

(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 11 月 28 日 (28.11.2019)



(10) 国际公布号
WO 2019/223153 A1

(51) 国际专利分类号:
G06F 16/35 (2019.01)

(21) 国际申请号: PCT/CN2018/102468

(22) 国际申请日: 2018 年 8 月 27 日 (27.08.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201810515743.0 2018 年 5 月 25 日 (25.05.2018) 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: 黄度新(HUANG, Duxin); 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)。 张川(ZHANG, Chuan); 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)。 金鑫(JIN, Xin); 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)。 王翼(WANG, Yi); 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)。

(54) Title: BIG DATA STRUCTURING METHOD, DEVICE, COMPUTER APPARATUS, AND STORAGE MEDIUM

(54) 发明名称: 海量数据结构化方法、装置、计算机设备及存储介质

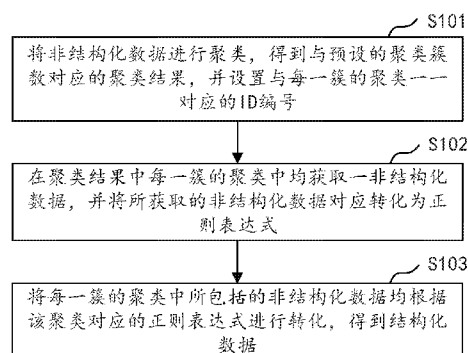


图 1

- S101 Cluster unstructured data, obtain a clustering result corresponding to a pre-determined number of clusters, and configure ID numbers in one-to-one correspondence to the clusters
- S102 Acquire one unstructured data item from each of the clusters in the clustering result, and correspondingly convert the acquired unstructured data items into a regular expression
- S103 Perform conversion on all the unstructured data contained in each cluster according to the regular expression corresponding to a given cluster, and obtain structured data

(57) Abstract: Embodiments of the present application disclose a big data structuring method, a device, a computer apparatus, and a storage medium. The method comprises: clustering unstructured data, obtaining a clustering result corresponding to a pre-determined number of clusters, and configuring ID numbers in one-to-one correspondence to the clusters; acquiring one unstructured data item from each of the clusters in the clustering result, and correspondingly converting the acquired unstructured data items into a regular expression; and converting all the unstructured data contained in each cluster according to the regular expression corresponding to a

(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, 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))。

given cluster, and obtaining structured data. In the method, a clustering algorithm is used to cluster a large volume of unstructured data, and a regular expression corresponding to each cluster is generated and applied to all data in the cluster, such that the large volume of unstructured data can be rapidly converted into structured data, thereby efficiently meeting a requirement of training data needed for deep learning.

(57) 摘要: 本申请实施例公开了一种海量数据结构化方法、装置、计算机设备及存储介质。该方法包括: 将非结构化数据进行聚类, 得到与预设的聚类簇数对应的聚类结果, 并设置与每一簇的聚类一一对应的ID编号; 在聚类结果中每一簇的聚类中均获取一非结构化数据, 并将所获取的非结构化数据对应转化为正则表达式; 将每一簇的聚类中所包括的非结构化数据均根据该聚类对应的正则表达式进行转化, 得到结构化数据。该方法实现由聚类算法将海量的非结构化数据做聚类, 每个簇对应生成一个正则表达式并应用到该簇的所有数据, 这样海量非结构化数据都能快速转化为结构化数据, 能快速的满足深度学习所需训练数据的需求。

发明名称：海量数据结构化方法、装置、计算机设备及存储介质

本申请要求于2018年5月25日提交中国专利局、申请号为201810515743.0、申请名称为“海量数据结构化方法、装置、计算机设备及存储介质”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

本申请涉及数据结构化技术领域，尤其涉及一种海量数据结构化方法、装置、计算机设备及存储介质。

背景技术

目前，深度学习的训练需要依靠大量标注好的数据或结构化数据，但是将非结构化数据变成结构化数据需要投入大量的人力成本以通过手动转化的方式来进行数据结构化处理，这就导致获取深度学习所需训练数据的效率低下。

