



(12)发明专利申请

(10)申请公布号 CN 107391650 A

(43)申请公布日 2017.11.24

(21)申请号 201710577110.8

(22)申请日 2017.07.14

(71)申请人 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

申请人 中科鼎富(北京)科技发展有限公司

(72)发明人 房平会 李德彦

(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 遂长明 许伟群

(51)Int.Cl.

G06F 17/30(2006.01)

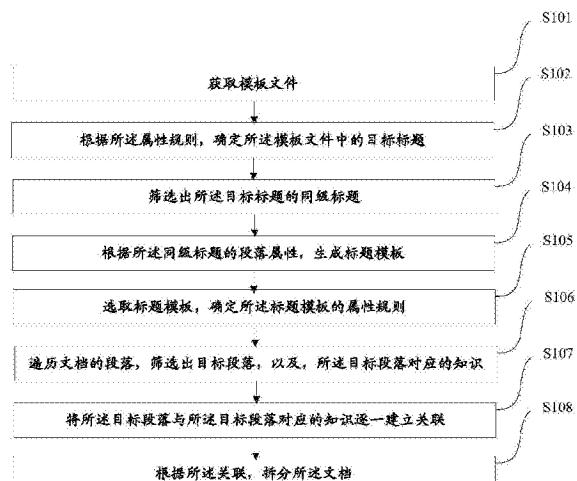
权利要求书2页 说明书12页 附图5页

(54)发明名称

一种文档的结构化拆分方法、装置及系统

(57)摘要

本申请实施例公开了一种文档的结构化拆分方法、装置及系统，所述方法在整篇文档中筛选出目标段落，所述目标段落为段落属性符合所述属性规则的段落；将所述目标段落与所述目标段落对应的知识逐一建立关联，此时，目标段落与目标段落对应的知识形成一个知识条目，进而将整篇文档拆分成多个知识条目。搜索系统在搜索有用知识的过程中，应用平台服务器仅需对知识条目进行分析，筛选出有用知识，缩小了搜索系统的搜索范围，进而缩短了搜索的时间，提高了系统带宽、数据库等资源的利用率。



1. 一种文档的结构化拆分方法,其特征在于,包括:

选取标题模板,确定所述标题模板的属性规则;

根据所述属性规则,遍历文档的段落,筛选出目标段落,以及,所述目标段落对应的知识,所述目标段落为段落属性符合所述属性规则的段落;

将所述目标段落与所述目标段落对应的知识逐一建立关联;

根据所述关联,拆分所述文档。

2. 根据权利要求1所述的方法,其特征在于,所述选取标题模板,确定所述标题模板的属性规则的步骤之前所述方法还包括:

获取模板文件;

确定所述模板文件中的目标标题;

筛选出所述目标标题的同级标题;

根据所述同级标题的段落属性,生成标题模板。

3. 根据权利要求2所述的方法,其特征在于,所述根据同级标题的段落属性,生成标题模板的步骤包括:

显示所述同级标题,以及,所述同级标题对应的知识;

判断所述同级标题对应的知识是否符合预置划分规则;

如果所述同级标题对应的知识符合预置划分规则,根据所述同级标题的段落属性,生成标题模板;

如果所述同级标题对应的知识不符合预置划分规则,调取所述同级标题的子标题;

根据所述同级标题的段落属性,以及,所述子级标题的段落属性,生成属性模板根据所述同级标题的段落属性,以及,所述子级标题的段落属性,生成属性模板,所述属性模板包括:根据同级标题的段落属性生成的同级标题模板,以及,根据所述子标题的段落属性生成的子级标题模板。

4. 根据权利要求1所述的方法,其特征在于,遍历所述文档的段落,筛选出目标段落的步骤包括:

遍历所述文档的段落,筛选出目标段落;

如果出现多于一个的目标段落,则增加一个正则表达式;

判断所述目标段落的内容是符合正则表达式;

如果所述目标段落的内容符合正则表达式,则保留所述目标段落;

如果所述目标段落的内容不符合正则表达式,则删除所述目标段落。

5. 根据权利要求1所述的方法,其特征在于,所述将目标段落与所述目标段落对应的知识逐一建立关联的步骤包括:

显示所述目标段落,以及,所述目标段落对应的知识;

判断所述目标段落对应的知识是否为有用知识;

如果所述目标段落对应的知识为有用知识,建立所述目标段落与所述目标段落对应的知识之间的关联;

如果所述目标段落对应的知识不是有用知识,删除所述目标段落,以及,所述目标段落对应的知识。

6. 根据权利要求1所述的方法,其特征在于,遍历所述文档的段落,筛选出目标段落的

步骤包括：

遍历所述文档的段落，确定所述段落的所属的属性级别；

如果所述段落的属性符合所述属性规则，所述段落为目标段落；

如果所述段落的属性符合所述标题模板的上一级标题的属性规则，分析所述段落对应的知识，得到分析结果；

根据所述分析结果生成一目标段落。

7. 根据权利要求1所述的方法，其特征在于，所述将目标段落与所述目标段落对应的知识逐一建立关联的步骤包括：

显示所述目标段落，以及，所述目标段落对应的知识；

如果所述目标段落对应的知识中包括图片，将所述图片以链接的形式存储在目标段落对应的知识中；

如果所述目标段落对应的知识中包括表格，将所述表格转化为成可以展示的格式存储在目标段落对应的知识中；

将所述目标段落与所述目标段落对应的知识逐一建立关联。

8. 根据权利要求1-7任意一项述的方法，其特征在于，所述属性规则包括：

字体的大小，字形、首行缩进距离，段前距离，段后距离中的一种或几种组合。

9. 一种文档的结构化生拆分装置，其特征在于，所述装置包括：

选取单元，用于选取标题模板，确定所述标题模板的属性规则；

遍历单元，用于根据所述属性规则，遍历文档的段落，筛选出目标段落，以及，所述目标段落对应的知识；

建立单元，用于将所述目标段落与所述目标段落对应的知识逐一建立关联；

拆分单元，用于根据所述关联，拆分所述文档。

10. 一种文档的结构化生拆分系统，其特征在于，所述系统包括：

应用平台服务器，以及，与其连接的数据存储服务器，所述数据存储服务器设置在所述应用平台服务器内部或独立设置，以及，与应用平台服务器通过互联网或移动互联网连接的终端；

所述应用平台服务器，用于选取标题模板，确定所述标题模板的属性规则；

用于根据所述属性规则，遍历文档的段落，筛选出目标段落，以及，所述目标段落对应的知识，所述目标段落为段落属性符合所述属性规则的段落；

用于将所述目标段落与所述目标段落对应的知识逐一建立关联；

用于根据所述关联，拆分所述文档；

所述终端用于向所述应用平台服务器发送文档，以及，用于接收才分后的文档；

所述数据存储服务器，用于相关数据的存储。

一种文档的结构化拆分方法,装置及系统

技术领域

[0001] 本申请实施例涉及文件搜索系统技术领域,特别涉及一种文档的结构化拆分方法,装置及系统。

背景技术

[0002] 随着互联网技术的发展,基于互联网的搜索系统也越来越多。典型的基于互联网的搜索系统如图1所示,这个系统一般有一个应用平台服务器1,以及与其连接的数据存储服务器2,该数据存储服务器2设置在平台服务器1内部或独立设置,以及,与应用平台服务器1通过互联网3或移动互联网3连接的终端4,通常,应用平台服务器1为终端4提供应用服务。

[0003] 信息搜索系统就是一个示例性的基于互联网的搜索系统。通常,用户在终端中输入想要了解的信息的“搜索词”,应用平台服务器1基于该“搜索词”,遍历所述存储服务器2中的文档,搜索出与所述“搜索词”相关联的有用知识,并将相关信息,发送至终端4进行显示。

