
20 值为 7;
- 如果该节点具有至少一个文本节点子节点或至少一个虚拟文本节点子节点, 并且还不具有块节点子节点, 则基于该节点的大小设置该块的 DoC 值如下:
 - o 如果该节点的宽度和高度都小于一第一阈值 (如, 100), 则设置该块
25 的 DoC 值为 8;
 - o 如果该节点的高度小于一第二阈值 (如, 200) 或者宽度小于一第三阈值 (如, 400), 则设置该块的 DoC 值为 7; 以及
 - o 如果该节点的面积 (如, 该节点的宽度和高度的乘积) 小于一第四阈值 (如, 100,000), 则设置该块的 DoC 值为 6; 以及
- 30 • 如果上述规则没有一个能够应用到具有一般 HTML 标签的块, 则设置该块

的 DoC 值为 5。

使用以下规则对具有<P>HTML 标签的可视块分配 DoC 值：

- 如果该节点的所有子节点是文本节点或虚拟文本节点，则设置块的 DoC 值为 10；
- 5 • 如果该节点具有至少一个文本节点子节点或者至少一个虚拟文本节点子节点，并且该节点的宽度或高度小于一阈值（如，150），则设置该块的 DoC 值为 8；
- 如果该节点不具有块节点子节点，则设置该块的 DoC 值为 7；以及
- 如果该节点的所有块节点子节点小于一阈值（如，宽度<150 且高度<100），则：
- 10 o 如果该节点具有至少一个文本节点子节点或者至少一个虚拟文本节点子节点，则设置该块的 DoC 值为 7；以及
- o 否则，设置该块的 DoC 值为 6。

使用以下规则对具有<TABLE> HTML 标签的可视块分配 DoC 值：

- 如果该节点的最大子节点小于一阈值（如，250*200），则设置该块的 DoC 15 值为 8；
- 如果该节点的所有子节点小于一阈值（如，宽度<150 且高度<100），则设置该块的 DoC 值为 7；以及
- 如果该节点的所有有效节点子节点大于一阈值（如，宽度>150 且高度>100），则设置该块的 DoC 值为 6。

20 使用以下示例规则对具有<TD> HTML 标签的可视块分配 DoC 值：

- 如果该节点的所有子节点是文本节点或虚拟文本节点，则设置该块的 DoC 值为 10；
- 如果该节点具有至少一个文本节点子节点或者至少一个虚拟文本节点子节点，并且该节点的宽度或高度小于一阈值（如，150），则设置该块的 DoC 值为 8；
- 25 • 如果该节点不具有块节点子节点，则设置该块的 DoC 值为 7；
- 如果该节点的所有块节点子节点小于一阈值（如，宽度<150 且高度<100），则设置该块的 DoC 值为 7；以及
- 如果该节点的最大子节点的大小小于一阈值（如，250*200），则设置该块的 DoC 值为 7。

30 使用以下示例规则对具有 HTML 标签的可视块分配 DoC 值：

- 如果该节点不具有块节点子节点，则设置该块的 DoC 值为 8;
 - 如果该节点的所有子节点具有 HTML 标签，则设置该块的 DoC 值为 8;
- 以及
- 如果该节点的所有子节点小于一阈值（如，宽度<150 且高度<100），则设置该块的 DoC 值为 8。

例如，可以使用具有 DivideDomtree 程序或函数（以下表 I 中示出），以及 Dividable 程序或函数（以下表 II 中示出）的算法来实现图 3 的规则和过程 170 的示例。在该算法中，值 pRoot 指当前选择的节点，值 nLevel 指 DOM 树中当前选择的节点的层，池指由该算法标识的可视块的组，顶块指 DOM 树或者本轮中评估的子树的顶节点（如，在第一轮中，顶块指 DOM 树的根节点），child 指当前选择的节点的子节点。以上描述了 Dividable 程序或函数的具体例程和启发式规则。每次调用 DivideDomtree 程序或函数时，向其传递 DOM 树的一个节点作为 pRoot 的值，并且传递 DOM 树中该节点的层的值作为 nLevel 的值。

15

表 I

算法 DivideDomtree(pRoot, nLevel)
{
IF(Dividable(pRoot, nLevel)==TRUE){
对 pRoot 的每一 child{
DivideDomtree(child, nLevel);
}
}ELSE{
将子树（pRoot）放入池中作为一个块;
}
}

表 II

算法 Dividable(pRoot, nLevel)
{
IF(pRoot 是顶块){
RETURN TRUE;
}ESLE{
对 TABLE、TD、P、UL 的特殊例程;
对一般标签的启发式规则;
}
}

可视分隔符检测

5 可视分隔符检测检测由可视块标识所标识的不同块之间的分隔符。这些分隔符可以是各种不同的形式，如文档中的线、文档中的空白区、对不同块的不同背景色等等。在几个实施例中，检测文档中水平和/或垂直方向上的分隔符。可视分隔符检测（如，图 2 的行动 146 中或由图 1 的可视分隔符检测器 106 实现）可以如下执行。

10 检测的分隔符是文档中不可视地与由可视块标识所标识的块的池中的任何块交叉的水平或垂直区域。这些分隔符用作用于区别文档内不同语义的标识符。分隔符由 2 元组(P_s, P_e)表示，其中， P_s 是该分隔符的开始像素， P_e 是该分隔符的终止像素。每一像素 P_s 和 P_e 被标识为 X, Y 坐标对，并且通过将 P_s 和 P_e 设置为矩形的对角（如，左下角和右上角、左上角和右下角等等）能够确定矩形区域。计算分隔符的宽度为 P_s 和 P_e 之间的差（如，在 X 方向）。

15 对水平和垂直方向作出参考。通常，在 X, Y 坐标系中，水平方向指 X 轴而垂直方向指 Y 轴。然而，也可以作为替换使用对水平和垂直方向的其它定向。

图 4 所示是用于执行可视分隔符检测的示例过程 200 的流程图。过程 200 以额外的细节说明了图 2 的行动 146。过程 200 可以以软件、固件、硬件或其组合实现。

20 最初，使用一个分隔符初始化分隔符列表（行动 202）。该一个分隔符包括可能潜在地为（多个）分隔符的文档的显示范围的所有区域。在几个实施例中，这一

初始分隔符包括该文档的整个范围，或者可选地，该分隔符可以是包括由虚拟块标识所标识的可视块的池中的所有块的单个矩形区域。

然后从标识的块的池中选择一个块（行动 204）。可以以任一多种方式从池中选择块（如，随机地、以它们添加到池的顺序、通过增加或减小大小等等）。然后

5 检查所选择的块是否包含在分隔符列表的分隔符中（行动 206）。如果一个块的整个范围都包括在分隔符的范围内，则认为该块包含在分隔符内。如果所选择的块包含在分隔符内，则将该分隔符分裂成两个分隔符（行动 208）。可以以不同的方式完成这一分裂，如，通过从分隔符列表中移除包含块的分隔符并向分隔符列表添加两个新的分隔符、修改该分隔符的 P_s 或 P_e 像素并添加一个新分隔符，等等。分隔

10 符的分裂导致先前为单个分隔符的分隔符变成该块的两侧的两个较小的分隔符。当检测水平分隔符时，两个较小的分隔符将在块的上方和下方，当检测垂直分隔符时，两个较小的分隔符将在块的左边和右边。

当在行动 208 分裂分隔符之后，或者如果所选择的块不包含在分隔符中，检查所选择的块是否跨越了分隔符（行动 210）。当检测水平分隔符时，如果一个块

15 的范围与分隔符的一部分交叉但是小于该分隔符的整个高度，则认为该块跨越该分隔符。当检测垂直分隔符时，如果一个块的范围与分隔符的一部分交叉但是小于该分隔符的整个宽度时，则认为该块跨越了该分隔符。

如果所选择的块跨越了分隔符，则更新该分隔符的参数，使该块不再跨越该分隔符（行动 212）。这一更新参数指修改该分隔符的 P_s 和/或 P_e 像素，使该块不

20 再跨越该分隔符。

在行动 212 中更新分隔符的参数之后，或者如果选择的块不跨越分隔符，检查所选择的块是否覆盖了分隔符（行动 214）。当检测水平分隔符时，如果一个块的范围与分隔符的至少一部分的整个高度交叉，认为该块覆盖了该分隔符。当检测垂直分隔符时，如果一个块的范围与分隔符的至少一部分的整个宽度交叉，则认为

25 该块覆盖了该分隔符。

如果所选择的块覆盖分隔符，则将该分隔符从分隔符列表中移除（行动 216）。在行动 216 中从分隔符列表中移除了该分隔符之后，或者如果选择的块不覆盖分隔符，则检查标识的块的池中是否还有尚未选择的另外的块（行动 218）。如果在块的池中有这样的块，则过程返回以在行动 204 中选择这些剩余块中的一个。然而，

30 如果在块的池中没有这样的块，则从分隔符列表中移除显示的边缘的四个分隔符

（如果存在这样的分隔符）（行动 202）。

执行分隔符检测来检测水平和垂直分隔符。可以首先检测水平分隔符然后检测垂直分隔符，或者可以首先检测垂直分隔符然后检测水平分隔符，或者可选地，可以并发地检测水平和垂直分隔符。

- 5 在几个实施例中，对水平和垂直分隔符执行图 4 的行动 202 到 218。由此，可以生成两个不同的分隔符列表。以多种方式组合这两个分隔符列表，如采用这两个分隔符列表的合并。例如，可以生成在两个列表之一中包括每一分隔符的最终分隔符列表。

