(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 107358208 B
(45)授权公告日 2018.07.13

(21)申请号 201710576556.9

(22)申请日 2017.07.14

(65)同一申请的已公布的文献号
申请公布号 CN 107358208 A

(43)申请公布日 2017.11.17

(73)专利权人 北京神州泰岳软件股份有限公司
地址 100089 北京市海淀区万泉庄路28号万柳新贵大厦A座601室
专利权人 中科鼎富(北京)科技发展有限公司

(72)发明人 徐龙 李德彦 杨宇

(74)专利代理机构 北京弘权知识产权代理事务所(普通合伙) 11363
代理人 逯长明 许伟群

(51)Int.Cl.
G06K 9/00(2006.01)

(56)对比文件
CN 101534306 A,2009.09.16,
CN 102855244 A,2013.01.02,
CN 106383817 A,2017.02.08,
US 2008244715 A1,2008.10.02,

审查员 孟繁杰

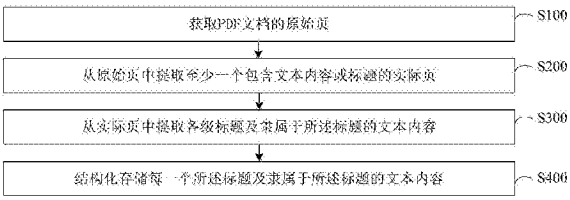
权利要求书2页 说明书10页 附图12页

(54)发明名称

一种PDF文档结构化信息提取方法及装置

(57)摘要

本申请实施例公开一种PDF文档结构化信息提取方法，所述方法包括：获取PDF文档的原始页；从所述原始页中提取至少一个包含文本内容或标题的实际页；从所述实际页中提取各级标题及隶属于所述标题的文本内容；结构化存储每一个所述标题及隶属于所述标题的文本内容。上述技术方案中的结构化信息提取方法能够把PDF文档中各级标题以及隶属于各级标题的相应文本内容提取出来，并结构化存储，从而得到结构化信息，使得PDF文档的结构化信息提取能够自动化实现，避免手工再处理，便捷高效。



CN 107358208 B

1.一种PDF文档结构化信息提取方法,其特征在于,所述方法包括:

获取PDF文档的原始页;

从所述原始页中提取至少一个包含文本内容或标题的实际页;

从所述实际页中提取各级标题及隶属于所述标题的文本内容;

结构化存储每一个所述标题及隶属于所述标题的文本内容;

从所述实际页中提取各级标题及隶属于所述标题的文本内容的步骤包括:

将每个第N级标题,及与该第N级标题所对应的内容,作为一个N级逻辑页,N取≥1的整数;

所述结构化存储每一个所述标题及隶属于所述标题的文本内容的步骤,包括:

结构化存储第1至第N+1级标题,及分别隶属于所述第1至第N+1级标题的文本内容,其中,隶属于第N+1级标题的文本内容为与该第N+1级标题对应的内容,隶属于第i级标题的文本内容为与该第i级标题对应的内容中除i+1级逻辑页之外的内容,i=1,2,…,N。

2.根据权利要求1所述的PDF文档结构化信息提取方法,其特征在于,从所述原始页中提取至少一个包含文本内容或标题的实际页的步骤,包括:

分别判断所述原始页中是否包含目录页、页眉和页脚;

将原始页中的目录页、页眉或页脚删除,得到至少一个实际页。

3.根据权利要求1所述的PDF文档结构化信息提取方法,其特征在于,从所述实际页中提取各级标题及隶属于所述标题的文本内容的步骤,还包括:

提取每个实际页中的第一级标题;

提取实际页中当前第一级标题与下一个第一级标题之间的内容,作为与当前第一级标题对应的内容;若当前第一级标题为实际页中最后一个第一级标题,提取该实际页中当前第一级标题之后的内容,作为与当前第一级标题对应的内容;

将每个第一级标题,及与该第一级标题所对应的内容,作为一个一级逻辑页;

若所述一级逻辑页中不存在下一级标题,所述结构化存储每一个所述标题及隶属于所述标题的文本内容的步骤,包括:

结构化存储每一个第一级标题及隶属于所述第一级标题的文本内容,其中,隶属于第一级标题的文本内容为与该第一级标题对应的内容。

4.根据权利要求3所述的PDF文档结构化信息提取方法,其特征在于,所述将每个第一级标题,及与该第一级标题所对应的内容,作为一个一级逻辑页的步骤之前,还包括以下步骤:

若当前实际页中没有第一级标题,将当前实际页的所有内容合并至上一个第一级标题对应的内容;

若当前实际页中的第一个第一级标题不在当前实际页的第一行,将所述当前实际页中第一个第一级标题之前的内容合并至上一个第一级标题对应的内容。

5.根据权利要求3所述的PDF文档结构化信息提取方法,其特征在于,从所述实际页中提取各级标题及隶属于所述标题的文本内容的步骤,还包括以下步骤:

分别从每一个N级逻辑页中提取第N+1级标题,及隶属于第N+1级标题的文本内容,N取≥1的整数。

6.根据权利要求5所述的PDF文档结构化信息提取方法,其特征在于,所述分别从每一

个N级逻辑页中提取第N+1级标题，及隶属于第N+1级标题的文本内容的步骤，包括：

提取每个N级逻辑页中的第N+1级标题；

提取当前第N+1级标题与下一个第N+1级标题之间的内容，作为与当前第N+1级标题对应的内容；若当前的第N+1级标题为N级逻辑页中最后一个第N+1级标题，提取该N级逻辑页中当前第N+1级标题之后的内容，作为与当前第N+1级标题对应的内容；

将每一个第N+1级标题，及与该第N+1级标题对应的内容，作为一个N+1级逻辑页。

7.根据权利要求5所述的PDF文档结构化信息提取方法，其特征在于，所述分别从每一个N级逻辑页中提取第N+1级标题，及隶属于第N+1级标题的文本内容的步骤包括：

确定每一个N级逻辑页中是否存在表格，若存在表格，将所述表格切分成表格区块，提取第N+1级标题及隶属于所述第N+1级标题的文本内容。

8.根据权利要求3-7任一项所述的PDF文档结构化信息提取方法，其特征在于，所述提取每个实际页中的第一级标题的步骤，包括：

获取实际页中的标题线及所述标题线在实际页中Y轴坐标；

若同一个实际页中当前标题线与下一个标题线的Y轴坐标之差小于3个Y轴单位时，将下一个标题线与当前标题线合并；

获取标题线之上离标题线最近的一行的文本内容作为实际页中的第一级标题。

9.一种PDF文档结构化信息提取装置，其特征在于，包括：

获取单元，用于获取PDF文档的原始页；

第一提取单元，用于从所述原始页中提取至少一个包含文本内容或标题的实际页；

第二提取单元，用于从所述实际页中提取各级标题及隶属于所述标题的文本内容；

存储单元，用于结构化存储每一个所述标题及隶属于所述标题的文本内容；

所述第二提取单元具体用于将每个第N级标题，及与该第N级标题所对应的内容，作为一个N级逻辑页，N取≥1的整数；

所述存储单元具体用于结构化存储第1至第N+1级标题，及分别隶属于所述第1至第N+1级标题的文本内容，其中，隶属于第N+1级标题的文本内容为与该第N+1级标题对应的内容，隶属于第i级标题的文本内容为与该第i级标题对应的内容中除i+1级逻辑页之外的内容，i＝1,2,…,N。

10.根据权利要求9所述的PDF文档结构化信息提取装置，其特征在于，所述第一提取单元，包括：

判断单元，用于分别判断所述原始页中是否包含目录页、页眉和页脚；

删除单元，用于将原始页中的目录页、页眉或页脚删除，得到至少一个实际页。

一种PDF文档结构化信息提取方法及装置

技术领域

[0001] 本申请涉及PDF文档信息提取领域，尤其涉及一种PDF文档结构化信息提取方法。此外，本申请还涉及一种PDF文档结构化信息提取装置。

背景技术

[0002] PDF(Portable Document Format，便携式文档格式)，是由Adobe Systems所发展出的文件格式，用于与应用程序、操作系统、硬件无关的方式进行文件交换，属于版式文档。PDF的页面之间相对独立，会忠实地再现原稿的每一个字符、颜色以及图象，但是PDF的存储是非结构化的数据存储格式，没有记录文档的逻辑结构，没有段落、表格等逻辑元素。

[0003] 提取PDF文档中的信息，通常采用OCR(Optical Character Recognition，光学字符识别)技术。但采用OCR技术所提取出来的PDF文档的信息，是以矢量的方式进行的渲染，每个字符之间是没有逻辑关系的(比如相邻、前后的关系)。提取出来的字符形成的文本仅是x、y、z三个坐标加上旋转量来渲染的矩阵。这样的文本存在格式和位置随意性大的问题，还需要手工再进行处理，才能得到具有明确层次结构的结构化信息。

[0004] 因此，采用现有方法提取PDF文档中的信息，提取到的文本中，文字格式和位置随意，无法便利地得到结构化信息，这是本领域技术人员亟待解决的问题。

发明内容

[0005] 本申请提供一种PDF文档结构化信息提取方法及一种PDF文档结构化信息提取装置，以解决通过现有技术无法便利地得到PDF文档结构化信息的问题。

[0006] 第一方面，本申请提供了一种PDF文档结构化信息提取方法，该方法包括：

[0007] 获取PDF文档的原始页；

[0008] 从所述原始页中提取至少一个包含文本内容或标题的实际页；

[0009] 从所述实际页中提取各级标题及隶属于所述标题的文本内容；

[0010] 结构化存储每一个所述标题及隶属于所述标题的文本内容。

[0011] 结合第一方面，在第一方面第一种可能的实现方式中，从所述原始页中提取至少一个包含文本内容或标题的实际页的步骤，包括：

[0012] 分别判断所述原始页中是否包含目录页、页眉和页脚；

[0013] 将原始页中的目录页、页眉或页脚删除，得到至少一个实际页。

[0014] 结合第一方面及上述可能的实现方式，在第一方面第二种可能的实现方式中，从所述实际页中提取各级标题及隶属于所述标题的文本内容的步骤，包括：

[0015] 提取每个实际页中的第一级标题；

[0016] 提取实际页中当前第一级标题与下一个第一级标题之间的内容，作为与当前第一级标题对应的内容；若当前第一级标题为实际页中最后一个第一级标题，提取该实际页中当前第一级标题之后的内容，作为与当前第一级标题对应的内容；

[0017] 将每个第一级标题，及与该第一级标题所对应的内容，作为一个一级逻辑页；

[0018] 若所述一级逻辑页中不存在下一级标题，所述结构化存储每一个所述标题及隶属于所述标题的文本内容的步骤，包括：

[0019] 结构化存储每一个第一级标题及隶属于所述第一级标题的文本内容，其中，隶属于第一级标题的文本内容为与该第一级标题对应的内容。

[0020] 结合第一方面及上述可能的实现方式，在第一方面第三种可能的实现方式中，所述将每个第一级标题，及与该第一级标题所对应的内容，作为一个一级逻辑页的步骤之前，还包括以下步骤：

[0021] 若当前实际页中没有第一级标题，将当前实际页的所有内容合并至上一个第一级标题对应的内容；

[0022] 若当前实际页中的第一个第一级标题不在当前实际页的第一行，将所述当前实际页中第一个第一级标题之前的内容合并至上一个第一级标题对应的内容。

[0023] 结合第一方面及上述可能的实现方式，在第一方面第四种可能的实现方式中，从所述实际页中提取各级标题及隶属于所述标题的文本内容的步骤，还包括以下步骤：

[0024] 分别从每一个N级逻辑页中提取第(N+1)级标题，及隶属于第(N+1)级标题的文本内容，N取≥1的整数。

[0025] 结合第一方面及上述可能的实现方式，在第一方面第五种可能的实现方式中，所述分别从每一个N级逻辑页中提取第(N+1)级标题，及隶属于第(N+1)级标题的文本内容的步骤，包括：

