(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 109558123 B
(45)授权公告日 2022.09.16

(21)申请号 201811463252.2
(22)申请日 2018.12.03
(65)同一申请的已公布的文献号
申请公布号 CN 109558123 A
(43)申请公布日 2019.04.02
(73)专利权人 掌阅科技股份有限公司
地址 100124 北京市朝阳区四惠大厦2029E
(72)发明人 吴馥江 黄鑫霞
(74)专利代理机构 北京市浩天知识产权代理事务所(普通合伙) 11276
专利代理师 宋菲 赵娅
(51)Int.Cl.
G06F 8/30(2018.01)
G06F 16/955(2019.01)
(56)对比文件
CN 101094135 A,2007.12.26
CN 102479216 A,2012.05.30
CN 108664511 A,2018.10.16
CN 105630780 A,2016.06.01
技术杂谈哈哈哈."如何用 Python 爬取网页制作电子书".《https://blog.csdn.net/GitChat/article/details/79141842》.2018,
技术杂谈哈哈哈."如何用 Python 爬取网页制作电子书".《https://blog.csdn.net/GitChat/article/details/79141842》.2018,
雀知安."scrapy抓取瀑布流模式图片".《https://www.jianshu.com/p/4ea0336f0ba5》.2017,
审查员 庄文龙

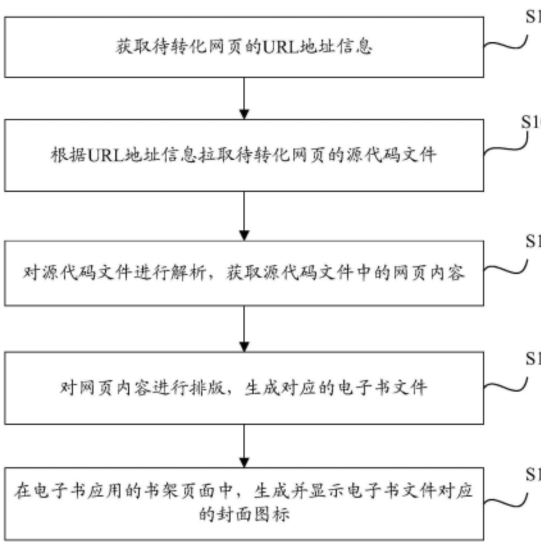
权利要求书6页 说明书10页 附图3页

(54)发明名称

网页转化电子书的方法、电子设备、存储介质

(57)摘要

本发明公开了一种网页转化电子书的方法、电子设备、存储介质,其方法包括:获取待转化网页的URL地址信息;根据URL地址信息拉取待转化网页的源代码文件;对源代码文件进行解析,获取源代码文件中的网页内容;对网页内容进行排版,生成对应的电子书文件;在电子书应用的书架页面中,生成并显示电子书文件对应的封面图标。通过本发明,根据待转化网页的URL地址信息,获取到网页内容,将网页内容生成对应的电子书文件,方便用户可以通过阅读电子书的方式来阅读网页内容,提高用户的阅读体验。进一步,生成的电子书文件显示在书架页面中,也方便用户查找,再次阅读等。



CN 109558123 B

1.一种网页转化电子书的方法,其包括:

基于在电子书应用的书架页面中的指定输入框获取用户输入内容,以获取待转化网页的URL地址信息;

根据所述URL地址信息拉取所述待转化网页的源代码文件;

对所述源代码文件进行解析,获取所述源代码文件中的网页内容;

对所述网页内容进行排版,生成对应的电子书文件,具体包括:

判断是否存在所述待转化网页关联的RSS页面;

若是,则拉取所述RSS页面的源代码文件;

对所述RSS页面的源代码文件进行解析,获取所述RSS页面的源代码文件中的网页内容;

根据所述待转化网页的网页内容与所述RSS页面的网页内容,生成对应的电子书文件;

在所述电子书应用的书架页面中,生成并显示所述电子书文件对应的封面图标;

获取电子书文件中用户添加的标注内容;

根据所述标注内容在所述电子书文件中的偏移位置,修改所述待转化网页的源代码文件;

基于修改后的源代码文件进行页面渲染,以得到在所述偏移位置对应网页位置处添加有所述标注内容的网页。

2.根据权利要求1所述的方法,其中,所述根据所述待转化网页的网页内容与所述RSS页面的网页内容,生成对应的电子书文件,包括:

将所述待转化网页的网页内容与所述RSS页面的网页内容进行整合排版,生成对应的电子书文件。

3.根据权利要求1所述的方法,其中,所述根据所述待转化网页的网页内容与所述RSS页面的网页内容,生成对应的电子书文件,包括:

对所述RSS页面的网页内容进行排版,生成所述RSS页面对应的电子书文件;

所述待转化网页对应的电子书文件进行排版,生成所述RSS页面对应的电子书文件;

相应地,所述在所述电子书应用的书架页面中,生成并显示所述电子书文件对应的封面图标,包括:

将所述RSS页面对应的电子书文件与所述待转化网页对应的电子书文件合并保存到关联账号的书架页面的同一文件夹内。

4.根据权利要求1所述的方法,其中,所述获取待转化网页的URL地址信息进一步包括:

根据用户在第三方浏览器中执行的网页保存操作,从系统粘贴板中读取待转化网页的URL地址信息;或者,获取用户在电子书应用的指定输入框内输入的待转化网页的URL地址信息。

5.根据权利要求1所述的方法,其中,所述对所述源代码文件进行解析,获取所述源代码文件中的网页内容进一步包括:

对所述源代码文件中的指定页面标签进行解析,提取指定页面标签以及所述指定页面标签标记的文字内容和/或图片。

6.根据权利要求1所述的方法,其中,所述对所述网页内容进行排版,生成对应的电子子书文件进一步包括:

依据默认格式信息或用户设置格式信息，对所述网页内容进行排版，生成对应的电子书文件；

或者，获取所述源代码文件对应的层叠样式表文件，依据所述层叠样式表文件对所述网页内容进行排版，生成对应的电子书文件。

7.根据权利要求5所述的方法，其中，所述生成对应的电子书文件还包括：根据指定页面标签标记的文字内容，生成所述电子书文件的目录。

8.根据权利要求1所述的方法，其中，所述在电子书应用的书架页面中，生成并显示所述电子书文件对应的封面图标进一步包括：

获取用户对待转化网页的命名设置信息，生成包含所述命名设置信息的封面图标；

或者，获取待转化网页的标题信息，生成包含所述标题信息的封面图标。

9.根据权利要求8所述的方法，其中，所述获取待转化网页的标题信息，生成包含所述标题信息的封面图标进一步包括：

获取待转化网页的标题信息，对所述标题信息进行分词处理得到分词结果；

从分词结果中提取关键词，生成包含所述关键词的封面图标。

10.根据权利要求1-9中任一项所述的方法，其中，所述在电子书应用的书架页面中，生成并显示所述电子书文件对应的封面图标进一步包括：

获取对所述源代码文件进行解析得到的至少一张图片，从至少一张图片中获取一张图片作为所述封面图标的配图；

或者，为用户提供封面编辑入口，获取用户通过封面编辑入口输入的图片或图片元素作为所述封面图标的配图；

或者，对待转化网页的URL地址信息进行解析得到所述待转化网页所属的站点，获取站点标识图标作为所述封面图标的配图。

11.根据权利要求10所述的方法，其中，所述根据所述URL地址信息拉取所述待转化网页的源代码文件进一步包括：

若所述待转化网页为批量网页的其中之一，则利用爬虫技术爬取批量网页的所有源代码文件；

若所述待转化网页为瀑布流页面的其中之一，则采用模拟操作的方式拉取所述瀑布流页面的所有源代码文件。

12.根据权利要求1所述的方法，其中，所述在电子书应用的书架页面中，生成并显示所述电子书文件对应的封面图标进一步包括：

根据所述网页URL地址信息，对所述电子书文件进行分类处理；

将属于同一分类的电子书文件保存到书架页面的同一文件夹内。

13.一种电子设备，包括：处理器、存储器、通信接口和通信总线，所述处理器、所述存储器和所述通信接口通过所述通信总线完成相互间的通信；

所述存储器用于存放至少一可执行指令，所述可执行指令使所述处理器执行以下操作：

基于在电子书应用的书架页面中的指定输入框获取用户输入内容，以获取待转化网页的URL地址信息；

根据所述URL地址信息拉取所述待转化网页的源代码文件；

对所述源代码文件进行解析，获取所述源代码文件中的网页内容；

对所述网页内容进行排版，生成对应的电子书文件，具体包括：

判断是否存在所述待转化网页关联的RSS页面；

若是，则拉取所述RSS页面的源代码文件；

对所述RSS页面的源代码文件进行解析，获取所述RSS页面的源代码文件中的网页内容；

根据所述待转化网页的网页内容与所述RSS页面的网页内容，生成对应的电子书文件；

在所述电子书应用的书架页面中，生成并显示所述电子书文件对应的封面图标；

获取电子书文件中用户添加的标注内容；

根据所述标注内容在所述电子书文件中的偏移位置，修改所述待转化网页的源代码文件；

基于修改后的源代码文件进行页面渲染，以得到在所述偏移位置对应网页位置处添加有所述标注内容的网页。

14.根据权利要求13所述的电子设备，所述可执行指令进一步使所述处理器执行以下操作：

将所述待转化网页的网页内容与所述RSS页面的网页内容进行整合排版，生成对应的电子书文件。

15.根据权利要求13所述的电子设备，所述可执行指令进一步使所述处理器执行以下操作：

对所述RSS页面的网页内容进行排版，生成所述RSS页面对应的电子书文件；

所述待转化网页对应的电子书文件进行排版，生成所述RSS页面对应的电子书文件；

相应地，所述在所述电子书应用的书架页面中，生成并显示所述电子书文件对应的封面图标，包括：

将所述RSS页面对应的电子书文件与所述待转化网页对应的电子书文件合并保存到关联账号的书架页面的同一文件夹内。

16.根据权利要求13所述的电子设备，所述可执行指令进一步使所述处理器执行以下操作：

根据用户在第三方浏览器中执行的网页保存操作，从系统粘贴板中读取待转化网页的URL地址信息；或者，获取用户在电子书应用的指定输入框内输入的待转化网页的URL地址信息。

17.根据权利要求13所述的电子设备，所述可执行指令进一步使所述处理器执行以下操作：

对所述源代码文件中的指定页面标签进行解析，提取指定页面标签以及所述指定页面标签标记的文字内容和/或图片。

18.根据权利要求13所述的电子设备，所述可执行指令进一步使所述处理器执行以下操作：

依据默认格式信息或用户设置格式信息，对所述网页内容进行排版，生成对应的电子书文件；

或者，获取所述源代码文件对应的层叠样式表文件，依据所述层叠样式表文件对所述

网页内容进行排版，生成对应的电子书文件。

19.根据权利要求17所述的电子设备，所述可执行指令进一步使所述处理器执行以下操作：

根据指定页面标签标记的文字内容，生成所述电子书文件的目录。

20.根据权利要求13所述的电子设备，所述可执行指令进一步使所述处理器执行以下操作：

获取用户对待转化网页的命名设置信息，生成包含所述命名设置信息的封面图标；

或者，获取待转化网页的标题信息，生成包含所述标题信息的封面图标。

21.根据权利要求20所述的电子设备，所述可执行指令进一步使所述处理器执行以下操作：

获取待转化网页的标题信息，对所述标题信息进行分词处理得到分词结果；

从分词结果中提取关键词，生成包含所述关键词的封面图标。

22.根据权利要求13-21中任一项所述的电子设备，所述可执行指令进一步使所述处理器执行以下操作：

获取对所述源代码文件进行解析得到的至少一张图片，从至少一张图片中获取一张图片作为所述封面图标的配图；

或者，为用户提供封面编辑入口，获取用户通过封面编辑入口输入的图片或图片元素作为所述封面图标的配图；

或者，对待转化网页的URL地址信息进行解析得到所述待转化网页所属的站点，获取站点标识图标作为所述封面图标的配图。

23.根据权利要求22所述的电子设备，所述可执行指令进一步使所述处理器执行以下操作：

若所述待转化网页为批量网页的其中之一，则利用爬虫技术爬取批量网页的所有源代码文件；

若所述待转化网页为瀑布流页面的其中之一，则采用模拟操作的方式拉取所述瀑布流页面的所有源代码文件。

24.根据权利要求13所述的电子设备，所述可执行指令进一步使所述处理器执行以下操作：

根据所述网页URL地址信息，对所述电子书文件进行分类处理；

将属于同一分类的电子书文件保存到书架页面的同一文件夹内。

25.一种计算机存储介质，所述存储介质中存储有至少一可执行指令，所述可执行指令使处理器执行以下操作：

基于在电子书应用的书架页面中的指定输入框获取用户输入内容，以获取待转化网页的URL地址信息；

根据所述URL地址信息拉取所述待转化网页的源代码文件；

对所述源代码文件进行解析，获取所述源代码文件中的网页内容；

对所述网页内容进行排版，生成对应的电子书文件，具体包括：

判断是否存在所述待转化网页关联的RSS页面；

若是，则拉取所述RSS页面的源代码文件；

对所述RSS页面的源代码文件进行解析，获取所述RSS页面的源代码文件中的网页内容；

根据所述待转化网页的网页内容与所述RSS页面的网页内容，生成对应的电子书文件；

在所述电子书应用的书架页面中，生成并显示所述电子书文件对应的封面图标；

获取电子书文件中用户添加的标注内容；

根据所述标注内容在所述电子书文件中的偏移位置，修改所述待转化网页的源代码文件；

基于修改后的源代码文件进行页面渲染，以得到在所述偏移位置对应网页位置处添加有所述标注内容的网页。

26.根据权利要求25所述的计算机存储介质，所述可执行指令进一步使处理器执行以下操作：

将所述待转化网页的网页内容与所述RSS页面的网页内容进行整合排版，生成对应的电子书文件。

27.根据权利要求25所述的计算机存储介质，所述可执行指令进一步使处理器执行以下操作：

对所述RSS页面的网页内容进行排版，生成所述RSS页面对应的电子书文件；

所述待转化网页对应的电子书文件进行排版，生成所述RSS页面对应的电子书文件；

相应地，所述在所述电子书应用的书架页面中，生成并显示所述电子书文件对应的封面图标，包括：

将所述RSS页面对应的电子书文件与所述待转化网页对应的电子书文件合并保存到关联账号的书架页面的同一文件夹内。

28.根据权利要求25所述的计算机存储介质，所述可执行指令进一步使处理器执行以下操作：

根据用户在第三方浏览器中执行的网页保存操作，从系统粘贴板中读取待转化网页的URL地址信息；或者，获取用户在电子书应用的指定输入框内输入的待转化网页的URL地址信息。