- 10 因此，可以从图 4 看到，可视分隔符检测从包括一个或多个可能分隔符的分隔符列表开始，并且改变这一可能分隔符列表直到检测过程完成，该列表包括检测的分隔符。改变分隔符列表来基于可视块是否与分隔符交迭以及它们如何与分隔符交迭（如，块包含在分隔符中、块跨越分隔符或者块覆盖分隔符）添加新分隔符、移除分隔符和/或修改分隔符的参数。

- 15 应当注意，当显示文档时，过程 200 检测的分隔符可以是多种不同的形式。例如，分隔符可以是文档中的空白区、文档中画出的一个或多个分隔符线、文档中的图像或其它图形、文档中以不同的颜色作出阴影的部分等等。

- 20 作为示例，可以在图 5a—5f 所示的示例中看到水平分隔符的检测。假定可视块标识在图 5a 的文档 248 中标识了四个可视块 240、242、244 和 246。使用斜交叉影线示出了可视块 240、242、244 和 246。最初，如图 5b 所示，分隔符列表中仅有分隔符 250。使用垂直线示出图 5b 到 5f 中的分隔符。分隔符 250 包括所有可视块 240、242、244 和 246。还假定可视块 240 是从可视块池中选择的第一块，可视块 242 是从可视块池中选择的第二块，可视块 244 是从可视块池中选择的第三块，可视块 246 是从可视块池中选择的第四块。

- 25 可视块 240 包含在分隔符 250 中，因此将分隔符 250 分裂成两个分隔符 252 和 254，如图 5c 所示。类似地，块 242 包含在分隔符 254 中，因此将分隔符 254 分裂成两个分隔符 256 和 258，如图 5d 所示。而且，块 244 包含在分隔符 258 中，因此将分隔符 258 分裂成两个分隔符 260 和 262，如图 5e 所示。

- 30 当分析可视块 246 时，确定可视块 246 跨越分隔符 256，并且也覆盖分隔符 260。因此，更新分隔符 256 的参数来反映较小的区域，并且从分隔符列表中移除分隔符 260，如图 5f 所示。

另外，过程 200 检测的每一分隔符被分配一个权值。分隔符的权值是当显示文档时分隔符可视到何种程度的指示（如，具有较高权值的分隔符比具有较低权值的分隔符更可视）。基于分隔符的两侧的可视块的各种特征分配分隔符的权值。分配给分隔符的权值用于比较分隔符（如，比较它们的权值）。分配给分隔符的权值通常不用于其它目的，因此权值能够虚拟地具有任何单位（只要它与同文档中其它分隔符的权值的单位一致的单位一致或者可以转化为该单位）。

应用一组规则来确定分隔符的权值。当涉及块时，规则涉及分隔符的两侧的最近的可视块（例如，在图 5f 中，最接近分隔符 256 的可视块在一侧是块 240，在另一侧是块 246）。以下是可以应用来确定分隔符的权值的一组示例规则：

10 • 当分隔符的不同侧的块之间的距离越大，分隔符的权值越高。距离可以是例如两个块的最近的边缘之间的距离。例如，对于水平分隔符，如果该距离小于或等于 10 像素，则设置分隔符的权值为 15，如果该距离大于 10 像素而小于或等于 20 像素，则设置分隔符的权值为 25，如果该距离大于 20 像素，则设置分隔符的权值为 35；对于垂直分隔符，如果该距离小于或等于 3 像素，则设置分隔符的权值为 25，如果距离大于 3 像素，则设置分隔符的权值为 45。

15 • 如果可视分隔符与 HTML 标签（如，<HR> HTML 标签）在同一位置，则分隔符的权值更高。例如，如果可视分隔符与<HR> HTML 在同一位置，则设置分隔符的权值为 50。

20 • 对于水平分隔符，如果分隔符不同侧的块的字体属性（如，字体大小和字体灰度）不同，则改变分隔符的权值。仅当对于块的每一个，块中所有文本具有同一属性时，或者如果块中的文本的至少一个阈值具有同一属性等等，应用这一规则。另外，如果分隔符之前的块中的字体大小大于分隔符之后的块中的字体大小，增加分隔符的权值。例如，如果分隔符之前（如，水平分隔符上方）的块中使用的字体大小大于分隔符之后（如，水平分隔符下方）的块中使用的字体大小，则基于分隔符之前的块中使用的字体大小设置分隔符的权值（如，如果字体大小大于或等于 36 磅，设置权值为 9；如果字体大小大于或等于 24 磅而小于 36 磅，设置权值为 8；如果字体大小大于或等于 18 磅而小于 24 磅，设置权值为 7；如果字体大小大于或等于 14 磅而小于 18 磅，设置权值为 6；如果字体大小大于 12 磅而小于 14 磅，设置权值为 5；如果字体大小小于或等于 12 磅，设置权值为零）。然而，如果分隔符之前的块中使用的字体大小小于分隔符之后的块中使用的字体大小，则基于分隔

25

30

符之后的块中使用的字体大小设置权值（如，如果字体大小大于或等于 36 磅，设置权值为 10；如果字体大小大于 24 磅而小于 36 磅，设置权值为 9；如果字体大小大于或等于 18 磅而小于 24 磅，设置权值为 8；如果字体大小大于或等于 14 磅而小于 18 磅，设置权值为 7；如果字体大小大于 12 磅而小于 14 磅，设置权值为 6；
5 如果字体大小小于或等于 12 磅，设置权值为零）。

- 如果分隔符不同侧的块的背景色不同，则增加分隔符的权值。例如，如果分隔符不同侧的背景色不同，则设置分隔符的权值为 40。

- 对于水平分隔符，当分隔符旁边的块的结构十分相似时，则减小分隔符的权值。例如，如果分隔符旁边的块都是具有相同的字体大小和相同的字体灰度的文本块，则设置分隔符的权值为零。

10

如果对具体情况应用这些规则的多个，则可以选择由这些规则设置的权值的一个，或者可以组合由这些规则设置的权值。例如，可以使用分配的权值的最小一个（或可选地，最大一个）作为该分隔符的权值、使用所有分配的权值的平均值作为该分隔符的权值、使用所有分配的权值的加权平均值作为该分隔符的权值等等。

15

内容结构构造

内容结构构造基于可视块标识中标识的块以及可视分隔符检测中检测的分隔符对文档生成内容结构。内容结构构造（如，图 2 的行动 148 中或由图 1 的内容结构构造器实现）可以如下执行。

20 内容结构是表示文档的语义内容的文档的一组一个或多个部分。在几个实施例中，内容结构是语义内容的分层树结构。

图 6 所示是用于执行内容结构构造的示例过程 280 的流程图。过程 280 以另外的细节说明了图 2 的行动 148。过程 280 可以以软件、固件、硬件或其组合执行。

最初，基于检测的分隔符和标识的可视块生成一个或多个虚拟块（行动 282）。
25 通过分析检测的分隔符生成虚拟块，从具有最低权值的分隔符开始。将检测的分隔符两侧的块合并成单个虚拟块。对具有下一较低权值的分隔符继续这一合并，并且继续直到遇到最大权值（这一最大权值可以定义为对这一特定标识的可视块组由上述可视分隔符检测计算的最大权值，或者可选地，可以是某一预定义或预编程的值）。

30 一旦生成虚拟块，对每一虚拟块确定一致度（行动 284）。基于合并来生成该

虚拟块的两个块之间的分隔符的权值计算这一一致度(这一分隔符可以在两个可视块之间或在两个虚拟块之间)。在特定的实施例中,依照以下示例规则确定虚拟块的一致度。在一些规则中,使用了两个变量 WEIGHT_TMP_TITLE2 和 WEIGHT_TMP_TITLE1。在检测所有的分隔符之后(如,图4的过程200完成之后)确定 WEIGHT_TMP_TITLE2 和 WEIGHT_TMP_TITLE1 变量的值。在两侧具有文本块的所有分隔符按权值储存。这些权值的最高值是赋值给 WEIGHT_TMP_TITLE1 变量的值,这些权值的第二较高值是赋值给 WEIGHT_TMP_TITLE2 变量的值。如果没有这类分隔符,则将 WEIGHT_TMP_TITLE1 和 WEIGHT_TMP_TITLE2 变量的值都设为值-1。如果仅有一个这样的分隔符,则该分隔符的权值用作 WEIGHT_TMP_TITLE1 的值,值-1用作 WEIGHT_TMP_TITLE2 的值。这些示例规则是:

- 如果两个合并的块之间的分隔符的权值小于或等于零,则设置一致度为 10;
- 如果两个合并的块之间的分隔符的权值小于或等于 WEIGHT_TMP_TITLE2,则设置一致度为 7;
- 如果两个合并的块之间的分隔符的权值小于或等于 WEIGHT_TMP_TITLE1,则设置一致度为 6;
- 如果两个合并的块之间的分隔符的权值小于或等于 9,则设置一致度为 5;
- 如果两个合并的块之间的分隔符的权值小于或等于 20,则设置一致度为 5;
- 如果两个合并的块之间的分隔符的权值小于或等于 40,则设置一致度为 4;
- 如果两个合并的块之间的分隔符的权值小于或等于 50,则设置一致度为 2;
- 如果不满足上述任一规则,则设置一致度为 1。