发明内容

本申请提供了一种海量数据结构化方法、装置、计算机设备及存储介质，旨在解决现有技术中将非结构化数据变成结构化数据需要投入大量的人力成本以通过手动转化的方式来进行数据结构化处理，导致获取深度学习所需训练数据的效率低下的问题。

第一方面，本申请提供了一种海量数据结构化方法，其包括：将非结构化数据进行聚类，得到与预设的聚类簇数对应的聚类结果，并设置与每一簇的聚类一一对应的ID编号；在聚类结果中每一簇的聚类中均获取一非结构化数据，并将所获取的非结构化数据对应转化为正则表达式；将每一簇的聚类中所包括的非结构化数据均根据该聚类对应的正则表达式进行转化，得到结构化数据。

第二方面，本申请提供了一种海量数据结构化装置，其包括：聚类单元，用于将非结构化数据进行聚类，得到与预设的聚类簇数对应的聚类结果，并设置与每一簇的聚类一一对应的ID编号；正则表达式获取单元，用于在聚类结果中每一簇的聚类中均获取一非结构化数据，并将所获取的非结构化数据对应转

化为正则表达式；数据结构化单元，用于将每一簇的聚类中所包括的非结构化数据均根据该聚类对应的正则表达式进行转化，得到结构化数据。

第三方面，本申请又提供了一种计算机设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现本申请提供的任一项所述的海量数据结构化方法。

第四方面，本申请还提供了一种存储介质，其中所述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时使所述处理器执行本申请提供的任一项所述的海量数据结构化方法。

本申请提供一种海量数据结构化方法、装置、计算机设备及存储介质。该方法实现由聚类算法将海量的非结构化数据做聚类，每个簇对应生成一个正则表达式并应用到该簇的所有数据，这样海量非结构化数据都能快速转化为结构化数据，能快速的满足深度学习所需训练数据的需求。

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为本申请实施例提供的一种海量数据结构化方法的示意图；

图2为本申请实施例提供的一种海量数据结构化方法的另一示意图；

图3是本申请实施例提供的一种海量数据结构化方法的子流程示意图；

图4为本申请实施例提供的一种海量数据结构化方法的另一子流程示意图；

图5为本申请实施例提供的一种海量数据结构化方法的另一子流程示意图；

图6为本申请实施例提供的一种海量数据结构化装置的示意性框图；

图7为本申请实施例提供的一种海量数据结构化装置的另一示意性框图；

图8为本申请实施例提供的一种海量数据结构化装置的子单元示意性框图；

图9为本申请实施例提供的一种海量数据结构化装置的另一子单元示意性框图；

图10为本申请实施例提供的一种海量数据结构化装置的另一子单元示意性框图；

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。

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。

请参阅图 1，图 1 是本申请实施例提供的一种海量数据结构化方法的示意流程图。该方法应用于台式电脑、手提电脑、平板电脑等终端中。如图 1 所示，该方法包括步骤 S101~S103。

S101、将非结构化数据进行聚类，得到与预设的聚类簇数对应的聚类结果，并设置与每一簇的聚类一一对应的 ID 编号。

在本实施例中，对非结构化数据进行聚类，是为了将具有相同或相似特征的非结构化数据归为一类。通过聚类进行分类后，能更加具有针对性的对每一聚类生成正则表达式，以实现后续的数据结构化处理。而且，对每一簇的聚类一一对应的 ID 编号是为了后续根据该 ID 编号对应建数据表格以存储所生成的结构化数据。

在一实施例中，如图 3 所示，所述步骤 S101 包括：

S1011、在多个非结构化数据中选取与预设的聚类簇数相同个数的非结构化数据，将所选取的非结构化数据作为每一簇的初始聚类中心；

S1012、根据多个非结构化数据与各初始聚类中心的相异值，将多个非结构化数据进行划分，得到初始聚类结果；

S1013、根据初始聚类结果，获取每一簇的调整后聚类中心；

S1014、根据调整后聚类中心，将多个非结构化数据进行划分，直至聚类结果保持相同的次数多于预设的次数，得到与预设的聚类簇数对应的聚类结果。

在本实施例中，对非结构化数据进行聚类时采用 k-means 算法进行，具体如下：

11) 从 n 个非结构化数据中任意选取 k 个非结构化数据，并作为 k 个簇的初始聚类中心；其中，非结构化数据的初始总个数为 n ，从其中任意选择 k 个数据 ($k < n$, k 是用户指定的参数，即所期望的簇的个数，也即预设的聚类簇数)，