29.根据权利要求25所述的计算机存储介质，所述可执行指令进一步使处理器执行以下操作：

对所述源代码文件中的指定页面标签进行解析，提取指定页面标签以及所述指定页面标签标记的文字内容和/或图片。

30.根据权利要求25所述的计算机存储介质，所述可执行指令进一步使处理器执行以下操作：

依据默认格式信息或用户设置格式信息，对所述网页内容进行排版，生成对应的电子书文件；

或者，获取所述源代码文件对应的层叠样式表文件，依据所述层叠样式表文件对所述网页内容进行排版，生成对应的电子书文件。

31.根据权利要求29所述的计算机存储介质，所述可执行指令进一步使处理器执行以下操作：

根据指定页面标签标记的文字内容，生成所述电子书文件的目录。

32.根据权利要求25所述的计算机存储介质，所述可执行指令进一步使处理器执行以下操作：

获取用户对待转化网页的命名设置信息，生成包含所述命名设置信息的封面图标；

或者，获取待转化网页的标题信息，生成包含所述标题信息的封面图标。

33.根据权利要求32所述的计算机存储介质，所述可执行指令进一步使处理器执行以下操作：

获取待转化网页的标题信息，对所述标题信息进行分词处理得到分词结果；

从分词结果中提取关键词，生成包含所述关键词的封面图标。

34.根据权利要求25-33中任一项所述的计算机存储介质，所述可执行指令进一步使处理器执行以下操作：

获取对所述源代码文件进行解析得到的至少一张图片，从至少一张图片中获取一张图片作为所述封面图标的配图；

或者，为用户提供封面编辑入口，获取用户通过封面编辑入口输入的图片或图片元素作为所述封面图标的配图；

或者，对待转化网页的URL地址信息进行解析得到所述待转化网页所属的站点，获取站点标识图标作为所述封面图标的配图。

35.根据权利要求34所述的计算机存储介质，所述可执行指令进一步使处理器执行以下操作：

若所述待转化网页为批量网页的其中之一，则利用爬虫技术爬取批量网页的所有源代码文件；

若所述待转化网页为瀑布流页面的其中之一，则采用模拟操作的方式拉取所述瀑布流页面的所有源代码文件。

36.根据权利要求25所述的计算机存储介质，所述可执行指令进一步使处理器执行以下操作：

根据所述网页URL地址信息，对所述电子书文件进行分类处理；

将属于同一分类的电子书文件保存到书架页面的同一文件夹内。

网页转化电子书的方法、电子设备、存储介质

技术领域

[0001] 本发明涉及软件领域，具体涉及一种网页转化电子书的方法、电子设备、存储介质。

背景技术

[0002] 用户可以使用浏览器浏览网页，当网页内容较多，用户一次没有阅读完成所有网页内容时，可以先保存网页以供后续阅读的需求。现有技术中，浏览器提供了网页保存功能，如收藏网页的URL地址，当用户需要再次浏览该网页时，点击URL地址向网络请求该网页即可。或者，浏览器还提供了将网页内容转化成图片的保存方式，将网页内容生成图片格式的文件保存在用户本地，用户通过浏览图片方便下次阅读。但使用浏览器对网页进行阅读时，其与利用电子书应用软件阅读电子书相比，用户体验不佳。如不同网页的显示格式五花八门，不同网页翻页操作方式不同，网页中包含过多广告等，都影响用户的阅读体验。而电子书应用软件可以统一电子书的阅读风格、翻页操作方式，且电子书应用软件中还提供了书架，展示用户阅读的电子书，方便用户直观的查看。

[0003] 因此，为提高用户阅读体验，急需一种将网页转化电子书的方法。

发明内容

[0004] 鉴于上述问题，提出了本发明以便提供一种克服上述问题或者至少部分地解决上述问题的网页转化电子书的方法、电子设备、存储介质。

[0005] 根据本发明的一个方面，提供了一种网页转化电子书的方法，其包括：

[0006] 获取待转化网页的URL地址信息；

[0007] 根据URL地址信息拉取待转化网页的源代码文件；

[0008] 对源代码文件进行解析，获取源代码文件中的网页内容；

[0009] 对网页内容进行排版，生成对应的电子书文件；

[0010] 在电子书应用的书架页面中，生成并显示电子书文件对应的封面图标。

[0011] 根据本发明的另一方面，提供了一种电子设备，包括：处理器、存储器、通信接口和通信总线，处理器、存储器和通信接口通过通信总线完成相互间的通信；

[0012] 存储器用于存放至少一可执行指令，可执行指令使处理器执行以下操作：

[0013] 获取待转化网页的URL地址信息；

[0014] 根据URL地址信息拉取待转化网页的源代码文件；

[0015] 对源代码文件进行解析，获取源代码文件中的网页内容；

[0016] 对网页内容进行排版，生成对应的电子书文件；

[0017] 在电子书应用的书架页面中，生成并显示电子书文件对应的封面图标。

[0018] 根据本发明的又一方面，提供了一种计算机存储介质，存储介质中存储有至少一可执行指令，可执行指令使处理器执行以下操作：

[0019] 获取待转化网页的URL地址信息；

[0020]　根据URL地址信息拉取待转化网页的源代码文件；

[0021]　对源代码文件进行解析，获取源代码文件中的网页内容；

[0022]　对网页内容进行排版，生成对应的电子书文件；

[0023]　在电子书应用的书架页面中，生成并显示电子书文件对应的封面图标。

[0024]　根据本发明提供的网页转化电子书的方法、电子设备、存储介质，获取待转化网页的URL地址信息；根据URL地址信息拉取待转化网页的源代码文件；对源代码文件进行解析，获取源代码文件中的网页内容；对网页内容进行排版，生成对应的电子书文件；在电子书应用的书架页面中，生成并显示电子书文件对应的封面图标。通过本发明，根据待转化网页的URL地址信息，获取到网页内容，将网页内容生成对应的电子书文件，方便用户可以通过阅读电子书的方式来阅读网页内容，提高用户的阅读体验。进一步，生成的电子书文件显示在书架页面中，也方便用户查找，再次阅读等。

[0025]　上述说明仅是本发明技术方案的概述，为了能够更清楚了解本发明的技术手段，而可依照说明书的内容予以实施，并且为了让本发明的上述和其它目的、特征和优点能够更明显易懂，以下特举本发明的具体实施方式。

附图说明

[0026]　通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本发明的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

[0027]　图1示出了根据本发明实施例一的网页转化电子书的方法的流程图；

[0028]　图2示出了根据本发明实施例二的网页转化电子书的方法的流程图；

[0029]　图3示出了根据本发明实施例四的一种电子设备的结构示意图。

具体实施方式

[0030]　下面将参照附图更详细地描述本公开的示例性实施例。虽然附图中显示了本公开的示例性实施例，然而应当理解，可以以各种形式实现本公开而不应被这里阐述的实施例所限制。相反，提供这些实施例是为了能够更透彻地理解本公开，并且能够将本公开的范围完整的传达给本领域的技术人员。