然后生成内容结构(行动286)。至少部分地基于粒度需求以及哪一虚拟块,如果有的话,满足该粒度需求来生成内容结构。如上所述,分析内容结构的每一叶节点来确定是否满足粒度需求。在几个实施例中,定义了容许一致度(PDoC)值,并且为满足粒度需求,需要内容结构的每一叶节点具有大于(或可选地,大于或等于)该 PDoC 值的 DoC 值。PDoC 值可以是预定义值(如,由系统设计者根据经验确定)。可以使用用于 PDoC 值的任一范围的值,较大的值通常导致具有更多但更小可视块的内容结构。PDoC 值的一个示例范围是从 6 到 9。

行动286中生成的内容结构可包括虚拟块和/或可视块。将行动282中标识的每一虚拟块的 DoC 值与 PDoC 值比较。对于其 DoC 值大于 PDoC 值的虚拟块,该

虚拟块的子块不作为内容结构中的块输出。

然后输出内容结构（行动 288）。可以以任一各种格式输出内容结构，在几个实施例中，以分层树格式输出，树的节点表示虚拟块和/或可视块。

图 7 所示是生成虚拟块和内容结构树的示例。在图 7 中，示出了六个可视块 300、302、304、306、308 和 310，由可视块标识所标识。另外，示出了块之间的分隔符 312、314、316、318 和 320，较粗线的分隔符（分隔符 314 和 318）比较细线的分隔符（分隔符 312、316 和 320）具有更大的权值。

每一可视块 300—310 具有内容结构树 322 中相应的内容结构块。另外，通过合并块 300 和 302 生成虚拟块 324，虚拟块 324 基于分隔符 312 的权值具有一 DoC 值。通过合并块 304 和 306 生成虚拟块 326，虚拟块 326 基于分隔符 316 的权值具有一 DoC 值。通过合并块 308 和 310 生成虚拟块 328，虚拟块 328 基于分隔符 320 的权值具有一 DoC 值。另外，通过合并虚拟块 324、326 和 328 生成虚拟块 330。分隔符 314 和 318 的权值相同，因此块 330 的 DoC 值可以基于分隔符 314 和 318 的任一个的权值。

图 7 所示的基于虚拟块的内容结构输出依赖于虚拟块 324、326、328 和 330 的 DoC 值。例如，如果虚拟块 324 的 DoC 值大于 PDoC 值，则虚拟块 324 作为内容结构中的块输出，而个别块 300 和 302 不作为内容结构中的块输出。然而，如果虚拟块 324 的 DoC 值不大于 PDoC 值，则虚拟块 324 不作为内容结构中的块输出，而个别块 300 和 302 作为内容结构中的块输出。作为另一示例，如果虚拟块 330 的 DoC 值大于 PDoC 值，则虚拟块 330 作为内容结构中的块输出，而个别块 324、326 和 328 不作为内容结构中的块输出，并且个别块 300、302、304、306、308 和 310 不作为内容结构中的块输出。

文档检索

可以使用这里描述的基于视觉的文档分割来使用文档检索技术。这类文档检索技术可单独依赖于基于视觉的文档分割，或者可选地与更多传统的搜索技术结合。

图 8 所示是用于使用这里描述的基于视觉的文档分割执行文档检索的示例过程 360 的流程图。过程 360 可以由例如图 1 的文档检索模块 110 实现。过程 360 可以以软件、固件、硬件或其组合执行。过程 360 可用于检索多种不同类型的文档

的任一个，包括，例如，web 页（如，在因特网和/或内联网可用）、个人书写的文件、摘要或总结等等。

最初，接收查询（行动 362）。通常，由用户输入查询，尽管可选地，可以从某一其它源（如，远程装置、软件应用等等）接收查询。访问从这里描述的基于视觉的文档分割技术获取的块的数据库（行动 364）。通常，将先执行基于视觉的文档分割过程并且所得的内容结构（包括可视块和/或虚拟块）已经储存在数据库或某一其它可访问位置中。可选地，基于视觉的文档分割过程可以响应于行动 362 中的查询接收，在一个或多个文档上执行。

一旦被访问，依照它们与查询标准匹配到何种程度对内容结构的块进行分级（行动 366）。可以在行动 366 中对可视块和/或虚拟块进行分级。行动 362 中接收的查询包括一个或多个搜索条件，也称为查询标准或查询条件。在几个实施例中，在行动 366，仅对包括至少一个搜索条件的块（或者可选地，仅对包括所有搜索条件的块）进行分级。可选地，可以在要分级的块的数量上加上限制（例如，可以配置过程 360 仅对 50 或 100 个块进行分级）。可以以任一多种不同的方式生成基于块的等级。例如，等级可以基于以下的一个或多个：块中包括了多少搜索条件、块中不同搜索条件的位置、块中搜索条件出现的频率等等。

然后基于块等级生成文档等级（行动 368）。可以以任一多种方式生成文档等级。在几个实施例中，来自文档的最高等级块的等级用作该文档的等级。可选地，可以组合该文档中所有块的等级（如，通过生成所有块的等级的平均值、通过生成所有块的等级的加权平均值等等）来生成该文档的等级。

然后将文档等级返回到例如请求者（行动 370）。可以以任一多种不同的方式返回文档等级，如每一文档的标识符（如，标题或统一资源定位器（URL））以及数字等级、依照其等级的每一文档的标识符的顺序显示（如，不显示每一文档的明确数字等级）、包括搜索标准的从文档的摘录、依照其等级顺序返回的实际文档等等。

应当理解，可以对过程 360 进行各种修改。例如，在行动 366 中，可以不对所有可访问文档进行分级或搜索。作为另一示例，可以在步骤 370 返回文档的块而不是整个文档。

图 9 所示是用于使用这里描述的基于视觉的文档分割执行文档检索的另一示例过程 400 的流程图。过程 400 可以例如由图 1 的文档检索模块实现。过程 400

可以以软件、固件、硬件或其组合执行。与图 8 的过程 360 类似，过程 400 可以用于检索各种不同类型的文档的任一个，包括，例如，web 页（如，因特网和/或内联网上可用）、个人书写的文件、摘要或总结等等。

过程 400 类似于过程 360，然而不在所有的可能文档上执行基于视觉的文档分割。相反，执行初始的搜索，然后使用来自初始搜索的文档执行基于视觉的文档分割过程。

最初，接收查询（行动 402）。与过程 360 类似，该查询可以是用户查询或来自某一其它源，并且该查询包括一个或多个搜索条件。在文档组上执行初始文档检索，并且对满足查询标准的文档进行分级（行动 404）。可以以任一多种方式执行这一初始检索。行动 404 中的这一初始检索过程的一个示例在 Robertson, S.E 的题为“Overview of Okapi Projects”，Journal of Documentation 杂志，卷 53，第一期，1997 年，3-7 页中有描述。行动 404 中也可以使用多种常规 web 搜索引擎的任一个。

过程 400 然后继续进行到在初始检索中检索到的文档上执行基于视觉的文档分割（行动 406）。然后使用查询标准（搜索条件）以及从基于视觉的文档分割中获取的块执行块检索（行动 408），类似于图 8 的行动 366。然后基于块的分级生成细化的文档等级（行动 410），与图 8 的行动 366 中的上述讨论类似。

然后通过组合来自行动 404 的初始文档等级和来自行动 410 的细化文档等级生成最终文档等级（行动 412）。可以以任一多种方式执行这一组合。例如，可以对来自行动 404 和 410 的等级共同求平均值，或者使用来自行动 404 和 410 的等级的加权平均值。可选地，可以使用来自行动 404 和 410 的特定文档的最高等级作为该文档的等级。在另一替代方法中，可以使用来自行动 410 的等级替代来自行动 404 的等级。

一旦生成最终文档等级，返回该文档等级（行动 414），类似于图 8 的行动 370。

图 10 所示是用于使用这里描述的基于视觉的文档分割执行文档检索的另一示例过程 440 的流程图。过程 440 可以如由图 1 的文档检索模块 110 实现。过程 440 可以以软件、固件、硬件或其组合执行。类似于图 8 的过程 360，过程 440 可以用于检索多种不同类型的文档的任一个，包括，例如，web 页（如，在因特网和/或内联网上可用）、由个人书写的文件、摘要或总结等等。

过程 440 类似于图 9 的过程 400，然而，使用基于视觉的文档分割来扩展查询

标准。

最初，接收查询（行动 442），类似于图 9 的行动 402。在文档组上执行初始文档检索，并且对满足查询标准的文档进行分级（行动 444），类似于图 9 的行动 404。然后在初始检索中检索到的文档上执行基于视觉的文档分割过程（行动 446），类似于图 9 的行动 406。

然而，与图 9 中完成的执行块检索相反，从基于视觉的文档分割获取的内容结构的块连同查询标准一起用来执行查询扩展（行动 448）。可以以任一多种方式执行查询扩展。

在几个实施例中，通过使用扩展条件选择的最高级块（如，具有最高等级的十个块，或具有最高等级的 10%的块）来执行查询扩展。依照它们满足原始查询条件到何种程度对块进行分级（类似于图 8 的行动 366）。以任一多种已知的方式选择扩展条件，然而，与传统的查询扩展不同，基于从基于视觉的文档分割过程获取的块选择扩展条件。例如，从基于视觉的文档分割过程获取的块中的所有条件（除原始查询条件之外）依照以下条件选择值 TSV 来加权：

$$TSV = w^{(1)} \times \frac{r}{R}$$

其中， r 表示包含条件的块的数量， R 表示从基于视觉的文档分割过程获取的块的总数， $w^{(1)}$ 表示 Q 中 T 的 Robertson/Spark Jones 权值，其中 Q 表示原始查询（行动 442 中接收）并且 T 表示原始查询中的搜索条件。

