

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局



(43) 国际公布日
2020 年 12 月 30 日 (30.12.2020)

(10) 国际公布号
WO 2020/258492 A1

(51) 国际专利分类号:
G06F 21/62 (2013.01)

(21) 国际申请号: PCT/CN2019/103028

(22) 国际申请日: 2019 年 8 月 28 日 (28.08.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910575551.3 2019年6月28日 (28.06.2019) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(72) 发明人: 唐志辉(TANG, Zhihui); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 北京市立方律师事务所(LIFANG & PARTNERS); 中国北京市东城区香河园街1号院信德京汇中心12层, Beijing 100028 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,

(54) Title: INFORMATION PROCESSING METHOD AND APPARATUS, STORAGE MEDIUM AND TERMINAL DEVICE

(54) 发明名称: 信息处理方法、装置、存储介质及终端设备

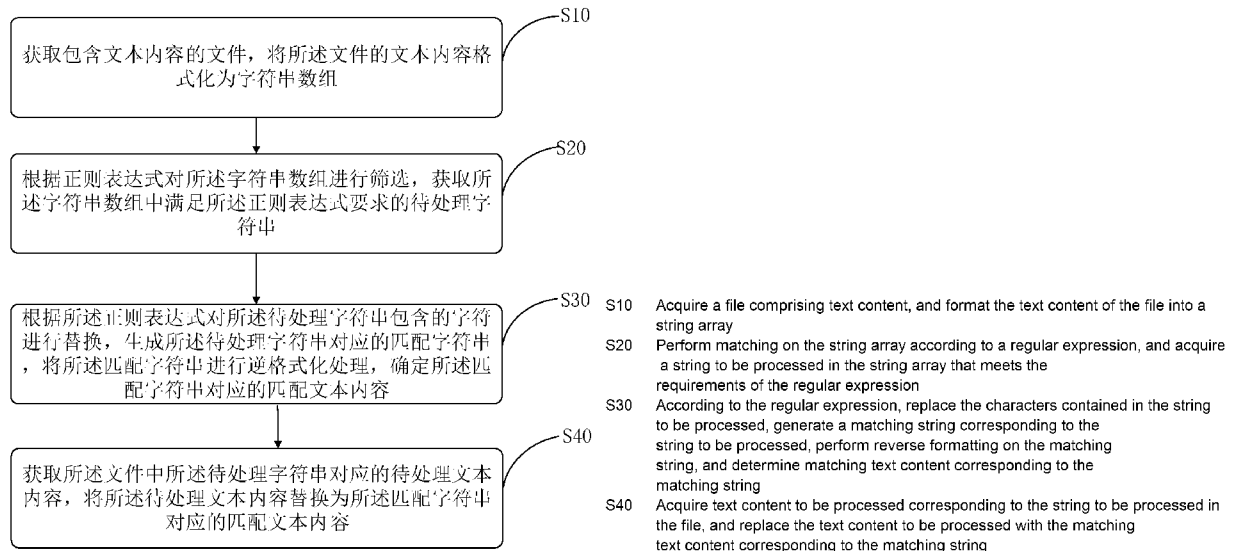


图 1

(57) Abstract: An information processing method and apparatus, a storage medium, and a terminal device, which relate to the field of data processing. The method comprises: acquiring a file comprising text content, and formatting the text content of the file into a string array; performing matching on the string array according to a regular expression, and acquiring a string to be processed in the string array that meets the requirements of the regular expression; according to the regular expression and the string to be processed, determining a matching string corresponding to the string to be processed; and acquiring text content to be processed corresponding to the string to be processed in the file, and replacing the text content to be processed with matching text content corresponding to the matching string. The described method may process text content in a file in batches, and increases processing efficiency for file content.



SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, ZA, ZM, ZW。

- (84) 指定国(除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

根据细则4.17的声明:

- 发明人资格(细则4.17(iv))

本国际公布:

- 包括国际检索报告(条约第21条(3))。

(57) 摘要: 一种信息处理方法、装置、存储介质及终端设备, 涉及数据处理领域, 所述方法包括: 获取包含文本内容的文件, 将所述文件的文本内容格式化为字符串数组; 根据正则表达式对所述字符串数组进行匹配, 获取所述字符串数组中满足正则表达式要求的待处理字符串; 根据所述正则表达式与所述待处理字符串, 确定所述待处理字符串对应的匹配字符串; 获取所述文件中所述待处理字符串对应的待处理文本内容, 将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容。该方法能够批量处理文件中的文本内容, 提高了文件内容的处理效率。

信息处理方法、装置、存储介质及终端设备

本申请要求于 2019 年 6 月 28 日提交中国专利局、申请号为
5 201910575551.3，发明名称为“信息处理方法、装置、存储介质及终端设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及数据处理领域，具体涉及一种信息处理方法、装置、存储
10 介质及终端设备。

背景技术

随着互联网信息的发展，公司信息、合同信息、个人信息及客户信息
等信息都记录在计算机设备中，而在不同时间段或面向不同的用户，需要
15 对记录在案的部分信息进行特殊处理，例如该些信息面向大众时，对重要、敏感的信息作屏蔽处理。而目前，发明人意识到对于此类信息，通常是对存储该些信息的文件直接作加密处理，或者通过人工一一修改文件中的重要、敏感信息，直接对文件进行加密处理无法让文件针对性地面向不同用户，同时，若面向不同的用户需要对文件进行不同处理，通过人工一一修
20 改文件需要耗费大量的资源，造成时间、人力、资源的浪费。

发明内容

为克服以上技术问题，特别是现有技术无法高效、低成本地进行文件
内容处理的问题，特提出以下技术方案：

25 第一方面，本申请提供了一种信息处理方法，包括：

获取包含文本内容的文件，将所述文件的文本内容格式化为字符串数
组；

根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中
满足所述正则表达式要求的待处理字符串；

根据所述正则表达式对所述待处理字符串包含的字符进行替换，生成所述待处理字符串对应的匹配字符串，将所述匹配字符串进行逆格式化处
理，确定所述匹配字符串对应的匹配文本内容；

5 获取所述文件中所述待处理字符串对应的待处理文本内容，将所述待
处理文本内容替换为所述匹配字符串对应的匹配文本内容。

第二方面，本申请提供一种信息处理装置，包括：

格式化模块：用于获取包含文本内容的文件，将所述文件的文本内容
格式化为字符串数组；

10 匹配模块：用于根据正则表达式对所述字符串数组进行筛选，获取所
述字符串数组中满足所述正则表达式要求的待处理字符串；

确定模块：用于根据所述正则表达式对所述待处理字符串包含的字符
进行替换，生成所述待处理字符串对应的匹配字符串，将所述匹配字符串
进行逆格式化处处理，确定所述匹配字符串对应的匹配文本内容；