[0026] 提取每个N级逻辑页中的第N+1级标题；

[0027] 提取当前第N+1级标题与下一个第N+1级标题之间的内容，作为与当前第N+1级标题对应的内容；若当前的第N+1级标题为N级逻辑页中最后一个第N+1级标题，提取该N级逻辑页中当前第N+1级标题之后的内容，作为与当前第N+1级标题对应的内容；

[0028] 将每一个第N+1级标题，及与该第N+1级标题对应的内容，作为一个N+1级逻辑页；

[0029] 所述结构化存储每一个所述标题及隶属于所述标题的文本内容的步骤，包括：

[0030] 结构化存储第1至第N+1级标题，及分别隶属于所述第1至第N+1级标题的文本内容，其中，隶属于第N+1级标题的文本内容为与该第N+1级标题对应的内容，隶属于第i级标题的文本内容为与该第i级标题对应的内容中除i+1级逻辑页之外的内容，i＝1,2,…,N。

[0031] 结合第一方面及上述可能的实现方式，在第一方面第六种可能的实现方式中，所述分别从每一个N级逻辑页中提取第N+1级标题，及隶属于第N+1级标题的文本内容的步骤包括：

[0032] 确定每一个N级逻辑页中是否存在表格，若存在表格，将所述表格切分成表格区块，提取第N+1级标题及隶属于所述第N+1级标题的文本内容。

[0033] 结合第一方面及上述可能的实现方式，在第一方面第七种可能的实现方式中，所述提取每个实际页中的第一级标题的步骤，包括：

[0034] 获取实际页中的标题线及所述标题线在实际页中Y轴坐标；

[0035] 若同一个实际页中当前标题线与下一个标题线的Y轴坐标之差小于3个Y轴单位时，将下一个标题线与当前标题线合并；

[0036] 获取标题线之上离标题线最近的一行的文本内容作为实际页中的第一级标题。

[0037] 第二方面，本申请还提供了一种PDF文档结构化信息提取装置，包括：

[0038] 获取单元，用于获取PDF文档的原始页；

[0039] 第一提取单元，用于从所述原始页中提取至少一个包含文本内容或标题的实际页；

[0040] 第二提取单元，用于从所述实际页中提取各级标题及隶属于所述标题的文本内容；

[0041] 存储单元，用于结构化存储每一个所述标题及隶属于所述标题的文本内容。

[0042] 结合第二方面，在第二方面第一种可能的实现方式中，所述第一提取单元，包括：

[0043] 判断单元，用于分别判断所述原始页中是否包含目录页、页眉和页脚；

[0044] 删除单元，用于将原始页中的目录页、页眉或页脚删除，得到至少一个实际页。

[0045] 与现有技术相比，该方法首先从PDF文档的原始页中去除可能对结构化信息的提取产生干扰的部分，例如目录页、页眉、页脚等，生成实际页，从而完成从原始页中提取实际页的步骤。然后从实际页中把各级标题以及隶属于各级标题的相应文本内容提取出来，结构化存储，从而得到结构化信息，使得PDF文档的结构化信息提取能够自动化实现，避免手工处理，便捷高效。

附图说明

[0046] 为了更清楚地说明本申请的技术方案，下面将对实施例中所需要使用的附图作简单地介绍，显而易见地，对于本领域普通技术人员而言，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0047] 图1至图7为本申请的PDF文档结构化信息本提取方法的一个具体实施方式的流程图；

[0048] 图8至图19为本申请的PDF文档结构化信息本提取方法的一个实施例中子步骤的效果示意图；

[0049] 图20为本申请的PDF文档结构化信息本提取装置的一个实施例的结构示意图。

具体实施方式

[0050] 为使本发明的目的、技术方案和优点更加清楚，下面将结合实施例及附图对本发明作进一步详细的描述。

[0051] 请参考图1，在一个具体实施方式中，PDF文档结构化信息本提取方法包括：

[0052] S100获取PDF文档的原始页。

[0053] S200从原始页中提取至少一个包含文本内容或标题的实际页。

[0054] S300从实际页中提取各级标题及隶属于所述标题的文本内容。

[0055] S400结构化存储每一个所述标题及隶属于所述标题的文本内容。

[0056] 结构化信息是指信息经过分析后可分解成多个互相关联的组成部分，各组成部分间有明确的层次结构。在本申请中，PDF文档结构化信息系指从PDF文档中提取出来的文本，文本中各级标题及隶属于标题的文本内容具有明确的层次结构。结构化信息后续可以通过html、word、txt等多种格式的文件展现出来。

[0057] 结构化存储是指把需要多个文件的内容按树形结构和层次保存到一个文件中去。在本申请中，结构化存储每一个所述标题及隶属于所述标题的文本内容，是指将各级标题，

以及隶属于各级标题的内容，按照树形结构和层次进行存储，从而获得PDF的文档的结构化信息。

[0058] 上述的方法，首先从PDF文档的原始页中去除可能对结构化信息的提取产生干扰的部分，例如目录页、页眉、页脚等，生成实际页，从而完成从原始页中提取实际页的步骤。然后从实际页中把各级标题以及隶属于各级标题的相应文本内容提取出来，结构化存储，从而得到结构化信息，使得PDF文档的结构化信息提取能够自动化实现，避免手工处理，便捷高效。

[0059] 下面对上述S100-S400的步骤进行详细的描述。

[0060] S100的步骤中，PDF文档的原始页可以通过用户输入而获取，也可以从存储介质上获取。

[0061] S200的步骤中，请参考图2，具体可以包括S210的步骤和S220的步骤。

[0062] S210判断原始页中是否包含目录页、页眉和页脚。

[0063] 在步骤S210中，包括以下步骤：

[0064] S211获取当前原始页的页码、当前原始页的字符及字符总行数；

[0065] S212将当前原始页的页码及字符与第一预设规则进行匹配，确定当前原始页是否为目录页。

[0066] 在步骤S211中，当前原始页的页码、当前原始页的字符及字符总行数，可以通过PDFBox、iText等工具来直接获取。其中，PDFBox是一个操作PDF文档的Java平台类库，是开源工具，任何人可以在其基础上进行编程，用来创建PDF文档、操作已经存在的文档和提取文档的文本信息。iText也是一个开源的用于生成PDF文档的一个java类库，通过iText不仅可以生成PDF或rtf的文档，而且可以将XML、Html文件转化为PDF文件。

[0067] 在步骤S212中，第一预设规则可以通过开发者或者用户来预先设定。例如，第一预设规则中，判断是否为目录页的规则包括：当前原始页的页码为第一页或第二页，且当前原始页上的标题序号所占的行数超过当前原始页的字符总行数的40％，当前原始页为目录页；或者，当前原始页的页码为第一页或第二页，且当前原始页的字符中，依次出现“中文、非中文连续符号、序号”形式的字符串所占的行数超过当前原始页的字符总行数的40％，当前原始页为目录页；或者，当前原始页的页码为第一页或第二页，且当前原始页上的字符中包含预设关键字，当前原始页为目录页。

[0068] 举例来说，如果当前原始页的页码为第一页或者第二页，并且原始页上的标题序号，比如“1.1”、“1.1.1”、“1、”、“2、”等，所占的行数超过当前原始页的字符总行数的40％，就确定当前原始页为目录页。或者，如果当前原始页的页码为第一页或者第二页，当前原始页的字符中，例如“签收本合同之日起10天(即犹豫期)内您若要求退保，本公司仅扣除工本费……1.4”、“第一章……15”等这样形式的字符串所占的行数超过当前原始页的字符总行数的40％，就确定当前原始页为目录页。还或者，请参考图8，如果当前原始页的页码为第一页或者第二页，并且当前原始页上包括“第一章”、“第一条”、“有限公司”、“目录”等这些预设关键字时，就确定当前原始页为目录页。

[0069] 在步骤S212的第一预设规则中，又例如，判断原始页中是否包含页眉的规则包括：若连续3-5页原始页上第一行字符相同，就确定原始页包含页眉。还例如，判断原始页中是否包含页脚的规则包括：若连续3-5页原始页上最后一行字符相同，就确定原始页包含页

脚。

[0070] S220将原始页中的目录页、页眉或页脚删除，得到至少一个实际页。

[0071] 具体地，若原始页中包含目录页，则将原始页中的目录页整页删除；若原始页中包含页眉，则将原始页中的页眉删除；若原始页中包含页脚，则将原始页中的页脚删除。从而去除可能对PDF文档的结构化文本提取产生干扰的原始页或原始页中的部分内容，得到至少一个实际页。

[0072] 在进行S300的步骤之前，可以先对实际页中处于同一行的字符进行合并，形成行文本，如图9所示，对同一行的字符进行合并，可以事先通过PDFBox等工具获取每一个实际页上字符的坐标信息，包括X轴坐标和Y轴坐标，将Y轴坐标相同或者差距在预设范围之内的字符合并，获得行文本。以行文本为单位进行遍历，来提取各级标题及隶属于所述标题的文本内容的步骤，例如，通过遍历实际页中的行文本，来提取第一级标题及隶属于该第一级标题的文本内容；通过遍历一级逻辑页，来提取一级逻辑页中第二级标题及隶属于该第二级标题的文本内容。

[0073] S300的步骤和相应的S400的步骤可以包括两种情况，一种是一级逻辑页中不存在下一级标题的情况，另一种是一级逻辑页中还存在下一级标题的情况。

[0074] 请参考图3、图4、图10至图14。图3为第一个实施例中S300-S400的流程图，图4为第一个实施例中S311步骤的流程图。图10为第一个实施例中S311的步骤的效果示意图；图11为第一个实施例中S312的步骤的效果示意图；图12为第一个实施例中S313的步骤的效果示意图；图13为第一个实施例中S314的步骤的效果示意图；图14为第一个实施例中S410的步骤的效果示意图。在第一个实施例中，S300的步骤包括：

[0075] S311提取每个实际页中的第一级标题；

[0076] S312提取实际页中当前第一级标题与下一个第一级标题之间的内容，作为与当前第一级标题对应的内容；若当前第一级标题为实际页中最后一个第一级标题，提取该实际页中当前第一级标题之后的内容，作为与当前第一级标题对应的内容；

[0077] S313若当前实际页中没有第一级标题，将当前实际页的所有内容合并至上一个第一级标题对应的内容；若当前实际页中的第一个第一级标题不在当前实际页的第一行时，将所述当前实际页中第一个第一级标题之前的内容合并至上一个第一级标题对应的内容；

[0078] S314将每个第一级标题，及与该第一级标题所对应的内容，作为一个一级逻辑页。

[0079] 若所述一级逻辑页中不存在下一级标题，相应的S400的步骤，包括：

[0080] S410结构化存储每一个第一级标题及隶属于所述第一级标题的文本内容，其中，隶属于第一级标题的文本内容为与该第一级标题对应的内容。

[0081] 在S311的步骤中，可以根据实际页中字体的大小、字体的样式、文字内容或者标题线等来提取实际页中的第一级标题；所述字体的大小、字体的样式、文字内容或者标题线都可以通过PDFBox、iText等工具来获取。

[0082] 通过实际页中的字体大小提取实际页中的第一级标题，例如，通过比较每个行文本的字体的大小，如果当前行文本的字体最大，则确定当前行文本为第一级标题。通过实际页中的字体样式提取实际页中的第一级标题，例如，通过行文本的字体样式与预设字体样式进行匹配，确定当前行文本为第一级标题。上述行文本的字体大小，可以采用当前行文本中第一个字符的大小作为该行文本的字体大小，也可以采用当前行文本中多个大小相同的