给定这些加权扩展条件，选择这些条件的若干个用作扩展条件。可以选择这些条件的不同个数，并且在几个实施例中，选择前十个条件作为扩展条件（即，具有最大条件选择值 TSV 的十个条件）。

然后基于扩展的查询标准和文档组执行最终检索，并且对满足扩展的查询标准的文档进行分级（行动 450）。行动 450 类似于行动 444 而执行，除了使用扩展的查询而不是接收的标准。扩展的查询包括原始查询条件以及扩展条件。在几个实施例中，对扩展查询中的条件进行加权。可以以多种方式对该条件进行加权，以当确定哪一文档匹配查询条件和/或当对文档进行分级时对原始查询条件给予优先选择。例如，可以如下对扩展查询的条件进行加权。对于原始条件（在行动 442 中作为查询的部分接收的条件），该条件的新权值为 $tf \times 2$ ，其中 tf 表示该条件在查询中的频率。对于每一扩展条件，该条件的权值设置如下：

$$\frac{1-(n-1)}{m}$$

其中, n 表示该条件的 TSV 等级值, m 是扩展条件的个数 (如, 特定实施例中为 10)。

然后返回从行动 450 的最终检索中获取的文档等级 (行动 452), 类似于图 9 的行动 414。

- 5 另外, 在几个实施例中, 上文参考图 8、图 9 或图 10 讨论的文档检索可采用将上述基于视觉的文档分割过程与固定长度方法组合的组合文档分割方法。在这一组合的文档分割方法中, 如上所述地获取可视块 (如, 参考图 2)。给定这些可视块, 使用重叠窗来将至少一些块划分成更小的单元。第一个窗从第一可视块的一个单词开始, 并且随后的窗与之前的窗重叠一个特定的数量 (如, 二分之一) 直到
- 10 块的最后。对应于这些窗的可视块的部分作为最终块输出。对于小于窗的长度 (窗通常是固定值) 的可视块, 可视块作为最终块输出, 而不进一步分割。然而, 通过将较大的块分割成等于窗的大小的较小的块, 减小了不同块的长度的方差。

使用这一组合文档分割方法, 在文档检索过程中使用组合文档分割方法输出的最终块来代替内容结构的可视和/虚拟块。

15

示例环境

- 图 11 说明了可用于实现这里描述的技术的一般环境 500 的一个示例。计算环境 500 仅为计算环境的一个示例, 并非建议对计算机和网络体系结构的使用或功能的范围的局限。也不应将计算环境 500 解释为具有关于示例计算环境 500 中说明的
- 20 任一组件或其组合的依赖或需求。

- 计算环境 500 包括以计算机 502 形式的通用计算装置。计算机 500 能够实现, 例如图 1 的可视块提取器 104、可视分隔符检测器 106、内容结构构造器 108 和/或文档检索模块 110。计算机 502 的组件包括但不限于, 一个或多个处理器 504、系统存储器 506 以及将各类系统组件包括处理器 504 耦合至系统存储器 506 的系统
- 25 总线 508。

- 系统总线 508 表示任一若干种总线结构类型的一个或多个, 包括存储器总线或存储器控制器、外围总线、加速图形端口以及使用各类总线结构的处理器或本地总线。作为示例, 这类结构包括工业标准体系结构 (ISA) 总线、微通道体系结构 (MCA) 总线、增强 ISA (EISA) 总线、视频电子标准协会 (VESA) 本地总线以
- 30 及外围部件互连 (PCI) 总线, 也称为 Mezzanine 总线。

计算机 502 通常包括各种计算机可读媒质。这类媒质可以是可由计算机 502 访问的任一可用媒质，包括易失和非易失、可移动和不可移动媒质。

系统存储器 506 包括以非易失存储器，如只读存储器（ROM）512 和/或非易失存储器，如随机存取存储器（RAM）514 形式的计算机可读媒质。基本输入/输出系统（BIOS）514，包含如在启动时协助在计算机 502 内的元件之间传输信息的基本例程，储存在 ROM 512 中。RAM 510 通常包含处理单元 504 立即可访问和/或当前正在操作的数据和/或程序模块。

计算机 502 也可包括其它可移动/不可移动、易失/非易失的计算机存储媒质。作为示例，图 11 示出了用于对不可移动、非易失磁媒质（未示出）进行读写的硬盘驱动器 516、用于对可移动、非易失磁盘 520（如，“软盘”）进行读写的磁盘驱动器 518 以及用于对可移动、非易失光盘 524 如 CD-ROM、DVD-ROM 或其它光媒质进行读写的光盘驱动器 522。硬盘驱动器 516、磁盘驱动器 518 以及光盘驱动器 522 的每一个通过一个或多个数据媒质接口 526 连接至系统总线 508。可选地，硬盘驱动器 516、磁盘驱动器 518 以及光盘驱动器 522 可以通过一个或多个接口（未示出）连接到系统总线 508。

驱动器及其相关的计算机可读媒质为计算机 502 提供了计算机可执行指令、数据结构、程序模块和其它数据的非易失存储。尽管该示例说明了硬盘 516、可移动磁盘 520 以及可移动光盘 524，应当理解，也可以使用能够储存可由计算机访问的数据的其它类型的计算机可读媒质，如盒式磁带或其它磁存储设备、闪存卡、CD-ROM、数字多功能盘（DVD）或其它光存储、随机存储存储器（RAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）等等，来实现该示例计算系统和环境。

多个程序模块可储存在硬盘 516、磁盘 520、光盘 524、ROM 512 和/或 RAM 510 中，作为示例，包括操作系统 526、一个或多个应用程序 528、其它程序模块 530 以及程序数据 532。这些操作系统 526、一个或多个应用程序 528、其它程序模块 530 以及程序数据 532 的每一个（或其组合）可实现支持分布式文件系统的常驻组件的所有或部分。

用户可以通过诸如键盘 534 和指向设备 536（如，“鼠标”）的输入设备向计算机 502 输入命令和信息。其它输入设备 538（未示出）可包括麦克风、操纵杆、游戏垫、圆盘式卫星天线、串行端口、扫描仪和/或其类似物。这些和其它输入设

备通过耦合至系统总线 508 的输入/输出接口 540 连接到处理单元 504。但是也可以通过其它接口和总线结构连接，如并行端口、游戏端口或通用串行总线（USB）。

监视器 542 或其它类型的显示设备也通过接口，如视频适配器 544 连接到系统总线 508。除监视器之外，其它外围输出设备可包括诸如扬声器（未示出）和打印机 546 的组件，可通过输入/输出接口 540 连接到计算机 502。

计算机 502 可以在使用到一个或多个远程计算机，如远程计算机装置 548 的逻辑连接的网络化环境中操作。作为示例，远程计算装置 548 可以是个人计算机、便携式计算机、服务器、路由器、网络计算机、对等设备或其它公用网络节点等等。示出远程计算装置 548 为可包括这里关于计算机 502 所描述的元件和特点的许多或全部的便携式计算机。

描述计算机 502 和远程计算机 548 之间的逻辑连接为局域网（LAN）550 和一般广域网（WAN）552。这类网络环境常见于办公室、企业范围计算机网络、内联网以及因特网。

当在 LAN 网络环境中实现时，计算机 502 通过网络接口或适配器 554 连接至局域网 550。当在 WAN 网络环境中实现时，计算机 502 通常包括调制解调器 556 或其它装置，用于通过广域网 552 建立通信。调制解调器 556 对计算机 502 可以是内置或外置的，通过输入/输出接口 540 或其它合适的机制连接至系统总线 508。可以理解，示出的网络连接是示例，也可以使用在计算机 502 和 548 之间建立通信链路的其它装置。

在网络化环境中，如与计算环境 500 示出的环境中，描述的与计算机 502 相关的程序模块或其部分可储存在远程存储器存储设备中。作为示例，远程应用程序 558 驻留在远程计算机 548 的存储器设备中。为说明目的，这里示出应用程序和诸如操作系统的其它可执行程序组件为离散块，尽管可以认识到，这类程序和组件在不同的时刻驻留在计算装置 502 的不同存储组件中，并且可由计算机的数据处理器执行。

总结

这里在计算机可执行指令的一般上下文中描述了各种模块和技术，计算机可执行指令如程序模块，由一个或多个计算机或其它装置执行。一般而言，程序模块可包括例程、程序、对象、组件、数据结构等等，执行特定的任务或实现特定的抽

象数据类型。通常，程序模块的功能可以如各种实施例中所描述的组合或分布。

这些模块和技术的一个实现可以储存在某一形式的计算机可读媒质中或通过其发送。计算机可读媒质可以是可由计算机访问的任一可用媒质。作为示例而非局限，计算机可读媒质可包括“计算机存储媒质”和“通信媒质”。

- 5 “计算机存储媒质”包括以用于储存信息的任一方法或技术实现的易失和非易失、可移动和不可移动媒质，信息如计算机可读指令、数据结构、程序模块或其它数据。计算机存储媒质包括但不限于，RAM、ROM、EEPROM、闪存或其它存储器技术、CD-ROM、数字多功能盘（DVD）或其它光盘存储、磁盒、磁带、磁盘存储或其它磁存储设备、或可以用来储存所期望的信息并可由计算机访问的任一
- 10 其它媒质。