15 替换模块：用于获取所述文件中所述待处理字符串对应的待处理文本
内容，将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内
容。

第三方面，本申请还提供了一种存储有计算机可读指令的非易失存储
介质，所述计算机可读指令被一个或多个处理器执行时，使得一个或多个
处理器执行一种信息处理方法，所述信息处理方法包括以下步骤：

20 获取包含文本内容的文件，将所述文件的文本内容格式化为字符串数
组；

根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中
满足所述正则表达式要求的待处理字符串；

25 根据所述正则表达式对所述待处理字符串包含的字符进行替换，生成
所述待处理字符串对应的匹配字符串，将所述匹配字符串进行逆格式化处
理，确定所述匹配字符串对应的匹配文本内容；

获取所述文件中所述待处理字符串对应的待处理文本内容，将所述待
处理文本内容替换为所述匹配字符串对应的匹配文本内容。。

第四方面，本申请还提供了一种终端设备，所述终端设备包括一个或

多个处理器、存储器、一个或多个计算机程序，其中所述一个或多个计算机程序用于执行一种信息处理方法，所述信息处理方法包括以下步骤：

获取包含文本内容的文件，将所述文件的文本内容格式化为字符串数组；

- 5 根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中满足所述正则表达式要求的待处理字符串；

根据所述正则表达式对所述待处理字符串包含的字符进行替换，生成所述待处理字符串对应的匹配字符串，将所述匹配字符串进行逆格式化处
理，确定所述匹配字符串对应的匹配文本内容；

- 10 获取所述文件中所述待处理字符串对应的待处理文本内容，将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容。

- 本申请提供了一种对文件中指定的文本内容进行处理的方法，在获取到文件后，将文件的文本内容进行处理，将文件的文本内容格式化为一个
15 一个的字符串，从而将文件的文本内容看作由一个个字符串组成的字符串数组，然后获取处理该文件文本内容的正则表达式，筛选所述字符串数组中满足正则表达式要求的待处理字符串，根据所述正则表达式对所述待处理字符串包含的字符进行替换，生成所述待处理字符串对应的匹配字符串，将所述匹配字符串进行逆格式化处
理，确定所述匹配字符串对应的匹配文本内容，在确定了待处理字符串跟各待处理字符串对应的匹配字符串
20 后，模拟打开所述文件，模拟执行替换操作，获取所述文件中所述待处理字符串对应的待处理文本内容，将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容，完成对文件中的待处理文本内容的转换，实现对文件中重要或敏感信息的屏蔽、修改或掩码处理，批量化的文件内容处理节省了时间，提高对文件中文本内容的处理效率。

25

附图说明

图 1 为本申请信息处理方法的一实施例流程示意图；

图 2 为本申请信息处理方法的另一实施例流程示意图；

图 3 为本申请信息处理装置的一实施例示意图；

图 4 为本申请终端设备的一实施例结构示意图。

具体实施方式

本领域技术人员应当理解，本申请所称的“应用”、“应用程序”、“计算机程序”以及类似表述的概念，是业内技术人员所公知的相同概念，是指由一系列计算机指令及相关数据资源有机构造的适于电子运行的计算机软件。除非特别指定，这种命名本身不受编程语言种类、级别，也不受其赖以运行的操作系统或平台所限制。理所当然地，此类概念也不受任何形式的终端所限制。

本申请实施例提供一种信息处理方法，如图 1 所示，所述方法包括以下步骤：

S10: 获取包含文本内容的文件，将所述文件的文本内容格式化为字符串数组。

本实施例中，在需要对文件中包含的内容进行处理时，例如文件中存储的公司信息，个人信息，客户信息，需要对特定的信息进行处理，所述处理包括掩码处理、替换处理、删改处理，首先获取包含文本内容的文件，然后对所述文件进行文本预处理，本实施例的一种实施方式，对所述文本内容进行格式化处理，具体的，将内容分析工具集合到本实施例的应用程序中，通过调用内容分析工具如 `tika` 将文件的文本内容进行处理，得到文件的 `metadata`, `content` 等内容，返回格式化信息，从而将文件的文本内容格式化为一个一个的字符串，本实施例将该些字符串的集合成定义为字符串数组。

S20: 根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中满足所述正则表达式要求的待处理字符串。

本实施例中，设置有一配置文件，配置文件中配置有正则表达式，一种实施方式，所述配置文件为 `xml` 配置文件，在 `xml` 配置文件中制定正则表达式，正则表达式是对字符串操作的一种逻辑公式，包括对普通字符（例如，`a` 到 `z` 之间的字母）和特殊字符（称为“元字符”）的操作，利用事先定义好的一些字符、及该些字符的组合组成一个“规则字符串”，通

过所述“规则字符串”对字符串进行过滤，具体的，根据 xml 配置文件的规则，利用正则表达式对所述字符串数组进行筛选，从所述字符串数组中查找到满足正则表达式规则的字符串，将该些字符串定义为待处理字符串，所述正则表达式中包括不同的条件规则，根据不同的条件规则匹配到不同的字符串，进一步的，还可以根据多个条件规则的组合匹配到满足多个条件规则的字符串，从而筛选出文件中指定的文本内容，例如，正则表达式包含的一条件规则为筛选出包含字符“aa”的字符串，根据该正则表达式对字符串数组中的字符串进行筛选，便能筛选出包含字符“aa”的字符串。

10 **S30:** 根据所述正则表达式对所述待处理字符串包含的字符进行替换，生成所述待处理字符串对应的匹配字符串，将所述匹配字符串进行逆格式化处理，确定所述匹配字符串对应的匹配文本内容。