多个字符的大小，作为该行文本的字体大小；上述行文本的字体样式，可以采用行文本中第一个字符的样式作为该行文本的字体样式，也可以采用当前行文本中多个样式相同的多个字符的样式，作为该行文本的字体样式。通过实际页中的文字内容提取实际页中的第一级标题，例如，通过文字内容与预设关键词进行匹配，如果文字内容中包含了“第一章”、“第二章”、“第一条”、“第一部分”等预设关键词，则确定当前行文本为第一级标题。

[0083] 针对一些第一级标题下设由标题线的PDF文档，还可以通过实际页中的标题线来提取第一级标题，请参考图4，具体地包括：

[0084] S3111获取实际页中的标题线及所述标题线在实际页中Y轴坐标；

[0085] S3112若同一个实际页中当前标题线与下一个标题线的Y轴坐标之差小于3个Y轴单位时，将下一个标题线与当前标题线合并；

[0086] S3113获取标题线之上离标题线最近的一行的文本作为实际页中的第一级标题。

[0087] 在S3111的步骤中，实际页中标题线的Y轴坐标可以通过PDFBox、iText等工具来获取。

[0088] 在S3113的步骤中，离标题线最近的一行，可以通过比较行文本的Y轴坐标与当前标题线的Y轴坐标之间的距离，来确定离标题线最近的一行，获取该行的文本作为实际页中的第一级标题。

[0089] 在从实际页中提取一级逻辑页的过程中，由于是按照实际页原有的顺序逐页进行提取，有可能会出现一种情况：理应作为同一个第一级标题对应的内容，却因为分别在前后两个实际页上而被拆开。通过上述S313的步骤，可以将实际页上这部分的内容合并到上一个一级标题对应的内容，从而保证每个一级逻辑页都能够包含完整的内容，克服普通的PDF文档信息获取方法中分页拆分的内容无法进行聚合的问题。

[0090] 请参考图5、图6、图15至图18，图5为第二个实施例中S300-S400的流程图，图6为第二个实施例中S320步骤的流程图。图15为第二个实施例中S321的步骤的效果示意图；图16为第二个实施例中S322的步骤的效果示意图；图17为第二个实施例中S323的步骤的效果示意图；图18为第二个实施例中S420中涉及隶属于第i级标题的文本内容的步骤的效果示意图。在第二个实施例中，S300的步骤包括：

[0091] S311提取每个实际页中的第一级标题；

[0092] S312提取实际页中当前第一级标题与下一个第一级标题之间的内容，作为与当前第一级标题对应的内容；若当前第一级标题为实际页中最后一个第一级标题，提取该实际页中当前第一级标题之后的内容，作为与当前第一级标题对应的内容；

[0093] S313若当前实际页中没有第一级标题，将当前实际页的所有内容合并至上一个第一级标题对应的内容；若当前实际页中的第一个第一级标题不在当前实际页的第一行时，将所述当前实际页中第一个第一级标题之前的内容合并至上一个第一级标题对应的内容；

[0094] S314将每个第一级标题，及与该第一级标题所对应的内容，作为一个一级逻辑页；

[0095] 若一级逻辑页中存在下一级标题，则还包括以下步骤，S320分别从每一个N级逻辑页中提取第(N+1)级标题，及隶属于第(N+1)级标题的文本内容，N取≥1的整数。这一步骤可以采用递归的过程，直到N级逻辑页中不包含第N+1级标题为止。具体地包括：

[0096] S321提取每个N级逻辑页中的第N+1级标题，N取≥1的整数；

[0097] S322提取当前第N+1级标题与下一个第N+1级标题之间的内容，作为与当前第N+1

级标题对应的内容;若当前的第N+1级标题为N级逻辑页中最后一个第N+1级标题,提取该N级逻辑页中当前第N+1级标题之后的内容,作为与当前第N+1级标题对应的内容;

[0098] S323将每一个第N+1级标题,及与该第N+1级标题对应的内容,作为一个N+1级逻辑页。

[0099] 相应地S400的步骤,包括:

[0100] S420结构化存储第1至第N+1级标题,及分别隶属于所述第1至第N+1级标题的文本内容,其中,隶属于第N+1级标题的文本内容为与该第N+1级标题对应的内容,隶属于第i级标题的文本内容为与该第i级标题对应的内容中除i+1级逻辑页之外的内容,i=1,2,…,N。

[0101] 此处需要说明的是,当N级逻辑页中包含第N+1级标题时,与第N级标题对应的内容,包含了隶属于第N级标题的文本内容,以及N+1级逻辑页。当N级逻辑页中不存在第N+1级标题时,与第N级标题对应的内容,就是隶属于第N级标题的文本内容。也就是说,在本申请中,与第N级标题对应的内容,和隶属于第N级标题的文本内容,二者之间是包含和被包含的关系。

[0102] 还需说明的是,若从实际页中提取到多个一级逻辑页,其中,部分一级逻辑页中不存在下一级标题,部分一级逻辑页中还存在下一级标题,则对于不存在下一级的一级逻辑页,结构化存储的步骤为第一个实施例中的结构化存储的步骤,对于还存在下一级标题的一级逻辑页,结构化存储的步骤为第二个实施例中的结构化存储的步骤,最后得到的PDF文档结构化信息中,包含了两个实施例中的结构化存储结果。

[0103] 针对一些第N级逻辑页中包含表格,且表格中有标题PDF文档,例如图19所示的PDF文档,则请参考图7和图19,图7为第三个实施例中S300-S400的流程图,图19为第三个实施例中320a中表格切分的示意图。在第三个实施例中,前述PDF文档结构化信息提取方法中,所述S320的步骤包括:

[0104] S320a确定每个N级逻辑页中是否存在表格,若存在表格,将所述表格切分成表格区块,提取第N+1级标题及隶属于所述第N+1级标题的文本内容。

[0105] 具体地,S320a的步骤中,“确定每个N级逻辑页中是否存在表格,若存在表格,将所述表格切分成表格区块”的步骤可以包括:

[0106] S320a1根据第二预设规则确定N级逻辑页中是否包含表格;所述第二预设规则包括:若第N级标题对应的内容中,同一行包括至少两个连续空格,且至少连续三行中所述空格的位置相同,确定当前N级逻辑页中存在表格,并以第一次出现至少两个连续空格的一行作为表格的起始行,最后一次出现至少两个连续空格的一行作为表格的结束行;

[0107] S320a2以表格的至少两个连续空格的位置为表格的纵向切割线,以表格中的空行为横向切割线,将表格切分成表格区块;

[0108] S320a3以从左往右,从上往下的顺序依次获取所述表格区块中的内容,与当前N级逻辑页中除该表格之外的内容一起,作为与当前N级逻辑页中的第N级标题所对应的内容。

[0109] S320a的步骤,通过将N级逻辑页中的表格进行切分,获取表格中的内容,代替了原有的表格,从而更新了原N级逻辑页中与第N级标题对应的内容,形成了新的N级逻辑页来替换原N级逻辑页。而在之后的步骤,也就是S321-323的从N级逻辑页中提取第N+1级标题及隶属于所述第N+1级标题的文本内容的步骤中,所述的N级逻辑页,是指新的N级逻辑页。

[0110] 需要说明的是,在处理一个PDF文档时的过程中,可能存在部分N级逻辑页存在表

格，部分N级逻辑页不存在表格的情况，此时，对于不存在表格的N级逻辑页，采用第二个实施例中S320的步骤来提取第N+1级标题及隶属于所述第N+1级标题的文本内容，对于存在含标题的表格的N级逻辑页，采用第三个实施例中S320a的步骤来提取第N+1级标题及隶属于所述第N+1级标题的文本内容。

[0111] 请参考图20，在另一个具体实施方式中，还提供一种PDF文档结构化信息提取装置，包括：

[0112] 获取单元1，用于获取PDF文档的原始页；

[0113] 第一提取单元2，用于从所述原始页中提取至少一个包含文本内容或标题的实际页；

[0114] 第二提取单元3，用于从所述实际页中提取各级标题及隶属于所述标题的文本内容；

[0115] 存储单元4，用于结构化存储每一个所述标题及隶属于所述标题的文本内容。

[0116] 可选地，第一提取单元2，包括：

[0117] 判断单元21，用于分别判断所述原始页中是否包含目录页、页眉和页脚；

[0118] 删除单元22，用于将原始页中的目录页、页眉或页脚删除，得到至少一个实际页。

[0119] 上述的PDF文档结构化信息提取装置，能够自动化提取PDF文档的结构化信息，避免手工处理，便捷高效。通过第一提取单元2删除了对PDF文档结构化信息提取有影响的目录页、页眉和页脚，从而进一步保证了结构化信息提取的准确性。

[0120] 可选地，第二提取单元3包括：

[0121] 第一级标题提取单元，用于提取每个实际页中的第一级标题；

[0122] 第一级内容提取单元，用于提取实际页中当前第一级标题与下一个第一级标题之间的内容，作为与当前第一级标题对应的内容；若当前第一级标题为实际页中最后一个第一级标题，提取该实际页中当前第一级标题之后的内容，作为与当前第一级标题对应的内容；

[0123] 一级逻辑页生成单元，用于将每个第一级标题，及与该第一级标题所对应的内容，作为一个一级逻辑页。

[0124] 存储单元4包括第一级存储单元，用于当所述一级逻辑页中不存在下一级标题时，结构化存储每一个第一级标题及隶属于所述第一级标题的文本内容，其中，隶属于第一级标题的文本内容为与该第一级标题对应的内容。

[0125] 可选地，第二提取单元3还包括合并单元，所述合并单元分别与第一级内容提取单元和一级逻辑页生成单元连接，用于若当前实际页中没有第一级标题，将当前实际页的所有内容合并至上一个第一级标题对应的内容；或用于若当前实际页中的第一个第一级标题不在当前实际页的第一行，将所述当前实际页中第一个第一级标题之前的内容合并至上一个第一级标题对应的内容。

[0126] 在从实际页中提取一级逻辑页的过程中，由于是按照实际页原有的顺序逐页进行提取，有可能会出现一种情况：理应作为同一个第一级标题对应的内容，却因为分别在前后两个实际页上而被拆开。通过上述的合并单元，可以将实际页上这部分的内容合并到上一个一级标题对应的内容，从而保证每个一级逻辑页都能够包含完整的内容，克服普通的PDF文档信息获取方法中分页拆分的内容无法进行聚合的问题。

[0127] 可选地，第二提取单元3还包括N级提取单元，用于分别从每一个N级逻辑页中提取第N+1级标题，及隶属于第N+1级标题的文本内容，N取≥1的整数。只有当N级逻辑页中存在第N+1级标题时，N级提取单元才运行，直到N级逻辑页中不存在第N+1级标题时，N级提取单元停止运行。

[0128] 可选地，N级提取单元包括：

[0129] 第N+1级标题提取单元，用于提取每个N级逻辑页中的第N+1级标题；

[0130] 第N+1级内容提取单元，用于提取当前第N+1级标题与下一个第N+1级标题之间的内容，作为与当前第N+1级标题对应的内容；若当前的第N+1级标题为N级逻辑页中最后一个第N+1级标题，提取该N级逻辑页中当前第N+1级标题之后的内容，作为与当前第N+1级标题对应的内容；

[0131] N+1级逻辑页生成单元，用于将每一个第N+1级标题，及与该第N+1级标题对应的内容，作为一个N+1级逻辑页。

[0132] 存储单元4还包括第N级存储单元，用于结构化存储第1至第N+1级标题，及分别隶属于所述第1至第N+1级标题的文本内容，其中，隶属于第N+1级标题的文本内容为与该第N+1级标题对应的内容，隶属于第i级标题的文本内容为与该第i级标题对应的内容中除i+1级逻辑页之外的内容，i＝1,2,…,N。N级存储单元只有当一级逻辑页中存在下一级标题时才运行，如果一级逻辑页中不存在下一级标题时，则第一级存储单元运行。