[0004] 但是,申请人发现现有技术提供的搜索系统在提供搜索有用知识的过程中存在搜索操作效率低以及搜索操作过多占用系统资源的问题。例如,用户想搜索“美食”相关的有用知识,此时,应用平台服务器1在遍历所述整篇word文档,搜索出与“美食”相关的有用知识。通常,整篇word文档是一个比较大的知识,应用平台服务器1在遍历所述整篇word文档的过程中,需要对整篇word文档的内容进行分析,然后,筛选出与“美食”相关的有用知识。在此过程中,应用平台服务器1,长时间的处于等待状态,降低了系统带宽、数据库等资源的利用率。

申请内容

[0006] 本申请的发明目的在于提一种文档的结构化拆分方法,装置及系统,以解决现有技术示出的搜索系统搜索效率低的技术问题。

[0007] 本申请实施例第一方面提供了一种文档的结构化拆分方法,包括:

[0008] 选取标题模板,确定所述标题模板的属性规则;

[0009] 根据所述属性规则,遍历文档的段落,筛选出目标段落,以及,所述目标段落对应的知识,所述目标段落为段落属性符合所述属性规则的段落;

[0010] 将所述目标段落与所述目标段落对应的知识逐一建立关联;

[0011] 根据所述关联,拆分所述文档。

[0012] 由以上技术方案可知,本申请实施例示出一种文档的结构化拆分方法,所述方法在整篇文档中筛选出目标段落,所述目标段落为段落属性符合所述属性规则的段落;将所述目标段落与所述目标段落对应的知识逐一建立关联,此时,目标段落与目标段落对应的知识形成一个知识条目,进而将整篇文档拆分成多个知识条目。搜索系统在搜索有用知识的过程中,应用平台服务器仅需对知识条目进行分析,筛选出有用知识,缩小了搜索系统的搜索范围,进而缩短了搜索的时间,提高了系统带宽、数据库等资源的利用率。

- [0013] 本申请实施例第二方面示出一种文档的结构化生拆分装置，所述装置包括：
- [0014] 选取单元，用于选取标题模板，确定所述标题模板的属性规则；
- [0015] 遍历单元，用于根据所述属性规则，遍历文档的段落，筛选出目标段落，以及，所述目标段落对应的知识；
- [0016] 建立单元，用于将所述目标段落与所述目标段落对应的知识逐一建立关联；
- [0017] 拆分单元，用于根据所述关联，拆分所述文档。
- [0018] 本申请实施例示出一种文档的结构化生拆分装置，所述装置在整篇文档中筛选出目标段落，所述目标段落为段落属性符合所述属性规则的段落；将所述目标段落与所述目标段落对应的知识逐一建立关联，此时，目标段落与目标段落对应的知识形成一个知识条目，进而将整篇文档拆分成多个知识条目。搜索系统在搜索有用知识的过程中，应用平台服务器仅需对知识条目进行分析，筛选出有用知识，缩小了搜索系统的搜索范围，进而缩短了搜索的时间，提高了系统带宽、数据库等资源的利用率。
- [0019] 本申请实施例第三方面示出一种文档的结构化生拆分系统，所述系统包括：
- [0020] 应用平台服务器，以及，与其连接的数据存储服务器，所述数据存储服务器设置在所述应用平台服务器内部或独立设置，以及，与应用平台服务器通过互联网或移动互联网连接的终端，
- [0021] 所述应用平台服务器，用于选取标题模板，确定所述标题模板的属性规则；
- [0022] 用于根据所述属性规则，遍历文档的段落，筛选出目标段落，以及，所述目标段落对应的知识，所述目标段落为段落属性符合所述属性规则的段落；
- [0023] 用于将所述目标段落与所述目标段落对应的知识逐一建立关联；
- [0024] 用于根据所述关联，拆分所述文档；
- [0025] 所述终端用于向所述应用平台服务器发送文档，以及，用于接收才分后的文档；
- [0026] 所述数据存储服务器，用于相关数据的存储。
- [0027] 本申请实施例示出一种文档的结构化生拆分系统，所述系统在整篇文档中筛选出目标段落，所述目标段落为段落属性符合所述属性规则的段落；将所述目标段落与所述目标段落对应的知识逐一建立关联，此时，目标段落与目标段落对应的知识形成一个知识条目，进而将整篇文档拆分成多个知识条目。搜索系统在搜索有用知识的过程中，应用平台服务器仅需对知识条目进行分析，筛选出有用知识，缩小了搜索系统的搜索范围，进而缩短了搜索的时间，提高了系统带宽、数据库等资源的利用率。

附图说明

[0028] 为了更清楚地说明本申请实施例或现有技术中的技术方案，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，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

- [0029] 图1为基于互联网的搜索系统的场景图；
- [0030] 图2为根据申请一优选的实施例示出的一种文档的结构化拆分方法的流程图；
- [0031] 图3为根据申请一优选的实施例示出的步骤104的详细的流程图；
- [0032] 图4为根据申请又一优选的实施例示出的步骤104的详细的流程图；

- [0033] 图5为根据申请一优选的实施例示出的步骤106的详细的流程图；
- [0034] 图6为根据申请一优选的实施例示出的步骤107的详细的流程图；
- [0035] 图7为根据申请又一优选的实施例示出的步骤107的详细的流程图；
- [0036] 图8为根据申请一优选的实施例示出的一种文档的结构化生拆分装置的结构框图；
- [0037] 图9-1为根据申请一优选的实施例示出的一种文档的结构化生拆分系统的结构框图；
- [0038] 图9-2为根据申请又一优选的实施例示出的一种文档的结构化生拆分系统的结构框图。

具体实施方式

[0039] 下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整的描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

[0040] 请参阅图2，本申请实施例示出一种文档的结构化拆分方法，所述方法包括以下的步骤：

[0041] S105选取标题模板，确定所述标题模板的属性规则；
[0042] 数据存储服务器中存储有多种标题模板，每种标题模板对应至少一种属性规则；
[0043] 例如：标题模板1对应的一级标题，标题模板1的属性规则为：段落标号-X；字体-黑体；字号-小三；首行缩进2字符；段前间距0.5行；加粗。

[0044] 标题模板2对应的二级标题，标题模板2的属性规则为：段落标号-X.X；字体-黑体；字号-四号；首行缩进2字符；段前间距0.5行；加粗。

[0045] 标题模板3对应的三级标题，标题模板3的属性规则为：段落标号-X.X.X；字体-黑体；字号-小四；首行缩进2字符；加粗。

[0046] 应用平台服务器根据实际需求选取标题模板，然后，确定所述标题模板的属性规则；

[0047] S106根据所述属性规则，遍历文档的段落，筛选出目标段落，以及，所述目标段落对应的知识；

[0048] 本申请实施例中，使用文字识别技术，分别对文档的每个段落的属性进行解析。

[0049] 通过对段落的属性的解析，能够得到文档中每个段落的段落属性，搜索出符合属性规则的段落。本申请实施例中的属性规则包括：字体名称、加粗字体、倾斜字体和划线字体等。

[0050] 对段落的属性进行解析的过程为：把每个段落的字体大小、缩进距离等段落属性提取出来，将所述段落属性与属性规则进行对比，如果符合属性规则要求就将该段落标注为目标段落，所述目标段落下边的文字即为所述目标段落对应的知识，直至下一个目标段落的出现。