为了对文件中特定的文本内容进行处理，包括对文本内容进行掩码处理、屏蔽处理、替换处理等，本实施例中，在获取到待处理字符串后，根据所述正则表达式对所述待处理字符串包含的字符进行替换，所述替换包括对字符串的字符进行更改，例如将字符 a 替换成字符 b；所述替换还包括对字符串的字符进行删除，即将字符 a 替换成空字符；所述替换还包括对字符串的字符进行增加，即将字符 a 替换成字符 ab；，对待处理字符串进行替换后，生成所述待处理字符串对应的匹配字符串，即将待处理字符串按照正则表达式的规则进行替换，替换的结果仍为字符串，本实施例将替换后的字符串定义为匹配字符串，所述匹配字符串为后续替换所述待处理字符串所用，例如，待处理的字符串 aaaa 对应的匹配字符串为 xxxx，待处理字符串 bbcc 对应的匹配字符串为 xyxy。字符串与文本内容存在映射关系，本实施例将生成的匹配字符串进行逆格式化处理，从而确定所述匹配字符串对应的匹配文本内容，并将匹配文本内容保存至配置文件中。

本实施例的一种实施方式，所述待处理字符串为文件的通用格式或标准格式的字符串，可以转化为任意格式的其他字符串。

S40: 获取所述文件中所述待处理字符串对应的待处理文本内容，将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容。

本实施例中，在 xml 配置文件中确定了文件中的待处理字符串及每个待处理字符串对应的匹配字符串后，模拟打开所述文件，具体的，获取所述文件的文件格式类型，然后根据所述文件格式类型获取模拟打开所述文件格式类型的方式，包括模拟打开.txt 格式文件、.doc 格式文件或.pdf 文件，然后再模拟执行文本内容替换操作，具体的，将所述待处理字符串转化为待处理文本内容，然后从文件中查找、定位所述待处理文本内容，再获取所述匹配字符串转化后对应的匹配文本内容，然后将所述待处理内容替换为所述的匹配文本内容，完成对文件中的所述待处理文本内容的转换，实现对文件中重要或敏感信息的屏蔽、修改或掩码处理。

本实施例提供了一种对文件中指定的文本内容进行处理的方法，在获取到文件后，将文件的文本内容进行处理，将文件的文本内容格式化为一个一个的字符串，从而将文件的文本内容看作由一个个字符串组成的字符串数组，然后获取处理该文件文本内容的正则表达式，筛选所述字符串数组中满足正则表达式要求的待处理字符串，根据所述正则表达式对所述待处理字符串包含的字符进行替换，生成所述待处理字符串对应的匹配字符串，将所述匹配字符串进行逆格式化处理，确定所述匹配字符串对应的匹配文本内容，在确定了待处理字符串跟各待处理字符串对应的匹配字符串后，获取所述文件中所述待处理字符串对应的待处理文本内容，将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容，完成对文件中的待处理文本内容的转换，实现对文件中重要或敏感信息的屏蔽、修改或掩码处理，批量化的文件内容处理节省了时间，提高对文件中文本内容的处理效率。

本申请的一种实施例，如图 2 所示，所述获取所述文件中所述待处理字符串对应的待处理文本内容，将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容之后，还包括：

S41： 获取所述待处理文本内容的原始文本格式，将所述匹配文本内容的文本格式设置为所述原始文本格式。

在实际应用中，由于不同文件具有不同的文本格式，甚至同一文件中不同文字内容亦具有不同的文本格式，包括粗体格式、不同字体格式、不

同颜色格式等，本实施例中，在将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容之后，获取所述待处理文本内容的原始文本格式，然后将替换掉所述待处理文本内容的匹配文本内容的文本格式设置为所述原始文本格式，设置后所述匹配文本内容的文本格式与待处理文本内容的文本格式相同，均为所述原始文本格式。例如，原先文本格式为粗体格式

5 格式的待处理文本内容，替换后文本内容的文本格式仍为粗体格式，原先文本格式为 5 号宋体的待处理文本内容，替换后文本内容的文本格式仍为 5 号宋体，在不修改原文件文本内容的文本格式下对文本内容进行替换或修改，从而达到不影响文件整体文本格式的优点。

10 本申请的一种实施例，所述根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中满足所述正则表达式要求的待处理字符串之后，还包括：

将所述正则表达式、待处理字符串及匹配字符串关联保存至配置文件；

15 接收对所述配置文件的正则表达式的修改，修改对应的匹配字符串，或

接收对所述配置文件的匹配字符串的修改，修改对应的正则表达式。

在实际应用中，可能不同部门对于文件的内容要求不同，对于文件中相同文本内容想要的处理结果亦不相同，例如，对于部门 S1，需要将文件 A 中的所有涉及人物姓名的文本内容替换为“***”的形式，对于部门 S2，需要将文件 A 中的所有涉及人物姓名的文本内容替换为“姓氏+**”的形式，本实施例中，将所述正则表达式、待处理字符串及匹配字符串关联保存至配置文件，例如保存至 xml 配置文件，然后当不同部门对于文件中相同文本内容的替换需求不同时，只需要修改所述 xml 配置文件，以修改

20 配置文件中的正则表达式或匹配字符串，便可根据不同的需求替换不同的文本内容。

25 本实施例的一种实施方式，接收对所述配置文件的正则表达式的修改，对正则表达式的替换条件进行修改，修改后根据待处理字符串及所述正则表达式的替换条件便能生成不同的匹配字符串，以修改待处理字符串

对应的匹配字符串，在需要将相同的待处理文本内容替换为不同的匹配文本内容时，可以通过修改配置文件中的正则表达式，从而修改相应的匹配字符串。本实施例的另一种实施方式，接收对所述配置文件的匹配字符串的修改，修改对应的正则表达式，当业务人员不具备修改正则表达式的能力时，所述配置文件中确定了待处理字符串后，然后通过修改所述匹配字符串，从而修改相应的正则表达式，保证后续能够将待处理字符串替换为所需的匹配字符串，实现实时修改所述配置文件，从而提高对文件中文本内容的处理效率，满足不同业务对文件内容处理的需求。

本申请的一种实施例，所述将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容之后，还包括：

根据替换后的匹配文本内容，生成文本内容替换文件；

统计所述文本内容替换文件的正确率，根据所述正确率确定目标正则表达式；

根据所述目标正则表达式，重新生成所述文本内容替换文件。

本实施例中，在将文件中的待处理文本内容替换为匹配文本内容后，根据替换后的匹配文本内容，生成文本内容替换文件，由于机器处理的开始阶段，可能存在文本内容替换错误率较高的现象，为了进一步提高机器处理的正确率，在一次处理后，统计所述文本内容替换文件的内容处理的正确率，当所述正确率低于预设值时，根据所述正确率调整此次替换时所使用的正则表达式，从而确定目标正则表达式，然后格局所述目标正则表达式，对所述文件再执行一次信息处理操作，重新生成文本内容替换文件。例如，在测试阶段，设置多个不同正则表达式应用于对同一文件的相同文本内容的处理，然后统计所述不同正则表达式的处理结果的正确率，进而根据正确率进行排序，在实际应用中，优先使用正确率高的正则表达式对文件进行处理，若实际应用中，该正则表达式的正确率不满足要求时，再更换为其它正则表达式，以提高对文件内容处理的正确率。