[0133] 需要说明的是，第二提取单元若从实际页中提取到多个一级逻辑页，其中，部分一级逻辑页中不存在下一级标题，部分一级逻辑页中还存在下一级标题，则对于不存在下一级的一级逻辑页，结构化存储采用第一级存储单元，对于还存在下一级标题的一级逻辑页，结构化存储采用第N级存储单元，在处理一个PDF文档时，两个存储单元可能都会使用到，也可以仅使用到其中一个存储单元。

[0134] 可选地，第二提取单元3还包括表格切分获取单元，用于确定每一个N级逻辑页中是否存在表格，若存在表格，将所述表格切分成表格区块，提取第N+1级标题及隶属于所述第N+1级标题的文本内容。当N级逻辑页中，与第N级标题对应的内容中包含表格，并且N+1级标题在表格中时，可以采用表格切分获取单元，直接切分表格，再提取第N+1级标题及隶属于所述第N+1级标题的文本内容。表格切分获取单元有时候可以单独使用，代替N级提取单元，有时候需要与N级提取单元配合使用。

[0135] 可选地，第一级标题提取单元可以包括：

[0136] 标题线获取单元，用于获取实际页中的标题线及所述标题线在实际页中Y轴坐标；

[0137] 标题线合并单元，用于当同一个实际页中当前标题线与下一个标题线的Y轴坐标之差小于3个Y轴单位时，将下一个标题线与当前标题线合并；

[0138] 第一级标题获取单元，用于获取标题线之上离标题线最近的一行的文本内容作为实际页中的第一级标题。

[0139] 本领域的技术人员可以清楚地了解到本发明实施例中的技术可借助软件加必需的通用硬件平台的方式来实现。基于这样的理解，本发明实施例中的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品可以存储在存储介质中，如ROM/RAM、磁碟、光盘等，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本发明各个实施例或者实施例的某些部分所

述的方法。

[0140] 本说明书中各个实施例之间相同相似的部分互相参见即可。以上所述的本发明实施方式并不构成对本发明保护范围的限定。

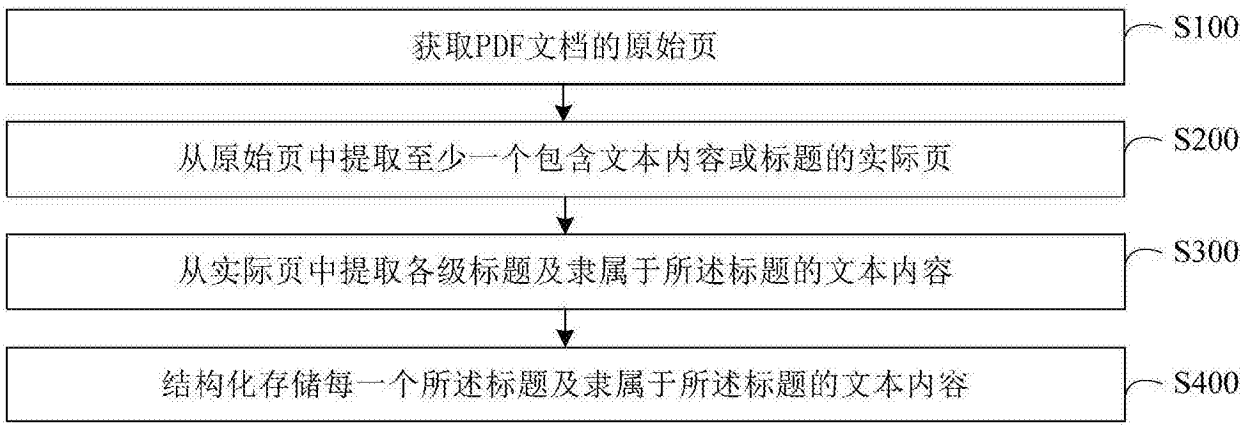

获取PDF文档的原始页
S100
从原始页中提取至少一个包含文本内容或标题的实际页
S200
从实际页中提取各级标题及隶属于所述标题的文本内容
S300
结构化存储每一个所述标题及隶属于所述标题的文本内容
S400

图1


分别判断原始页中是否包含目录页、页眉和页脚
S210
将原始页中的目录页、页眉或页脚删除，得到至少一个实际页
S220

图2

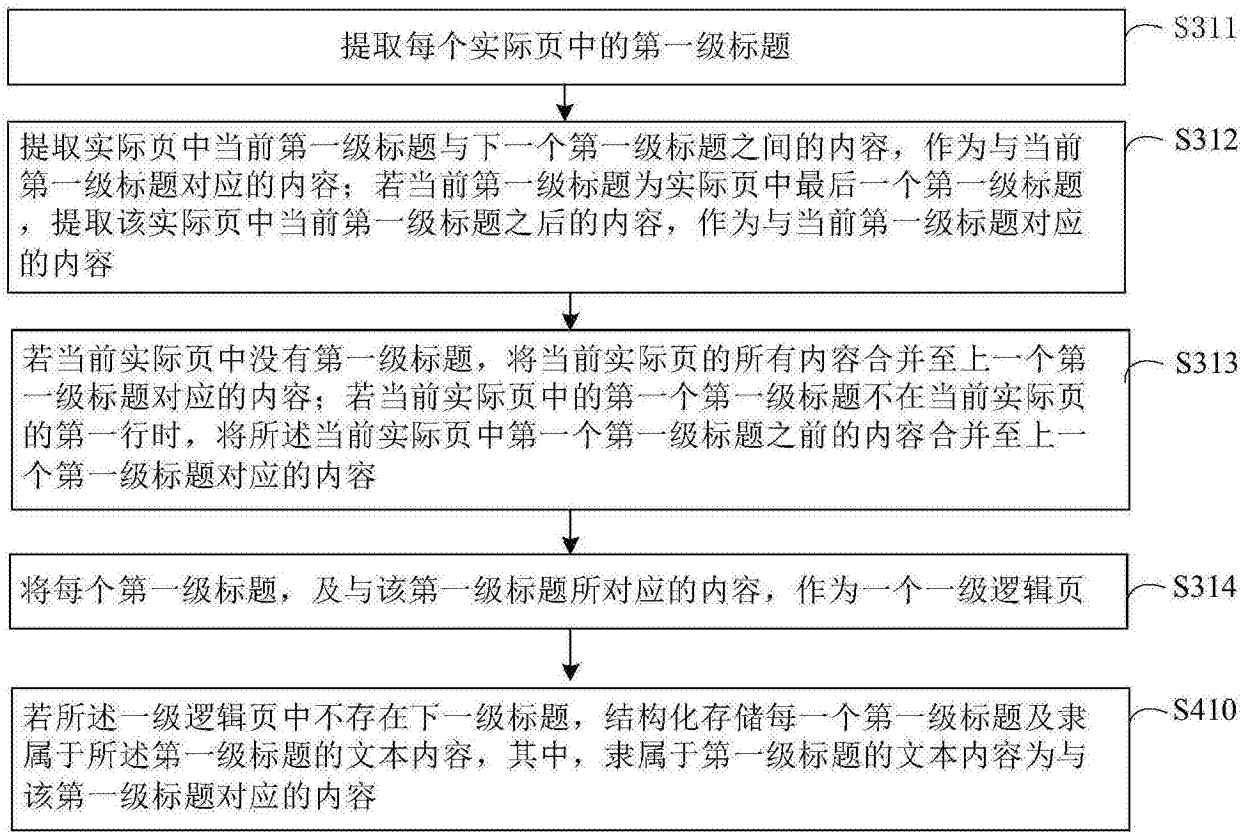

提取每个实际页中的第一级标题
S311
提取实际页中当前第一级标题与下一个第一级标题之间的内容，作为与当前第一级标题对应的内容；若当前第一级标题为实际页中最后一个第一级标题，提取该实际页中当前第一级标题之后的内容，作为与当前第一级标题对应的内容
S312
若当前实际页中没有第一级标题，将当前实际页的所有内容合并至上一个第一级标题对应的内容；若当前实际页中的第一个第一级标题不在当前实际页的第一行时，将所述当前实际页中第一个第一级标题之前的内容合并至上一个第一级标题对应的内容
S313
将每个第一级标题，及与该第一级标题所对应的内容，作为一个一级逻辑页
S314
若所述一级逻辑页中不存在下一级标题，结构化存储每一个第一级标题及隶属于所述第一级标题的文本内容，其中，隶属于第一级标题的文本内容为与该第一级标题对应的内容
S410

图3

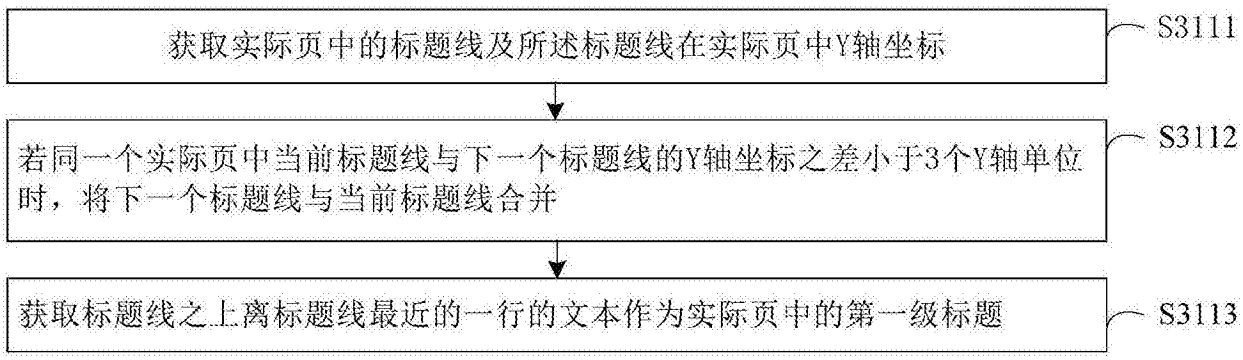

获取实际页中的标题线及所述标题线在实际页中Y轴坐标
S3111
若同一个实际页中当前标题线与下一个标题线的Y轴坐标之差小于3个Y轴单位时，将下一个标题线与当前标题线合并
S3112
获取标题线之上离标题线最近的一行的文本作为实际页中的第一级标题
S3113