[0051] S107将所述目标段落与所述目标段落对应的知识逐一建立关联；

[0052] 每个目标段落与所述目标段落对应的知识组成一个知识条目。

[0053] S108根据所述关联，拆分所述文档。

[0054] 本申请实施例示出一种文档的结构化拆分方法，所述方法在整篇文档中筛选出目标段落，所述目标段落为段落属性符合所述属性规则的段落；将所述目标段落与所述目标段落对应的知识逐一建立关联，此时，目标段落与目标段落对应的知识形成一个知识条目，将整篇文档拆分成多个知识条目。搜索系统在搜索有用知识的过程中，应用平台服务器，仅需通过对知识条目进行分析，筛选出有用知识，缩小了搜索系统的搜索范围，进而缩短了搜索的时间，提高了系统带宽、数据库等资源的利用率。

[0055] 实施例1：

[0056] 文档1：

[0057] 1液芯光波导；

[0058] 液芯光波导对应的知识。

[0059] 1.1液芯光波导的发展历程；

[0060] 液芯光波导的发展历程对应的知识。

[0061] 1.2液芯光波导的传光原理；

[0062] 液芯光波导的传光原理对应的知识。

[0063] 1.3液芯光波导的特点；

[0064] 液芯光波导的特点对应的知识。

[0065] 1.4液芯光波导在分析领域中的应用；

[0066] 液芯光波导在分析领域中的应用对应的知识。

[0067] 1.4.1液芯光波导在萃取方向上的应用；

[0068] 液芯光波导在萃取方向上的应用对应的知识。

[0069] 1.4.2液芯光波导在传感方向上的应用；

[0070] 液芯光波导在传感方向上的应用对应的知识。

[0071] 2离子液体；

[0072] 离子液体对应的知识。

[0073] 2.1离子液体的发展历程；

[0074] 离子液体的发展历程对应的知识。

[0075] 2.2离子液体性质及组成；

[0076] 离子液体性质及组成对应的知识。

[0077] 2.3离子液体在萃取分离中的应用；

[0078] 离子液体在萃取分离中的应用对应的知识。

[0079] 用户需要了解离子液体的性质及其组成的相关内容；

[0080] 应用平台服务器根据用户上传的文档1，以及，选择标题模板，确定用户选取的标题模板对应的属性规则为：段落标号-X.X；字体-黑体；字号-四号；首行缩进2字符；段前间距0.5行；加粗。

[0081] 应用平台服务器任务启动后用poi加载文档1，将整篇文档1切分成段落列表，然后，遍历段落列表，筛选出目标段落：

[0082] 1.1液芯光波导的发展历程；

[0083] 1.2液芯光波导的传光原理；

- [0084] 1.3液芯光波导的特点；
[0085] 1.4液芯光波导在分析领域中的应用；
[0086] 2.1离子液体的发展历程；
[0087] 2.2离子液体性质及组成；
[0088] 2.2离子液体性质及组成；
[0089] 2.3离子液体在萃取分离中的应用；
[0090] 将上述目标段落与所述目标段落对应的知识逐一建立关联，根据所述关联，拆分所述文档1，将整篇文档1拆分成多个知识条目。
[0091] 搜索系统在搜索“离子液体性质及其组成”相关的有用知识过程中，应用平台服务器，仅需对“2.2离子液体性质及组成”与对应的组成的知识条目进行分析，筛选出有用知识；在此过程中，降低了搜索系统的搜索范围，缩短了搜索系统的搜索的时间，提高了系统带宽、数据库等资源的利用率。
[0092] 请继续参阅图2所述方法还包括：
[0093] S101获取模板文件；
[0094] 首先上传模板文件，系统会将模板文件保存到数据存储服务器内，然后添加相应标题，例如标题为：离子液体；用户在标题示例输入域中输入“离子液体”；
[0095] S102确定所述模板文件中的目标标题；
[0096] 应用平台服务器会根据输入内容“离子液体”扫描文档中的每一个段落，所有包含“离子液体”的标题和段落均会被罗列出来，用户任意选择一个标题作为目标标题，应用平台服务器根据用户的选择，确定模板文件中的目标标题。
[0097] S103筛选出所述目标标题的同级标题；
[0098] S104根据所述同级标题的段落属性，生成标题模板。
[0099] 应用平台服务器扫描整篇文档，对文档中的每一个段落进行属性比对，发现与目标标题的属性一致的段落便会记录下来，所述记录下来的标题便为所述目标标题的同级标题，将所述同级标题，以及，目标标题显示在同级标题列表中，用户可以通过查看同级标题列表来确认自己选择的标题能不能正确拆分文档，若果可以正确的拆分文档，则标题模板生成。
[0100] 实施例2：
[0101] 用户在标题示例输入域中输入“离子液体”；应用平台服务器会根据输入内容“离子液体”扫描每一个段落，然后，将包含输入内容的标题和段落全部罗列出来：
[0102] 2离子液体；
[0103] 2.1离子液体的发展历程；
[0104] 2.2离子液体性质及组成；
[0105] 2.3离子液体在萃取分离中的应用；
[0106] 用户根据需求选择“2.1离子液体的发展历程”作为目标标题，应用平台服务器根据用户的选择，确定“2.1离子液体的发展历程”为目标标题。
[0107] 应用平台服务器扫描整篇文档，对段落的属性的解析，筛选出目标标题的同级标题，并显示在同级标题列表中；
[0108] 显示的内容为：

- [0109] 1.1液芯光波导的发展历程;
- [0110] 液芯光波导的发展历程对应的知识。
- [0111] 1.2液芯光波导的传光原理;
- [0112] 液芯光波导的传光原理对应的知识。
- [0113] 1.3液芯光波导的特点;
- [0114] 液芯光波导的特点对应的知识。
- [0115] 1.4液芯光波导在分析领域中的应用;
- [0116] 液芯光波导在分析领域中的应用对应的知识。
- [0117] 2.1离子液体的发展历程;
- [0118] 离子液体的发展历程对应的知识。
- [0119] 2.2离子液体性质及组成;
- [0120] 离子液体性质及组成对应的知识。
- [0121] 2.3离子液体在萃取分离中的应用;
- [0122] 离子液体在萃取分离中的应用对应的知识。
- [0123] 用户可以通过查看同级标题列表来确认自己选择的目标标题能不能正确拆分文档模板,如果自己选的择的目标标题可以正确拆分文档模板,则标题模板生成。
- [0124] 在一些情况下用户选择的目标标题对文档的拆分不够细致,还可以继续添加二级目录甚至三级目录,具体添加到几级标题模板,可以根据具体需求进行定制。
- [0125] 具体的,请参阅图4,本申请实施例中所述步骤S104包括以下步骤:
- [0126] S10411显示所述同级标题,以及,所述同级标题对应的知识;
- [0127] S10412判断所述同级标题对应的知识是否符合预置划分规则;
- [0128] 所述预置划分规则为对一篇文档的划分细度。
- [0129] 例如:在文档1中,预置划分规则是对文档1的划分细度为“液芯光波导的传光原理”与“液芯光波导的发展历程”。
- [0130] 如果最初“1液芯光波导和2离子液体”为所述同级标题,然后,分液芯光波导对应的知识和离子液体对应的知识。其中,芯光波导对应的知识,包含“液芯光波导的传光原理”与“液芯光波导的发展历程”两部分内容,显然采用“1液芯光波导和2离子液体”作为所述同级标题来划分文档1划分细度不符合预置划分规则。
- [0131] 如果所述同级标题对应的知识符合预置划分规则,执行S10413根据所述同级标题的段落属性,生成标题模板;
- [0132] 如果所述同级标题对应的知识不符合预置划分规则,执行S10414调取所述同级标题的子标题;
- [0133] S10415根据所述同级标题的段落属性,以及,所述子级标题的段落属性,生成属性模板,所述属性模板包括:根据同级标题的段落属性生成的同级标题模板,以及,根据所述子标题的段落属性生成的子级标题模板。
- [0134] 实施例3:
- [0135] 用户在标题示例输入域中输入“液芯光波导”;应用平台服务器会根据输入内容“液芯光波导”扫描每一个段落,将包含“液芯光波导”的标题和段落会罗列出来:
- [0136] 显示内容:

- [0137] 1液芯光波导；
- [0138] 液芯光波导对应的知识。
- [0139] 1.1液芯光波导的发展历程；
- [0140] 液芯光波导的发展历程对应的知识。
- [0141] 1.2液芯光波导的传光原理；
- [0142] 液芯光波导的传光原理对应的知识。
- [0143] 1.3液芯光波导的特点；
- [0144] 液芯光波导的特点对应的知识。
- [0145] 1.4液芯光波导在分析领域中的应用；
- [0146] 液芯光波导在分析领域中的应用对应的知识。
- [0147] 1.4.1液芯光波导在萃取方向上的应用；
- [0148] 液芯光波导在萃取方向上的应用对应的知识。
- [0149] 1.4.2液芯光波导在传感方向上的应用；
- [0150] 液芯光波导在传感方向上的应用对应的知识。
- [0151] 用户根据需求选择“1液芯光波导”作为目标标题，应用平台服务器根据用户的选择，确定“1液芯光波导”为目标标题。
- [0152] 应用平台服务器扫描整篇文档，对每一个段落的属性的解析，发现和“1液芯光波导”的属性一致的段落会记录下来，显示在同级标题列表中。
- [0153] 显示内容：
- [0154] 1液芯光波导；
- [0155] 2离子液体；
- [0156] 应用平台服务器分析“1液芯光波导和2离子液体；”的段落属性，根据“1液芯光波导和2离子液体；”的段落属性生成一级标题模板；此时一级标题模板将整篇文档拆分成两个知识条目；如果用户觉得一级标题模板对文档拆分的不够细致，应用平台服务器调取所述“1液芯光波导和2离子液体”的子标题；
- [0157] 所述子标题包括：
- [0158] 1.1液芯光波导的发展历程；
- [0159] 1.2液芯光波导的传光原理；
- [0160] 1.3液芯光波导的特点；
- [0161] 1.4液芯光波导在分析领域中的应用；
- [0162] 2.1离子液体的发展历程；
- [0163] 2.2离子液体性质及组成；
- [0164] 2.3离子液体在萃取分离中的应用；
- [0165] 根据上述子标题的属性生成的子级标题模板。
- [0166] 可见本申请实施例示出的标题模板包括：根据同级标题的段落属性生成的同级标题模板，以及，根据所述子标题的段落属性生成的子级标题模板。
- [0167] 通过本申请实施例示出的方法可以根据用户的需求生成一级标题模板，二级标题模板，以及，三级标题模板等等。具体生成几级标题模板，可以根据具体需求进行定制。
- [0168] 在一些情况下，用户输入的标题内容对应不同级别的标题，此时，无法唯一确定目

标标题的同级标题,在此情况下,在用户输入的标题的内容上增加一个正则表达式,以进一步限定目标标题的内容,进而使得目标标题唯一确定。

[0169] 请参阅图4,本申请实施例中步骤S104包括以下步骤:

[0170] S10421遍历所述文档的段落,筛选出目标段落;

[0171] S10422如果出现多于一个的目标段落,则增加一个正则表达式;

[0172] S10423判断所述目标段落的内容是符合正则表达式;

[0173] S10424如果所述目标段落的内容符合正则表达式,则保留所述目标段落;

[0174] S10425如果所述目标段落的内容不符合正则表达式,则删除所述目标段落。

[0175] 具体的,实施例4:

[0176] 用户在标题示例输入域中输入“发展历程”,应用平台服务器会根据输入内容“发展历程”扫描每一个段落,包含“发展历程”的标题和段落会罗列出来:

[0177] 显示内容:

[0178] 1.1液芯光波导的发展历程;

[0179] 2.1离子液体的发展历程;

[0180] 此时,在搜索的过程中,出现两个目标标题“1.1液芯光波导的发展历程和2.1离子液体的发展历程”,在此情况下,本申请实施例示出的方法,添加一个正则表达式(离子液体),此时,“2.1离子液体的发展历程”符合正则表达式,被重新定义为目标标题而被保留。“1.1液芯光波导的发展历程”会被删除。

[0181] 请参阅图5,本申请实施例示中S106包括以下步骤:

[0182] S1061遍历所述文档的段落,确定所述段落的所属的属性级别;

[0183] S1062如果所述段落的属性符合所述属性规则,则所述段落为目标段落;

[0184] S1063如果所述段落的属性符合所述标题模板的上一级标题的属性规则,分析所述段落对应的知识,得到分析结果;

[0185] S1064根据所述分析结果生成一目标段落。

[0186] 实施例5:

[0187] 用户选择的标题模板是二级标题,首先,上传文档并选择标题模板,根据用户选择的标题模板,确定所述标题模板对应的属性规则为段落标号-X.X;字体-黑体;字号-四号;首行缩进2字符;段前间距0.5行;加粗。

[0188] 任务启动后首先用poi加载文档,将整篇文档切分成段落列表,遍历段落列表,显示目标段落,所述目标段落的同级段落,以及,上述段落对应的知识,

[0189] 1.1液芯光波导的发展历程;

[0190] 液芯光波导的发展历程对应的知识。

[0191] 1.2液芯光波导的传光原理;

[0192] 液芯光波导的传光原理对应的知识。

[0193] 1.3液芯光波导的特点;

[0194] 液芯光波导的特点对应的知识。

[0195] 1.4液芯光波导在分析领域中的应用;

[0196] 液芯光波导在分析领域中的应用对应的知识。

[0197] 2.1离子液体的发展历程;

- [0198] 离子液体的发展历程对应的知识。
- [0199] 2.2离子液体性质及组成；
- [0200] 离子液体性质及组成对应的知识。
- [0201] 2.3离子液体在萃取分离中的应用；
- [0202] 离子液体在萃取分离中的应用对应的知识。
- [0203] 在搜索的过程中，应用平台服务器判断“1液芯光波导”的属性符合所述标题模板的上一级标题的属性规则；此时，应用平台服务器分析“液芯光波导对应的知识”，得到“液芯光波导对应的知识”的内容为液芯光波导的原理简介，应用平台服务器根据液芯光波导的原理简介生成一个新的目标段落“1.5液芯光波导的原理简介”，并将该目标段落与对应的知识生成一条知识条目，将生成的知识条目存储在solr(独立的企业级搜索应用服务器)中。
- [0204] 在将整篇文档拆分成多个知识条目并显示的过程中，不乏存在一些知识条目，所述知识条目中包含的知识未涉及到任何有用知识，这些知识条目称之为无用知识条目，在此条件下本申请实施例示出的方法，将所述无用知识条目删除，进而减少系统的内存占用率，提高系统的搜索效率。
- [0205] 请参阅图6，本申请实施例示中S107包括以下步骤：
- [0206] S10711显示所述目标段落，以及，所述目标段落对应的知识；
- [0207] S10712判断所述目标段落对应的知识是否为有用知识；
- [0208] 如果所述目标段落对应的知识为有用知识，则执行S10713建立所述目标段落与所述目标段落对应的知识之间的关联；
- [0209] 如果所述目标段落对应的知识不是有用知识，则执行S10714删除所述目标段落，以及，所述目标段落对应的知识。
- [0210] 实施例6：
- [0211] 用户搜索“光波导”相关的知识，首先，上传文档并选择标题模板，确定所述标题模板对应的属性规则为段落标号-X.X；字体-黑体；字号-四号；首行缩进2字符；段前间距0.5行；加粗。任务启动后首先用poi加载文档，将整篇文档切分成段落列表，遍历段落列表，显示所述目标段落，所述目标段落的同级段落，以及上述段落对应的知识；
- [0212] 1.1液芯光波导的发展历程；
- [0213] 液芯光波导的发展历程对应的知识。
- [0214] 1.2液芯光波导的传光原理；
- [0215] 液芯光波导的传光原理对应的知识。
- [0216] 1.3液芯光波导的特点；
- [0217] 液芯光波导的特点对应的知识。
- [0218] 1.4液芯光波导在分析领域中的应用；
- [0219] 液芯光波导在分析领域中的应用对应的知识。
- [0220] 2.1离子液体的发展历程；
- [0221] 离子液体的发展历程对应的知识。
- [0222] 2.2离子液体性质及组成；
- [0223] 离子液体性质及组成对应的知识。