本申请的一种实施例，所述根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中满足所述正则表达式要求的待处理字符串之前，包括：

获取所述文件的属性信息，根据预设规则匹配出与所述文件的属性信息对应的正则表达式。

在实际应用中，不同文件随着业务发展的不同阶段，文件所需要处理的内容不尽相同，本实施例中，通过不同的正则表达式实现对不同阶段的文件进行不同的处理，具体的，在根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中满足所述正则表达式要求的待处理字符串之前，获取所述文件的属性信息，然后根据预设规则匹配出与所述文件的属性信息对应的正则表达式，从而确定不同属性信息的文件此时所需的正则表达式，在一种实施方式中，所述属性信息包括文件的创建时间，然后判断当前时间与所述文件创建时间之间的时长间隔，获取该时长间隔对应的正则表达式，然后再根据不同的正则表达式对不同创建时间的文件进行相应的处理。例如，若文件的创建时间与当前时间的间隔时长大于 6 个月，匹配的正则表达式 A，基于正则表达式 A 将文件中的文本内容 X 及 Y 作掩码处理，而当所述时间间隔小于 6 个月，匹配的正则表达式 B，基于正则表达式 B 将文件中的文本内容 Y 作掩码处理，对于文本内容 X 不作掩码处理。本实施例的一种应用场景，对于申请专利的交底材料，判断该交底材料的创建时间是否小于 12 个月，若是则屏蔽技术信息和发明人信息；若该交底材料的创建时间大于 12 个月后，该交底材料的专利文献一般已公开，不需要屏蔽该交底材料的技术信息，只屏蔽发明人信息。

本申请的一种实施例，所述根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中满足所述正则表达式要求的待处理字符串之前，包括：

判断所述文件的文件格式是否为预置格式，若否，将所述文件格式转换为预置格式；

获取所述预置格式对应的正则表达式。

在实际中，待处理的文件包括多种文件格式，如.doc 格式、.txt 格式、.pdf 格式，为了减少对不同格式的文件开发不同的处理逻辑代码而带来二额外的工作量，本实施例中，在确定正则表达式时，判断所述文件的文件格式是否为预置格式，若否，将所述文件格式转换为预置格式，预置

格式的文件则有相应的正则表达式进行匹配，然后获取所述预置格式对应的正则表达式，从而实现对不同文件格式的文件进行处理，提高对不同文件格式的文件的处理效率。

如图 3 所示，在另一种实施例中，本申请提供了一种信息处理装置，
5 包括：

格式化模块 10：用于获取包含文本内容的文件，将所述文件的文本内容格式化为字符串数组；

匹配模块 20：用于根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中满足所述正则表达式要求的待处理字符串；

10 确定模块 30：用于根据所述正则表达式对所述待处理字符串包含的字符进行替换，生成所述待处理字符串对应的匹配字符串，将所述匹配字符串进行逆格式化处理，确定所述匹配字符串对应的匹配文本内容；

15 替换模块 40：用于获取所述文件中所述待处理字符串对应的待处理文本内容，将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容。

本申请的一种实施例，所述替换模块 40 还包括执行；

获取所述待处理文本内容的原始文本格式，将所述匹配文本内容的文本格式设置为所述原始文本格式。

本申请的一种实施例，所述装置还包括：

20 配置模块：用于将所述正则表达式、待处理字符串及匹配字符串关联保存至配置文件；接收对所述配置文件的正则表达式的修改，修改对应的匹配字符串，或接收对所述配置文件的匹配字符串的修改，修改对应的正则表达式。

本申请的一种实施例，所述装置还包括：

25 调整模块：用于根据替换后的匹配文本内容，生成文本内容替换文件；统计所述文本内容替换文件的正确率，根据所述正确率确定目标正则表达式；根据所述目标正则表达式，重新生成所述文本内容替换文件。

本申请的一种实施例，所述匹配模块 20 还包括执行：

获取所述文件的属性信息，根据预设规则匹配出与所述文件的属性信

息对应的正则表达式。

本申请的一种实施例，所述匹配模块 20 还包括执行：

判断所述文件的文件格式是否为预置格式，若否，将所述文件格式转换为预置格式；获取所述预置格式对应的正则表达式。

- 5 在另一种实施例中，本申请提供了一种存储有计算机可读指令的非易失存储介质，计算机可读存储介质上存储有计算机程序，该计算机程序被处理器执行时实现上述实施例所述的信息处理方法。其中，所述计算机可读存储介质包括但不限于任何类型的盘（包括软盘、硬盘、光盘、CD-ROM、和磁光盘）、ROM（Read-Only Memory，只读存储器）、RAM（Random
10 AcceSS Memory，随即存储器）、EPROM（EraSable Programmable Read-Only Memory，可擦写可编程只读存储器）、EEPROM（Electrically EraSable Programmable Read-Only Memory，电可擦可编程只读存储器）、闪存、磁性卡片或光线卡片。也就是，存储设备包括由设备（例如，计算机、手机）以能够读的形式存储或传输信息的任何介质，可以是只读存储器，磁盘或
15 光盘等。

- 本申请实施例提供的一种存储有计算机可读指令的非易失存储介质，可实现获取包含文本内容的文件，将所述文件的文本内容格式化为字符串数组；根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中满足所述正则表达式要求的待处理字符串；根据所述正则表达式对所述
20 待处理字符串包含的字符进行替换，生成所述待处理字符串对应的匹配字符串，将所述匹配字符串进行逆格式化处理，确定所述匹配字符串对应的匹配文本内容；获取所述文件中所述待处理字符串对应的待处理文本内容，将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容。通过提供一种对文件中指定的文本内容进行处理的方法，在获取到文件
25 后，将文件的文本内容进行处理，将文件的文本内容格式化为一个一个的字符串，从而将文件的文本内容看作由一个个字符串组成的字符串数组，然后获取处理该文件文本内容的正则表达式，筛选所述字符串数组中满足正则表达式要求的待处理字符串，根据所述正则表达式对所述待处理字符串包含的字符进行替换，生成所述待处理字符串对应的匹配字符串，将所

