(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
2019 年 9 月 26 日 (26.09.2019)

WIPO | PCT

(10) 国际公布号
WO 2019/179012 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01)

(21) 国际申请号: PCT/CN2018/100930

(22) 国际申请日: 2018 年 8 月 17 日 (17.08.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201810241226.9 2018年3月22日 (22.03.2018) 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: 张雨嘉 (ZHANG, Yujia); 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)。
倪振 (NI, Zhen); 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: 深圳市精英专利事务所 (SHENZHEN TALENT PATENT SERVICE); 中国广东省深圳市福田区深南中路6009号绿景广场B栋20层B, Guangdong 518000 (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,

WO 2019/179012 A1

(54) **Title:** METHOD, DEVICE, APPARATUS AND COMPUTER READABLE STORAGE MEDIUM FOR PROCESSING TEXT DATA

(54) **发明名称:** 文本数据处理方法、装置、设备及计算机可读存储介质

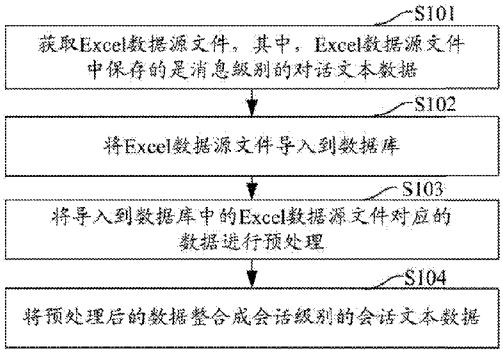

图 1

S101 OBTAIN AN EXCEL DATA SOURCE FILE, WHEREIN THE EXCEL DATA SOURCE FILE STORES MESSAGE-LEVEL DIALOGUE TEXT DATA

S102 IMPORT THE EXCEL DATA SOURCE FILE INTO A DATABASE

S103 PRE-PROCESS DATA CORRESPONDING TO THE EXCEL DATA SOURCE FILE IMPORTED IN THE DATABASE

S104 INTEGRATE THE PRE-PROCESSED DATA INTO CONVERSATION-LEVEL CONVERSATION TEXT DATA

(57) **Abstract:** Embodiments of the present application provide a method, a device, an apparatus, and a computer readable storage medium for processing text data. The method comprises: obtaining an Excel data source file, wherein the Excel data source file stores message-level dialogue text data; importing the Excel data source file into a database; pre-processing data corresponding to the Excel data source file imported in the database; and integrating the pre-processed data into conversation-level conversation text data.

[见续页]

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, 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)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要：本申请实施例提供一种文本数据处理方法、装置、设备及计算机可读存储介质。所述方法包括：获取Excel数据源文件，其中，所述Excel数据源文件中保存的是消息级别的对话文本数据；将所述Excel数据源文件导入到数据库；将导入到数据库中的所述Excel数据源文件对应的数据进行预处理；将预处理后的数据整合成会话级别的会话文本数据。

发明名称：文本数据处理方法、装置、设备及计算机可读存储介质

本申请要求于 2018 年 3 月 22 日提交中国专利局、申请号为 201810241226.9、发明名称为“文本数据处理方法、装置、设备及计算机可读存储介质”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技术领域，尤其涉及一种文本数据处理方法、装置、设备及计算机可读存储介质。

背景技术

在坐席销售的过程中，可能会与客户产生大量的对话文本数据，这些对话文本数据会保存在坐席销售的平台中。若要想对对话文本数据进行分析，如分析坐席是否有说错、是否有不礼貌行为或者通过坐席说错的地方来查看坐席哪一方面的知识比较欠缺以更好地为坐席制定培训计划等，目前采用的方法是先随机抽取一定条数的消息文本内容，再通过人工的方法进行分析。而由于对话文本数据量较大，因此分析人员拿到的对话文本数量大，且为消息级别，散乱无序、无上下文关系、无人员关系等，给分析工作造成很大的不便。

发明内容

本申请实施例提供一种文本数据处理方法、装置、设备及计算机可读存储介质，可将散乱无序、无上下文关系、无人员关系的消息级别的对话文本数据整合为以预定格式显示的会话级别的会话文本数据，整合的效率高，且方便分析人员进行更一步的分析。

第一方面，本申请实施例提供了一种文本数据处理方法，该方法包括：

获取 Excel 数据源文件，其中，所述 Excel 数据源文件中保存的是消息级别的对话文本数据；

将所述 Excel 数据源文件导入到数据库；

将导入到数据库中所述 Excel 数据源文件对应的数据进行预处理；

将预处理后的数据整合成会话级别的会话文本数据。

第二方面，本申请实施例提供了一种文本数据处理装置，该装置包括用于执行上述第一方面所述的文本数据处理方法的单元。

第三方面，本申请实施例提供了一种计算机设备，所述计算机设备包括存储器，以及与所述存储器相连的处理器；

所述存储器用于存储实现文本数据处理的计算机程序，所述处理器用于运行所述存储器中存储的计算机程序，以执行上述第一方面所述的文本数据处理的方法。

第四方面，本申请实施例提供了一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序包括程序指令，所述程序指令被处理器执行时，实现上述第一方面所述的文本数据处理的方法。

本申请实施例可将散乱无序、无上下文关系、无人员关系的消息级别的对话文本数据整合为以预定格式显示的会话级别的会话文本数据，整合的效率高，且方便分析人员进行更一步的分析。

附图说明

为了更清楚地说明本申请实施例技术方案，下面将对实施例描述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。

图 1 是本申请实施例提供的一种文本数据处理方法的流程示意图；

图 2 是本申请实施例提供的一种文本数据处理方法的子流程示意图；

图 3 是本申请实施例提供的一种文本数据处理方法的另一子流程示意图；

图 4 是本申请实施例提供的一种文本数据处理方法的另一子流程示意图；

图 5 是本申请另一实施例提供的一种文本数据处理方法的流程示意图；

图 6 是本申请实施例提供的一种文本数据处理装置的示意性框图；

图 7 是本申请实施例提供的预处理单元的示意性框图；

图 8 是本申请实施例提供的整合单元的示意性框图；

图 9 是本申请实施例提供的导入单元的示意性框图；

图 10 是本申请另一实施例提供的文本数据处理装置的示意性框图；

图 11 是本申请实施例提供的一种计算机设备的示意性框图。

具体实施方式

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、完整地描述，显然，所描述的实施例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。

应当理解，当在本说明书和所附权利要求书中使用时，术语“包括”和“包含”指示所描述特征、整体、步骤、操作、元素和/或组件的存在，但并不排除一个或多个其它特征、整体、步骤、操作、元素、组件和/或其集合的存在或添加。