[0031]　实施例一

[0032]　图1示出了根据本发明实施例一的网页转化电子书的方法的流程图，如图1所示，网页转化电子书的方法具体包括如下步骤：

[0033]　步骤S101，获取待转化网页的URL地址信息。

[0034]　在对网页进行转化时，需要先获取待转化网页的URL地址信息。在一个可选的实施方式中，用户在第三方浏览器中执行网页保存操作时，根据网页保存操作会在内存中系统粘贴板添加待转化网页的URL地址信息，当调起电子书应用时，电子书应用可以直接从系统粘贴板中读取到待转化网页的URL地址信息。或者，在另一个可选的实施方式中，提供给用户指定待转化网页的URL地址信息的操作入口。如在电子书应用的书架页面中为用户提供指定输入框，用户可以在指定输入框输入待转化网页的URL地址信息，获取用户在指定输入框内输入的内容，获取到待转化网页的URL地址信息。

[0035] 步骤S102,根据URL地址信息拉取待转化网页的源代码文件。

[0036] 根据URL地址信息,可以通过支持HTTP协议的客户端模拟浏览器对URL地址进行网页请求,根据接收到待转化网页的响应结果,从中拉取得到待转化网页的源代码文件。

[0037] 步骤S103,对源代码文件进行解析,获取源代码文件中的网页内容。

[0038] 源代码文件中包含了用于标记网页显示样式的页面标签,如P标签、a标签、h1-h6标签、span标签等页面标签,还包含了网页显示的文字内容、图片等网页内容。通过对网页的源代码文件进行解析,可以得到页面标签以及页面标签所标记的文字内容、图片等网页内容。

[0039] 在一个可选的实施方式中,由于源代码文件中包含了多种页面标签,有些页面标签如body标签其用于定义网页的主体,不是对具体的文字内容、图片等网页内容进行标记。因此,可以在对源代码文件进行解析时,为提高解析速度,对源代码文件中的指定页面标签进行解析,提取指定页面标签以及指定页面标签标记的文字内容和/或图片。指定页面标签包括如h1-h6标签等用于标记文字内容的页面标签,或者如a标签等用于标记图片、标记图片链接的页面标签,以获取生成电子书文件所需要的文字内容和图片。

[0040] 步骤S104,对网页内容进行排版,生成对应的电子书文件。

[0041] 由于获取到的网页内容没有进行排版,直接将其生成电子书文件会导致用户阅读体验不佳,因此,对获取到的网页内容进行排版,使得生成的电子书文件更符合用户的阅读习惯。具体的,可以依据电子书文件默认格式信息(如从电子书文件应用程序获取显示电子书文件时默认的字体、字号、背景色等格式信息)对网页内容进行排版,生成对应的电子书文件。或者,还可以根据用户设置的格式信息(如从电子书文件应用程序中获取用户设置的字体、字号、背景色等格式信息,或提供用户设置格式信息的入口,获取用户设置的字体、字号、背景色等格式信息),对网页内容进行排版,生成对应的电子书文件。或者,在获取源代码文件的同时,获取源代码文件对应的层叠样式表文件(css文件,其包含了对网页内容显示格式的设置),依据层叠样式表文件对网页内容进行排版,还原网页内容显示格式,生成对应的电子书文件。

[0042] 进一步,在生成电子书文件时,还可以根据指定页面标签标记的文字内容,生成电子书文件的目录。如指定页面标签为h2-h6标签,这些页面标签标记的文字内容一般为文字内容中每段落前的文字内容,因此,可以提取这些页面标签标记的文字内容生成电子书文件的目录。

[0043] 步骤S105,在电子书应用的书架页面中,生成并显示电子书文件对应的封面图标。

[0044] 在生成电子书文件后,可以在电子书应用的书架页面中,生成并显示电子书文件对应的封面图标,以便将电子书文件保存到用户的关联账号的书架中,方便用户随时翻阅。封面图标包含了标题信息、关键词、配图等信息。标题信息可以通过获取用户对待转化网页的命名设置信息,即提供命名入口,由用户来命名电子书文件的标题信息;或者,获取待转化网页的标题信息作为电子书文件的标题信息,待转化网页中的标题信息一般通过页面标签h1标签标记,将h1标签标记的文字内容作为标题信息。在获取到待转化网页的标题信息后,还可以对标题信息进行分词处理得到分词结果;从分词结果中可以提取得到关键词。关键词包括如作者、地址、时间、事件等信息关键词,可以根据标题信息、关键词等生成包含标题信息、关键词等封面图标。考虑到文字形式的封面图标辨识度有限,还可以设置封面图标

的配图。具体的，获取对源代码文件进行解析得到的至少一张图片，从至少一张图片中获取一张图片作为封面图标的配图，即利用待转化网页自己的图片作为封面图标的配图；或者，为用户提供封面编辑入口，获取用户通过封面编辑入口输入的图片或图片元素作为封面图标的配图，满足用户个性化需求；或者，对待转化网页的URL地址信息进行解析得到待转化网页所属的站点，获取站点标识图标作为封面图标的配图，可以使用户清楚的了解到电子书文件的来源。生成的封面图标可以包含上述的标题信息、关键词、配图等多个信息，也可以仅包含其中一个或两个信息，此次不做限定。在电子书应用中，当接收到用户对书架页面中的访问请求时，呈现包含电子书文件对应的封面图标的书架页面，方便用户点击阅读。

[0045] 进一步，在书架页面中，还可以根据网页URL地址信息，对生成的电子书文件进行分类处理。如根据网页URL地址信息获取URL地址中的域名字段，根据域名字段确定网页所属站点，将属于同一站点的网页转化生成的电子书文件归为同一分类，在保存时，将属于同一分类的电子书文件保存到书架页面的同一文件夹内，方便用户浏览。

[0046] 根据本发明提供的网页转化电子书的方法，获取待转化网页的URL地址信息；根据URL地址信息拉取待转化网页的源代码文件；对源代码文件进行解析，获取源代码文件中的网页内容；对网页内容进行排版，生成对应的电子书文件；在电子书应用的书架页面中，生成并显示电子书文件对应的封面图标。通过本发明，根据待转化网页的URL地址信息，获取到网页内容，将网页内容生成对应的电子书文件，方便用户可以通过阅读电子书的方式来阅读网页内容，提高用户的阅读体验。进一步，生成的电子书文件显示在书架页面中，也方便用户查找，再次阅读等。

[0047] 实施例二

[0048] 图2示出了根据本发明实施例二的网页转化电子书的方法的流程图，如图2所示，网页转化电子书的方法包括如下步骤：

[0049] 步骤S201，获取待转化网页的URL地址信息。

[0050] 用户确定待转化网页后，可以从内存系统粘贴板或根据用户输入获取待转化网页的URL地址信息。

[0051] 步骤S202，根据URL地址信息拉取待转化网页的源代码文件。

[0052] 考虑到实际网页显示时，存在分页显示的情况，如待转化网页可能为批量网页其中之一、待转化网页为瀑布流页面的其中之一等，仅对待转化网页进行转化，得到的不是完整的电子书文件；将每个分页单独制作成电子书文件，使用户在阅读时需要翻阅多本电子书文件才能阅读完成网页，不符合用户的阅读习惯，且占用大量书架的展示位资源，造成浪费。