述匹配字符串进行逆格式化处理，确定所述匹配字符串对应的匹配文本内容，在确定了待处理字符串跟各待处理字符串对应的匹配字符串后，获取所述文件中所述待处理字符串对应的待处理文本内容，将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容，完成对文件中的待处理
5 文本内容的转换，实现对文件中重要或敏感信息的屏蔽、修改或掩码处理，批量化的文件内容处理节省了时间，提高对文件中文本内容的处理效率。

本申请实施例提供的存储有计算机可读指令的非易失存储介质可以实现上述信息处理方法的实施例，具体功能实现请参见方法实施例中的说明，在此不再赘述。

10 此外，在又一种实施例中，本申请还提供一种终端设备，如图 4 所示，所述终端设备包括处理器 403、存储器 405、输入单元 407 以及显示单元 409 等器件。本领域技术人员可以理解，图 4 示出的结构器件并不构成对所有终端设备的限定，可以包括比图示更多或更少的部件，或者组合某些部件。所述存储器 405 可用于存储计算机程序 401 以及各功能模块，所述
15 处理器 403 运行存储在存储器 405 的计算机程序 401，从而执行设备的各种功能应用以及数据处理。所述存储器 405 可以是内存储器或外存储器，或者包括内存储器 and 外存储器两者。内存储器可以包括只读存储器 (ROM)、可编程 ROM(PROM)、电可编程 ROM(EPROM)、电可擦写可编程 ROM(EEPROM)、快闪存储器、或者随机存储器。外存储器可以包括硬
20 盘、软盘、ZIP 盘、U 盘、磁带等。本申请所公开的存储器包括但不限于这些类型的存储器。本申请所公开的存储器 405 只作为例子而非作为限定。

输入单元 407 用于接收信号的输入及接收用户的输入，输入单元 407 可包括触控面板以及其它输入设备，触控面板可收集用户在其上或附近的
25 触摸操作（比如用户使用手指、触笔等任何适合的物体或附件在触控面板上或在触控面板附近的操作），并根据预先设定的程序驱动相应的连接装置；其它输入设备可以包括但不限于物理键盘、功能键（比如播放控制按键、开关按键等）、轨迹球、鼠标、操作杆中的一种或多种。显示单元 409 可用于显示用户输入的信息或提供给用户的信息以及计算机设备的各

种菜单。显示单元 409 可采用液晶显示器、有机发光二极管等形式。处理器 403 是计算机设备的控制中心，利用各种接口和线路连接整个电脑的各个部分，通过运行或执行存储在存储器 403 内的软件程序和/或模块，以及调用存储在存储器内的数据，执行各种功能和处理数据。

- 5 在一种实施方式中，所述终端设备包括一个或多个处理器 403，以及一个或多个存储器 405，一个或多个计算机程序 401，其中所述一个或多个计算机程序 401 被存储在存储器 405 中并被配置为由所述一个或多个处理器 403 执行，所述一个或多个计算机程序 401 配置用于执行以上实施例所述的文件文本内容处理方法。图 4 中所示的一个或多个处理器 403 能够
10 执行、实现图 3 中所示的格式化模块 10、匹配模块 20、确定模块 30、替换模块 40 的功能。

- 本申请实施例提供的一种终端设备，可实现获取包含文本内容的文件，将所述文件的文本内容格式化为字符串数组；根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中满足所述正则表达式要求的
15 待处理字符串；根据所述正则表达式对所述待处理字符串包含的字符进行替换，生成所述待处理字符串对应的匹配字符串，将所述匹配字符串进行逆格式化处
理，确定所述匹配字符串对应的匹配文本内容；获取所述文件中所述待处理字符串对应的待处理文本内容，将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容。通过提供一种对文件中指定的文
20 本内容进行处理的方法，在获取到文件后，将文件的文本内容进行处理，将文件的文本内容格式化为一个一个个的字符串，从而将文件的文本内容看作由一个个字符串组成的字符串数组，然后获取处理该文件文本内容的正则表达式，筛选所述字符串数组中满足正则表达式要求的待处理字符串，根据所述正则表达式对所述待处理字符串包含的字符进行替换，生成所述
25 待处理字符串对应的匹配字符串，将所述匹配字符串进行逆格式化处
理，确定所述匹配字符串对应的匹配文本内容，在确定了待处理字符串跟各待处理字符串对应的匹配字符串后，获取所述文件中所述待处理字符串对应的待处理文本内容，将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容，完成对文件中的待处理文本内容的转换，实现对文件中重

要或敏感信息的屏蔽、修改或掩码处理，批量化的文件内容处理节省了时间，提高对文件中文本内容的处理效率。

本申请实施例提供的终端设备可以实现上述提供的信息处理方法的实施例，具体功能实现请参见方法实施例中的说明，在此不再赘述。

- 5 以上所述仅是本申请的部分实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本申请原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本申请的保护范围。

权 利 要 求 书

1.一种信息处理方法，包括：

5 获取包含文本内容的文件，将所述文件的文本内容格式化为字符串数组；

根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中满足所述正则表达式要求的待处理字符串；

10 根据所述正则表达式对所述待处理字符串包含的字符进行替换，生成所述待处理字符串对应的匹配字符串，将所述匹配字符串进行逆格式化处
理，确定所述匹配字符串对应的匹配文本内容；

获取所述文件中所述待处理字符串对应的待处理文本内容，将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容。

15 2.根据权利要求 1 所述的方法，所述获取所述文件中所述待处理字符串对应的待处理文本内容，将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容之后，所述方法还包括：

获取所述待处理文本内容的原始文本格式，将所述匹配文本内容的文本格式设置为所述原始文本格式。

20 3.根据权利要求 1 所述的方法，所述根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中满足所述正则表达式要求的待处理字符串之后，还包括：

将所述正则表达式、待处理字符串及匹配字符串关联保存至配置文件；

接收对所述配置文件的正则表达式的修改，修改对应的匹配字符串，或，

25 接收对所述配置文件的匹配字符串的修改，修改对应的正则表达式。

4.根据权利要求 1 所述的方法，所述将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容之后，还包括：

根据替换后的匹配文本内容，生成文本内容替换文件；

统计所述文本内容替换文件的正确率，根据所述正确率确定目标正则

表达式；

根据所述目标正则表达式，重新生成所述文本内容替换文件。

5 根据权利要求 1 所述的方法，所述根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中满足所述正则表达式要求的待处理字符串之前，包括：