图 1 为本申请实施例提供的一种文本数据处理方法的流程示意图。该方法包括以下步骤 S101-S104。

S101，获取 Excel 数据源文件，其中，Excel 数据源文件中保存的是消息级别的对话文本数据。Excel 数据源文件中保存的是坐席与客户之间的对话文本数据。该对话文本数据从坐席销售的平台中获取，获取后保存在 Excel 文件中。该对话文本数据属于消息级别，可以理解为对话文本数据是以坐席与客户之间发送的消息为单位保存的数据，该对话文本数据由众多的消息文本数据组成，每一条消息文本数据包括发送人、接收人、具体的消息内容、发送消息的时间等。由于可能同时有多个坐席同时跟不同的客户有交流，如此会导致 Excel 数据源中保存的数据散乱无序、无上下文关系、无人员关系等。对于同一个坐席与同一个客户之间的消息文本数据，也可能也会出现重复保存，散乱无序等情况。

S102，将 Excel 数据源文件导入到数据库。其中，数据库可以是 My Sql 数据库、Sql Server 数据库、Oracle 数据库等可以处理大量数据的数据库之中的一种。由于每天会产生大量的 Excel 源文件数据，需要及时将该 Excel 源文件数据

保存到数据库，保存到数据库中的 Excel 数据源文件中的数据可供后续使用。在将 Excel 数据源文件导入到数据库之前，建立与数据库之间的连接，并在结束导入之前保持与数据库之间的连接。

将 Excel 数据源文件导入到数据库时需使用开放源码函式库。其中，开放源码函式库指的是 Apache POI（Poor Obfuscation Implementation），其是用 Java 编写的免费开源的跨平台的 Java API，它允许程序员使用 Java 程序创建，修改和显示 MS Office 文件。可以理解地，Apache POI 提供对 MIcrosoft Office 格式文件读和写的功能。Apache POI 也可以简称为 POI。具体地，将 Excel 数据源文件导入到数据库，包括：通过 POI 读取 Excel 数据源文件，并将读取的 Excel 数据源文件保存到数据库。

将 Excel 数据源文件导入到数据库中之后，保存的仍然是消息级别的对话文本数据，该对话文本数据由众多的消息文本数据组成，每一条消息文本数据包括发送人、接收人、具体的消息内容、发送消息的时间等。将 Excel 数据源文件导入到数据库后，可以显示如表 1 所示的形式。需要注意的是，表 1 所示只是一个示例。从表 1 中可以看到，保存的对话文本数据共有 12 条消息文本数据，每条消息文本数据中包括消息编号、发送人、接收人、发送消息的时间、具体的消息内容。也可以看出保存的对话文本数据是杂乱无序的，如发送时间在后面的消息文本数据却显示在发送时间在前面的消息文本数据之前，如消息内容“希望赵大哥抓住机会，是个好产品”发送时间在消息内容“公司推出一款新产品，有兴趣看看？”之后，却显示在消息内容“公司推出一款新产品，有兴趣看看？”前面。

消息编号	发送人	接收人	发送消息的时间	消息内容
1	王五	赵六	2017-01-01 10:01:01	赵大哥，你好
2	赵六	王五	2017-01-01 10:01:02	早上好
3	王五	赵六	2017-01-01 10:01:04	希望赵大哥抓住机会，是个好产品
4	王五	赵六	2017-01-01 10:01:03	公司推出一款新产品，有兴趣看看？
5	赵六	王五	2017-01-01 10:01:06	发来看看
6	赵六	王五	2017-01-01 10:01:05	好的
7	张三	李四	2017-01-01 12:01:02	李老师，在吗？
8	李四	张三	2017-01-01 12:01:03	在的，小张。

7	李四	张三	2017-01-01 12:01:07	行，没有问题。
10	张三	李四	2017-01-01 12:01:06	好的，有需要和我说。
11	张三	李四	2017-01-01 12:01:04	上次我和你提及的产品，是否有关注？
12	李四	张三	2017-01-01 12:01:05	我还在观望，不过还不确定。

表 1 保存到数据库中的对话文本数据

S103，将导入到数据库中的数据进行预处理。预处理的方法包括去重、筛选等，以去掉重复的无价值的数据，以及筛选出符合分析需求的数据，避免其他的数据对进一步分析造成影响，如降低分析的效率等。

具体地，如图 2 所示，将导入到数据库中的数据进行预处理，包括以下步骤 S201-S202。S201，将导入到数据库中的数据去重。由于保存到数据库中的数据是消息级别的对话文本数据，可能在保存的时候对某些消息文本数据进行了重复保存，那么需要对导入到数据库中的数据去重，以去掉重复的无价值的数据。S202，从去重后的数据中筛选出预设消息类型的消息文本数据。其中，筛选包括两种类型的筛选，一，将具体的消息内容为空的值删除；二，从众多的消息类型中筛选出符合预设消息类型的消息文本。其中，消息类型有很多，如位置、分享小程序、红包、加好友请求、文字、图片、语音、小视频、动图等消息类型。可以根据分析人员具体分析的意图来确定预设消息类型，预设消息类型由分析人员预先设定。如分析坐席是否有说错、是否有不礼貌行为或者通过坐席说错的地方来查看坐席哪一方面的知识比较欠缺以更好地为坐席制定培训计划等意图，预设消息类型可以包括：文字、图片、语音、小视频、动图等。可以理解地，文字、图片、语音、小视频、动图等预设消息类型存在分析的必要性。

S104，将预处理后的数据整合成会话级别的会话文本数据。会话级别的会话文本数据理解为以坐席与客户之间的一个对话（会话）为单位保存的数据，即会话文本数据中保存的坐席与客户之间的多个对话数据。每个对话数据中对应有多条消息文本数据，该多条消息文本数据对应的发送消息的时间是连续的，连续指前后两条消息文本数据的发送消息的时间间隔不超过一定时间，如 5 分钟等。可以理解地，同一个坐席和同一个客户前一天和后一天的对话，属于两个不同的对话。将预处理后的数据整合成会话级别的会话文本数据，可以根据

发送人、接收人来作为一个分组单元，将预处理后的数据分成多组，这意味发送人和接收人这两者相同的数据分为一组，根据分组将同一组中的消息文本数据按照消息发送的时间做正向排序，再按照分组将对应组中的排序后的消息文本数据进行显示。

具体地，如图 3 所示，将预处理后的数据整合成会话级别的会话文本数据，包括以下步骤 S301-S303。S301，从预处理后的数据中查找每条消息文本数据中的发送人和接收人，将发送人和接收人作为一个集合。S302、按照集合对消息文本数据进行分组。具体地，将集合相同的消息文本数据分成一组，如此就分成了多组的数据。这意味着分成一组的发送人和接收人是同一个对话中的两个人，不同发送人和接收人的对话分成了不同的组。如表 1 中张三和李四在一组，王五和赵六在一组。S303、将每组中的消息文本数据按照预定格式显示。其中，先将每组中的消息文本数据按照发送消息时间的先后顺序排序，将排序后的消息文本数据按照预定格式显示，其中预定格式可以为：发送消息的时间[空格]发送人[冒号]具体的消息内容。预定格式也可以为其他的格式。