- [0224] 2.3离子液体在萃取分离中的应用；
- [0225] 离子液体在萃取分离中的应用对应的知识。
- [0226] 应用平台服务器判断所述目标段落对应的知识是否为有用知识；其中，“离子液体的发展历程对应的知识；离子液体性质及组成对应的知识；离子液体在萃取分离中的应用对应的知识”均为无用知识，应用平台服务器将“2.1离子液体的发展历程，2.2离子液体性质及组成，以及，2.3离子液体在萃取分离中的应用”以及，上述段落对应的知识删除；将“1.1液芯光波导的发展历程，1.2液芯光波导的传光原理，1.3液芯光波导的特点，1.4液芯光波导在分析领域中的应用”与这些段落对应的知识逐一建立关联，根据所述关联，拆分所述整篇文档，将整篇文档拆分成多个知识条目。
- [0227] 本申请实施例示出的方法，将所述无用知识条目删除，进而减少系统的内存占用率，提高系统的搜索效率。
- [0228] 请参阅图7，本申请实施例示中S107包括以下步骤：
- [0229] S10721显示所述目标段落，以及，所述目标段落对应的知识；
- [0230] S10722如果所述目标段落对应的知识中包括图片，将所述图片以链接的形式存储在目标段落对应的知识中，或，如果所述目标段落对应的知识中包括表格，将所述表格转化为成可以展示的格式存储在目标段落对应的知识中；
- [0231] S10723将所述目标段落与所述目标段落对应的知识逐一建立关联；
- [0232] 图片和表格是通过一个公用的poi插件来进行判断：
- [0233] 其中，以下代码表示该段落有表格：
- ```
[0234] Body Element Type.TABLE.equals (be.get Element Type ())
```
- [0235] 以下代码能识别出图片：
- ```
[0236] XWPFParagraphparagraph=be.getBody () .getParagraphArray (i) ;List<XWPFRun>xwpfRuns=paragraph.getRuns () ;for (XWPFRunxwpfRun:xwpfRuns) {Stringctr=xwpfRun.getCTR () .toString ();if (null!=ctr&&ctr.indexOf ("图片") !=-1) {picflag=true;}}
```
- [0237] S10724将所述目标段落与所述目标段落对应的知识逐一建立关联。
- [0238] 遍历段落列表，判断段落内容里是不是有图片，如果有图片则将图片用文件流抽取出以特定的方式命名存储在该文档对应的路径下；
- [0239] 判断段落内容里是不是有表格，如果有表格就将表格里面的单元格利用| |
| --- |
|， 标签进行处理，存储成页面可以展示的内容。 |
- [0240] 可选择的，所述属性规则包括：字体的大小，字形、首行缩进距离，段前距离，断后距离中的一种或几种组合。
- [0241] 对于二级标题而言，段落属性对应为：段落标号-X.X；字体-黑体；字号-四号；首行缩进2字符；段前间距0.5行；加粗；
- [0242] 对于文档1而言中的标而言；如果确定一个段落的字号-四号，则无需对于其他的段落属性进行识别，便可确定该段落为二级标题。
- [0243] 对于三级标题而言，由于文档中正文对应的字号-小四，三级标题对应的字号也为小四，此时需要确定该段落的字体是为黑体，如果字体为黑体，则确定该段落为三级标题段落，如果该段落的字体为楷体，则确定该段落为正文段落。

- [0244] 请参阅图8,本申请实施例第二方面示出一种文档的结构化拆分装置,所述装置包括:
- [0245] 选取单元21,用于选取标题模板,确定所述标题模板的属性规则;
- [0246] 22遍历单元22,用于根据所述属性规则,遍历文档的段落,筛选出目标段落,以及,所述目标段落对应的知识;
- [0247] 建立单元23,用于将所述目标段落与所述目标段落对应的知识逐一建立关联;
- [0248] 24拆分单元,用于根据所述关联,拆分所述文档。
- [0249] 本申请实施例示出一种文档的结构化生拆分装置,所述装置在整篇文档中筛选出目标段落,所述目标段落为段落属性符合所述属性规则的段落;将所述目标段落与所述目标段落对应的知识逐一建立关联,此时,目标段落与目标段落对应的知识形成一个知识条目,进而将整篇文档拆分成多个知识条目。搜索系统在搜索有用知识的过程中,应用平台服务器仅需对知识条目进行分析,筛选出有用知识,缩小了搜索系统的搜索范围,进而缩短了搜索的时间,提高了系统带宽、数据库等资源的利用率。
- [0250] 本申请实施例第三方面示出一种文档的结构化生拆分系统,所述系统包括:
- [0251] 应用平台服务器31,以及,与其连接的数据存储服务器32,所述数据存储服务器32设置在所述应用平台服务器31内部或独立设置,以及,与应用平台服务器31通过互联网或移动互联网连接的终端33;
- [0252] 所述应用平台服务器31,用于选取标题模板,确定所述标题模板的属性规则;
- [0253] 用于根据所述属性规则,遍历文档的段落,筛选出目标段落,以及,所述目标段落对应的知识,所述目标段落为段落属性符合所述属性规则的段落;
- [0254] 用于将所述目标段落与所述目标段落对应的知识逐一建立关联;
- [0255] 用于根据所述关联,拆分所述文档;
- [0256] 所述终端33用于向所述应用平台服务器发送文档,以及,用于接收才分后的文档;
- [0257] 所述数据存储服务器32,用于相关数据的存储。
- [0258] 本申请实施例示出一种文档的结构化生拆分系统,所述系统在整篇文档中筛选出目标段落,所述目标段落为段落属性符合所述属性规则的段落;将所述目标段落与所述目标段落对应的知识逐一建立关联,此时,目标段落与目标段落对应的知识形成一个知识条目,进而将整篇文档拆分成多个知识条目。搜索系统在搜索有用知识的过程中,应用平台服务器仅需对知识条目进行分析,筛选出有用知识,缩小了搜索系统的搜索范围,进而缩短了搜索的时间,提高了系统带宽、数据库等资源的利用率。
- [0259] 由以上技术方案可知,本申请实施例公开了一种文档的结构化拆分方法,装置及系统,所述方法在整篇文档中筛选出目标段落,所述目标段落为段落属性符合所述属性规则的段落;将所述目标段落与所述目标段落对应的知识逐一建立关联,此时,目标段落与目标段落对应的知识形成一个知识条目,进而将整篇文档拆分成多个知识条目。搜索系统在搜索有用知识的过程中,应用平台服务器仅需对知识条目进行分析,筛选出有用知识,缩小了搜索系统的搜索范围,进而缩短了搜索的时间,提高了系统带宽、数据库等资源的利用率。
- [0260] 本发明可用于众多通用或专用的计算系统环境或配置中,例如:个人计算机、服务器计算机、手持设备或便携式设备、平板型设备、多处理器系统、基于微处理器的系统、置顶