将初始选择的 k 个数据作为初始聚类中心。

12) 分别计算剩下的非结构化数据到 k 个簇初始聚类中心的相异度, 将剩下的非结构化数据分别划归到相异度最低的簇, 得到初始聚类结果; 即是剩下的每一非结构化数据选择距其距离最近的初始聚类中心, 并与该初始聚类中心归为一类; 这样就以初始选择的初始聚类中心将海量的非结构化数据划分为 k 簇, 每一簇数据都有一个初始聚类中心。

13) 根据初始聚类结果, 重新计算 k 个簇各自的聚类中心; 具体计算方法是取每一簇中所有非结构化数据的各自维度的算术平均数, 例如每一的非结构化数据具有 3 个维度, 分别是商品编码, 品牌, 型号, 就针对每一簇分别计算这三个维度的聚类中心, 选择一个三个维度都符合算数平均值的非结构化数据作为新的聚类中心, 重新选择该簇数据中更优的聚类中心。

14) 将 n 个非结构化数据中全部元素按照新的中心重新聚类;

15) 重复第 14) 步, 直到聚类结果不再变化, 得到与预设的聚类簇数对应的聚类结果。

在完成了聚类分类之后, 再对每一最终聚类结果一一设置 ID 编号, 即可实现快速的将海量的非结构化数据进行分组。

S102、在聚类结果中每一簇的聚类中均获取一非结构化数据, 并将所获取的非结构化数据对应转化为正则表达式。

在本实施例中, 在聚类结果中每一簇的聚类中均获取一非结构化数据时, 可以是随机从每一簇聚类中获取一非结构化数据, 也可以直接选择该簇聚类的聚类中心所对应的非结构化数据 (例如 k 个簇的聚类则需要从每一簇聚类中都选择一个, 一共选择 k 个非结构化数据)。完成对每一簇的聚类中均获取一非结构化数据时, 则将所选中的 k 个非结构化数据均转化为正则表达式。

在一实施例中, 如图 4 所示, 所述步骤 S102 包括:

S1021、将所获取的与每一簇的聚类一一对应的非结构化数据根据字符类型进行划分, 得到与所获取的非结构化数据一一对应的划分结果; 其中, 每一划分结果均包括多个子字符串;

S1022、根据每一划分结果所包括子字符串, 生成与每一划分结果对应的多个子正则表达式;

S1023、将同一划分结果内对应的多个子正则表达式串接, 得到该划分结果

对应的正则表达式。

例如，有一非结构化的初始数据如下：

8471701000 固态硬盘 品牌:Crucial|型号:CT1024M550SSD1|容量:1TB 无缓存|无转速|用于计算机，非移动硬盘,接口类型：SATA III，传输速率：6Gbps

将非结构化的数据进行字符串进行划分后，再对每个字符串依次生成子正则表达式，将子正则表达式串接，得到与该非结构化的数据对应的正则表达式；

例如，可以根据 ABC 生成对应的正则表达式**\bABC\b**；可以根据 ABC123 生成对应的正则表达式**\bABC\b\d{3}**；根据 ABC123-生成对应的正则表达式**\bABC\b\d{3}-**；根据 ABC123-1 生成对应的正则表达式**\bABC\b\d{3}-\d**；根据 ABC123-1-生成对应的正则表达式**\bABC\b\d{3}-\d-**；根据 ABC121-1-1 生成对应的正则表达式**\bABC\b\d{3}-\d-\d**；根据 ABC-1 生成对应的正则表达式**\bABC\b-\d**。