整合后的会话文本可以显示为如表 2 所示的形式。需要注意的是，表 2 中所示仅仅是一个示例。从表 2 中可以看出，整合后的会话文本数据有 2 个对话数据，每个对话数据中包括对话编号、对话内容。每个对话内容中对应有多条消息文本数据，该多条消息文本数据是按照对应的发送消息的时间是顺序排列的。每条消息文本数据按照预定格式：发送消息的时间[空格]发送人[冒号]具体的消息内容，进行显示，如：2017-01-01 10:01:01 王五：赵大哥，你好。如此，把消息级别的对话文本数据整合成为会话级别的会话文本数据，整合后的会话文本一眼就可以看出是两个人之间的对话，方便分析人员查看和审阅文本内容。

对话编号	对话内容
1	2017-01-01 12:01:02 张三：李老师，在吗？ 2017-01-01 12:01:03 李四：在的，小张。 2017-01-01 12:01:04 张三：上次我和你提及的产品，是否有关注？ 2017-01-01 12:01:05 李四：我还在观望，不过还不确定。 2017-01-01 12:01:06 张三：好的，有需要和我说。 2017-01-01 12:01:07 李四：行，没有问题。

2	2017-01-01 10:01:01 王五：赵大哥，你好 2017-01-01 10:01:02 赵六：早上好 2017-01-01 10:01:03 王五：公司推出一款新产品，有兴趣看看？ 2017-01-01 10:01:04 王五：希望赵大哥抓住机会，是个好产品 2017-01-01 10:01:05 赵六：好的 2017-01-01 10:01:06 赵六：发来看看

表 2 整合后的会话文本

以上实施例可将散乱无序、无上下文关系、无人员关系的消息级别的对话文本数据整合为以预定格式显示的会话级别的会话文本数据，方便了分析人员查看和审阅文本内容，以进行更一步的分析。

在一实施例中，所述 Excel 数据源文件是压缩过的 XML 格式文件，需要注意的是，这里的 Excel 特指包含 Office 2007 及后续版本，Office 2007 及后续版本的单个 Sheet 可以支持 1048576 行数据，数据量非常大，因此为了方便保存和数据传输，将 Excel 数据源文件导入到数据库之前，需要将 Excel 文件进行处理转换为压缩过的 XML 格式文件。具体地，获取 Excel 数据源文件，包括：获取 Excel 文件，将获取的 Excel 文件进行处理转换为压缩过的 XML 格式文件，将压缩过的 XML 格式文件称为 Excel 数据源文件。

其中，将 Excel 文件进行处理转换为压缩过的 XML 格式文件，具体地，先完成 Excel 文件到 XML 的文件映射，即将 Excel 文件中的列映射成 XML 文件中的属性。如新建一个 XML 格式的文件，在 XML 格式的文件中，编辑写入两个以上的节点对，每个节点对中有与 Excel 文件中列数一致的子节点，子节点相当于 XML 文件中的属性。其中，在本实施例中，列数一致可以理解为：子节点的个数与 Excel 数据源文件中列数相同，子节点的名称可以与 Excel 数据源中的列名称相同，也可以不同。在 XML 格式的文件中，编辑写入两个以上的节点对，是为了避免 Excel 文件中的数据导入到 XML 文件中，只导出一列的数据。再打开 Excel 文件，点击开发工具--源--XML 映射，添加新建的 XML 文件，会出现 XML 文件的属性列表，将 XML 文件的每个属性拖到 Excel 对应列上，以完成映射。然后点击 XML 选项下的导出，以将 Excel 文件导出为 XML 文件。将导出的 XML 文件在进行压缩，以方便保存和数据传输等。

如图 4 所示，将 Excel 数据源文件导入到数据库，即步骤 S102，包括以下

步骤 S401-S403。

S401，利用开放源码函式库读取并解压 Excel 数据源文件以得到 XML 格式文件。具体地，利用 POI 读取经过压缩的 XML 文件；再将 XML 文件解压缩以得到 XML 格式文件。

S402，将所述 XML 格式文件解析成多行的数据。

譬如，下面 XML 数据对应一行 Excel 数据：

<note>

<to>George</to>

<from>John</from>

<heading>Reminder</heading>

<body>Don't forget the meeting!</body>

</note>

POI 先读取<note>，标示为行首节点，然后读取<to>，识别为一个子节点的开始，再读取 George，识别为子节点的值，最后读取</to>，标示为子节点的结束。因为 XML 规定 </> 中的 / 为节点结束符，<> 和 </> 中间为节点值，如果中间没有任何字符，则认为没有值。依次读取其他子节点，最后读取</note>，标示为行尾节点。这样一行数据解析完毕。如此可以解析出多行数据。具体地，可以通过开放源码函式库进行解析；也可以通过其他解析方式进行解析，如 SAX（Simple API for XML）等。

以上步骤 S401-S402 利用 POI 读取并解压数据源文件，其中数据源文件是经过压缩的 XML 文件，并将解压后的 XML 格式文件解析成多行的数据，比起直接读取 Excel 文件中的数据，速度要快很多。

S403，利用开放源码函式库，通过连接池将解析后的多行数据保存到数据库。

其中，连接池是与数据库连接的通道。通过连接池将 Java 和数据库连接，其中，Java 指的是使用 Java 语言编写的连接数据库的代码所在的设备。其中，连接池可使用 Hikari Java 数据库连接池。数据库连接池负责分配、管理、释放数据库连接，它保证应用程序可以重复使用同一个连接而不需要每次都建立数据库连接，如果数据库连接时间超过设置的最长数据库连接时间会自动释放链

接，为了避免因为没有释放链接而导致的数据库连接遗漏，因此，数据库连接池可以明显的提高数据库的连接性能。数据库连接池在初始化的时候会放入一定数量的连接，这个连接是由最小连接数决定的，就算没有用到这些连接，这个连接也会放在连接池中。如果连接数超过最大连接数，那么会放入队列中等待释放链接再使用。数据库连接是非常占用资源的，尤其是在高并发的情况下，如果每次都去建立数据库连接就会有性能问题，也会影响一个应用程序的延展性。数据库连接池避免了数据库连接频繁建立、关闭的开销，提高了数据库连接效率。使用 Hikari Java 数据库连接池，连接速度快，稳定性也非常好。利用开放源码函式库，通过连接池将解析后的多行数据保存到数据库。

以上步骤 S403 利用开放源码函式库，通过 Hikari Java 数据库连接池将解析后的数据保存到数据库，使用 Hikari Java 数据库连接池可以提高数据库连接的速度，如此也提高了保存数据的速度。

图 4 所示的实施例通过将 Excel 数据源文件解压得到 XML 格式文件，解析 XML 格式文件中的数据，并将解析后的数据，通过高效连接池保存到数据库，可以快速将大量的 Excel 数据源文件导入到数据库，提高了 Excel 数据源文件导入到数据库的效率。

图 5 是本申请另一实施例提供的一种文本数据处理方法的流程示意图。该方法包括步骤 S501-S506。其中，步骤 S501-S504 请参看图 1 实施例的所述的部分，在此不再赘述。下面将详细描述步骤 S505-S506。