- “通信媒质”通常在诸如载波或其它传输机制的已调制数据信号中包含计算机可读指令、数据结构、程序模块或其它数据。术语“已调制数据信号”指以对信号中的信息进行编码的方式设置或改变其一个或多个特征的信号。作为示例而非局限，通信媒质包括有线媒质，如有线网络或直接连线连接，以及无线媒质，如声学、
- 15 RF、红外和其它无线媒质。上述任一的组合也应当包括在计算机可读媒质的范围之内。

这里描述的并在附图中示出了各种流程图。这些流程图中的行动的顺序仅为示例—可以改变这些顺序，以不同的顺序和/或并发地执行这些行动。

- 另外，这里给出了许多具体数字示例（如，具体的阈值、具体的大小、具体的
- 20 的权值等等）。这些具体数字示例仅为示例，可以可选地使用其它值。

尽管上述描述使用了对结构化特点和/或方法行动特定的语言，应当理解，所附权利要求书中所定义的本发明不限于这里描述的特定的特点或行动。揭示了特定的特点和行动作为实现本发明的示例性形式。

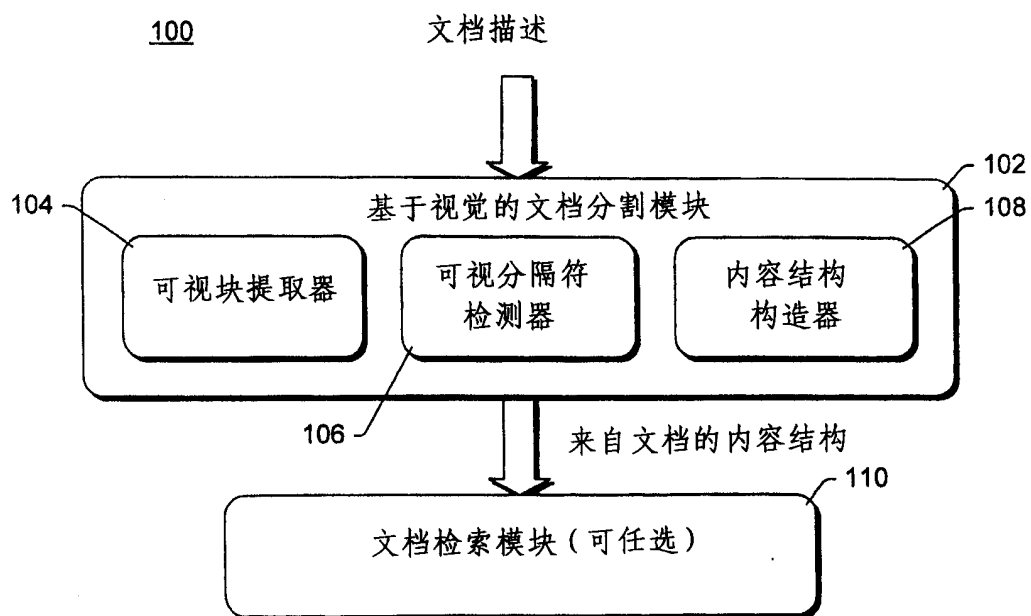


图 1

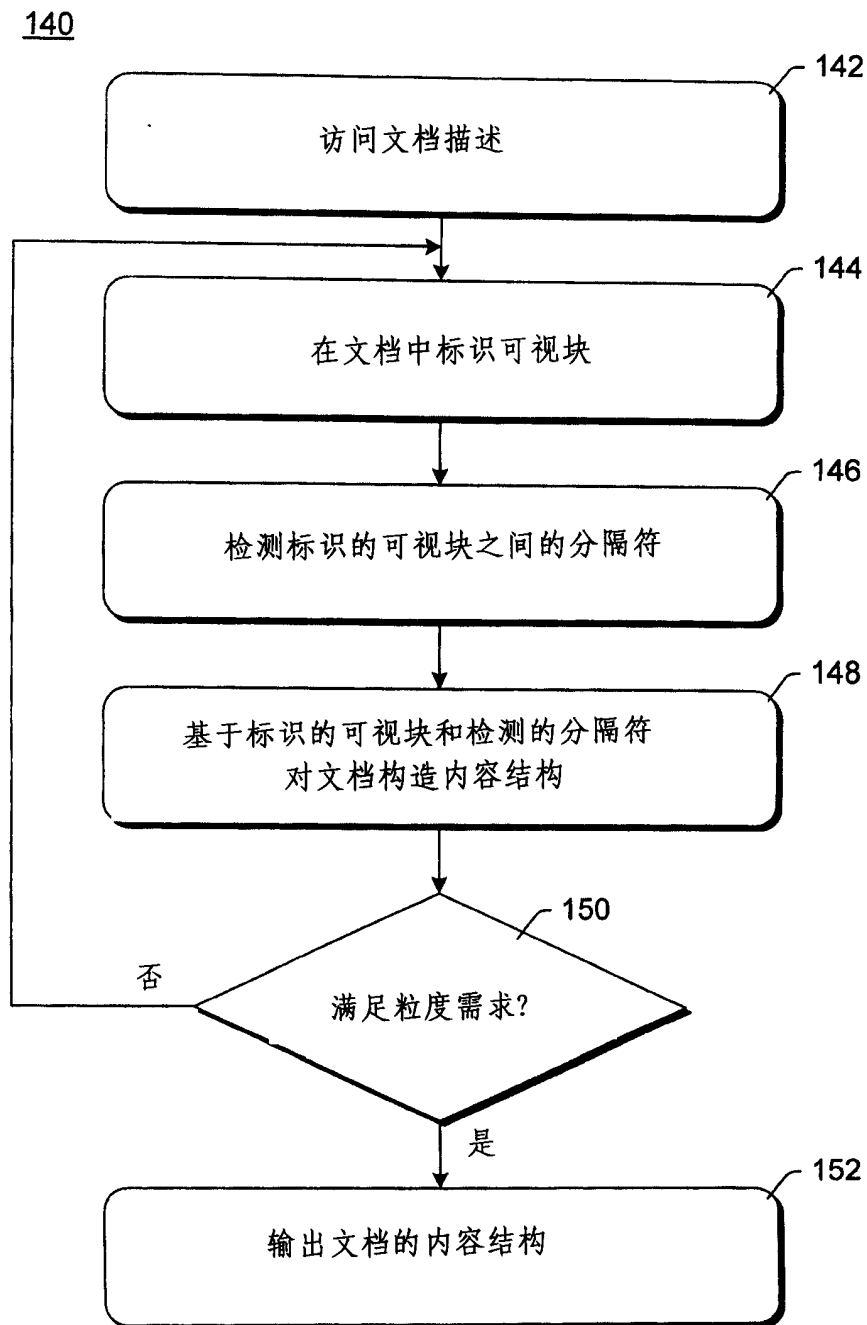


图 2

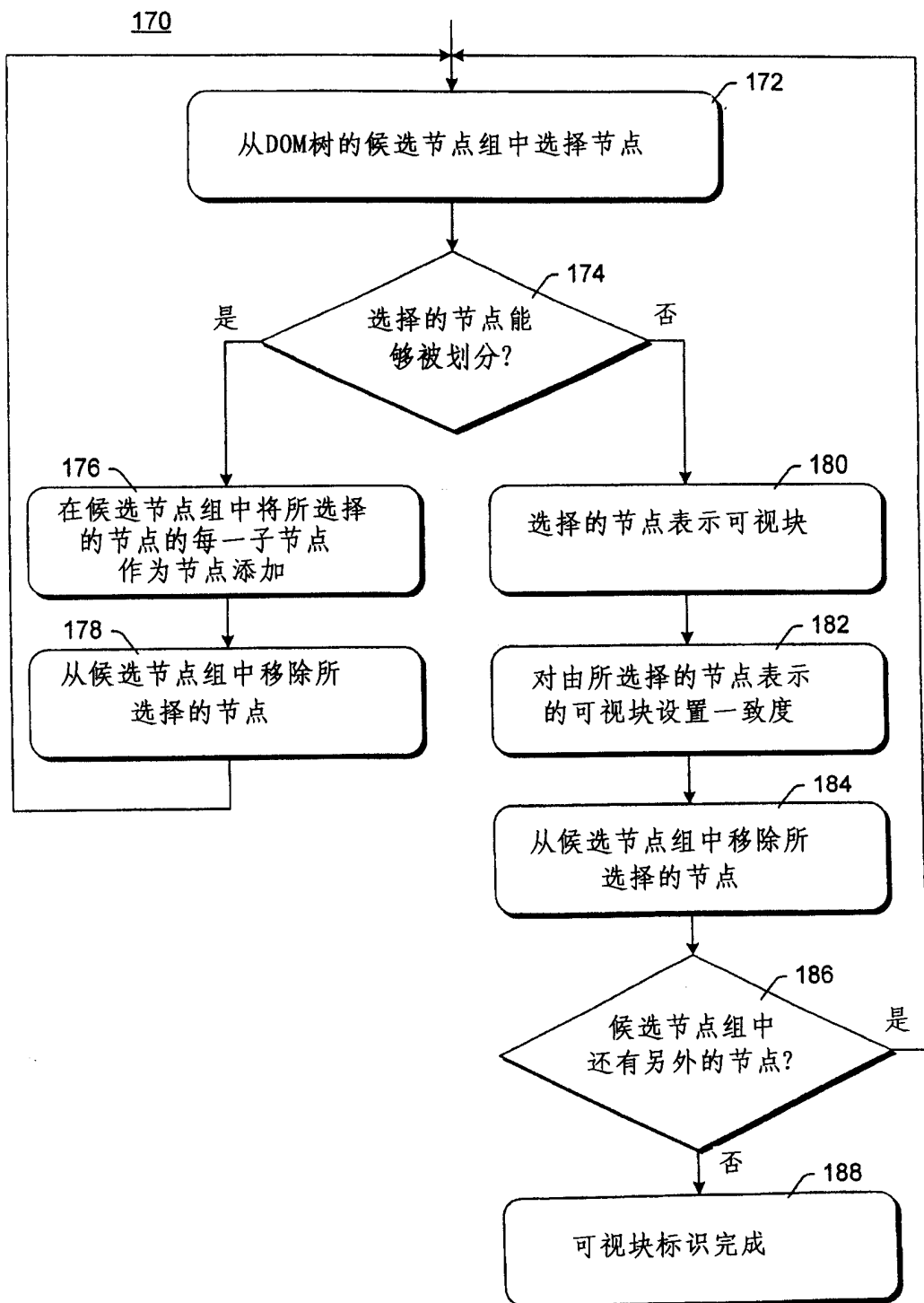


图 3

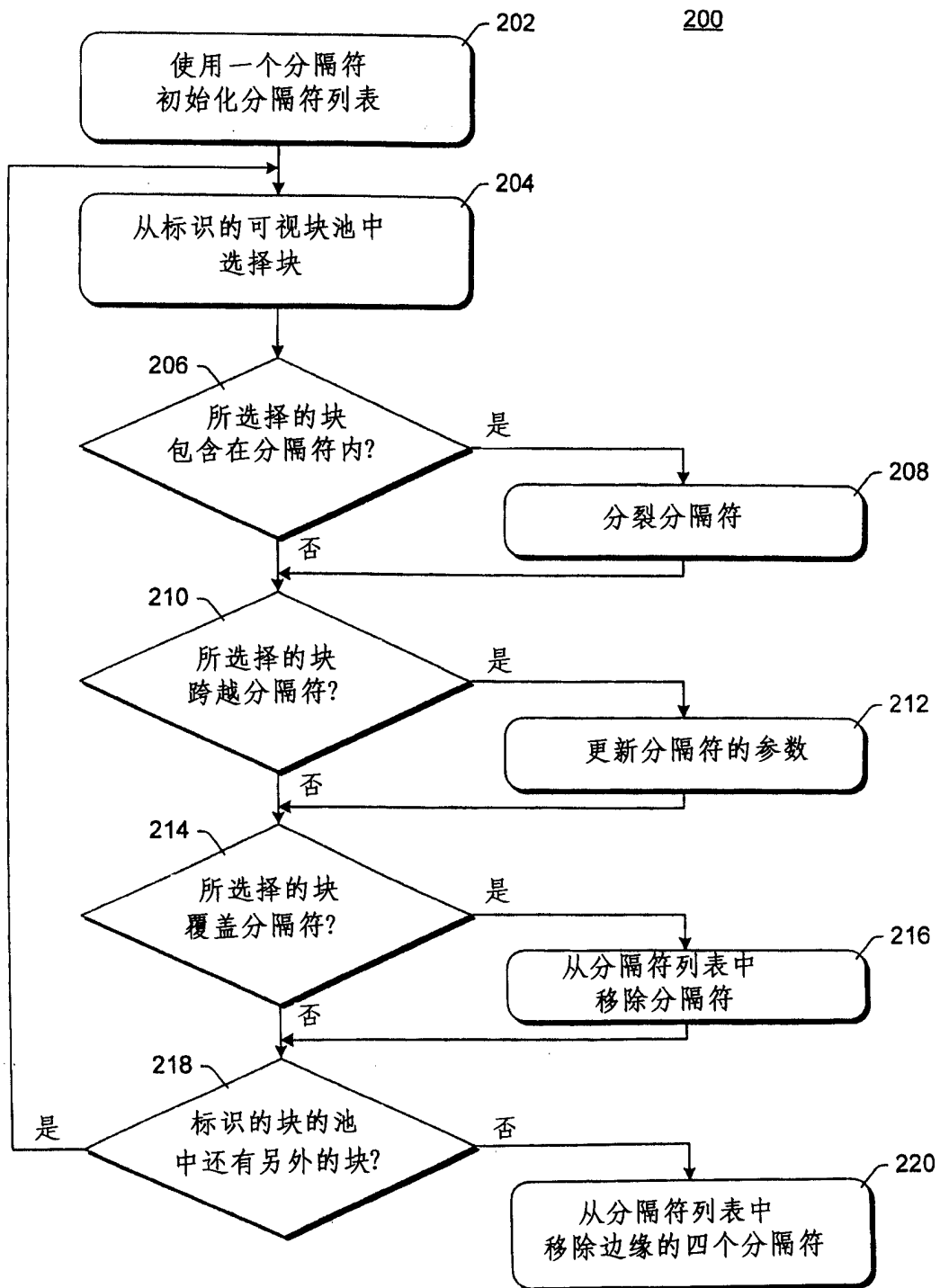


图 4

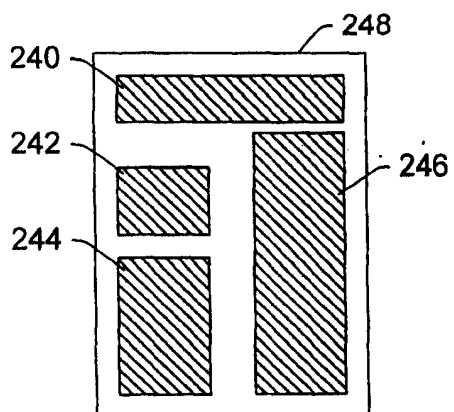


图 5a