[0053] 为转化得到完整的电子书文件，需要获取网页的所有源代码文件。在一个可选的实施方式中，在拉取待转化网页的源代码文件后，可以根据源代码文件中的翻页链接信息或页码信息等判断待转化网页是否为批量网页的其中之一，若待转化网页为批量网页的其中之一，根据当前待转化页面的URL地址信息，以及预设爬取策略，可以获取其他批量网页的URL地址信息，进而爬取到批量网页。如利用爬虫技术爬取批量网页，在爬取时，结合批量网页中URL地址信息中的页码信息，对应的将页码信息进行替换，得到其他批量网页的URL地址信息，进而可以爬取到其他批量网页的源代码文件。按照页码信息的顺序，整合批量网页，得到批量网页的所有源代码文件。在另一个可选的实施方式中，在拉取待转化网页的源

代码文件后，可以根据源代码文件是否完结等情况判断待转化网页是否为瀑布流页面的其中之一，若待转化网页为瀑布流页面的其中之一，则采用模拟操作的方式，如在当前待转化网页右侧拖动条的预设位置处模拟向下拉动操作，触发瀑布流页面下一页的刷新操作，模拟实现瀑布流页面下一页的网页请求，可以拉取到瀑布流页面下一页的源代码文件。在模拟操作拉取到瀑布流各个页面的源代码文件后，按照页码顺序，整合瀑布流网页，得到瀑布流网页的所有源代码文件。

[0054] 步骤S203，对源代码文件进行解析，获取源代码文件中的网页内容。

[0055] 解析源代码文件，得到网页中的文字内容、图片等网页内容。

[0056] 步骤S204，判断是否存在待转化网页关联的RSS页面。

[0057] RSS页面与待转化网页的网页内容会具有关联性或相似性，为更好的完善生成的电子书文件，也可以将RSS页面中的网页内容完善至电子书文件中。因此，可以根据待转化网页的源代码文件进行判断，判断是否存在待转化网页关联的RSS页面，如判断待转化网页的源代码文件中是否有RSS订阅功能，若有，则判断存在待转化网页关联的RSS页面，执行步骤S205；否则，判断不存在待转化网页关联的RSS页面，执行步骤S207。

[0058] 步骤S205，拉取RSS页面的源代码文件。

[0059] 从待转化网页的源代码文件中获取与RSS页面相关的数据，从中得到RSS页面的URL地址信息，根据RSS页面的URL地址信息，请求RSS页面，拉取得到RSS页面的源代码文件。

[0060] 步骤S206，对RSS页面的源代码文件进行解析，获取RSS页面的源代码文件中的网页内容。

[0061] 解析RSS页面的源代码文件，得到RSS页面的文字内容、图片等网页内容。

[0062] 步骤S207，对网页内容进行排版，生成对应的电子书文件。

[0063] 网页内容包括了待转化网页的源代码文件中的网页内容，对待转化网页的源代码文件中的网页内容进行排版，生成对应的电子书文件，具体参考实施例一中步骤S104的描述，此处不再赘述。

[0064] 若存在与待转化网页关联的RSS页面，则网页内容还包括RSS页面的源代码文件中的网页内容。在一个可选实施方式中，对于RSS页面的源代码文件中的网页内容，在生成电子书文件时，可以将待转化网页的网页内容与RSS页面的网页内容进行整合，对整合后的网页内容排版，生成对应的电子书文件。即生成的电子书文件包括了待转化网页的网页内容与RSS页面的网页内容。或者，在另一个可选实施方式中，在将待转化网页的源代码文件中的网页内容生成对应的电子书文件后，对于RSS页面的源代码文件中的网页内容进行排版，生成RSS页面对应的电子书文件。将RSS页面对应的电子书文件与待转化网页对应的电子书文件合并保存到关联账号的书架页面的同一文件夹内，作为关联的电子书文件进行展示。

[0065] 步骤S208，在电子书应用的书架页面中，生成并显示电子书文件对应的封面图标。

[0066] 在生成电子书文件后，在电子书应用的书架页面中，生成并显示电子书文件对应的封面图标，方便用户点击对应的封面图标翻阅电子书文件。

[0067] 步骤S209，获取电子书文件中用户添加的标注内容。

[0068] 电子书文件提供给用户添加标注内容的入口，用户可以在阅读电子书文件的过程中，针对电子书文件中的文字内容、图片等内容进行批注、添加书签等操作。这些标注内容同电子书文件的内容还可以一起转化为网页，满足用户将电子书文件转化为网页的需求。

具体的，需要先从电子书文件中获取用户添加的标注内容，以及获取到标注内容在电子书文件中的偏移位置，如标注内容1在XXX文字右侧，标注内容2在X图片下方等。

[0069] 步骤S210，根据标注内容在电子书文件中的偏移位置，修改待转化网页的源代码文件。

[0070] 根据标注内容在电子书文件中的偏移位置，对应的找到待转化网页的源代码文件中对应的网页内容的位置，修改待转化网页的源代码文件，在其中添加标注内容。此处，待转化网页的源代码文件可以为通过备份步骤S202所得到的源代码文件。

[0071] 步骤S211，基于修改后的源代码文件进行页面渲染，以得到在偏移位置对应网页位置处添加有标注内容的网页。

[0072] 基于修改后的源代码文件进行页面渲染，展示得到的网页除显示网页内容外，在与偏移位置对应网页位置处还显示了添加的标注内容，方便用户以网页形式查看。

[0073] 根据本发明提供的网页转化电子书的方法，可以将待转化网页以及与待转化网页关联的RSS页面一起转化为电子书文件，方便用户在阅读电子书文件时，更好的了解其内容。进一步，为满足用户将电子书文件重新转化为网页的需求，本发明还提供了电子书文件转化为网页的方法，将电子书文件以及用户在阅读过程中添加的标注内容一起转化为网页，方便用户阅读网页、以网页形式展示或转发网页进行分享等操作。

[0074] 实施例三

[0075] 本申请实施例三提供了一种非易失性计算机存储介质，计算机存储介质存储有至少一可执行指令，该计算机可执行指令可执行上述任意方法实施例中的网页转化电子书的方法。

[0076] 可执行指令具体可以用于使得处理器执行以下操作：

[0077] 获取待转化网页的URL地址信息；根据URL地址信息拉取待转化网页的源代码文件；对源代码文件进行解析，获取源代码文件中的网页内容；对网页内容进行排版，生成对应的电子书文件；在电子书应用的书架页面中，生成并显示电子书文件对应的封面图标。

[0078] 在一种可选的实施方式中，可执行指令进一步使处理器执行以下操作：根据用户在第三方浏览器中执行的网页保存操作，从系统粘贴板中读取待转化网页的URL地址信息；或者，获取用户在电子书应用的指定输入框内输入的待转化网页的URL地址信息。

[0079] 在一种可选的实施方式中，可执行指令进一步使处理器执行以下操作：对源代码文件中的指定页面标签进行解析，提取指定页面标签以及指定页面标签标记的文字内容和/或图片。