图4

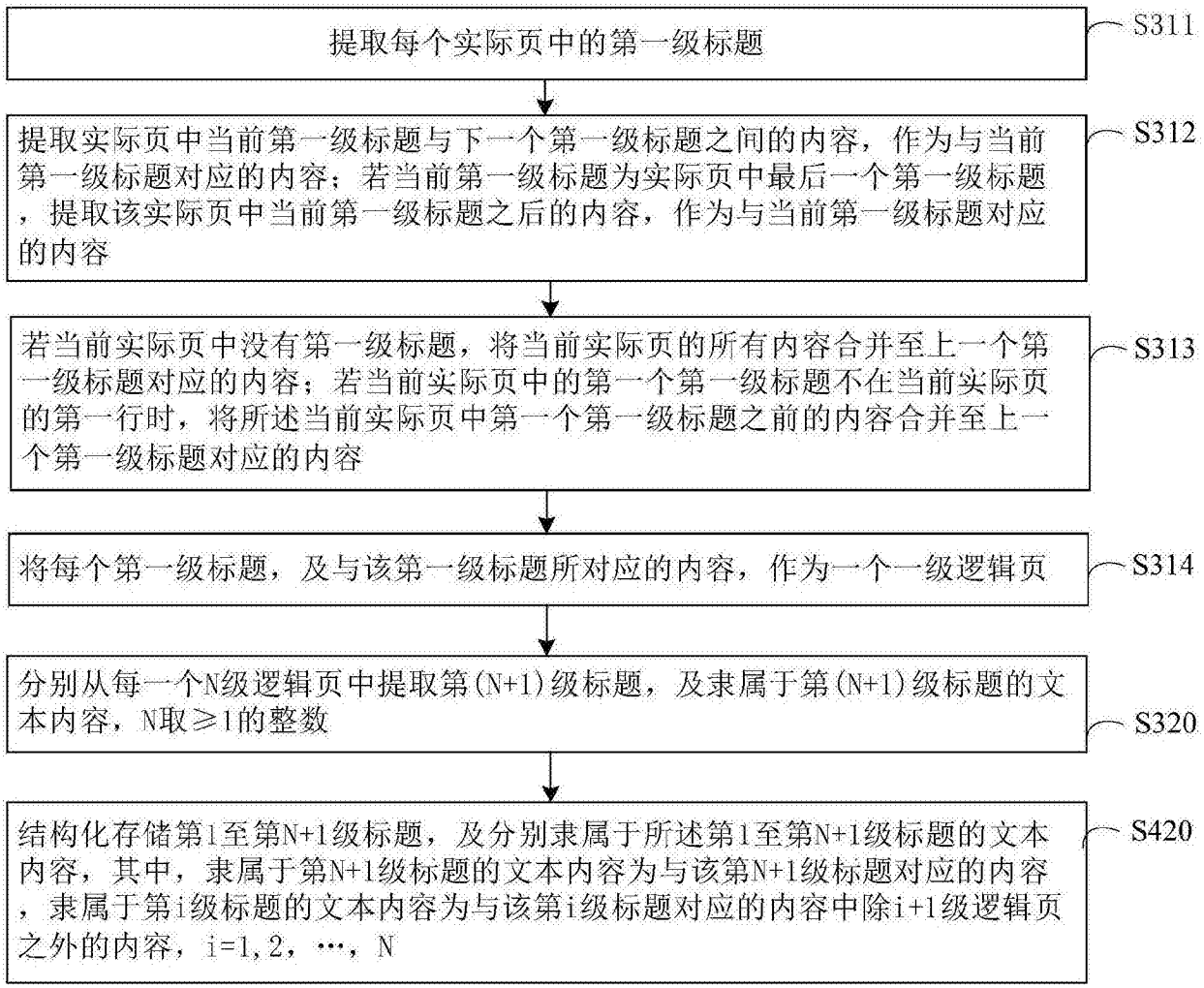

提取每个实际页中的第一级标题
S311
提取实际页中当前第一级标题与下一个第一级标题之间的内容，作为与当前第一级标题对应的内容；若当前第一级标题为实际页中最后一个第一级标题，提取该实际页中当前第一级标题之后的内容，作为与当前第一级标题对应的内容
S312
若当前实际页中没有第一级标题，将当前实际页的所有内容合并至上一个第一级标题对应的内容；若当前实际页中的第一个第一级标题不在当前实际页的第一行时，将所述当前实际页中第一个第一级标题之前的内容合并至上一个第一级标题对应的内容
S313
将每个第一级标题，及与该第一级标题所对应的内容，作为一个一级逻辑页
S314
分别从每一个N级逻辑页中提取第(N+1)级标题，及隶属于第(N+1)级标题的文本内容，N取≥1的整数
S320
结构化存储第1至第N+1级标题，及分别隶属于所述第1至第N+1级标题的文本内容，其中，隶属于第N+1级标题的文本内容为与该第N+1级标题对应的内容，隶属于第i级标题的文本内容为与该第i级标题对应的内容中除i+1级逻辑页之外的内容，i=1,2，…，N
S420

图5

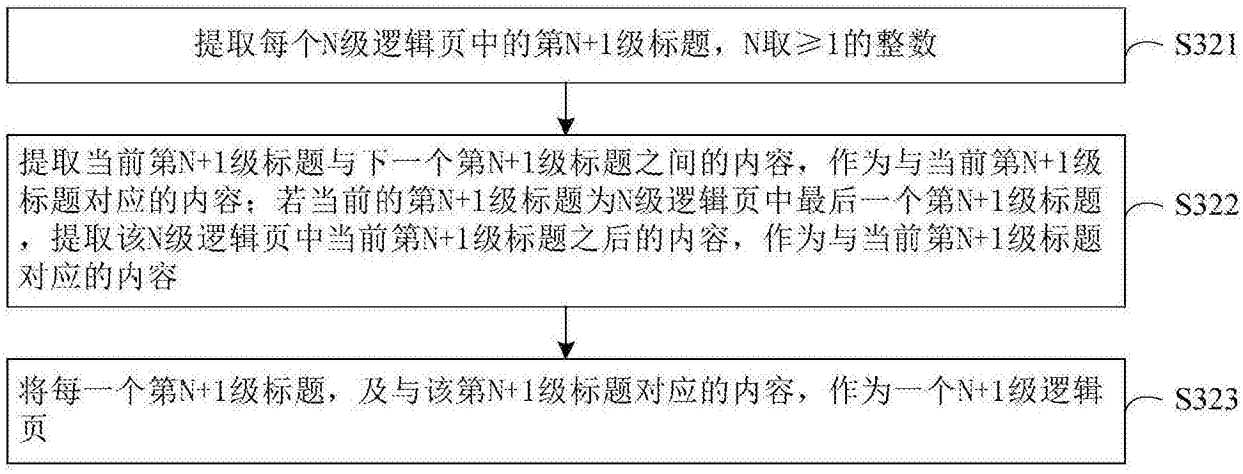

提取每个N级逻辑页中的第N+1级标题，N取≥1的整数
S321
提取当前第N+1级标题与下一个第N+1级标题之间的内容，作为与当前第N+1级标题对应的内容；若当前的第N+1级标题为N级逻辑页中最后一个第N+1级标题，提取该N级逻辑页中当前第N+1级标题之后的内容，作为与当前第N+1级标题对应的内容
S322
将每一个第N+1级标题，及与该第N+1级标题对应的内容，作为一个N+1级逻辑页
S323

图6

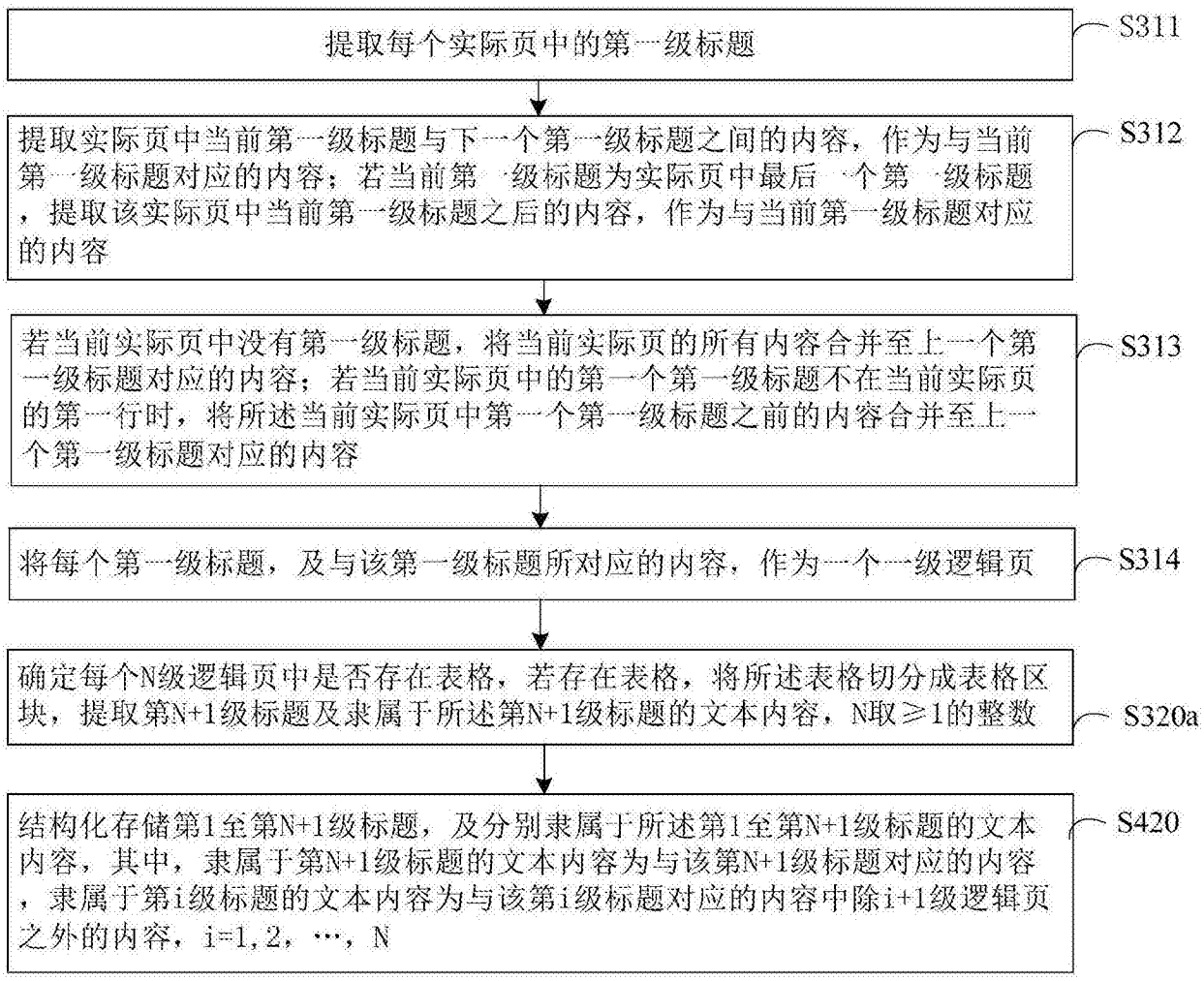


图7

目 录

图8

第一条 附加保险合同的构成

1.1 [illegible]

1.2 [illegible]

第二条 保险合同成立与生效

2.1 [illegible]

2.2 [illegible] 24 [illegible]

第三条 承保范围

3.1 [illegible] 30 [illegible] 55 [illegible]

第四条 保险责任

4.1 [illegible]