S505，对会话级别的会话文本数据建立倒排索引。其中，可使用全文搜索引擎来实现，如 ElasticSearch 全文搜索引擎。具体地，将会话级别的会话文本数据进行分词；统计分成的词在对会话文本数据中出现的次数和位置，如统计词“分红”在会话文本数据中出现的次数和位置，其中，位置包括在哪个会话文本数据表、哪段会话（用对话编号表示）等，需要注意的是，整合后的会话文本数据的数据量很大，因此会将会话文本数据放在不同的数据表中，或者放在不同终端上的不同表中，以减少一个终端损坏而带来的损失，同时也减少后续因大量查询工作而带来的压力；将分成的词根据出现的次数和位置进行倒排索引。通过该倒排索引可以根据分成的词快速获取包含这个词的对话列表，即哪些对话中出现了该词。

S506，根据接收到的查询关键字，利用建立的倒排索引，从会话文本数据中筛选出与所述查询关键字匹配的会话文本数据。其中，查询关键字可以由用户输入，检测并接收用户输入的查询关键字。根据该查询关键字，从建立的倒排索引中进行查询，即根据查询关键字与倒排索引中分成的词进行匹配，并返回与该查询关键字匹配成功的词所在的对话编号，根据该对话编号找到对应的对话内容，并返回对应的对话内容。如此为分析人员提供方便，以加快分析人员的分析速度。

该实施例通过对会话文本数据建立倒排索引，根据接收到的查询关键字，利用建立的倒排索引，从会话文本数据中筛选出与所述查询关键字匹配的会话文本数据，如此可以提供对会话文本数据的查询功能，方便分析人员进行进一步的分析。同时对会话文本数据建立倒排索引，加快了查询的速度，进一步提高了分析人员的分析效率。

图 6 是本申请实施例提供的一种文本数据处理装置的示意性框图。如图 6 所示，该装置 60 包括获取单元 601、导入单元 602、预处理单元 603、整合单元 604。

获取单元 601，用于获取 Excel 数据源文件，其中，Excel 数据源文件中保存的是消息级别的对话文本数据。

导入单元 602，用于将 Excel 数据源文件导入到数据库。

预处理单元 603，用于将导入到数据库中 Excel 数据源文件对应的的数据进行预处理。

具体地，如图 7 所示，将导入到数据库中的 Excel 数据源文件对应的数据进行预处理，即预处理单元 603 包括去重单元 701、筛选单元 702。去重单元 701，用于将导入到数据库中的数据去重。筛选单元 702，用于从去重后的数据中筛选出预设消息类型的消息文本数据。

整合单元 604，将预处理后的数据整合成会话级别的会话文本数据。

具体地，如图 8 所示，将预处理后的数据整合成会话级别的会话文本数据，即整合单元 604 包括集合形成单元 801、分组单元 802、显示单元 803。集合形成单元 801，用于从预处理后的数据中查找每条消息文本数据中的发送人和接收人，将发送人和接收人作为一个集合。分组单元 802，用于按照集合对消息

文本数据进行分组。显示单元 803，用于将每组中的消息文本数据按照预定格式显示。

在一实施例中，所述 Excel 数据源文件是压缩过的 XML 格式文件，需要注意的是，这里的 Excel 特指包含 Office 2007 及后续版本，Office 2007 及后续版本的单个 sheet 可以支持 1048576 行数据，数据量非常大，因此为了方便保存和数据传输，将 Excel 数据源文件导入到数据库之前，需要将 Excel 文件进行处理转换为压缩过的 XML 格式文件。具体地，获取单元，用于获取 Excel 文件，将获取的 Excel 文件进行处理转换为压缩过的 XML 格式文件。将压缩过的 XML 格式文件称为 Excel 数据源文件。

如图 9 所示，将 Excel 数据源文件导入到数据库，即导入单元 602 包括解压单元 901、解析单元 902、保存单元 903。

解压单元 901，用于利用开放源码函式库读取并解压 Excel 数据源文件以得到 XML 格式文件。

解析单元 902，用于将所述 XML 格式文件解析成多行的数据。具体地解析方法请参看对应方法实施例中的具体描述。

以上解压单元 901、解析单元 902，利用 POI 读取并解压数据源文件，其中数据源文件是经过压缩的 XML 文件，并将解压后的 XML 格式文件解析成多行的数据，比起直接读取 Excel 中的数据，速度要快很多。

保存单元 903，用于利用开放源码函式库，通过连接池将解析后的多行数据保存到数据库。

以上保存单元 903，利用开放源码函式库，通过 Hikari Java 数据库连接池将解析后的数据保存到数据库，使用 Hikari Java 数据库连接池可以提高数据库连接的速度，如此也提高了保存数据的速度。

图 10 是本申请另一实施例提供的一种文本数据处理装置的示意性框图。该装置 100 包括获取单元 101、导入单元 102、预处理单元 103、整合单元 104、索引单元 105、查询单元 106。其中，获取单元 101、导入单元 102、预处理单元 103、整合单元 104 请参看图 6 实施例的所述的部分，在此不再赘述。下面将详细描述索引单元 105、查询单元 106。

索引单元 105，用于对会话级别的会话文本数据建立倒排索引。其中，可

使用全文搜索引擎来实现，如 ElasticSearch 全文搜索引擎。

查询单元 106，用于根据接收到的查询关键字，利用建立的倒排索引，从会话文本数据中筛选出与所述查询关键字匹配的会话文本数据。

上述装置实施例的具体工作过程和达到的有益效果，请参看前述方法实施例对应的实施过程和有益效果，在此不再赘述。

上述装置可以实现为一种计算机程序的形式，计算机程序可以在如图 11 所示的计算机设备上运行。

图 11 为本申请实施例提供的一种计算机设备的示意性框图。该计算机设备 110 可以是手机、pad等便携式设备，也可以是台式机等非便携式设备，该计算机设备 110 也可以是以服务器的形式存在。该设备 110 包括通过系统总线 111 连接的处理器 112、存储器和网络接口 113，其中，存储器可以包括非易失性存储介质 114 和内存储器 115。

该非易失性存储介质 114 可存储操作系统 1141 和计算机程序 1142。该计算机程序 1142 被执行时，可使得处理器 112 执行一种文本数据处理方法。该处理器 112 用于提供计算和控制能力，支撑整个设备 110 的运行。该内存储器 115 为非易失性存储介质中的计算机程序的运行提供环境，该计算机程序被处理器 112 执行时，可使得处理器 112 执行一种文本数据处理方法。该网络接口 113 用于进行网络通信，如接收指令等。本领域技术人员可以理解，图 11 中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的设备 110 的限定，具体的设备 110 可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

其中，所述处理器 112 用于运行存储在存储器中的计算机程序，以实现前述文本数据处理方法的任一实施例。

应当理解，在本申请实施例中，所称处理器 112 可以是中央处理单元 (Central Processing Unit，CPU)，该处理器还可以是其他通用处理器、数字信号处理器 (Digital Signal Processor，DSP)、专用集成电路 (Application Specific Integrated Circuit，ASIC)、现成可编程门阵列 (Field-Programmable Gate Array，FPGA) 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