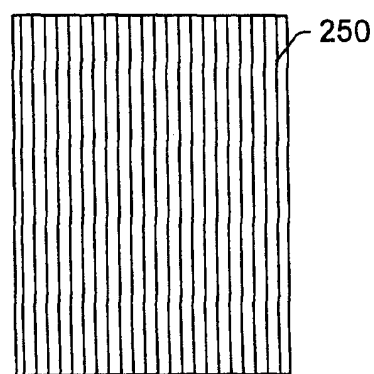


图 5b

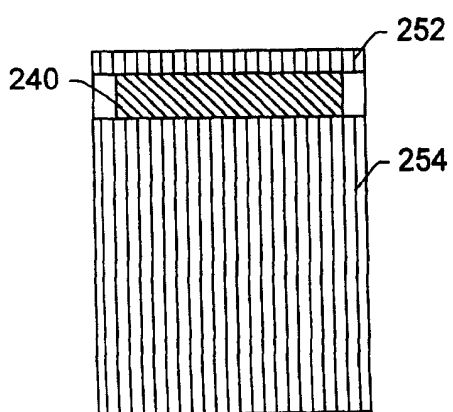


图 5c

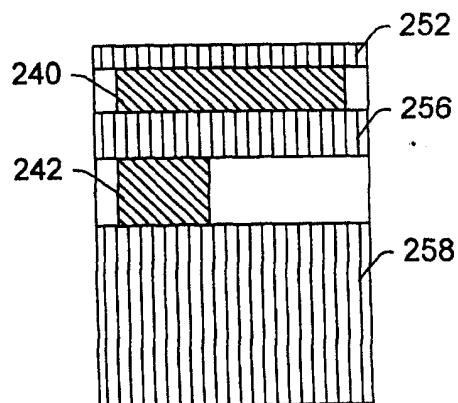


图 5d

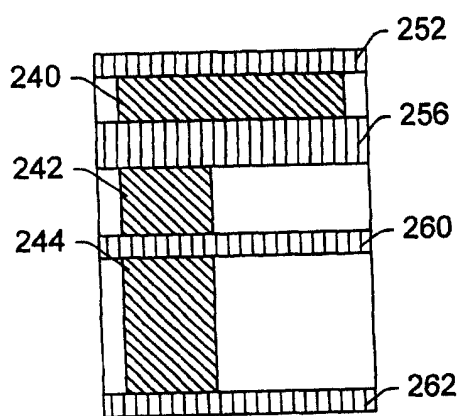


图 5e

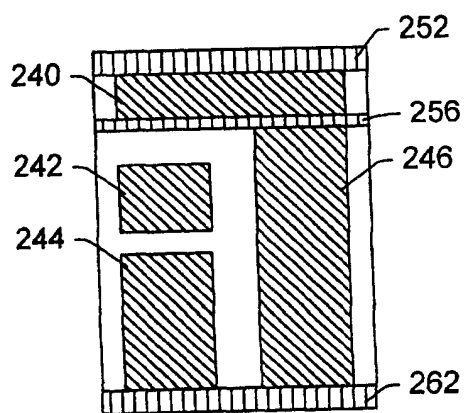


图 5f

280

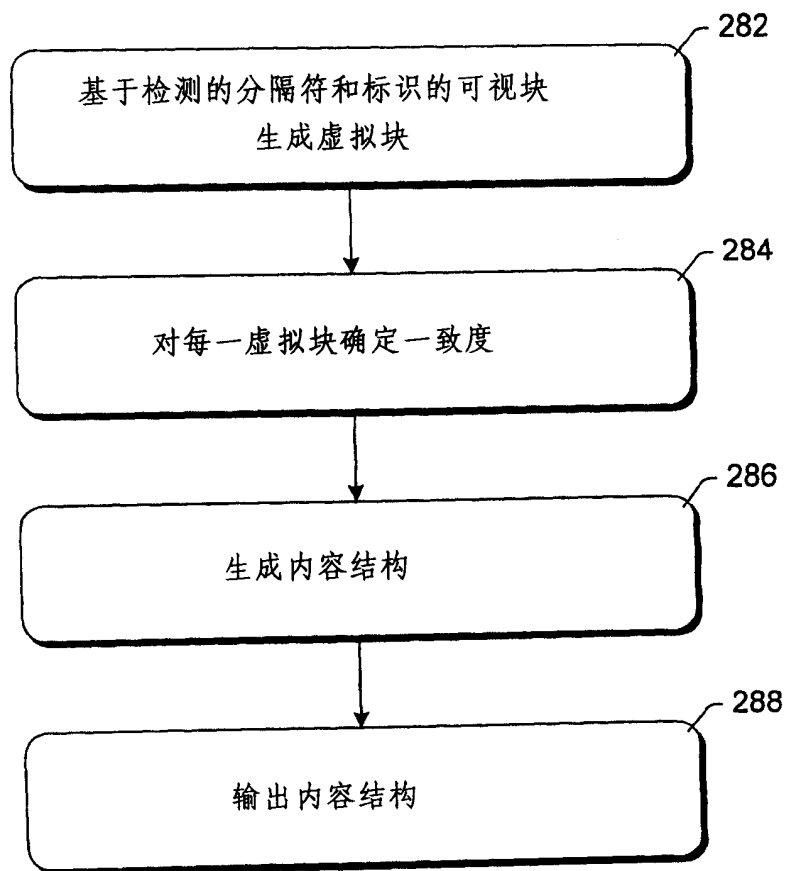
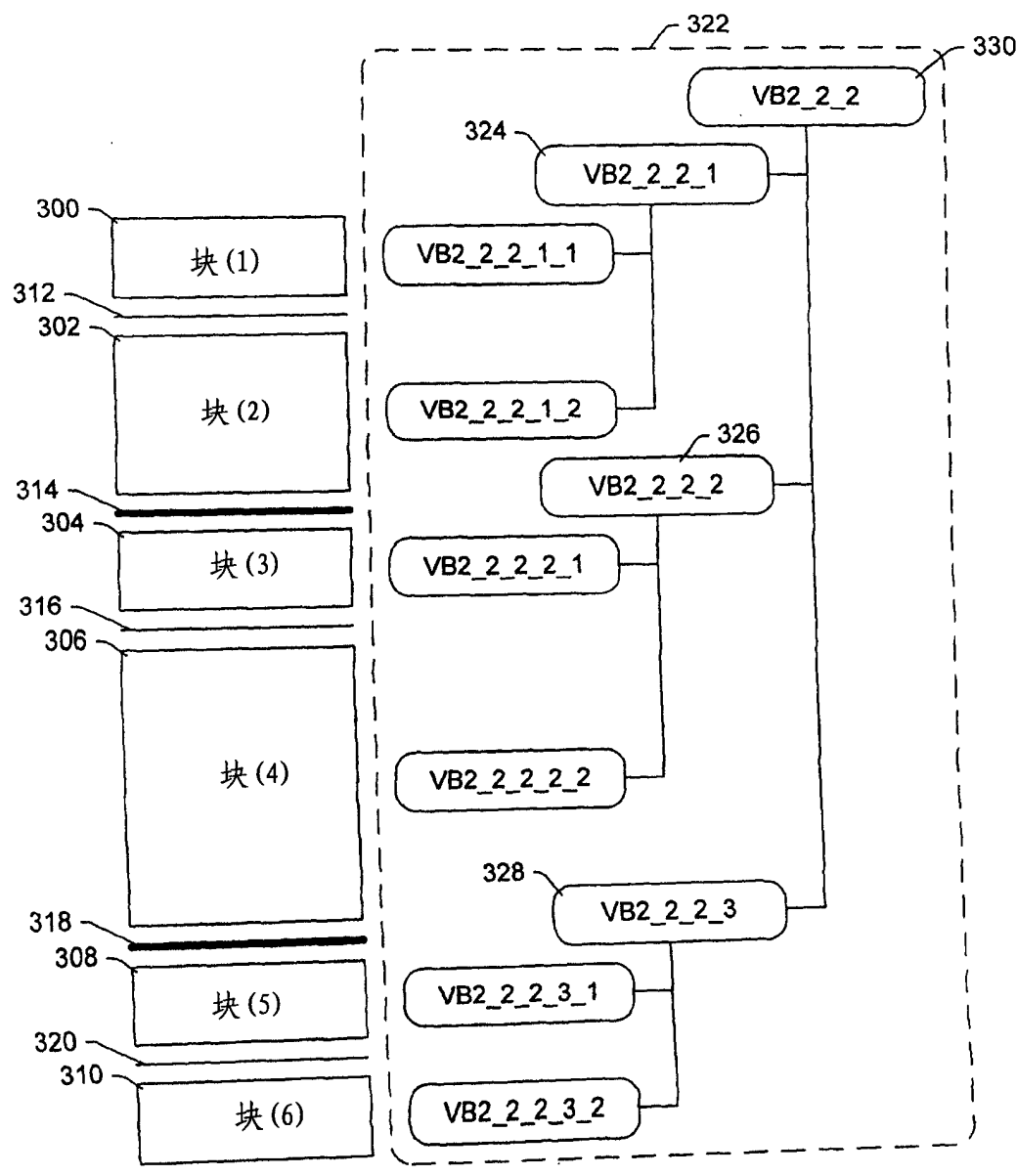