[0080] 在一种可选的实施方式中，可执行指令进一步使处理器执行以下操作：依据默认格式信息或用户设置格式信息，对网页内容进行排版，生成对应的电子书文件；或者，获取源代码文件对应的层叠样式表文件，依据层叠样式表文件对网页内容进行排版，生成对应的电子书文件。

[0081] 在一种可选的实施方式中，可执行指令进一步使处理器执行以下操作：根据指定页面标签标记的文字内容，生成电子书文件的目录。

[0082] 在一种可选的实施方式中，可执行指令进一步使处理器执行以下操作：获取用户对待转化网页的命名设置信息，生成包含命名设置信息的封面图标；或者，获取待转化网页的标题信息，生成包含标题信息的封面图标。

[0083] 在一种可选的实施方式中，可执行指令进一步使处理器执行以下操作：获取待转化网页的标题信息，对标题信息进行分词处理得到分词结果；从分词结果中提取关键词，生成包含关键词的封面图标。

[0084] 在一种可选的实施方式中，可执行指令进一步使处理器执行以下操作：获取对源代码文件进行解析得到的至少一张图片，从至少一张图片中获取一张图片作为封面图标的配图；或者，为用户提供封面编辑入口，获取用户通过封面编辑入口输入的图片或图片元素作为封面图标的配图；或者，对待转化网页的URL地址信息进行解析得到待转化网页所属的站点，获取站点标识图标作为封面图标的配图。

[0085] 在一种可选的实施方式中，可执行指令进一步使处理器执行以下操作：若待转化网页为批量网页的其中之一，则利用爬虫技术爬取批量网页的所有源代码文件；若待转化网页为瀑布流页面的其中之一，则采用模拟操作的方式拉取瀑布流页面的所有源代码文件。

[0086] 在一种可选的实施方式中，可执行指令进一步使处理器执行以下操作：根据网页URL地址信息，对电子书文件进行分类处理；将属于同一分类的电子书文件保存到书架页面的同一文件夹内。

[0087] 在一种可选的实施方式中，可执行指令进一步使处理器执行以下操作：判断是否存在待转化网页关联的RSS页面；若是，则拉取RSS页面的源代码文件；对RSS页面的源代码文件进行解析，获取RSS页面的源代码文件中的网页内容；对RSS页面的网页内容进行排版，生成RSS页面对应的电子书文件，将RSS页面对应的电子书文件与待转化网页对应的电子书文件合并保存到关联账号的书架页面的同一文件夹内。

[0088] 在一种可选的实施方式中，可执行指令进一步使处理器执行以下操作：判断是否存在待转化网页关联的RSS页面；若是，则拉取RSS页面的源代码文件；对RSS页面的源代码文件进行解析，获取RSS页面的源代码文件中的网页内容；将待转化网页的网页内容与RSS页面的网页内容进行整合排版，生成对应的电子书文件。

[0089] 在一种可选的实施方式中，可执行指令进一步使处理器执行以下操作：获取电子书文件中用户添加的标注内容；根据标注内容在电子书文件中的偏移位置，修改待转化网页的源代码文件；基于修改后的源代码文件进行页面渲染，以得到在偏移位置对应网页位置处添加有标注内容的网页。

[0090] 实施例四

[0091] 图3示出了根据本发明实施例四的一种电子设备的结构示意图，本发明具体实施例并不对电子设备的具体实现做限定。

[0092] 如图3所示，该电子设备可以包括：处理器(processor)302、通信接口(Communications Interface)304、存储器(memory)306、以及通信总线308。

[0093] 其中：

[0094] 处理器302、通信接口304、以及存储器306通过通信总线308完成相互间的通信。

[0095] 通信接口304，用于与其它设备比如客户端或其它服务器等的网元通信。

[0096] 处理器302，用于执行程序310，具体可以执行上述网页转化电子书的方法实施例中的相关步骤。

[0097] 具体地，程序310可以包括程序代码，该程序代码包括计算机操作指令。

[0098] 处理器302可能是中央处理器CPU,或者是特定集成电路ASIC(Application Specific Integrated Circuit),或者是被配置成实施本发明实施例的一个或多个集成电路。服务器包括的一个或多个处理器,可以是同一类型的处理器,如一个或多个CPU;也可以是不同类型的处理器,如一个或多个CPU以及一个或多个ASIC。

[0099] 存储器306,用于存放程序310。存储器306可能包含高速RAM存储器,也可能还包括非易失性存储器(non-volatile memory),例如至少一个磁盘存储器。

[0100] 程序310具体可以用于使得处理器302执行以下操作:

[0101] 在一种可选的实施方式中,程序310用于使得处理器302获取待转化网页的URL地址信息;根据URL地址信息拉取待转化网页的源代码文件;对源代码文件进行解析,获取源代码文件中的网页内容;对网页内容进行排版,生成对应的电子书文件;在电子书应用的书架页面中,生成并显示电子书文件对应的封面图标。

[0102] 在一种可选的实施方式中,程序310用于使得处理器302根据用户在第三方浏览器中执行的网页保存操作,从系统粘贴板中读取待转化网页的URL地址信息;或者,获取用户电子书应用的在指定输入框内输入的待转化网页的URL地址信息。

[0103] 在一种可选的实施方式中,程序310用于使得处理器302对源代码文件中的指定页面标签进行解析,提取指定页面标签以及指定页面标签标记的文字内容和/或图片。

[0104] 在一种可选的实施方式中,程序310用于使得处理器302依据默认格式信息或用户设置格式信息,对网页内容进行排版,生成对应的电子书文件;或者,获取源代码文件对应的层叠样式表文件,依据层叠样式表文件对网页内容进行排版,生成对应的电子书文件。

[0105] 在一种可选的实施方式中,程序310用于使得处理器302根据指定页面标签标记的文字内容,生成电子书文件的目录。

[0106] 在一种可选的实施方式中,程序310用于使得处理器302获取用户对待转化网页的命名设置信息,生成包含命名设置信息的封面图标;或者,获取待转化网页的标题信息,生成包含标题信息的封面图标。

[0107] 在一种可选的实施方式中,程序310用于使得处理器302获取待转化网页的标题信息,对标题信息进行分词处理得到分词结果;从分词结果中提取关键词,生成包含关键词的封面图标。

[0108] 在一种可选的实施方式中,程序310用于使得处理器302获取对源代码文件进行解析得到的至少一张图片,从至少一张图片中获取一张图片作为封面图标的配图;或者,为用户提供封面编辑入口,获取用户通过封面编辑入口输入的图片或图片元素作为封面图标的配图;或者,对待转化网页的URL地址信息进行解析得到待转化网页所属的站点,获取站点标识图标作为封面图标的配图。

[0109] 在一种可选的实施方式中,若待转化网页为批量网页的其中之一,程序310用于使得处理器302利用爬虫技术爬取批量网页的所有源代码文件;若待转化网页为瀑布流页面的其中之一,程序310用于使得处理器302采用模拟操作的方式拉取瀑布流页面的所有源代码文件。

[0110] 在一种可选的实施方式中,程序310用于使得处理器302根据网页URL地址信息,对电子书文件进行分类处理;将属于同一分类的电子书文件保存到书架页面的同一文件夹内。