在本申请的另一实施例中提供了一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序包括程序指令，所述程序指令当被处理器执行，以实现前述文本数据处理方法的任一实施例。

所述计算机可读存储介质可以是前述任一实施例所述的终端的内部存储单元，例如终端的硬盘或内存。所述计算机可读存储介质也可以是所述终端的外部存储设备，例如所述终端上配备的插接式硬盘，智能存储卡（Smart Media Card, SMC），安全数字（Secure Digital, SD）卡等。进一步地，所述计算机可读存储介质还可以既包括所述终端的内部存储单元也包括外部存储设备。

在本申请所提供的几个实施例中，应该理解到，所揭露的终端和方法，可以通过其它的方式实现。例如，以上所描述的终端实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式。所属领域的技术人员可以清楚地了解到，为了描述的方便和简洁，上述描述的终端和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到各种等效的修改或替换，这些修改或替换都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以权利要求的保护范围为准。

权 利 要 求 书

1.一种文本数据处理方法，其特征在于，所述方法包括：

获取 Excel 数据源文件，其中，所述 Excel 数据源文件中保存的是消息级别的对话文本数据；

将所述 Excel 数据源文件导入到数据库；

将导入到数据库中的所述 Excel 数据源文件对应的数据进行预处理；

将预处理后的数据整合成会话级别的会话文本数据。

2.根据权利要求 1 所述的方法，其特征在于，所述 Excel 数据源文件是压缩过的 XML 格式文件，所述将所述 Excel 数据源文件导入到数据库，包括：

利用开放源码函式库读取并解压 Excel 数据源文件以得到 XML 格式文件；

将所述 XML 格式文件解析成多行的数据；

利用开放源码函式库，通过连接池将解析后的多行的数据保存到数据库。

3.根据权利要求 1 所述的方法，其特征在于，所述将导入到数据库中的数据进行预处理，包括：

将导入到数据库中的数据去重；

从去重后的数据中筛选出预设消息类型的消息文本数据。

4.根据权利要求 1 所述的方法，其特征在于，所述将预处理后的数据整合成会话级别的会话文本数据，包括：

从预处理后的数据中查找每条消息文本数据中的发送人和接收人，将发送人和接收人作为一个集合；

按照集合对消息文本数据进行分组；

将每组中的消息文本数据按照预定格式显示，以形成会话级别的会话文本数据。

5.根据权利要求 1 所述的方法，其特征在于，所述方法还包括：

对会话级别的会话文本数据建立倒排索引；

根据接收到的查询关键字，利用建立的倒排索引，从所述会话文本数据中筛选出与所述查询关键字匹配的会话文本数据。

6.一种文本数据处理装置，其特征在于，所述文本数据处理装置包括：

获取单元，用于获取 Excel 数据源文件，其中，所述 Excel 数据源文件中保存的是消息级别的消息文本数据；

导入单元，用于将所述 Excel 数据源文件导入到数据库；

预处理单元，用于将导入到数据库中的所述 Excel 数据源文件对应的数据进行预处理；

整合单元，用于将预处理后的数据整合成会话级别的会话文本数据。

7.根据权利要求 6 所述的装置，其特征在于，所述 Excel 数据源文件是压缩过的 XML 格式文件，所述导入单元，包括：

解压单元，用于利用开放源码函式库读取并解压 Excel 数据源文件以得到 XML 格式文件；

解析单元，用于将所述 XML 格式文件解析成多行的数据；

保存单元，用于利用开放源码函式库，通过连接池将解析后的多行的数据保存到数据库。

8.根据权利要求 6 所述的装置，其特征在于，所述预处理单元包括：

去重单元，用于将导入到数据库中的数据去重；

筛选单元，用于从去重后的数据中筛选出预设消息类型的消息文本数据。

9.根据权利要求 6 所述的装置，其特征在于，所述整合单元，包括：

集合形成单元，用于从预处理后的数据中查找每条消息文本数据中的发送人和接收人，将发送人和接收人作为一个集合；

分组单元，用于按照集合对消息文本数据进行分组；

显示单元，用于将每组中的消息文本数据按照预定格式显示，以形成会话级别的会话文本数据。

10.根据权利要求 6 所述的装置，其特征在于，所述装置还包括：

索引单元，用于对会话级别的会话文本数据建立倒排索引；

查询单元，用于根据接收到的查询关键字，利用建立的倒排索引，从所述会话文本数据中筛选出与所述查询关键字匹配的会话文本数据。

11.一种计算机设备，其特征在于，所述计算机设备包括存储器，以及与所述存储器相连的处理器；

所述存储器用于存储实现文本数据处理的计算机程序；所述处理器用于运

行所述存储器中存储的计算机程序，以执行如下步骤：

获取 Excel 数据源文件，其中，所述 Excel 数据源文件中保存的是消息级别的对话文本数据；

将所述 Excel 数据源文件导入到数据库；

将导入到数据库中的所述 Excel 数据源文件对应的数据进行预处理；

将预处理后的数据整合成会话级别的会话文本数据。

12.根据权利要求 11 所述的计算机设备，其特征在于，所述 Excel 数据源文件是压缩过的 XML 格式文件，所述处理器在执行所述将所述 Excel 数据源文件导入到数据库时，具体执行如下步骤：

利用开放源码函式库读取并解压 Excel 数据源文件以得到 XML 格式文件；

将所述 XML 格式文件解析成多行的数据；

利用开放源码函式库，通过连接池将解析后的多行的数据保存到数据库。

13.根据权利要求 11 所述的计算机设备，其特征在于，所述处理器在执行所述将导入到数据库中的数据进行预处理时，具体执行如下步骤：

将导入到数据库中的数据去重；

从去重后的数据中筛选出预设消息类型的消息文本数据。

14.根据权利要求 11 所述的计算机设备，其特征在于，所述处理器在执行所述将预处理后的数据整合成会话级别的会话文本数据时，具体执行如下步骤：

从预处理后的数据中查找每条消息文本数据中的发送人和接收人，将发送人和接收人作为一个集合；

按照集合对消息文本数据进行分组；

将每组中的消息文本数据按照预定格式显示，以形成会话级别的会话文本数据。

15.根据权利要求 11 所述的计算机设备，其特征在于，所述处理器还执行如下步骤：

对会话级别的会话文本数据建立倒排索引；

根据接收到的查询关键字，利用建立的倒排索引，从所述会话文本数据中筛选出与所述查询关键字匹配的会话文本数据。

16.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质存储

有计算机程序，所述计算机程序包括程序指令，所述程序指令被处理器执行时，实现如下步骤：

获取 Excel 数据源文件，其中，所述 Excel 数据源文件中保存的是消息级别的对话文本数据；

将所述 Excel 数据源文件导入到数据库；

将导入到数据库中的所述 Excel 数据源文件对应的数据进行预处理；

将预处理后的数据整合成会话级别的会话文本数据。

17.根据权利要求 16 所述的计算机可读存储介质，其特征在于，所述 Excel 数据源文件是压缩过的 XML 格式文件，所述处理器在执行所述将所述 Excel 数据源文件导入到数据库时，具体实现如下步骤：