图 6



280

图 7

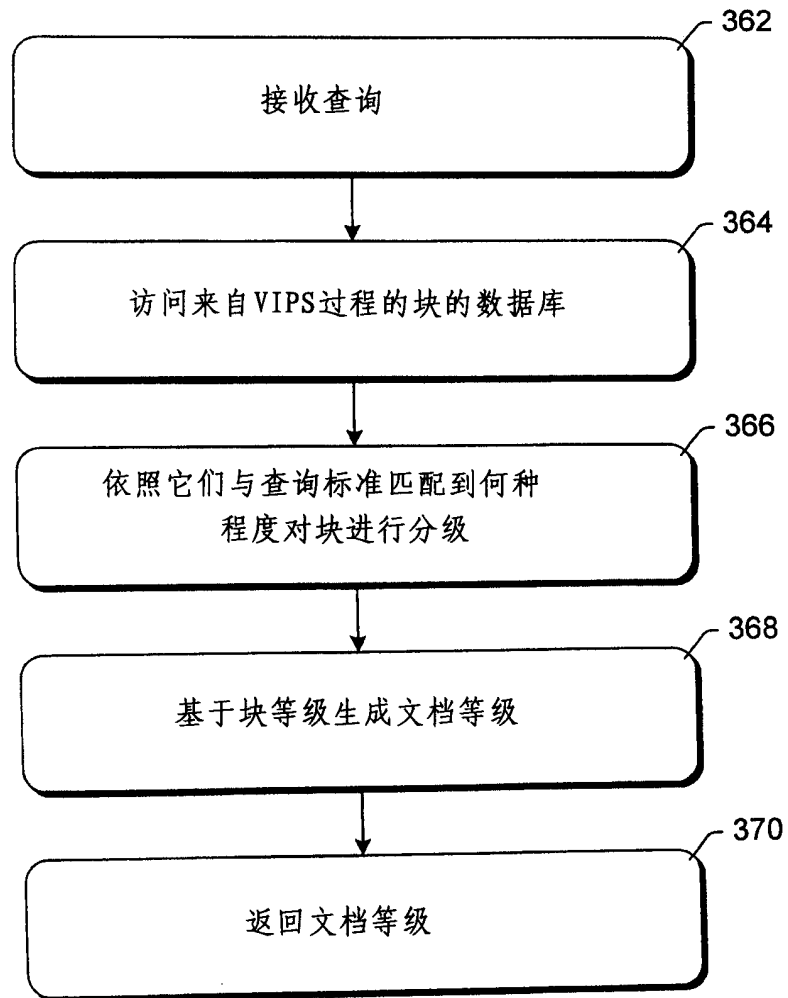
360

图 8

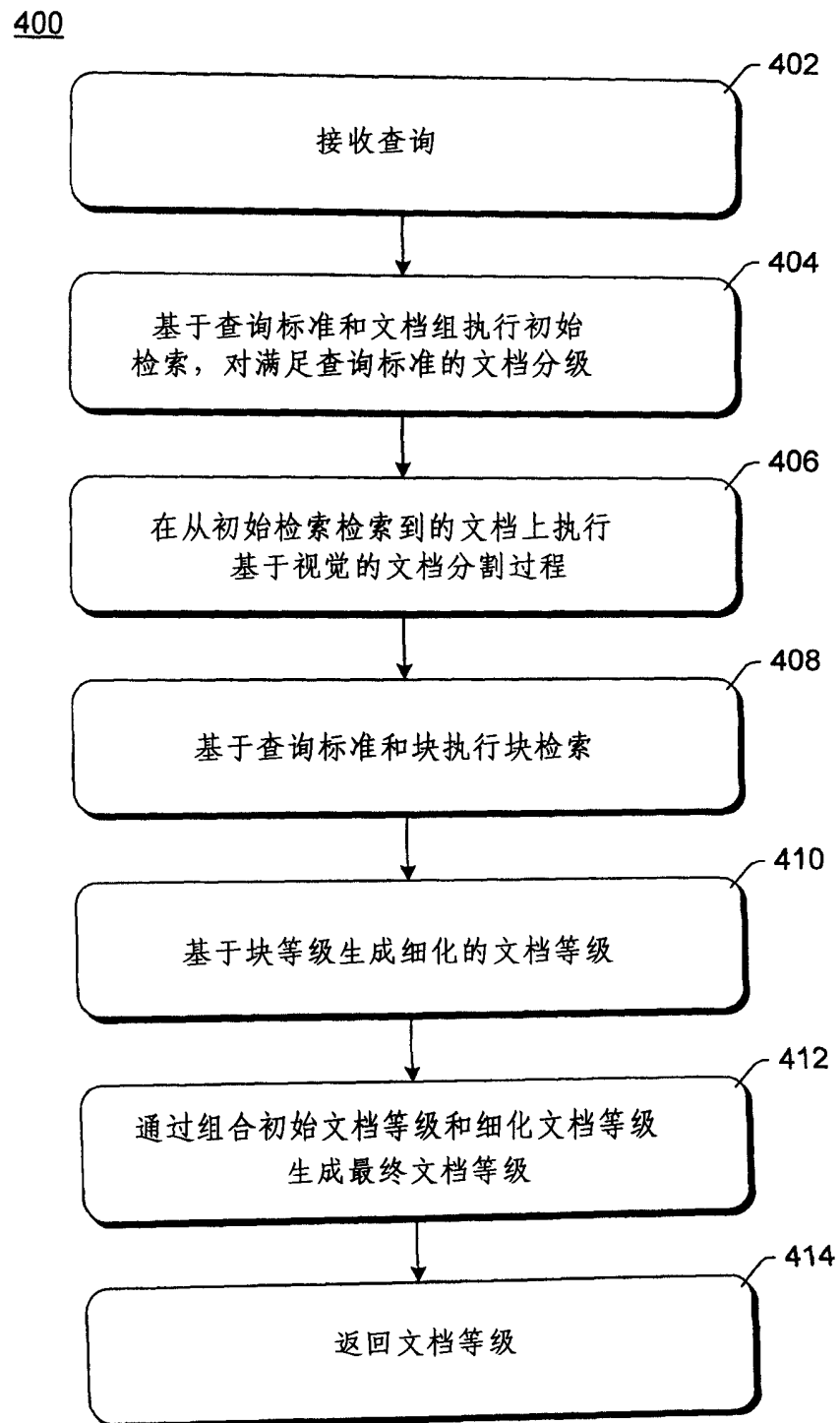


图 9

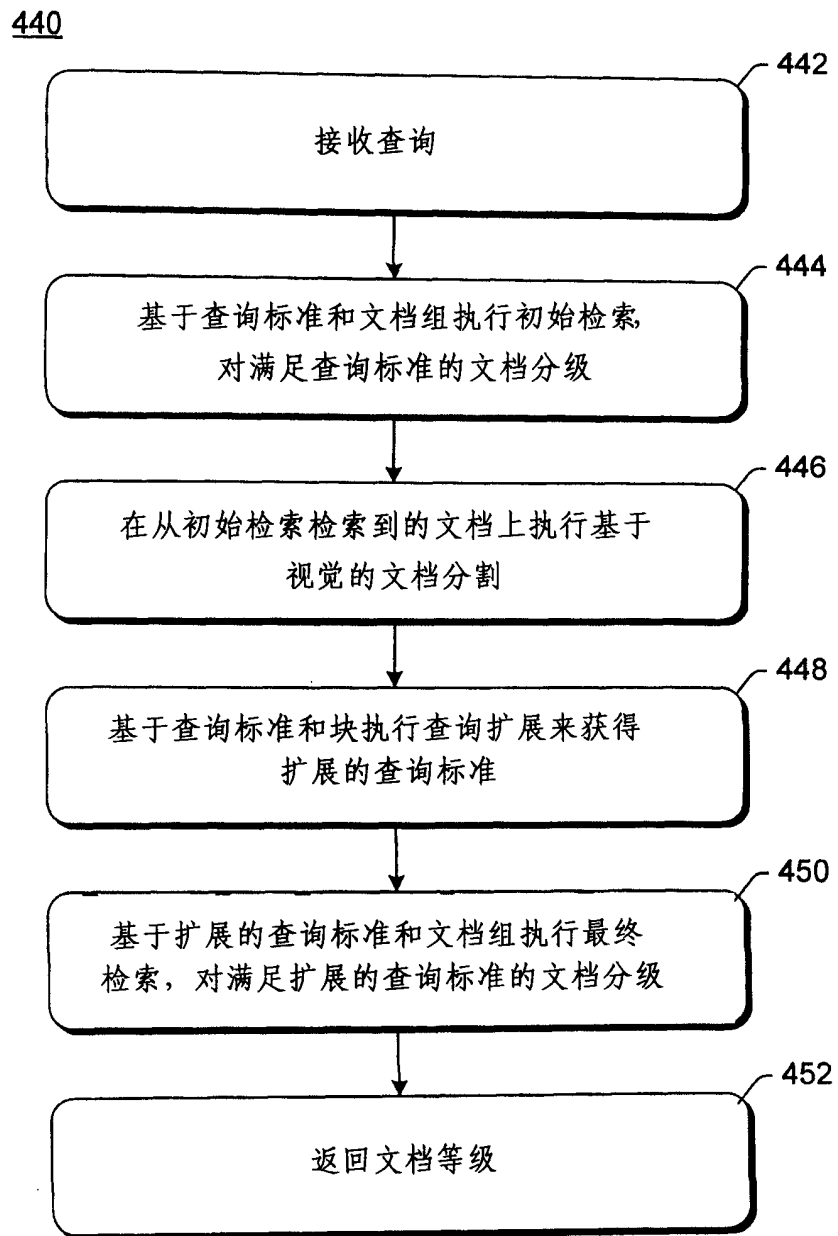


图 10

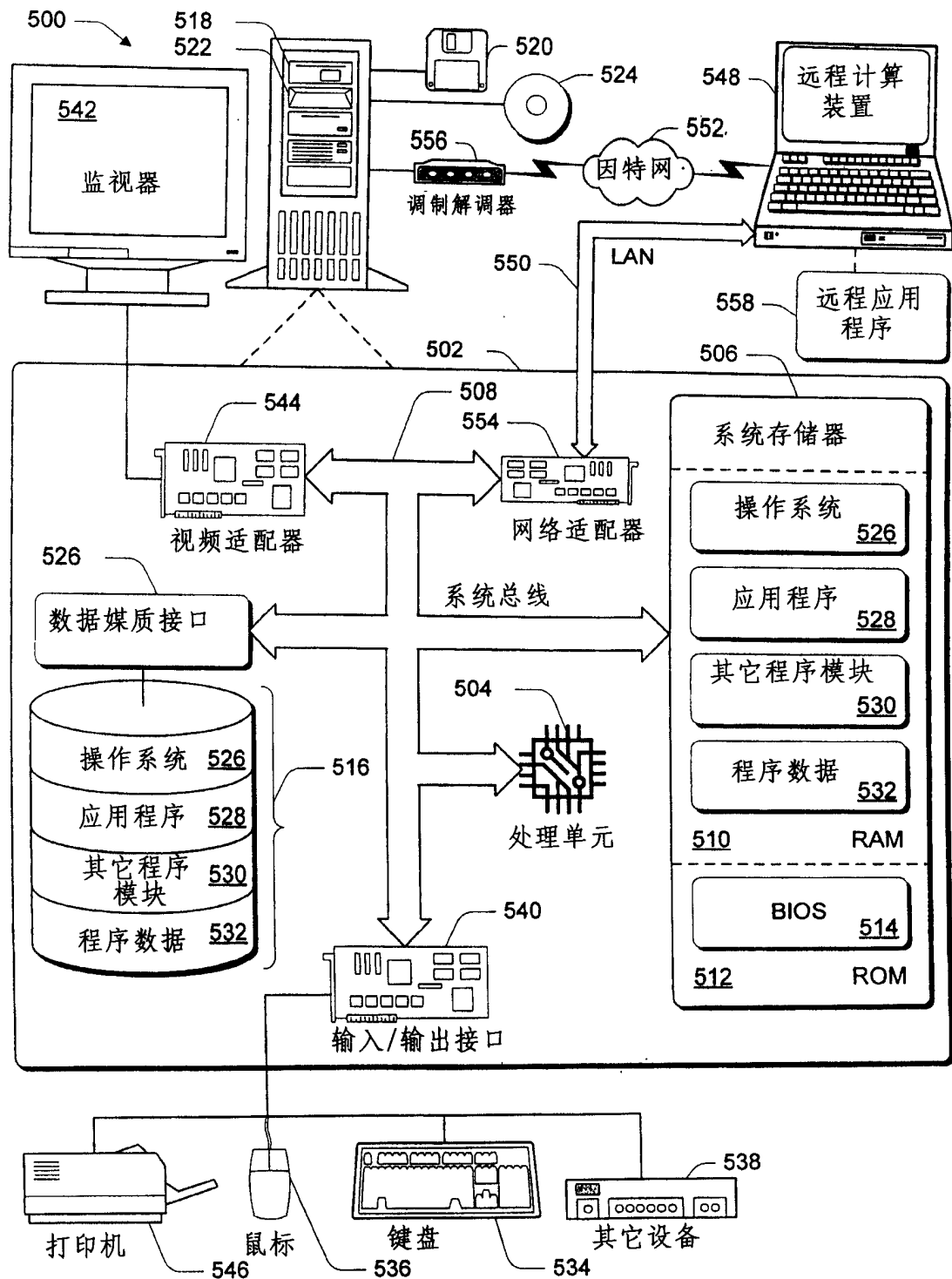


图 11