图9

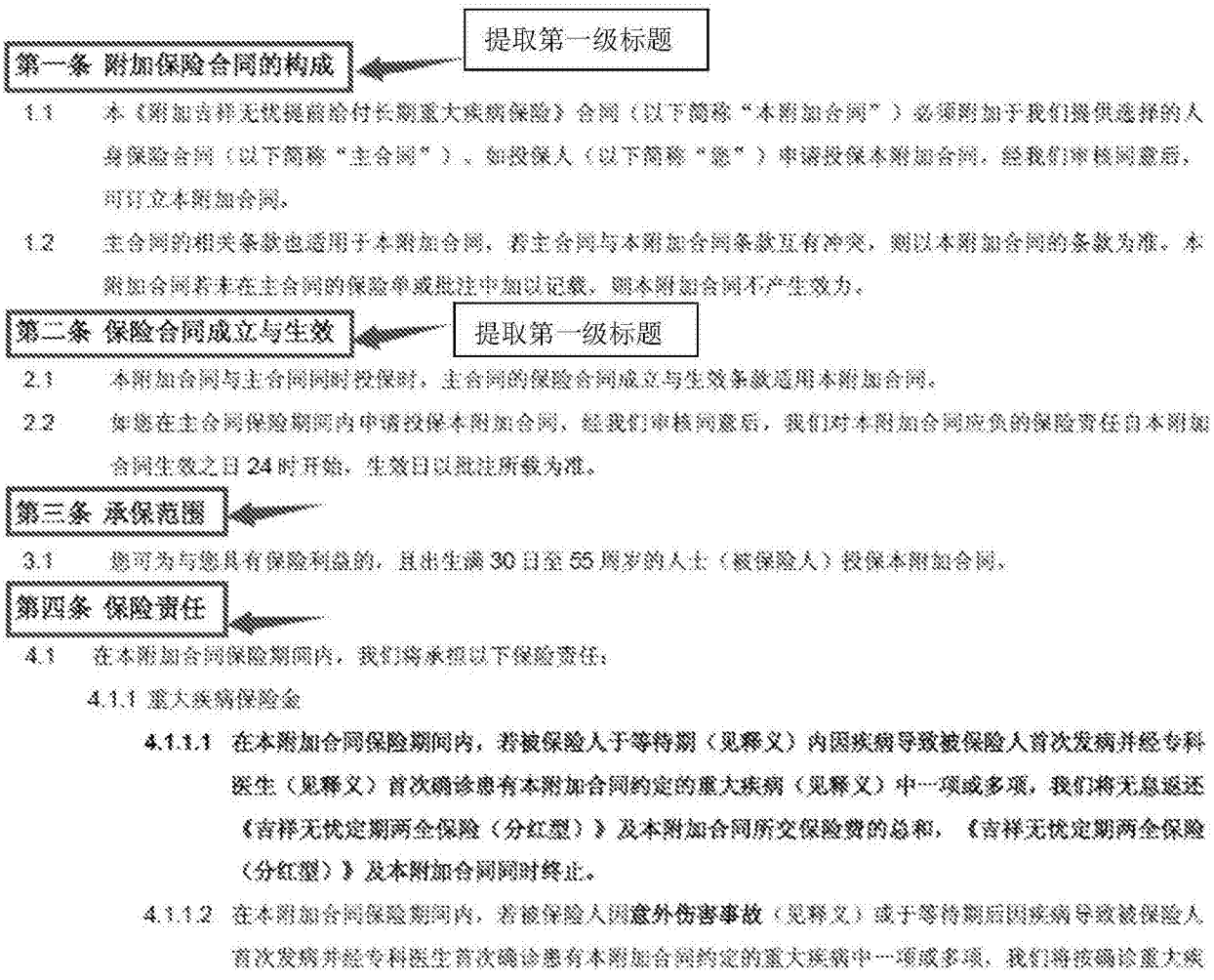

提取第一级标题
第一条 附加保险合同的构成
1.1 本《附加吉祥无忧提前给付长期重大疾病保险》合同（以下简称“本附加合同”）必须附加于我们提供选择的人身保险合同（以下简称“主合同”）。如投保人（以下简称“您”）申请投保本附加合同，经我们审核同意后，可订立本附加合同。
1.2 主合同的相关条款也适用于本附加合同，若主合同与本附加合同条款互有冲突，则以本附加合同的条款为准。本附加合同若未在主合同的保险单或批注中加以记载，则本附加合同不产生效力。
提取第一级标题
第二条 保险合同成立与生效
2.1 本附加合同与主合同同时投保时，主合同的保险合同成立与生效条款适用本附加合同。
2.2 如您在主合同保险期间内申请投保本附加合同，经我们审核同意后，我们对本附加合同应负的保险责任自本附加合同生效之日 24 时开始，生效日以批注所载为准。
第三条 承保范围
3.1 您可为与您具有保险利益的，且出生满 30 日至 55 周岁的人士（被保险人）投保本附加合同。
第四条 保险责任
4.1 在本附加合同保险期间内，我们将承担以下保险责任：
4.1.1 重大疾病保险金
4.1.1.1 在本附加合同保险期间内，若被保险人于等待期（见释义）内因疾病导致被保险人首次发病并经专科医生（见释义）首次确诊患有本附加合同约定的重大疾病（见释义）中一项或多项，我们将无息返还《吉祥无忧定期两全保险（分红型）》及本附加合同所交保险费的总和，《吉祥无忧定期两全保险（分红型）》及本附加合同同时终止。
4.1.1.2 在本附加合同保险期间内，若被保险人因意外伤害事故（见释义）或于等待期后因疾病导致被保险人首次发病并经专科医生首次确诊患有本附加合同约定的重大疾病中一项或多项，我们将按确诊重大疾

图10

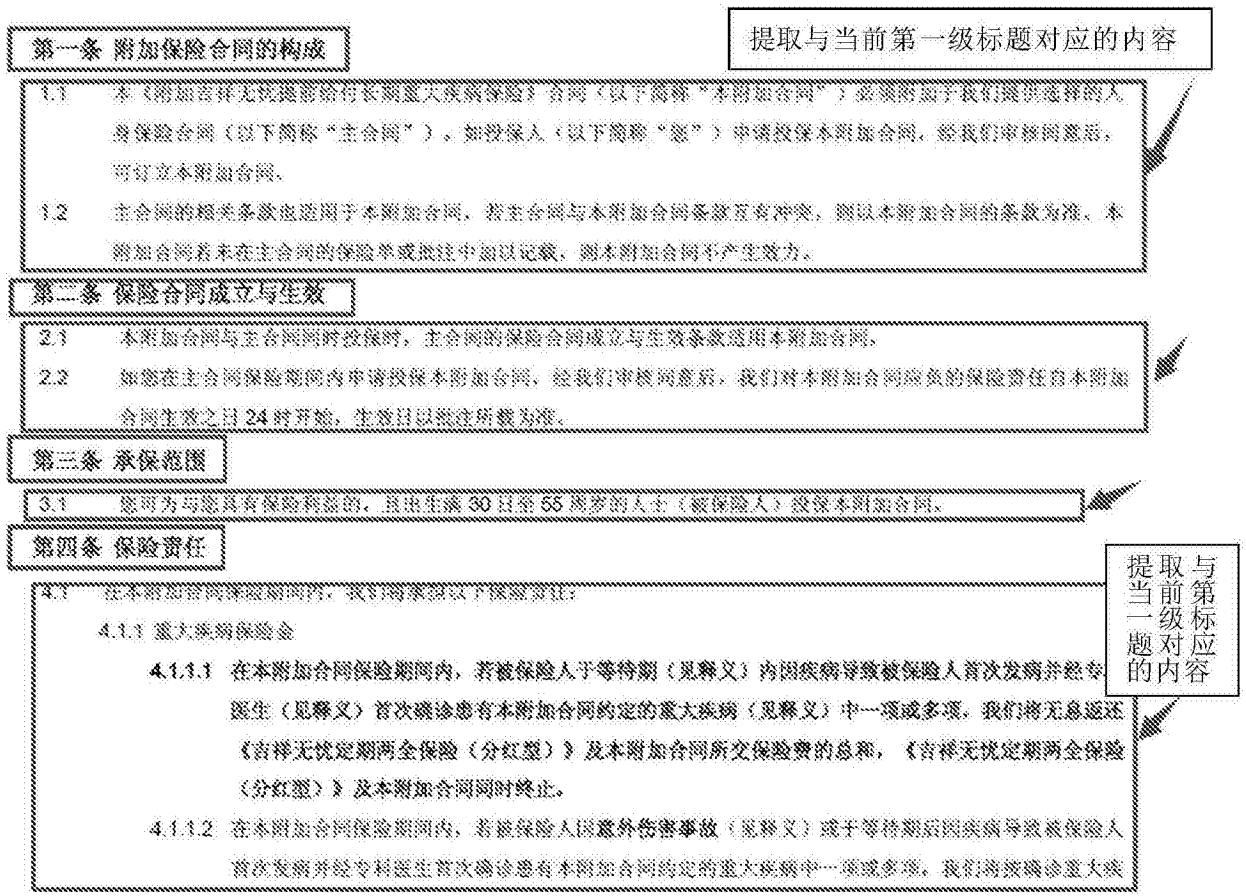

提取与当前第一级标题对应的内容
第一条 附加保险合同的构成
1.1 本《附加吉祥无忧提前给付长期重大疾病保险》合同（以下简称“本附加合同”）必须附属于我们提供选择的人身保险合同（以下简称“主合同”），如投保人（以下简称“您”）申请投保本附加合同，经我们审核同意后，可订立本附加合同。
1.2 主合同的相关条款也适用于本附加合同，若主合同与本附加合同条款互有冲突，则以本附加合同的条款为准。本附加合同若未在主合同的保险单或批注中加以记载，则本附加合同不产生效力。
第二条 保险合同成立与生效
2.1 本附加合同与主合同同时投保时，主合同的保险合同成立与生效条款适用本附加合同。
2.2 如您在主合同保险期间内申请投保本附加合同，经我们审核同意后，我们对本附加合同应负的保险责任自本附加合同生效之日 24 时开始，生效日以批注所载为准。
第三条 承保范围
3.1 您可为与您具有保险利益的、自出生满 30 日至 55 周岁的人士（被保险人）投保本附加合同。
第四条 保险责任
4.1 在本附加合同保险期间内，我们将承担以下保险责任：
4.1.1 重大疾病保险金
4.1.1.1 在本附加合同保险期间内，若被保险人于等待期（见释义）内因疾病导致被保险人首次发病并经专
医生（见释义）首次确诊患有本附加合同约定的重大疾病（见释义）中一项或多项，我们将无息返还
《吉祥无忧定期两全保险（分红型）》及本附加合同所交保险费的总和，《吉祥无忧定期两全保险
（分红型）》及本附加合同同时终止。
4.1.1.2 在本附加合同保险期间内，若被保险人因意外伤害事故（见释义）或于等待期后因疾病导致被保险人
首次发病并经专科医生首次确诊患有本附加合同约定的重大疾病中一项或多项，我们将确诊重大疾
提取与当前第一级对应的内容