利用开放源码函式库读取并解压 Excel 数据源文件以得到 XML 格式文件；

将所述 XML 格式文件解析成多行的数据；

利用开放源码函式库，通过连接池将解析后的多行的数据保存到数据库。

18.根据权利要求 16 所述的计算机可读存储介质，其特征在于，所述处理器在执行所述将导入到数据库中的数据进行预处理时，具体实现如下步骤：

将导入到数据库中的数据去重；

从去重后的数据中筛选出预设消息类型的消息文本数据。

19.根据权利要求 16 所述的计算机可读存储介质，其特征在于，所述处理器在执行所述将预处理后的数据整合成会话级别的会话文本数据时，具体实现如下步骤：

从预处理后的数据中查找每条消息文本数据中的发送人和接收人，将发送人和接收人作为一个集合；按照集合对消息文本数据进行分组；将每组中的消息文本数据按照预定格式显示，以形成会话级别的会话文本数据。

20.根据权利要求 16 所述的计算机可读存储介质，其特征在于，所述程序指令被处理器执行时，所述处理器还实现如下步骤：

对会话级别的会话文本数据建立倒排索引；

根据接收到的查询关键字，利用建立的倒排索引，从所述会话文本数据中筛选出与所述查询关键字匹配的会话文本数据。

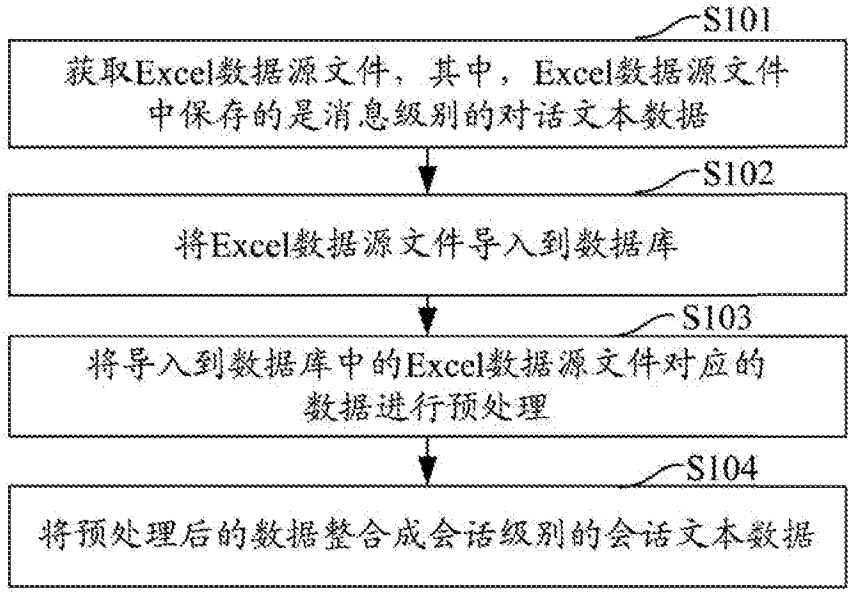
S101
获取Excel数据源文件，其中，Excel数据源文件中保存的是消息级别的对话文本数据
S102
将Excel数据源文件导入到数据库
S103
将导入到数据库中的Excel数据源文件对应的数据进行预处理
S104
将预处理后的数据整合成会话级别的会话文本数据

图 1

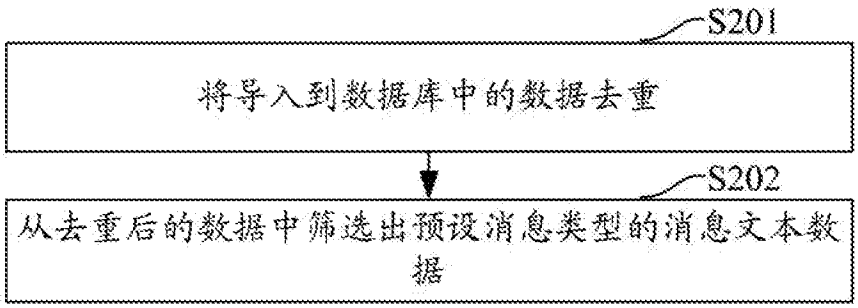
S201
将导入到数据库中的数据去重
S202
从去重后的数据中筛选出预设消息类型的消息文本数据

图 2

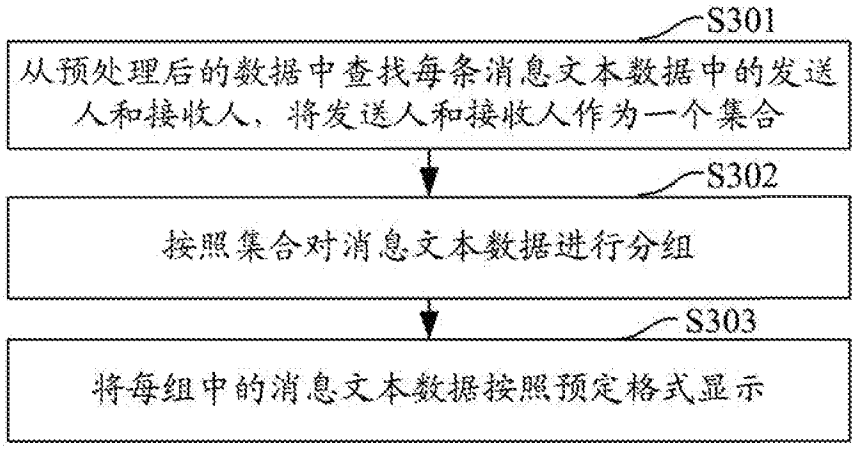
S301
从预处理后的数据中查找每条消息文本数据中的发送人和接收人，将发送人和接收人作为一个集合
S302
按照集合对消息文本数据进行分组
S303
将每组中的消息文本数据按照预定格式显示

图 3

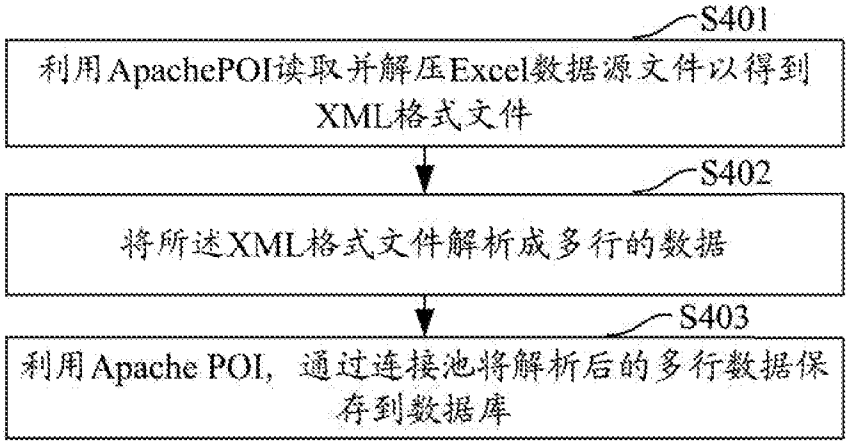
S401
利用ApachePOI读取并解压Excel数据源文件以得到XML格式文件
S402
将所述XML格式文件解析成多行的数据
S403
利用Apache POI，通过连接池将解析后的多行数据保存到数据库