盒、可编程的消费电子设备、网络PC、小型计算机、大型计算机、包括以上任何系统或设备的分布式计算环境等等。

[0261] 本发明可以在由计算机执行的计算机可执行指令的一般上下文中描述,例如程序模块。一般地,程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等等。也可以在分布式计算环境中实践本发明,在这些分布式计算环境中,由通过通信网络而被连接的远程处理设备来执行任务。在分布式计算环境中,程序模块可以位于包括存储设备在内的本地和远程计算机存储介质中。

[0262] 需要说明的是,在本文中,诸如“第一”和“第二”等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来,而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者设备所固有的要素。

[0263] 本领域技术人员在考虑说明书及实践这里公开的申请后,将容易想到本申请的其它实施方案。本申请旨在涵盖本申请的任何变型、用途或者适应性变化,这些变型、用途或者适应性变化遵循本申请的一般性原理并包括本申请未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的,本申请的真正范围和精神由下面的权利要求指出。

[0264] 应当理解的是,本申请并不局限于上面已经描述并在附图中示出的精确结构,并且可以在不脱离其范围进行各种修改和改变。本申请的范围仅由所附的权利要求来限制。

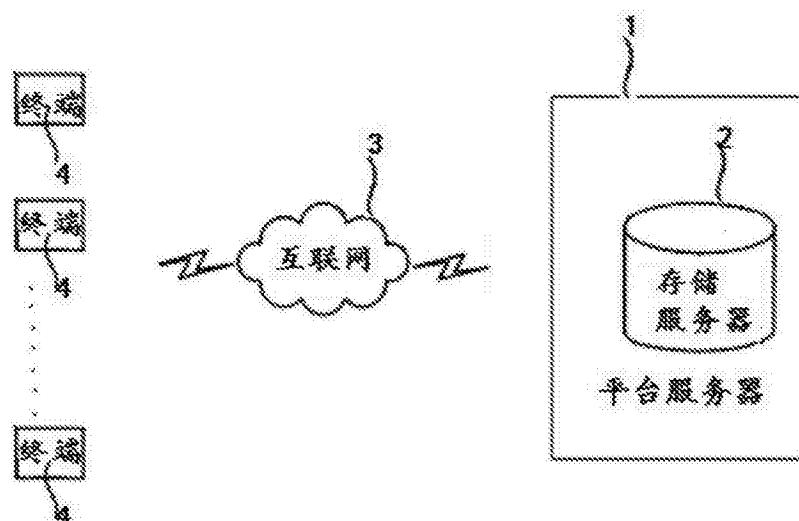


图1

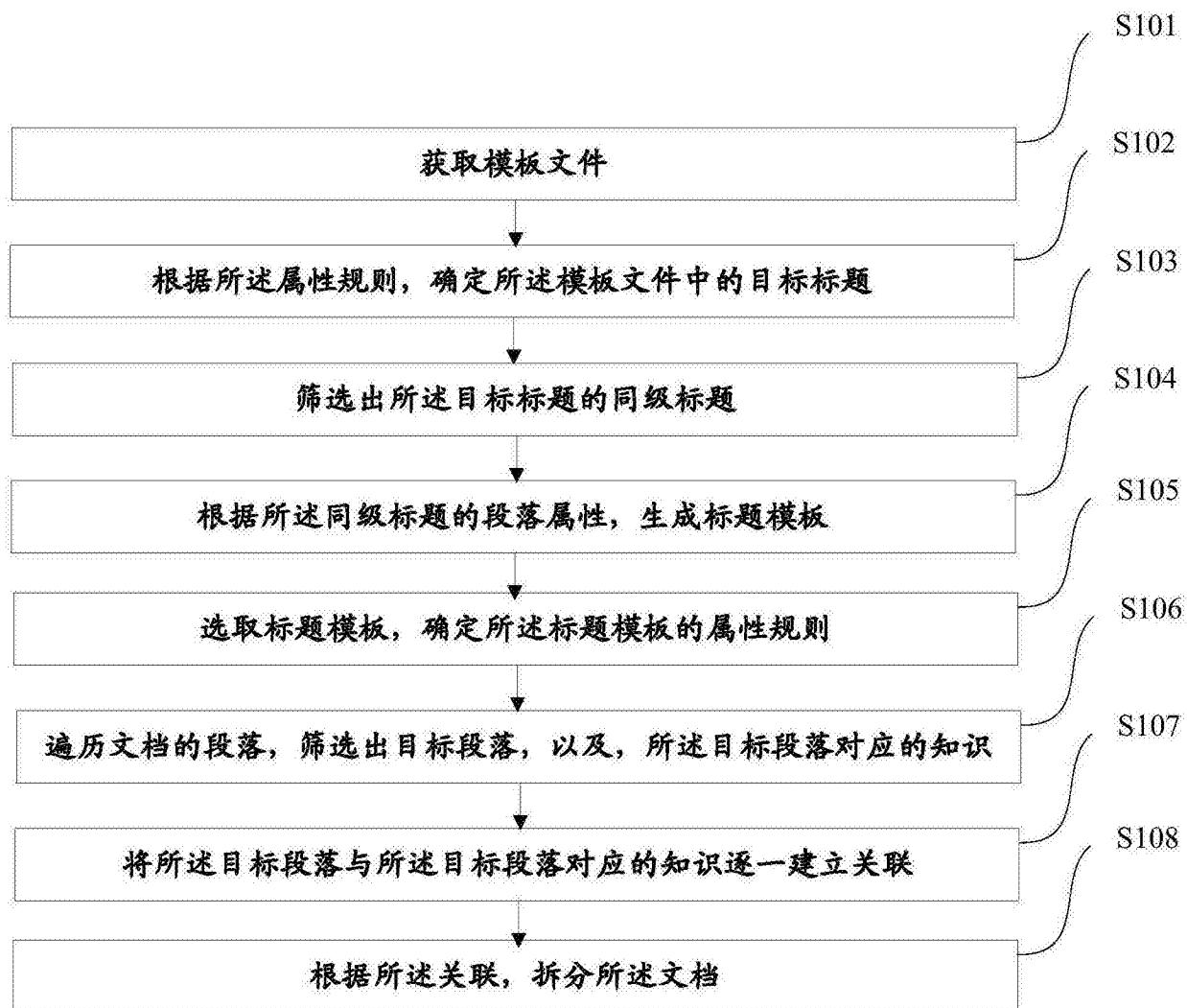