获取所述文件的属性信息，根据预设规则匹配出与所述文件的属性信息对应的正则表达式。

10 6.根据权利要求 1 所述的方法，所述根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中满足所述正则表达式要求的待处理字符串之前，包括：

判断所述文件的文件格式是否为预置格式，若否，将所述文件格式转换为预置格式；

获取所述预置格式对应的正则表达式。

7.一种信息处理装置，包括：

15 格式化模块：用于获取包含文本内容的文件，将所述文件的文本内容格式化为字符串数组；

匹配模块：用于根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中满足所述正则表达式要求的待处理字符串；

20 确定模块：用于根据所述正则表达式对所述待处理字符串包含的字符进行替换，生成所述待处理字符串对应的匹配字符串，将所述匹配字符串进行逆格式化处理，确定所述匹配字符串对应的匹配文本内容；

替换模块：用于获取所述文件中所述待处理字符串对应的待处理文本内容，将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容。

25 8.一种存储有计算机可读指令的非易失性存储介质，所述计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行一种信息处理方法，所述信息处理方法包括以下步骤：

获取包含文本内容的文件，将所述文件的文本内容格式化为字符串数组；

根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中满足所述正则表达式要求的待处理字符串；

根据所述正则表达式对所述待处理字符串包含的字符进行替换，生成所述待处理字符串对应的匹配字符串，将所述匹配字符串进行逆格式化处
5 理，确定所述匹配字符串对应的匹配文本内容；

获取所述文件中所述待处理字符串对应的待处理文本内容，将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容。

9.根据权利要求 8 所述的非易失性存储介质，所述获取所述文件中所述待处理字符串对应的待处理文本内容，将所述待处理文本内容替换为所
10 述匹配字符串对应的匹配文本内容之后，所述方法还包括：

获取所述待处理文本内容的原始文本格式，将所述匹配文本内容的文本格式设置为所述原始文本格式。

10.根据权利要求 8 所述的非易失性存储介质，所述根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中满足所述正则表达式
15 要求的待处理字符串之后，还包括：

将所述正则表达式、待处理字符串及匹配字符串关联保存至配置文件；

接收对所述配置文件的正则表达式的修改，修改对应的匹配字符串，
或，

20 接收对所述配置文件的匹配字符串的修改，修改对应的正则表达式。

11.根据权利要求 8 所述的非易失性存储介质，所述将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容之后，还包括：

根据替换后的匹配文本内容，生成文本内容替换文件；

统计所述文本内容替换文件的正确率，根据所述正确率确定目标正则
25 表达式；

根据所述目标正则表达式，重新生成所述文本内容替换文件。

12.根据权利要求 8 所述的非易失性存储介质，所述根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中满足所述正则表达式要求的待处理字符串之前，包括：

获取所述文件的属性信息，根据预设规则匹配出与所述文件的属性信息对应的正则表达式。

13.根据权利要求 8 所述的非易失性存储介质，所述根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中满足所述正则表达式要求的待处理字符串之前，包括：

判断所述文件的文件格式是否为预置格式，若否，将所述文件格式转换为预置格式；

获取所述预置格式对应的正则表达式。

14.一种终端设备，包括：

10 一个或多个处理器；

存储器；

一个或多个计算机程序，其中所述一个或多个计算机程序被存储在所述存储器中并被配置为由所述一个或多个处理器执行，所述一个或多个计算机程序配置用于执行一种信息处理方法，所述信息处理方法包括以下步骤：

15 获取包含文本内容的文件，将所述文件的文本内容格式化为字符串数组；

根据正则表达式对所述字符串数组进行筛选，获取所述字符串数组中满足所述正则表达式要求的待处理字符串；

20 根据所述正则表达式对所述待处理字符串包含的字符进行替换，生成所述待处理字符串对应的匹配字符串，将所述匹配字符串进行逆格式化处

理，确定所述匹配字符串对应的匹配文本内容；

25 15.根据权利要求 14 所述的终端设备，所述获取所述文件中所述待处理字符串对应的待处理文本内容，将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容之后，所述方法还包括：

获取所述待处理文本内容的原始文本格式，将所述匹配文本内容的文本格式设置为所述原始文本格式。

16.根据权利要求 14 所述的终端设备,所述根据正则表达式对所述字符串数组进行筛选,获取所述字符串数组中满足所述正则表达式要求的待处理字符串之后,还包括:

5 将所述正则表达式、待处理字符串及匹配字符串关联保存至配置文件;

接收对所述配置文件的正则表达式的修改,修改对应的匹配字符串,或,

接收对所述配置文件的匹配字符串的修改,修改对应的正则表达式。

10 17.根据权利要求 14 所述的终端设备,所述将所述待处理文本内容替换为所述匹配字符串对应的匹配文本内容之后,还包括:

根据替换后的匹配文本内容,生成文本内容替换文件;

统计所述文本内容替换文件的正确率,根据所述正确率确定目标正则表达式;

根据所述目标正则表达式,重新生成所述文本内容替换文件。

15 18.根据权利要求 14 所述的终端设备,所述根据正则表达式对所述字符串数组进行筛选,获取所述字符串数组中满足所述正则表达式要求的待处理字符串之前,包括:

获取所述文件的属性信息,根据预设规则匹配出与所述文件的属性信息对应的正则表达式。

20 19.根据权利要求 14 所述的终端设备,所述根据正则表达式对所述字符串数组进行筛选,获取所述字符串数组中满足所述正则表达式要求的待处理字符串之前,包括:

判断所述文件的文件格式是否为预置格式,若否,将所述文件格式转换为预置格式;

25 获取所述预置格式对应的正则表达式。

20.根据权利要求 14 所述的终端设备,所述根据正则表达式对所述字符串数组进行筛选,获取所述字符串数组中满足所述正则表达式要求的待处理字符串之前,包括:

判断所述文件的文件格式是否为预置格式,若否,将所述文件格式转

换为预置格式；

获取所述预置格式对应的正则表达式。

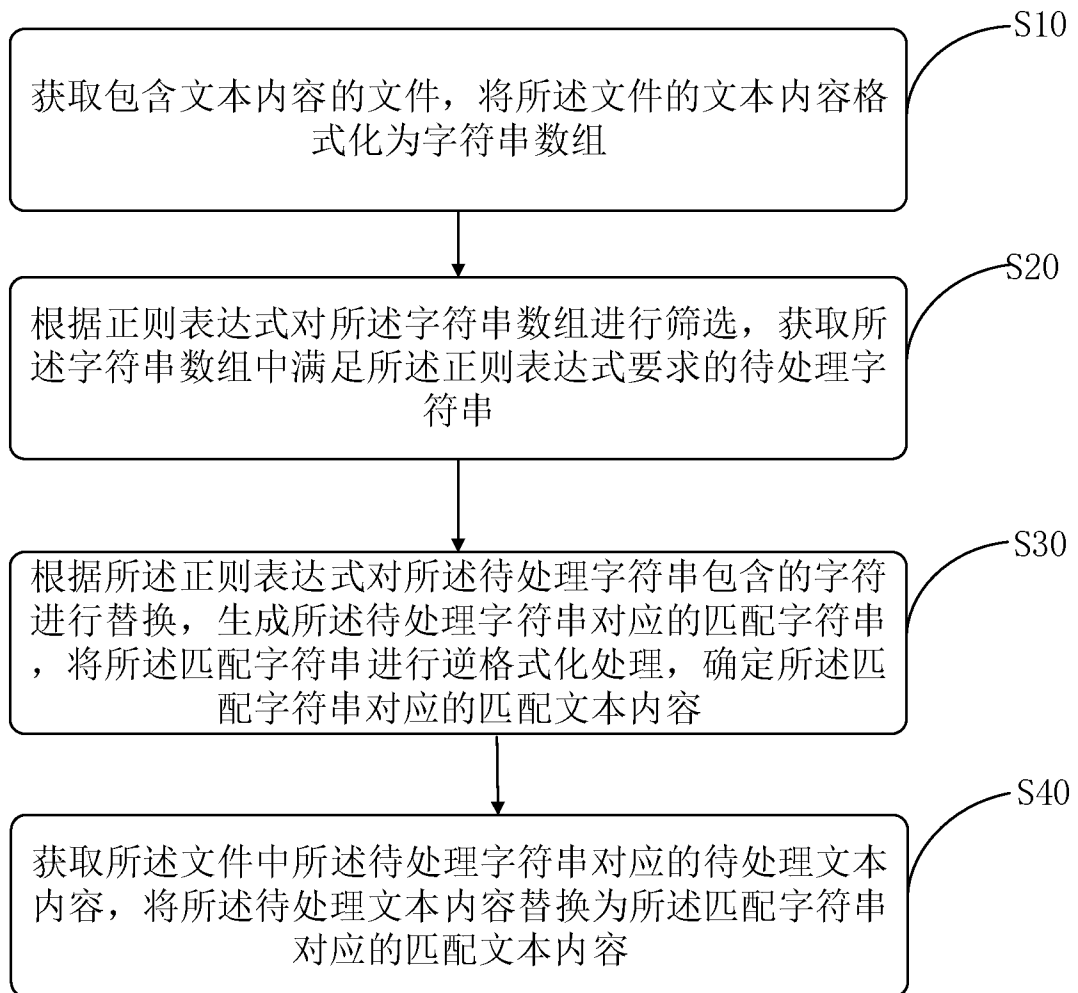


图 1

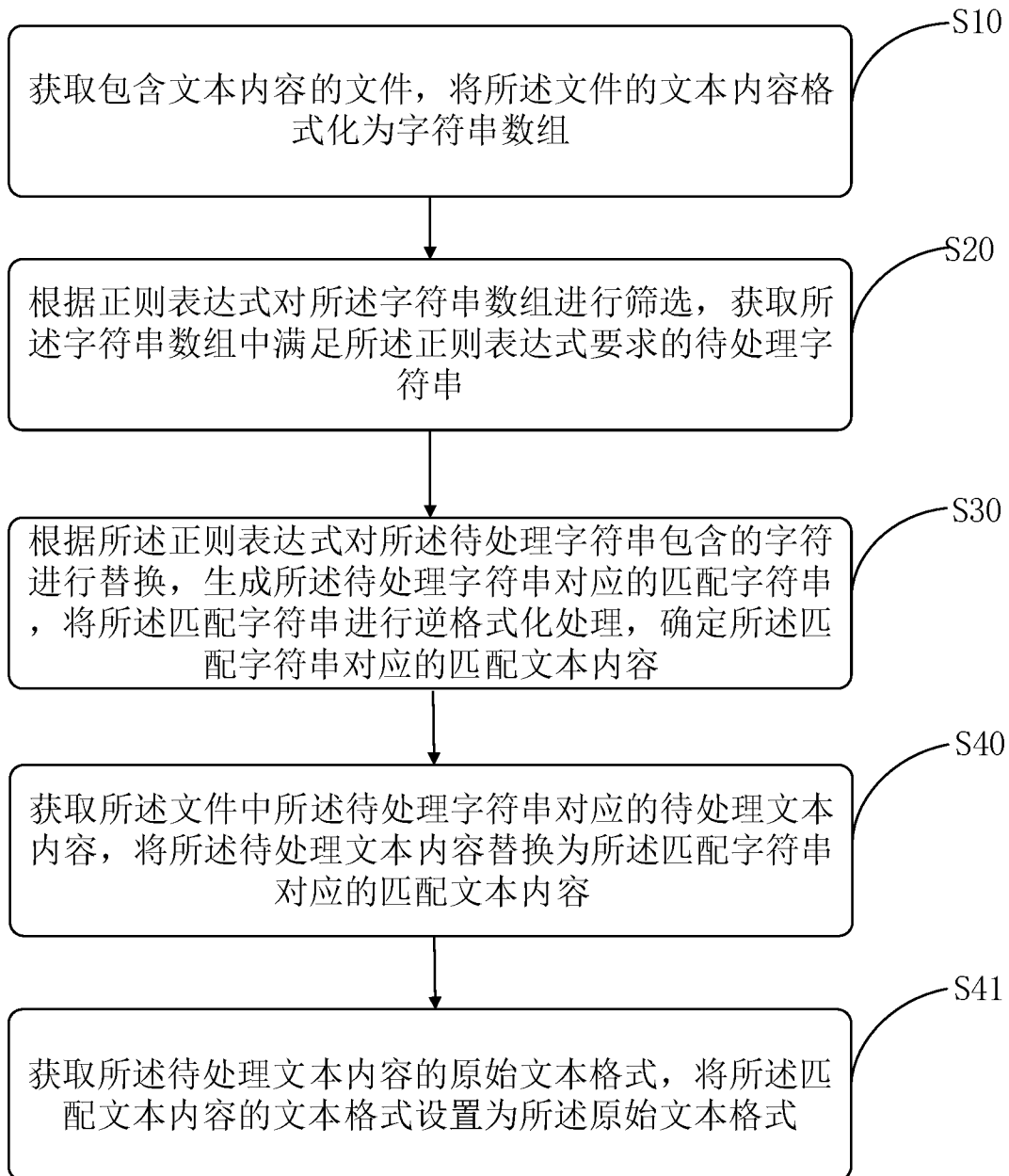


图 2

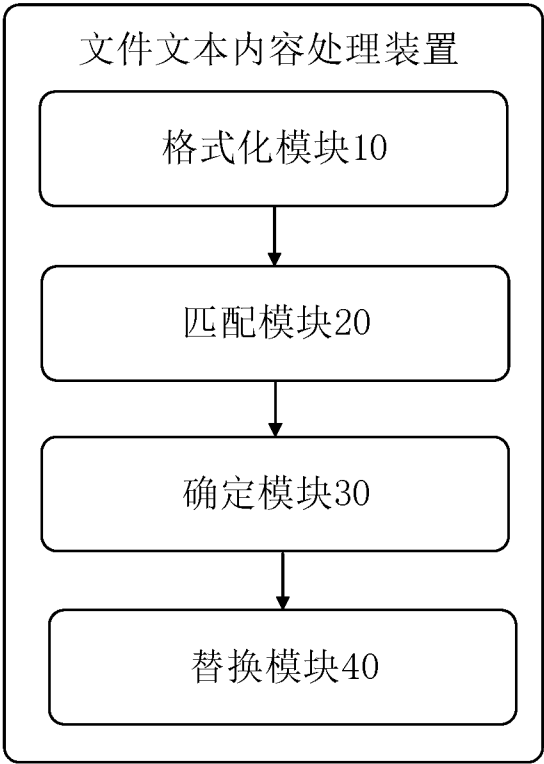


图 3

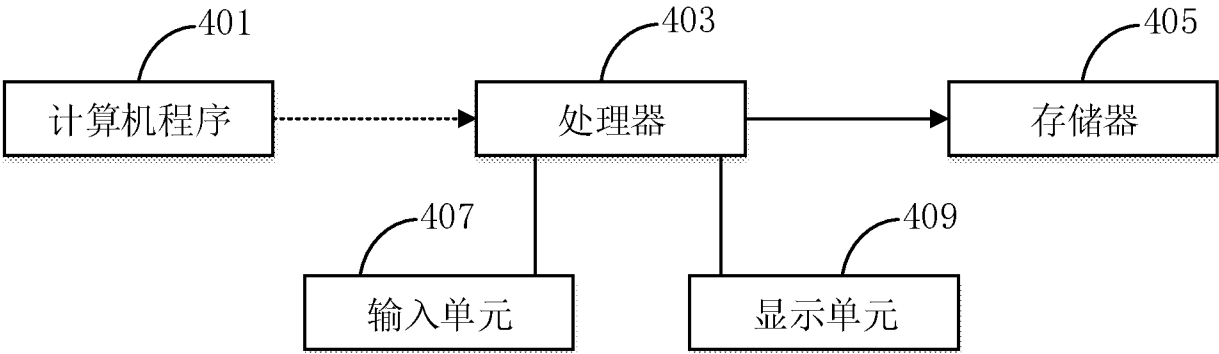


图 4

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/103028

A. CLASSIFICATION OF SUBJECT MATTER

G06F 21/62(2013.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI, EPODOC, CNPAT, CNKI, IEEE: 文本, 字符串, 筛选, 正则表达式, 匹配, 关联, 替换, 敏感, 安全, 加密, text, string, filter, regular w expression, match, associate, replace, sensitive, secure, encrypted

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 109829328 A (SHANGHAI ZAMPLUS RONGXUAN TECHNOLOGY CO., LTD.) 31 May 2019 (2019-05-31) description, paragraphs [0028]-[0060], [0072] and [0078]	1-20
A	CN 109376547 A (BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 22 February 2019 (2019-02-22) entire document	1-20
A	CN 107329957 A (NETEASE (HANGZHOU) NETWORK CO., LTD.) 07 November 2017 (2017-11-07) entire document	1-20
A	CN 105701074 A (BEIJING JINGDONG SHANGKE INFORMATION TECHNOLOGY CO., LTD. et al.) 22 June 2016 (2016-06-22) entire document	1-20
A	US 2006265357 A1 (POTTS, Matthew P.) 23 November 2006 (2006-11-23) entire document	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

29 February 2020

Date of mailing of the international search report

26 March 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/103028