图11

7.2 在我们调整保险费率后，您当自调整后的首个保险单周年日起交付续期的保险费率所计算的保险费。

7.3 若本附加合同保险费率发生调整，我们将及时通知您。

第八条 受益人

8.1 除本附加合同另有约定外，重大疾病保险金受益人为被保险人本人。

8.2 受益人为多人时，可以确定受益顺序和受益份额；如果没有确定份额，各受益人按照相等份额享有受益权。

8.3 被保险人为无民事行为能力人或限制民事行为能力人的，可以由其监护人指定受益人。

8.4 您或者被保险人可以变更受益人并书面通知我们。我们收到变更受益人的书面通知后，在保险单或其他保险凭证上批注或附贴批单。

将所述当前实际页中第一个第一级标题之前的内容合并至上一个第一级标题对应的内容

8.5 您在指定和变更受益人时，必须经过被保险人同意。

8.6 被保险人身故后，有下列情形之一的，保险金作为被保险人的遗产，由我们依照《中华人民共和国继承法》的规定履行给付保险金的义务：

(1) 没有指定受益人，或者受益人指定不明无法确定的；

(2) 受益人先于被保险人身故，没有其他受益人的；

(3) 受益人依法丧失受益权或者放弃受益权，没有其他受益人的。

8.7 受益人与被保险人在同一事件中身故，且不能确定身故先后顺序的，推定受益人身故在先。

8.8 受益人故意造成被保险人身故、伤残、疾病，或者故意杀害被保险人未遂的，该受益人丧失受益权。

第九条 保险金申请

9.1 重大疾病保险金的申请人为重大疾病保险金受益人。在申请重大疾病保险金时，申请人须填写保险金给付申请书，并提供下列证明和资料：

(1) 保险合同；

(2) 申请人的有效身份证件；

图12

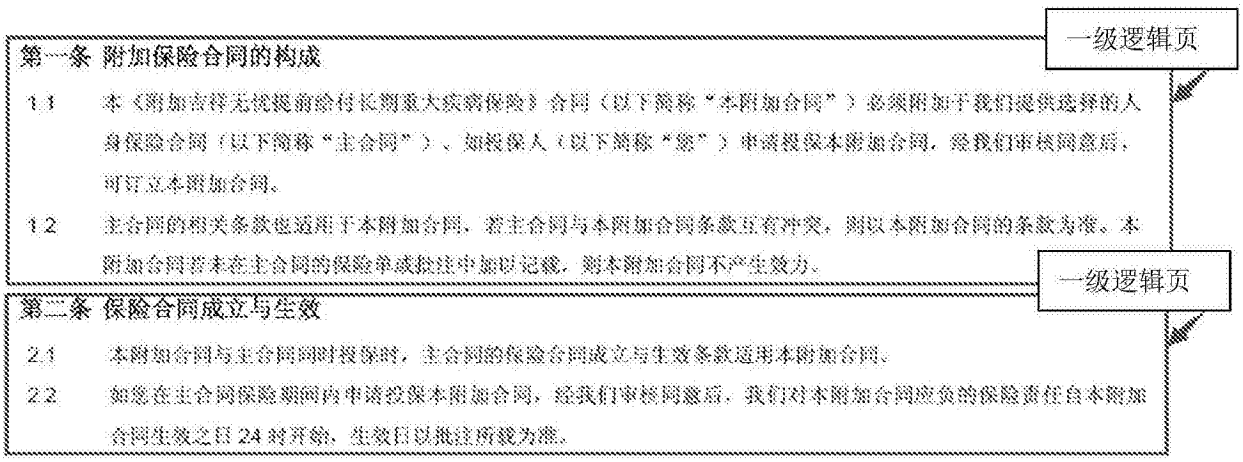

一级逻辑页
第一条 附加保险合同的构成
1.1 本《附加吉祥无忧提前给付长期重大疾病保险》合同（以下简称“本附加合同”）必须附加于我们提供选择的人身保险合同（以下简称“主合同”）、如投保人（以下简称“您”）申请投保本附加合同，经我们审核同意后，可订立本附加合同。
1.2 主合同的相关条款也适用于本附加合同，若主合同与本附加合同条款互有冲突，则以本附加合同的条款为准。本附加合同若未在主合同的保险单或批注中加以记载，则本附加合同不产生效力。
一级逻辑页
第二条 保险合同成立与生效
2.1 本附加合同与主合同同时投保时，主合同的保险合同成立与生效条款适用本附加合同。
2.2 如您在主合同保险期间内申请投保本附加合同，经我们审核同意后，我们对本附加合同应负的保险责任自本附加合同生效之日 24 时开始，生效日以批注所载为准。

图13

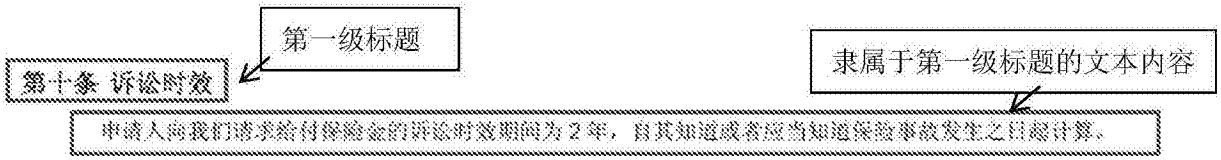

第一级标题
第十条 诉讼时效
隶属于第一级标题的文本内容
申请人向我们请求给付保险金的诉讼时效期间为 2 年，自其知道或者应当知道保险事故发生之日起计算。

图14

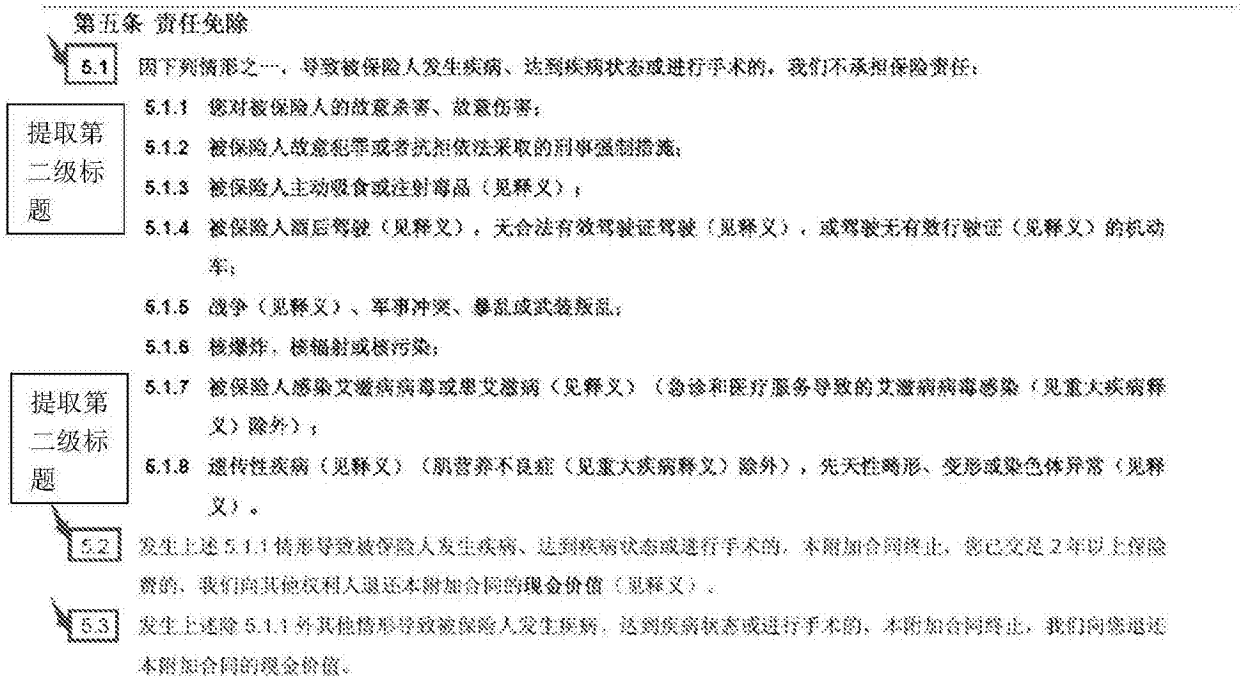

第五条 责任免除
5.1 因下列情形之一，导致被保险人发生疾病、达到疾病状态或进行手术的，我们不承担保险责任：
提取第二级标题
5.1.1 您对被保险人的故意杀害、故意伤害；
5.1.2 被保险人故意犯罪或者抗拒依法采取的刑事强制措施；
5.1.3 被保险人主动吸食或注射毒品（见释义）；
5.1.4 被保险人醉后驾驶（见释义），无合法有效驾驶证驾驶（见释义），或驾驶无有效行驶证（见释义）的机动车；
5.1.5 战争（见释义）、军事冲突、暴乱或武装叛乱；
5.1.6 核爆炸、核辐射或核污染；
提取第二级标题
5.1.7 被保险人感染艾滋病病毒或患艾滋病（见释义）（急诊和医疗服务导致的艾滋病病毒感染（见重大疾病释义）除外）；
5.1.8 遗传性疾病（见释义）（肌营养不良症（见重大疾病释义）除外），先天性畸形、变形或染色体异常（见释义）。
5.2 发生上述 5.1.1 情形导致被保险人发生疾病、达到疾病状态或进行手术的，本附加合同终止，您已交足 2 年以上保险费的，我们向其他权利人退还本附加合同的现金价值（见释义）。
5.3 发生上述除 5.1.1 外其他情形导致被保险人发生疾病、达到疾病状态或进行手术的，本附加合同终止，我们向您退还本附加合同的现金价值。

图15

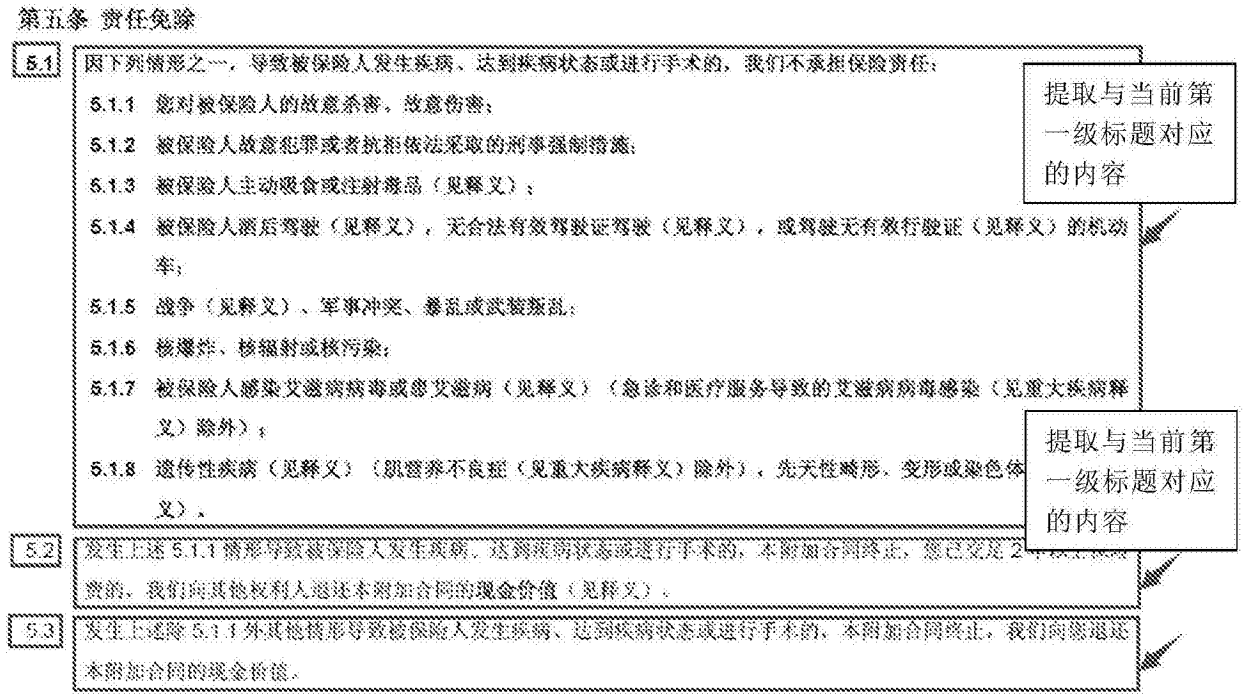

第五条 责任免除
5.1
因下列情形之一，导致被保险人发生疾病、达到疾病状态或进行手术的，我们不承担保险责任：
5.1.1 您对被保险人的故意杀害、故意伤害；
5.1.2 被保险人故意犯罪或者抗拒依法采取的刑事强制措施；
5.1.3 被保险人主动吸食或注射毒品（见释义）；
5.1.4 被保险人酒后驾驶（见释义），无合法有效驾驶证驾驶（见释义），或驾驶无有效行驶证（见释义）的机动车；
5.1.5 战争（见释义）、军事冲突、暴乱或武装叛乱；
5.1.6 核爆炸、核辐射或核污染；
5.1.7 被保险人感染艾滋病病毒或患艾滋病（见释义）（急诊和医疗服务导致的艾滋病病毒感染（见重大疾病释义）除外）；
5.1.8 遗传性疾病（见释义）（肌营养不良症（见重大疾病释义）除外），先天性畸形、变形或染色体异常（见释义）。
5.2
发生上述5.1.1情形导致被保险人发生疾病，达到疾病状态或进行手术的，本附加合同终止，您已交足2年以上保险费的，我们向其他权利人退还本附加合同的现金价值（见释义）。
5.3
发生上述除5.1.1外其他情形导致被保险人发生疾病、达到疾病状态或进行手术的，本附加合同终止，我们向您退还本附加合同的现金价值。
提取与当前第一级标题对应的内容
提取与当前第一级标题对应的内容