上述举例的示例，包括以下步骤：

21) 通过自动识别将非结构化的数据中的字符以字符类型（如数字类字符划为第一类、字母类字符划为第二类、符号类字符划分为第三类）将该示例划分为多段字符串；例如，将 ABC121-1-1 进行划分时，A、B、C 是三个字母，C 后面是 1 则是数字，此时作自动划分时则在 C 和 1 之间划分开；同样的，121 中最后一位的 1 和第一个-之间划分开，第一个-与之后的第一个 1 之间划分开，第一个-与之后的第一个 1 与第二个-之间划分开，第二个-与最后一位 1 之间划分开，即按字符类型进行划分后得到了以下几段 ABC、121、-、1、-、1；

22) 选取非结构化数据划分得到的多段字符串中的每一段，并对应生成正则表达式。

通过上述转化，使得每一簇的聚类均得到一个正则表达式，通过该正则表达式的替换、检索等功能，将非结构化数据转化为结构化数据，实现了对海量的非结构化数据批量转化处理，提高了转化效率。

S103、将每一簇的聚类中所包括的非结构化数据均根据该聚类对应的正则表达式进行转化，得到结构化数据。

在本实施例中，由于每一聚类中被选出的非结构化数据都生成了对应的正则表达式，再根据正则表达式将每一聚类中剩余的非结构化数据进行转化，生

成结构化数据。例如上述示例中所列举的非结构化数据与正则表达式匹配后，即可对应标准化为结构化数据，其格式如下：

商品名称：固态硬盘

商品编码：8471701000

品牌：Crucial

型号：CT1024M550SSD1

用途：用于计算机

功能：存储器功能

容量：1TB。

通过正则表达式将海量的非结构化数据转化为结构化数据，提供数据规范化效率，节约开发时间。

在一实施例中，所述步骤 S103 包括：

获取每一簇的聚类所对应的正则表达式所包括的多个子正则表达式，将每一簇的聚类中所包括的非结构化数据均与该聚类对应的多个子正则表达式进行转化，得到与每一非结构化数据对应的多个初始转化结果；

将同一非结构化数据对应的多个初始转化结果进行串接，得到该非结构化数据对应的结构化数据。

在本实施例中，由于每一簇的聚类所对应的正则表达式均是包括多个子正则表达式，故该簇聚类对应的非结构化数据均可以根据该簇聚类对应的正则表达式所包括多个子正则表达式分段进行匹配和转化，得到与每一非结构化数据对应的多个初始转化结果。此时，再与由子正则表达式串接形成正则表达式过程一样，同样通过串接的方式将同一非结构化数据对应的多个初始转化结果进行处理，得到该非结构化数据对应的结构化数据。通过分段转化的方式，能确保转化结果更加准确。

在一实施例中，将同一非结构化数据对应的多个初始转化结果进行串接，得到该非结构化数据对应的结构化数据中，是将同一非结构化数据对应的多个初始转化结果尾部均增加一转行符后再进行串接，得到该非结构化数据对应的结构化数据。