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	109829328	A	31 May 2019	None	
CN	109376547	A	22 February 2019	None	
CN	107329957	A	07 November 2017	None	
CN	105701074	A	22 June 2016	None	
US	2006265357	A1	23 November 2006	None	

国际检索报告

国际申请号

PCT/CN2019/103028

A. 主题的分类

G06F 21/62 (2013.01) i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

WPI, EPDOC, CNPAT, CNKI, IEEE: 文本, 字符串, 筛选, 正则表达式, 匹配, 关联, 替换, 敏感, 安全, 加密, text, string, filter, regular w expression, match, associate, replace, sensitive, secure, encrypted

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 109829328 A (上海晶赞融宣科技有限公司) 2019年 5月 31日 (2019 - 05 - 31) 说明书[0028]-[0060], [0072], [0078]段	1-20
A	CN 109376547 A (北京邮电大学) 2019年 2月 22日 (2019 - 02 - 22) 全文	1-20
A	CN 107329957 A (网易杭州网络有限公司) 2017年 11月 7日 (2017 - 11 - 07) 全文	1-20
A	CN 105701074 A (北京京东尚科信息技术有限公司 等) 2016年 6月 22日 (2016 - 06 - 22) 全文	1-20
A	US 2006265357 A1 (POTTS, Matthew P.) 2006年 11月 23日 (2006 - 11 - 23) 全文	1-20

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2020年 2月 29日

国际检索报告邮寄日期

2020年 3月 26日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

马鑫

电话号码 86-(10)-53961354

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/103028

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	109829328	A	2019年 5月 31日	无	
CN	109376547	A	2019年 2月 22日	无	
CN	107329957	A	2017年 11月 7日	无	
CN	105701074	A	2016年 6月 22日	无	
US	2006265357	A1	2006年 11月 23日	无	