图2

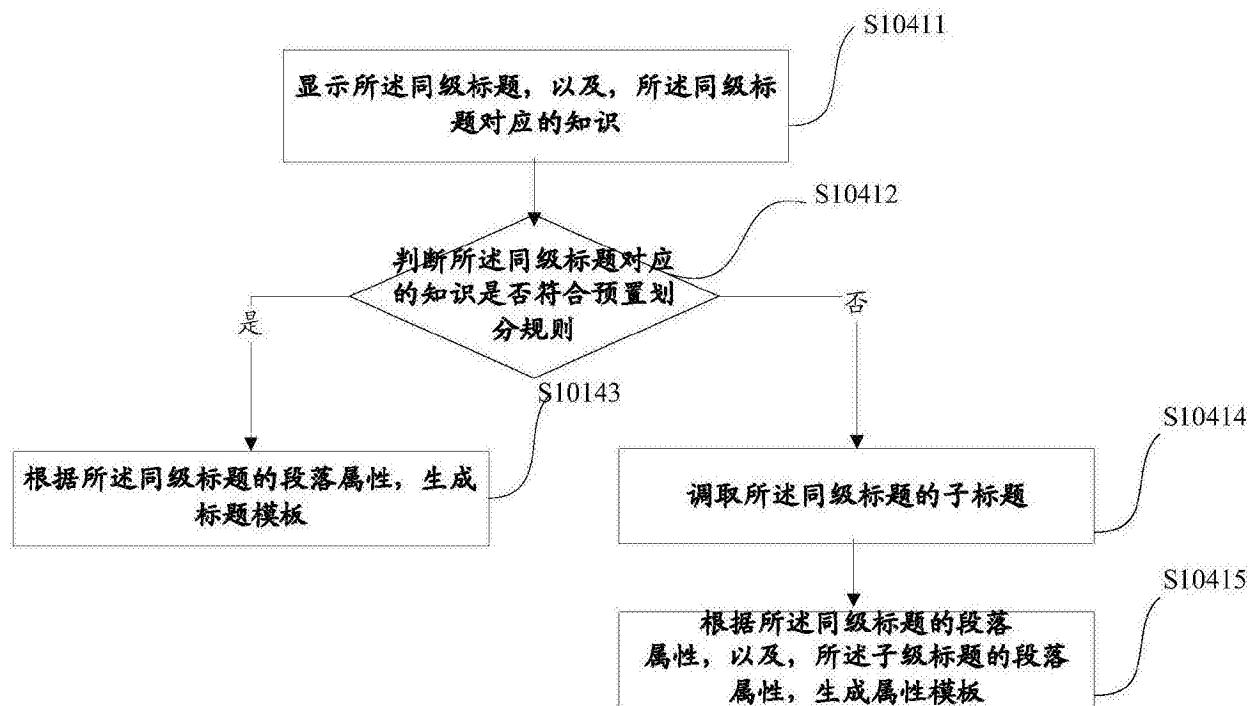


图3

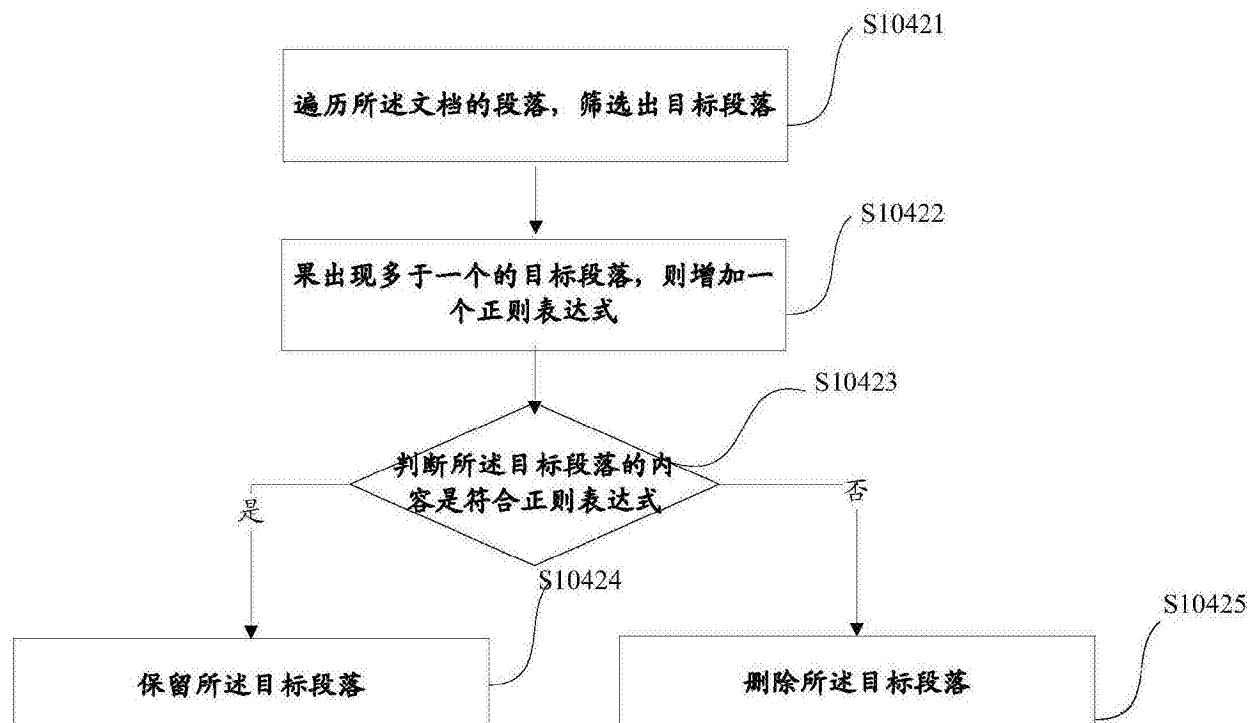


图4

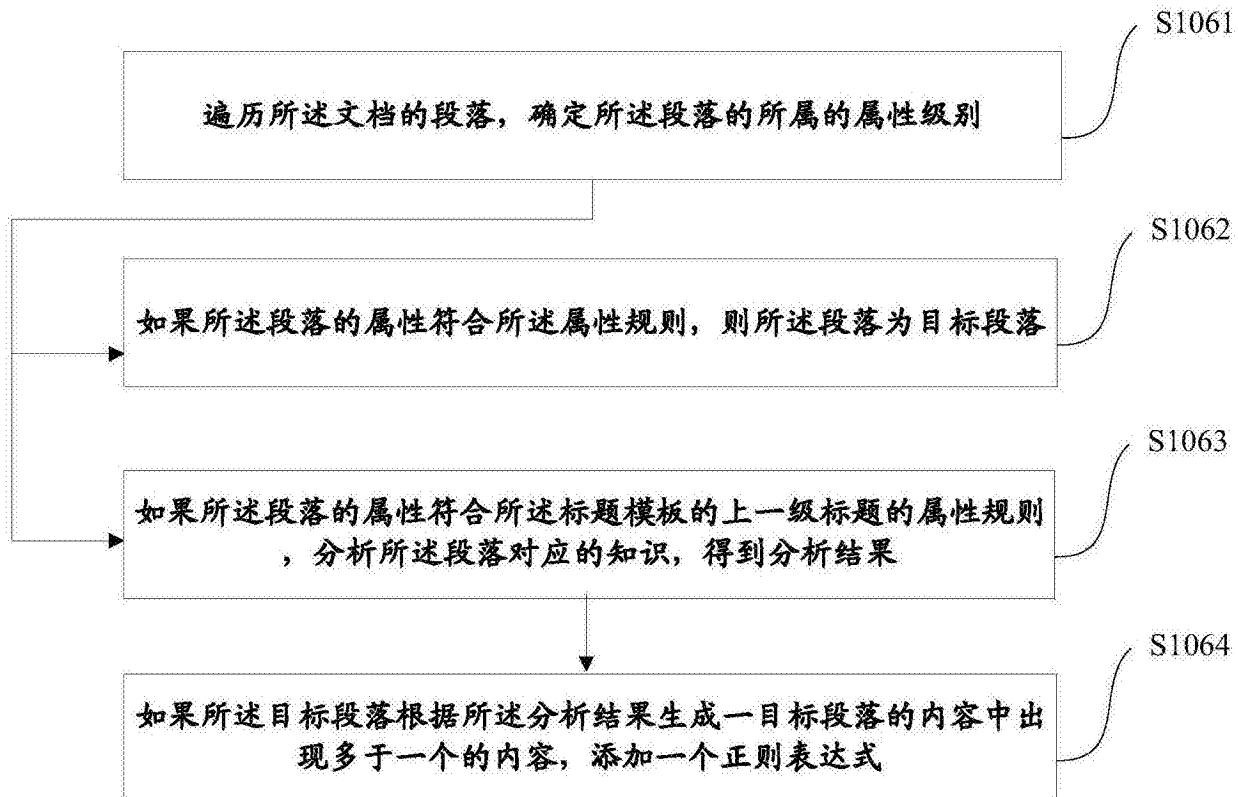


图5

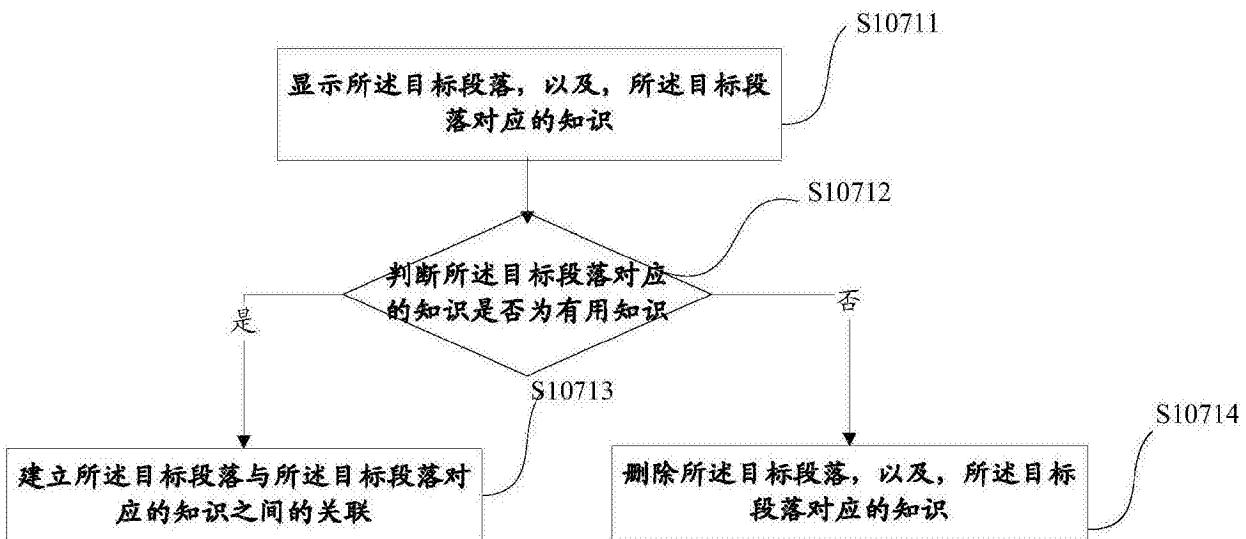


图6

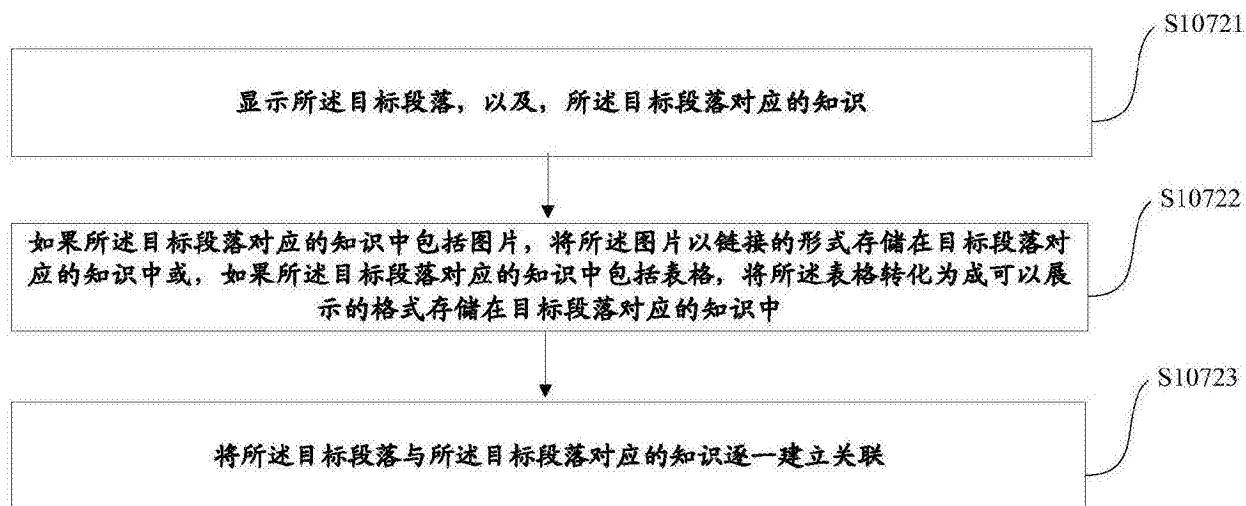


图7

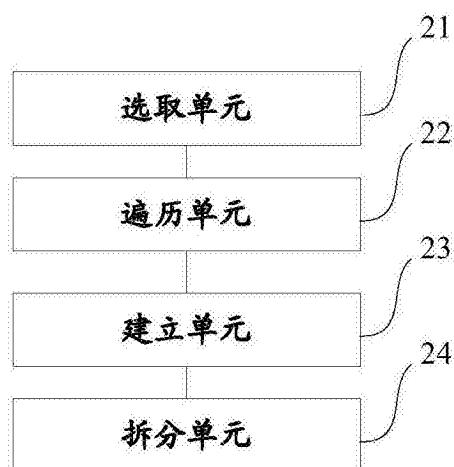


图8

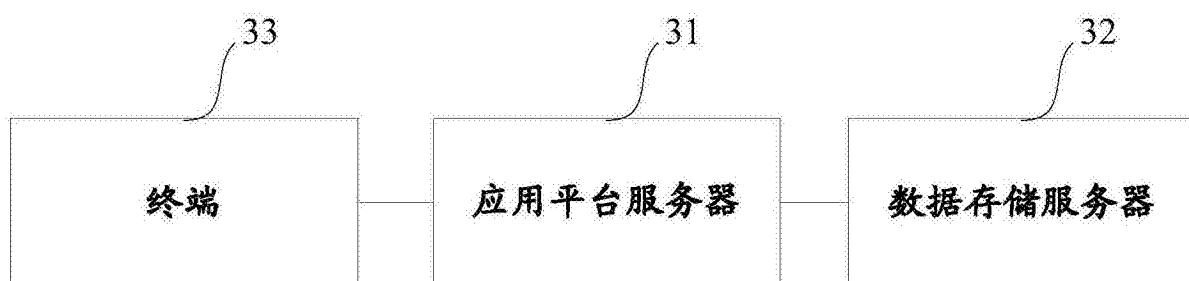


图9-1

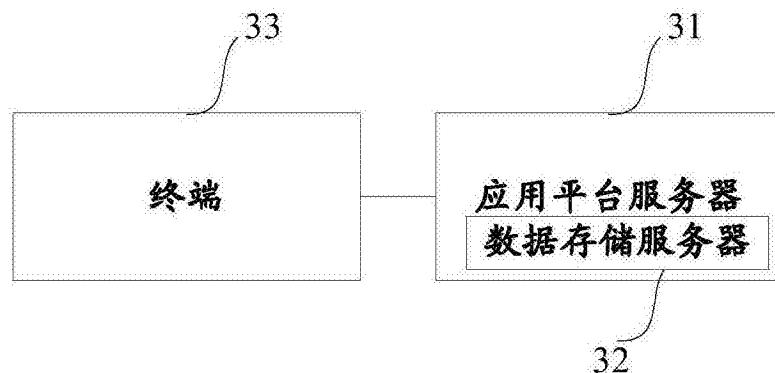


图9-2