[0111] 在一种可选的实施方式中，程序310用于使得处理器302判断是否存在待转化网页关联的RSS页面；若是，则拉取RSS页面的源代码文件；对RSS页面的源代码文件进行解析，获取RSS页面的源代码文件中的网页内容；对RSS页面的网页内容进行排版，生成RSS页面对应的电子书文件，将RSS页面对应的电子书文件与待转化网页对应的电子书文件合并保存到关联账号的书架页面的同一文件夹内。

[0112] 在一种可选的实施方式中，程序310用于使得处理器302判断是否存在待转化网页关联的RSS页面；若是，则拉取RSS页面的源代码文件；对RSS页面的源代码文件进行解析，获取RSS页面的源代码文件中的网页内容；将待转化网页的网页内容与RSS页面的网页内容进行整合排版，生成对应的电子书文件。

[0113] 在一种可选的实施方式中，程序310用于使得处理器302获取电子书文件中用户添加的标注内容；根据标注内容在电子书文件中的偏移位置，修改待转化网页的源代码文件；基于修改后的源代码文件进行页面渲染，以得到在偏移位置对应网页位置处添加有标注内容的网页。

[0114] 程序310中各步骤的具体实现可以参见上述网页转化电子书的实施例中的相应步骤中对应的描述，在此不赘述。所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的设备和模块的具体工作过程，可以参考前述方法实施例中的对应过程描述，在此不再赘述。

[0115] 通过本实施例提供的方案，获取待转化网页的URL地址信息；根据URL地址信息拉取待转化网页的源代码文件；对源代码文件进行解析，获取源代码文件中的网页内容；对网页内容进行排版，生成对应的电子书文件；在电子书应用的书架页面中，生成并显示电子书文件对应的封面图标。通过本发明，根据待转化网页的URL地址信息，获取到网页内容，将网页内容生成对应的电子书文件，方便用户可以通过阅读电子书的方式来阅读网页内容，提高用户的阅读体验。进一步，生成的电子书文件显示在书架页面中，也方便用户查找，再次阅读等。

[0116] 在此处所提供的说明书中，说明了大量具体细节。然而，能够理解，本发明的实施例可以在没有这些具体细节的情况下实践。在一些实例中，并未详细示出公知的方法、结构和技术，以便不模糊对本说明书的理解。

[0117] 类似地，应当理解，为了精简本公开并帮助理解各个发明方面中的一个或多个，在上面对本发明的示例性实施例的描述中，本发明的各个特征有时被一起分组到单个实施例、图、或者对其的描述中。然而，并不应将该公开的方法解释成反映如下意图：即所要求保护的本发明要求比在每个权利要求中所明确记载的特征更多的特征。更确切地说，如权利要求书所反映的那样，发明方面在于少于前面公开的单个实施例的所有特征。因此，遵循具体实施方式的权利要求书由此明确地并入该具体实施方式，其中每个权利要求本身都作为本发明的单独实施例。

[0118] 本领域那些技术人员可以理解，可以对实施例中的设备中的模块进行自适应性地改变并且把它们设置在与该实施例不同的一个或多个设备中。可以把实施例中的模块或单元或组件组合成一个模块或单元或组件，以及此外可以把它们分成多个子模块或子单元或子组件。除了这样的特征和/或过程或者单元中的至少一些是相互排斥之外，可以采用任何组合对本说明书(包括伴随的权利要求、摘要和附图)中公开的所有特征以及如此公开的任

何方法或者设备的所有过程或单元进行组合。除非另外明确陈述，本说明书(包括伴随的权利要求、摘要和附图)中公开的每个特征可以由提供相同、等同或相似目的的替代特征来代替。

[0119] 此外，本领域的技术人员能够理解，尽管在此所述的一些实施例包括其它实施例中所包括的某些特征而不是其它特征，但是不同实施例的特征的组合意味着处于本发明的范围之内并且形成不同的实施例。例如，在权利要求书中，所要求保护的实施例的任意之一都可以以任意的组合方式来使用。

[0120] 应该注意的是上述实施例对本发明进行说明而不是对本发明进行限制，并且本领域技术人员在不脱离所附权利要求的范围的情况下可设计出替换实施例。在权利要求中，不应将位于括号之间的任何参考符号构造成对权利要求的限制。单词“包含”不排除存在未列在权利要求中的元件或步骤。位于元件之前的单词“一”或“一个”不排除存在多个这样的元件。本发明可以借助于包括有若干不同元件的硬件以及借助于适当编程的计算机来实现。单词第一、第二、以及第三等的使用不表示任何顺序。可将这些单词解释为名称。

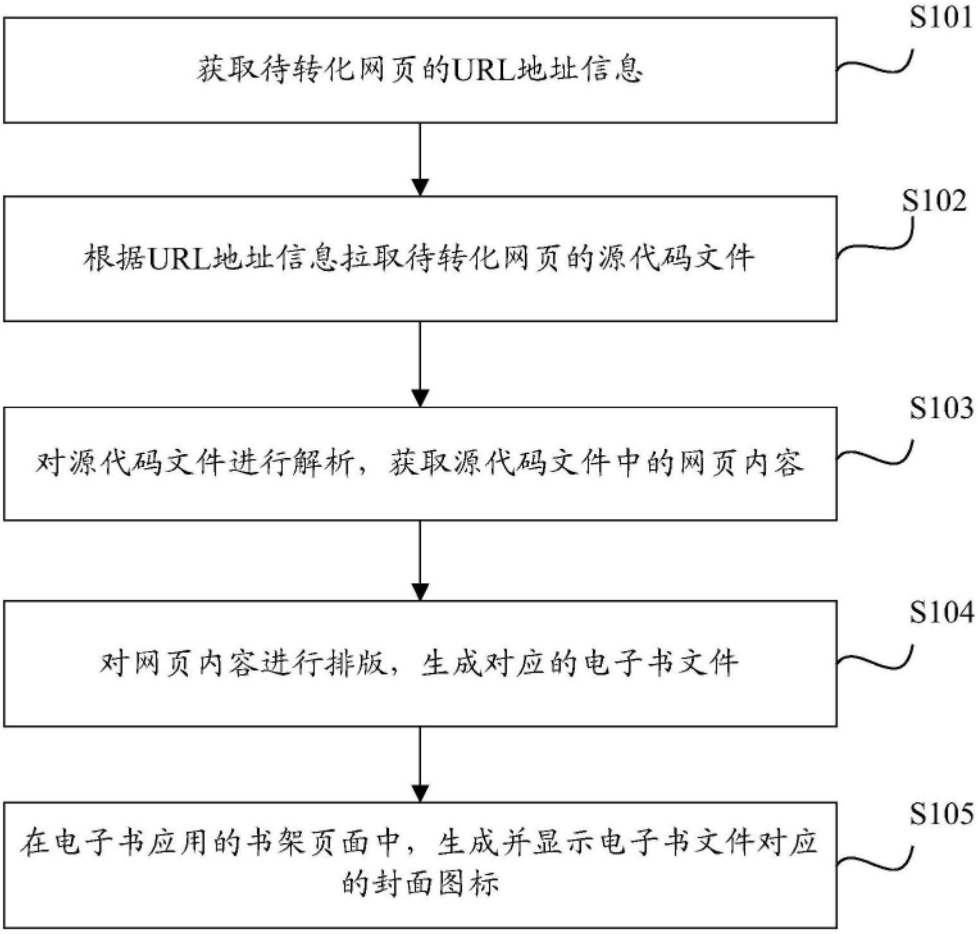

获取待转化网页的URL地址信息
S101
根据URL地址信息拉取待转化网页的源代码文件
S102
对源代码文件进行解析，获取源代码文件中的网页内容
S103
对网页内容进行排版，生成对应的电子书文件
S104
在电子书应用的书架页面中，生成并显示电子书文件对应的封面图标
S105