图 4

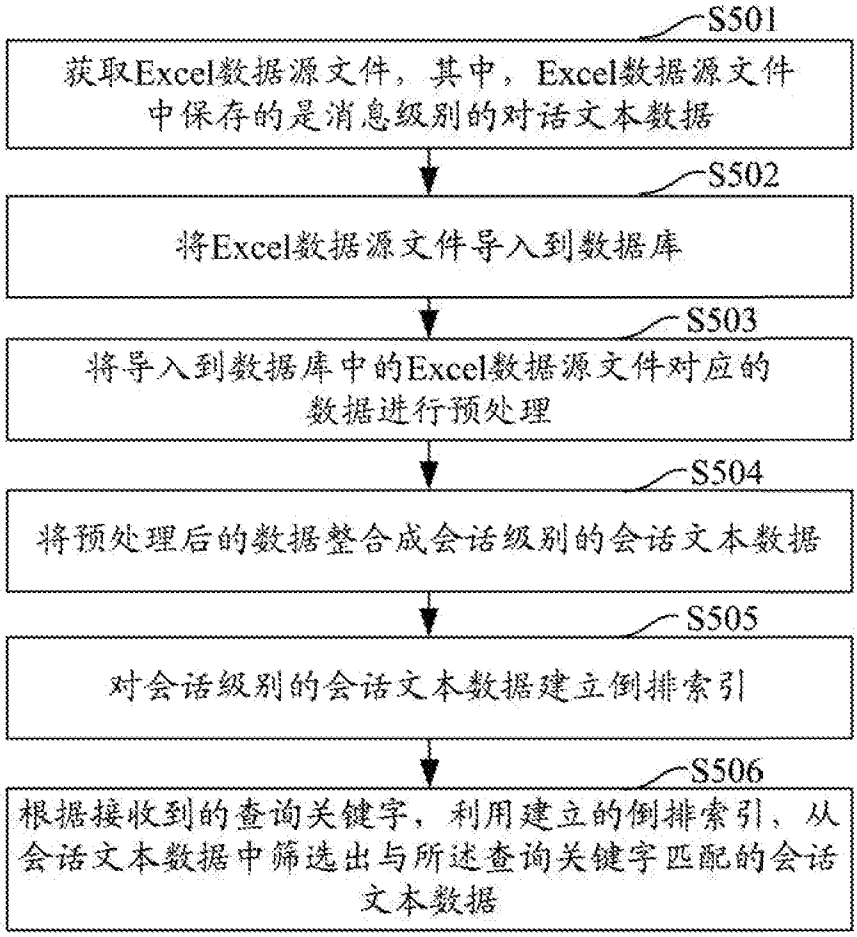
S501
获取Excel数据源文件，其中，Excel数据源文件中保存的是消息级别的对话文本数据
S502
将Excel数据源文件导入到数据库
S503
将导入到数据库中的Excel数据源文件对应的数据进行预处理
S504
将预处理后的数据整合成会话级别的会话文本数据
S505
对会话级别的会话文本数据建立倒排索引
S506
根据接收到的查询关键字，利用建立的倒排索引、从会话文本数据中筛选出与所述查询关键字匹配的会话文本数据

图 5

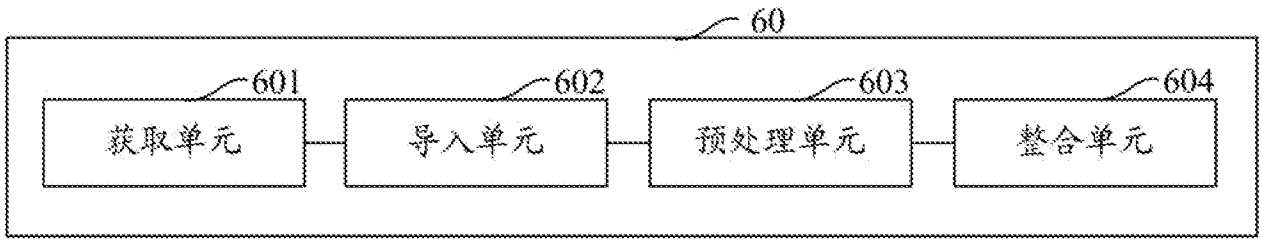
60
601
获取单元
602
导入单元
603
预处理单元
604
整合单元

图 6

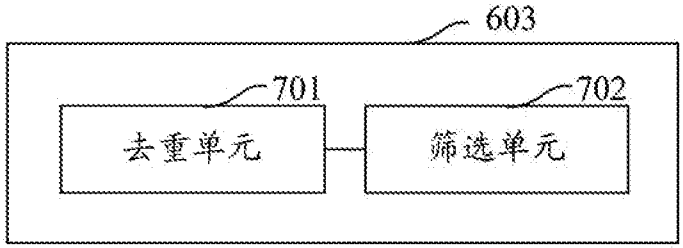
603
701
去重单元
702
筛选单元

图 7

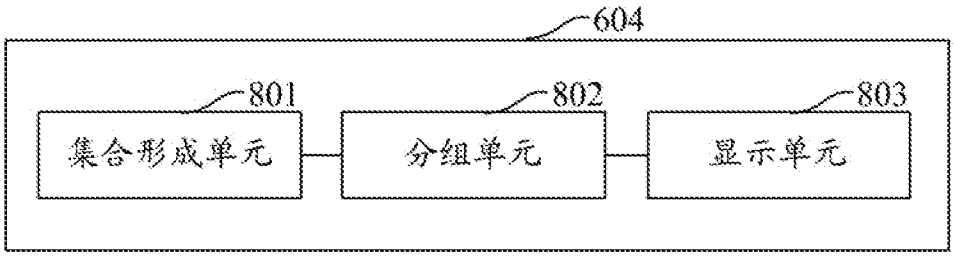
604
801
集合形成单元
802
分组单元
803
显示单元

图 8


602
901
解压单元
902
解析单元
903
保存单元

图 9

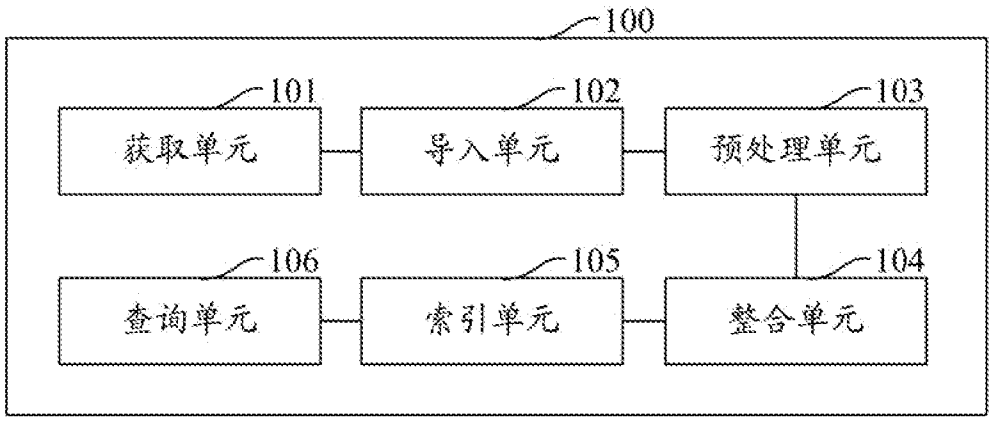
100
101
获取单元
102
导入单元
103
预处理单元
106
查询单元
105
索引单元
104
整合单元