图16

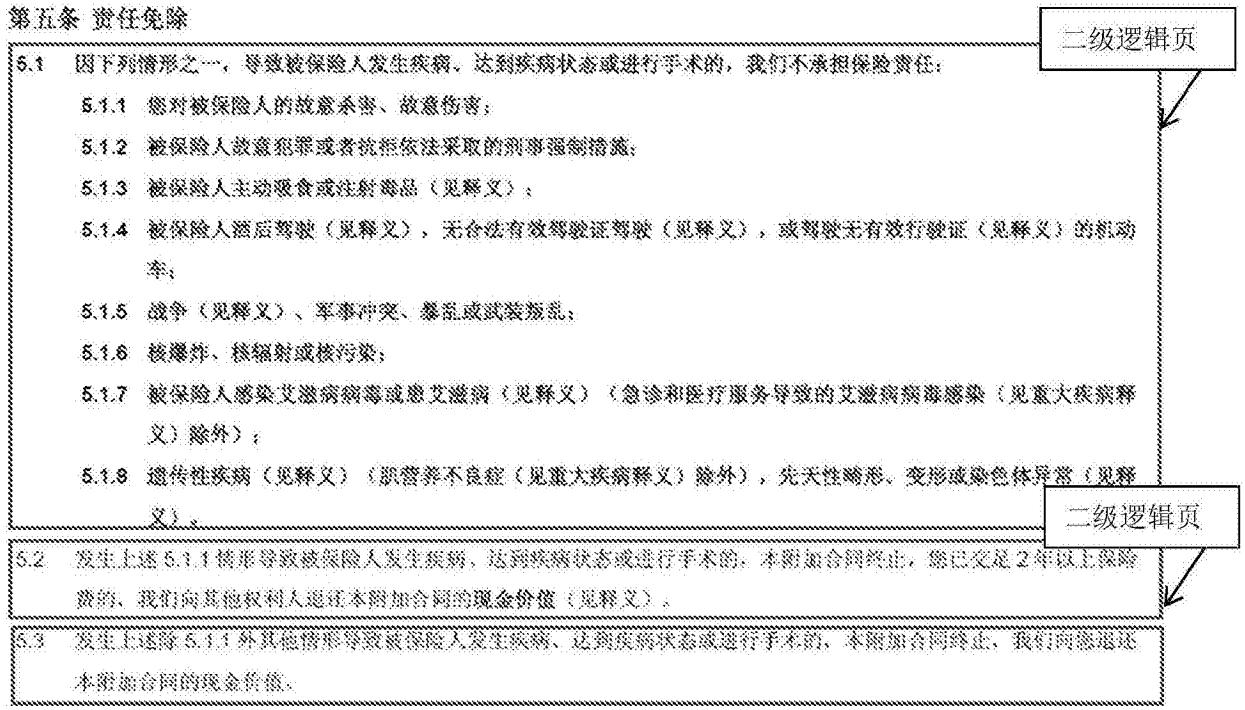

第五条 责任免除
5.1 因下列情形之一，导致被保险人发生疾病、达到疾病状态或进行手术的，我们不承担保险责任：
5.1.1 您对被保险人的故意杀害、故意伤害；
5.1.2 被保险人故意犯罪或者抗拒依法采取的刑事强制措施；
5.1.3 被保险人主动吸食或注射毒品（见释义）；
5.1.4 被保险人酒后驾驶（见释义），无合法有效驾驶证驾驶（见释义），或驾驶无有效行驶证（见释义）的机动车；
5.1.5 战争（见释义）、军事冲突、暴乱或武装叛乱；
5.1.6 核爆炸、核辐射或核污染；
5.1.7 被保险人感染艾滋病病毒或患艾滋病（见释义）（急诊和医疗服务导致的艾滋病病毒感染（见重大疾病释义）除外）；
5.1.8 遗传性疾病（见释义）（肌营养不良症（见重大疾病释义）除外），先天性畸形、变形或染色体异常（见释义）。
5.2 发生上述5.1.1情形导致被保险人发生疾病，达到疾病状态或进行手术的，本附加合同终止，您已交足2年以上保险费的，我们向其他权利人退还本附加合同的现金价值（见释义）。
5.3 发生上述除5.1.1外其他情形导致被保险人发生疾病、达到疾病状态或进行手术的，本附加合同终止，我们向您退还本附加合同的现金价值。
二级逻辑页
二级逻辑页

图17

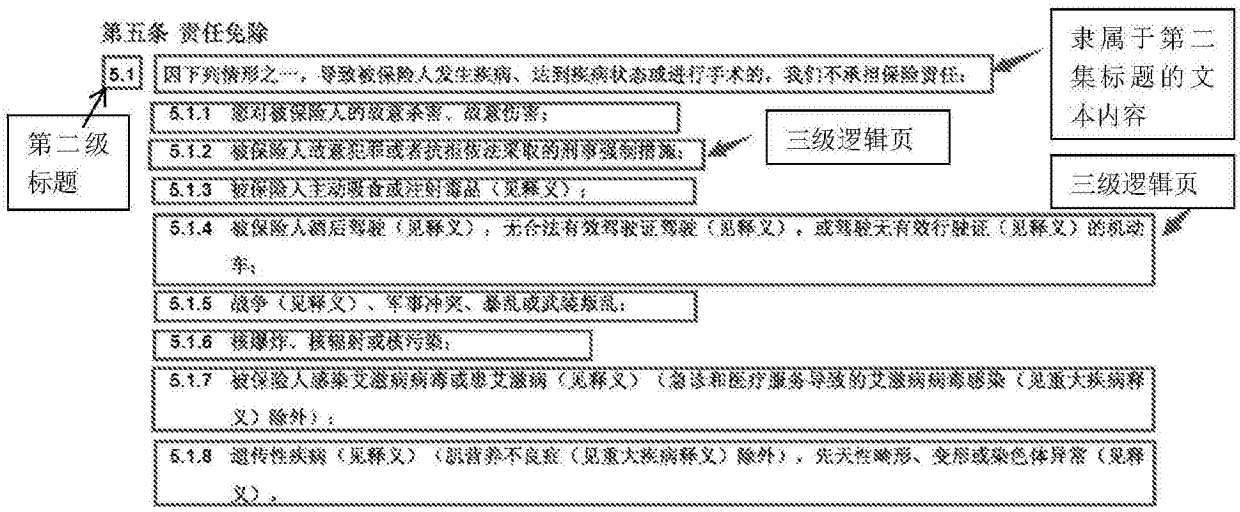

第五条 责任免除
5.1
因下列情形之一，导致被保险人发生疾病、达到疾病状态或进行手术的，我们不承担保险责任：
5.1.1 您对被保险人的故意杀害、故意伤害；
5.1.2 被保险人故意犯罪或者抗拒依法采取的刑事强制措施；
5.1.3 被保险人主动吸食或注射毒品（见释义）；
5.1.4 被保险人醉后驾驶（见释义），无合法有效驾驶证驾驶（见释义），或驾驶无有效行驶证（见释义）的机动车；
5.1.5 战争（见释义）、军事冲突、暴乱或武装叛乱；
5.1.6 核爆炸、核辐射或核污染；
5.1.7 被保险人感染艾滋病病毒或患艾滋病（见释义）（输血和医疗服务导致的艾滋病病毒感染（见重大疾病释义）除外）；
5.1.8 遗传性疾病（见释义）（胎儿营养不良症（见重大疾病释义）除外），先天性畸形、变形或染色体异常（见释义）。
第二级标题
隶属于第二集标题的文本内容
三级逻辑页
三级逻辑页

图18

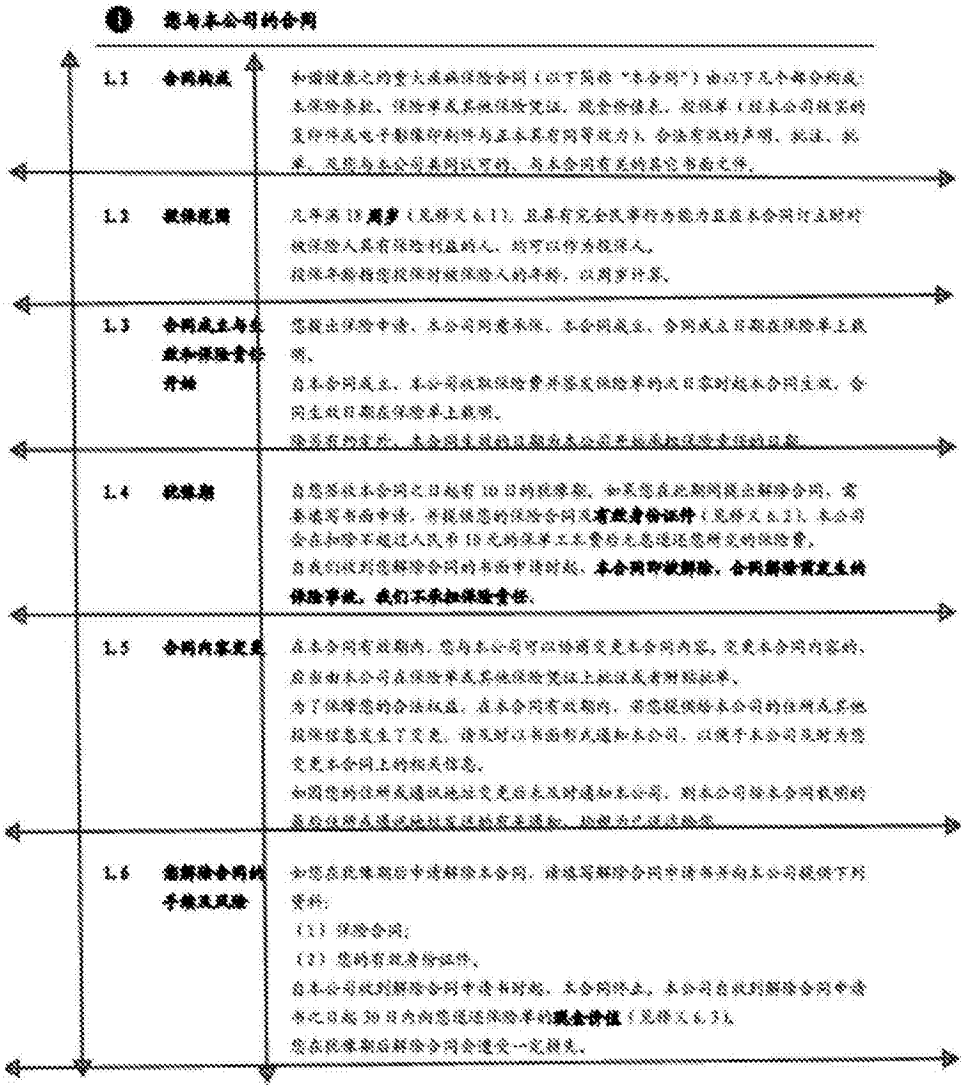

1 您与本公司的合同
1.1 合同构成
1.2 犹豫期
1.3 合同成立与生效和保险责任开始
1.4 撤销权
1.5 合同内容变更
1.6 您解除合同时手续及风险

图19

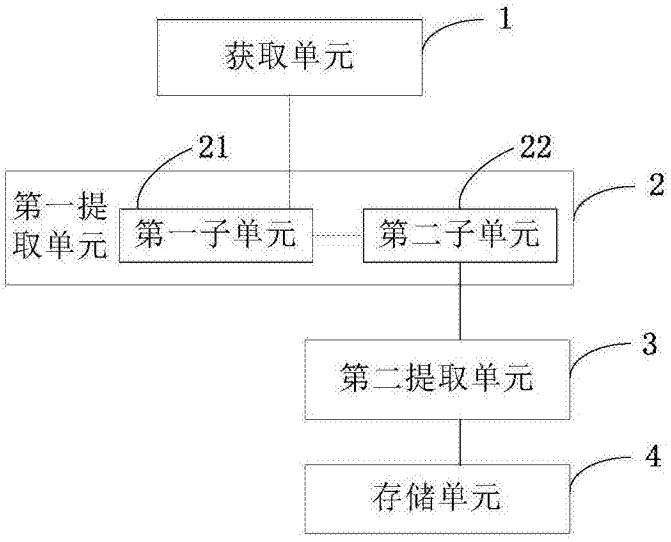


图20