图1

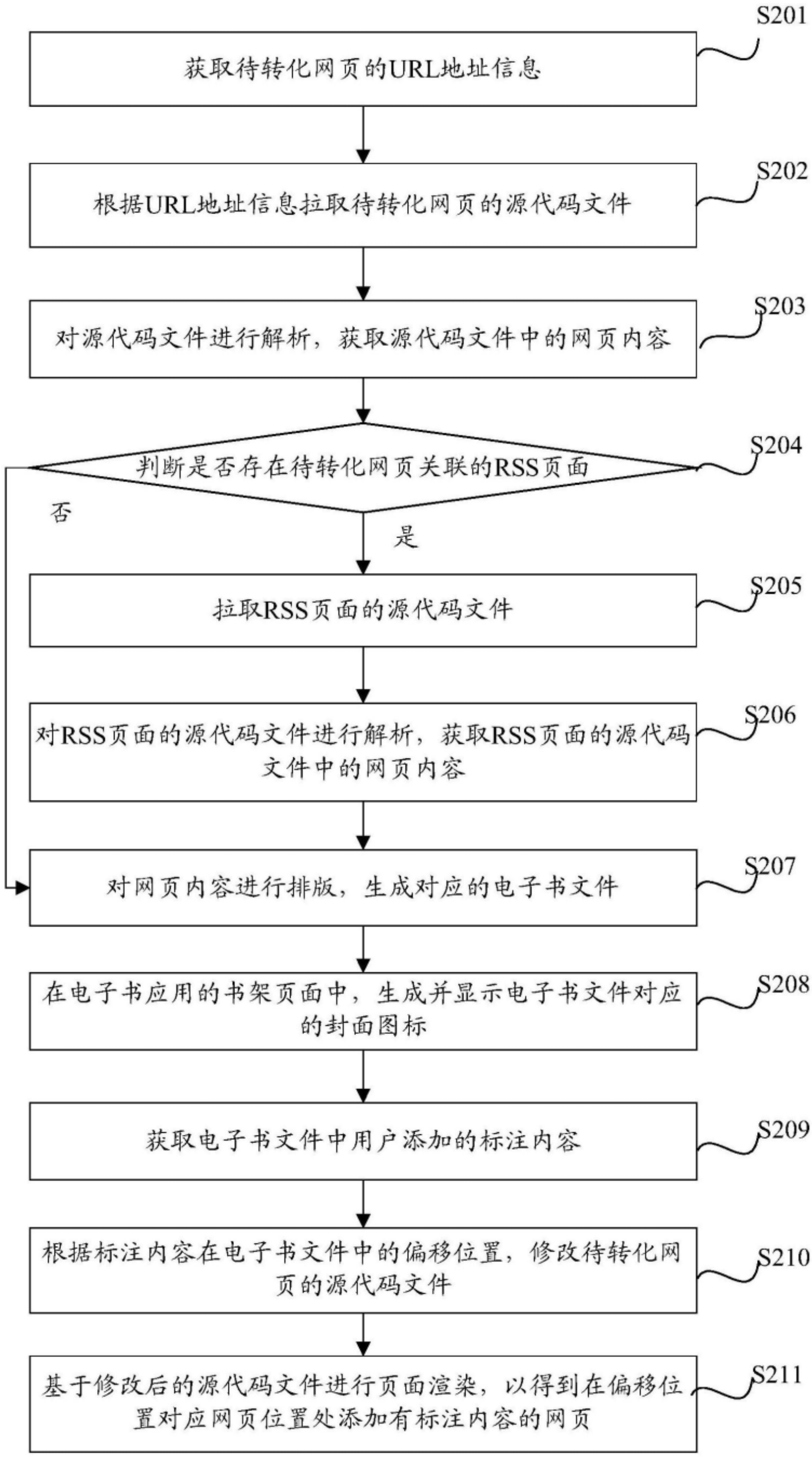

S201
获取待转化网页的URL地址信息
S202
根据URL地址信息拉取待转化网页的源代码文件
S203
对源代码文件进行解析，获取源代码文件中的网页内容
S204
判断是否存在待转化网页关联的RSS页面
否
是
S205
拉取RSS页面的源代码文件
S206
对RSS页面的源代码文件进行解析，获取RSS页面的源代码文件中的网页内容
S207
对网页内容进行排版，生成对应的电子书文件
S208
在电子书应用的书架页面中，生成并显示电子书文件对应的封面图标
S209
获取电子书文件中用户添加的标注内容
S210
根据标注内容在电子书文件中的偏移位置，修改待转化网页的源代码文件
S211
基于修改后的源代码文件进行页面渲染，以得到在偏移位置对应网页位置处添加有标注内容的网页

图2

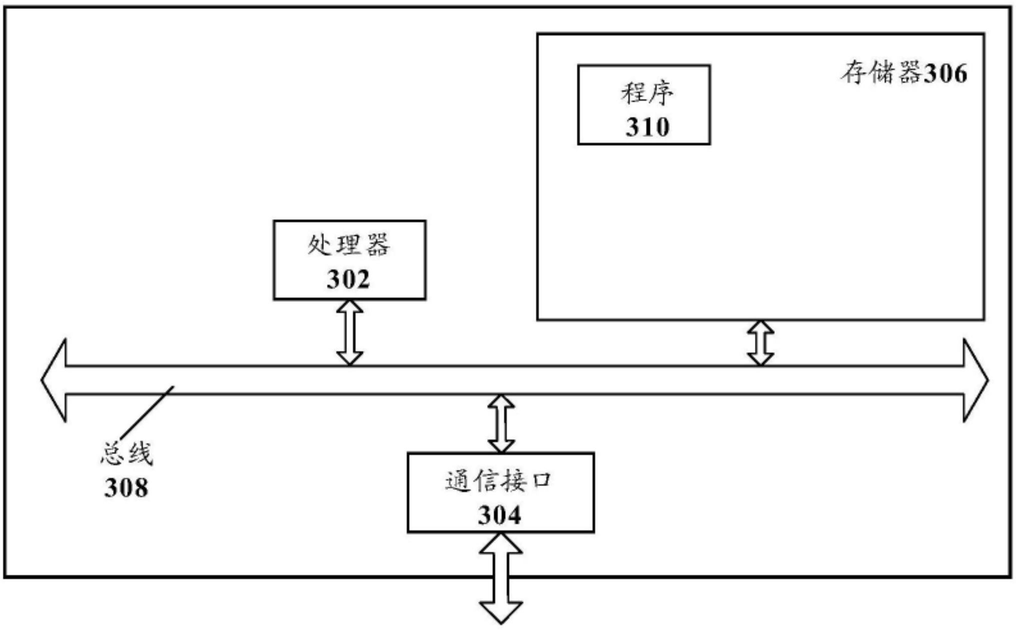


图3