图 10

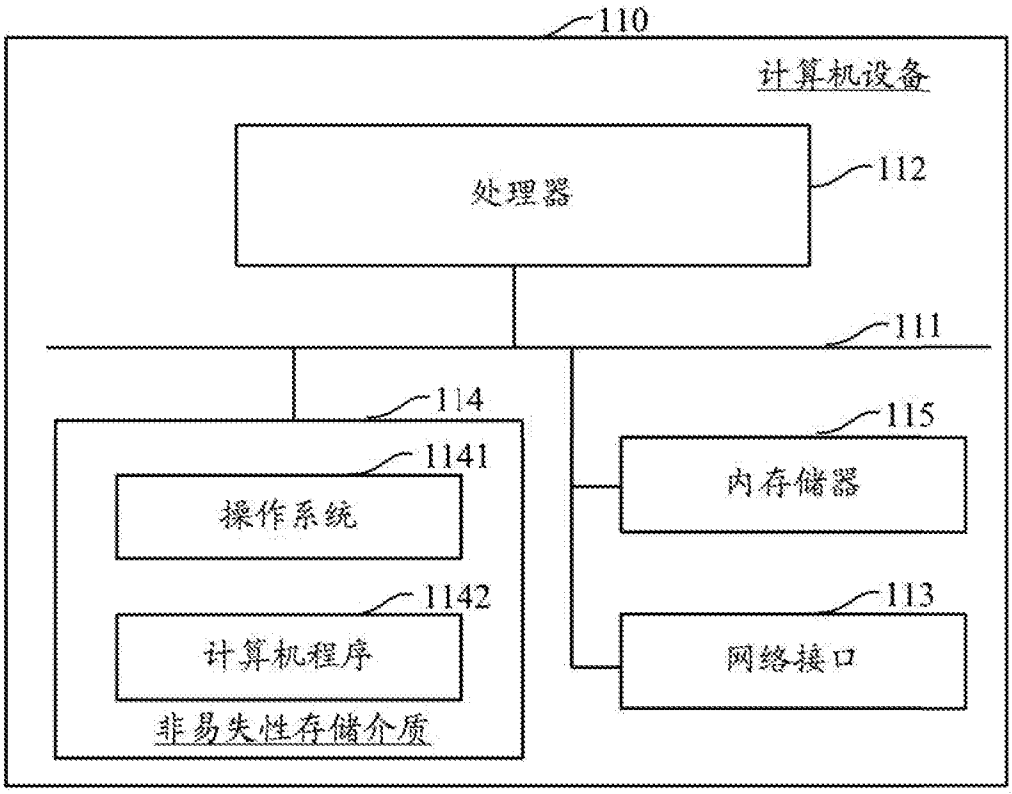
110
计算机设备
处理器
112
111
114
1141
操作系统
1142
计算机程序
非易失性存储介质
115
内存储器
113
网络接口

图 11

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/100930

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30(2006.01)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 17/-

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)

DWPI, CNTXT, USTXT, SIPOABS, CPRSABS, CNKI: 文本, 数据, excel, 对话, 会话, 导入, 数据库, 分析, 信息, 消息, 索引, 用户, 客户, 上下文, 客服, 服务, 整理, text, data, dialog, session, , database, analysis, message, information, index, user, context, service, customer, manage, management

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 102866990 A (BEIJING SOGOU INFORMATION SERVICE CO., LTD. ET AL.) 09 January 2013 (2013-01-09) entire document	1-20
A	CN 104428764 A (THOMSON REUTERS GLOBAL RESOURCES) 18 March 2015 (2015-03-18) entire document	1-20
A	KR 20090053174 A (KT CORP. ET AL.) 27 May 2009 (2009-05-27) 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	Date of mailing of the international search report
21 December 2018	**27 December 2018**

Name and mailing address of the ISA/CN	Authorized officer
State Intellectual Property Office of the P. R. China **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing** **100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/100930

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	102866990	A	09 January 2013	CN	102866990	B	03 August 2016
CN	104428764	A	18 March 2015	CN	104428764	B	26 September 2017
				JP	2015527641	A	17 September 2015
				WO	2014001915	A3	26 June 2014
				US	9218344	B2	22 December 2015
				US	2014006424	A1	02 January 2014
				JP	6403340	B2	10 October 2018
				WO	2014001915	A2	03 January 2014
				EP	2867798	A2	06 May 2015
KR	20090053174	A	27 May 2009		None		

国际检索报告

国际申请号 PCT/CN2018/100930

A. 主题的分类

G06F 17/30(2006.01)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 17/-

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））

DWPI, CNTXT, USTXT, SIPOABS, CPRSABS, CNKI: 文本，数据，excel，对话，会话，导入，数据库，分析，信息，消息，索引，用户，客户，上下文，客服，服务，整理，text, data, dialog, session, , database, analysis, message, information, index, user, context, service, customer, manage, management

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
A	CN 102866990 A (北京搜狗信息服务有限公司等) 2013年 1月 9日 (2013-01-09) 全文	1-20
A	CN 104428764 A (汤姆森路透社全球资源公司) 2015年 3月 18日 (2015-03-18) 全文	1-20
A	KR 20090053174 A (KT CORP 等) 2009年 5月 27日 (2009-05-27) 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	国际检索报告邮寄日期
2018年 12月 21日	2018年 12月 27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 田竞 电话号码 62089417

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/100930

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	102866990	A	2013年 1月 9日	CN	102866990	B	2016年 8月 3日
CN	104428764	A	2015年 3月 18日	CN	104428764	B	2017年 9月 26日
				JP	2015527641	A	2015年 9月 17日
				WO	2014001915	A3	2014年 6月 26日
				US	9218344	B2	2015年 12月 22日
				US	2014006424	A1	2014年 1月 2日
				JP	6403340	B2	2018年 10月 10日
				WO	2014001915	A2	2014年 1月 3日
				EP	2867798	A2	2015年 5月 6日
KR	20090053174	A	2009年 5月 27日		无		