在本实施例中，通过对同一非结构化数据对应的多个初始转化结果尾部均增加一转行符后再进行串接，使得在后续的存储阶段，更容易的识别结构化数

据的字段，一一对应建立数据表格的字段。

在一实施例中，如图 2 所示，所述步骤 S103 之后，还包括：

S104、将每一簇的聚类中的每一结构化数据根据文本分类算法填充至该聚类对应的数据表格中进行存储。

在一实施例中，如图 5 所示，所述步骤 S104 包括：

S1041、根据每一簇的聚类的 ID 编号，一一对应建立数据表格；

S1042、根据每一簇的聚类中所选定结构化数据的字段，一一对应建立数据表格的字段；

S1043、将每一簇的聚类中所包括结构化数据根据 Fasttext 文本分类算法填充至该聚类对应的数据表格的对应字段中。

在本实施例中，由于将非结构化的数据通过正则表达式转化后，需填充至数据库中的数据表格中进行存储，具体可以如下：

41) 根据每一聚类的 ID 编号对应建立数据表格；

例如有 N 个聚类的 ID 编号，则对应建立 N 个数据表格，每一数据表格是相应的根据其中一个聚类的 ID 编号建立；

42) 根据每一簇的聚类中所选定结构化数据的字段，一一对应建立数据表格的字段；

例如，在 ID 编号为 1 的聚类中，选定的结构化数据如：

商品名称：固态硬盘

商品编码：8471701000

品牌：Crucial

型号：CT1024M550SSD1

用途：用于计算机

功能：存储器功能

容量：1TB。

则在 ID 编号为 1 对应的数据表格中建立商品名称、商品编码、品牌、型号、用途、功能、容量的字段。

43) 将聚类的 ID 编号所对应的每一条结构化数据按 Fasttext 文本分类算法填充至数据表格的对应字段中；

其中，Fasttext 文本分类算法的模型如下：

$$-\frac{1}{N} \sum_{n=1}^N y_n \log (f(BAx_n))$$

其中, $\langle x_n, y_n \rangle$ 是一条训练样本, y_n 是训练目标, x_n 是 normalized bag of features。矩阵参数 A 是基于 word 的 look-up table, 也就是 A 是词的 embedding 向量。 Ax_n 矩阵运算的数学意义是将 word 的 embedding 向量找到后相加或者取平均, 得到 hidden 向量。矩阵参数 B 是函数 f 的参数, 函数 f 是一个多分类问题, 所以 $f(BAx_n)$ 是一个多分类的线性函数。优化目标是使的这个多分类问题的似然越大越好。

也即, 通过 Fasttext 文本分类算法可以将每一结构化数据所包括的多个字段的数据对应填充至数据表中的对应字段中。

可见, 该方法实现由聚类算法将海量的非结构化数据做聚类, 每个簇对应生成一个正则表达式并应用到该簇的所有数据, 这样海量非结构化数据都能快速转化为结构化数据, 能快速的满足深度学习所需训练数据的需求。

本申请实施例还提供一种海量数据结构化装置, 该海量数据结构化装置用于执行前述海量数据结构化方法的任一实施例。具体地, 请参阅图 6, 图 6 是本申请实施例提供的一种海量数据结构化装置的示意性框图。海量数据结构化装置 100 可以配置于台式电脑、平板电脑、手提电脑、等终端中。

如图 6 所示, 海量数据结构化装置 100 包括聚类单元 101、正则表达式获取单元 102、数据结构化单元 103。

聚类单元 101, 用于将非结构化数据进行聚类, 得到与预设的聚类簇数对应的聚类结果, 并设置与每一簇的聚类一一对应的 ID 编号。

在一实施例中, 如图 8 所示, 所述聚类单元 101 包括:

初始中心选择单元 1011, 用于在多个非结构化数据中选取与预设的聚类簇数相同个数的非结构化数据, 将所选取的非结构化数据作为每一簇的初始聚类中心; 初始聚类单元 1012, 用于根据多个非结构化数据与各初始聚类中心的相异值, 将多个非结构化数据进行划分, 得到初始聚类结果; 聚类中心调整单元 1013, 用于根据初始聚类结果, 获取每一簇的调整后聚类中心; 聚类调整单元 1014, 用于根据调整后聚类中心, 将多个非结构化数据进行划分, 直至聚类结果保持相同的次数多于预设的次数, 得到与预设的聚类簇数对应的聚类结果。

正则表达式获取单元 102, 用于在聚类结果中每一簇的聚类中均获取一非结

构化数据，并将所获取的非结构化数据对应转化为正则表达式。

在一实施例中，如图 9 所示，所述正则表达式获取单元 102 包括：

字符串划分单元 1021，用于将所获取的与每一簇的聚类一一对应的非结构化数据根据字符类型进行划分，得到与所获取的非结构化数据一一对应的划分结果；其中，每一划分结果均包括多个子字符串；

分段生成单元 1022，用于根据每一划分结果所包括子字符串，生成与每一划分结果对应的多个子正则表达式；

串接单元 1023，用于将同一划分结果内对应的多个子正则表达式串接，得到该划分结果对应的正则表达式。

数据结构化单元 103，用于将每一簇的聚类中所包括的非结构化数据均根据该聚类对应的正则表达式进行转化，得到结构化数据。

在一实施例中，所述数据结构化单元 103 包括：

初始转化结果获取单元，用于获取每一簇的聚类所对应的正则表达式所包括的多个子正则表达式，将每一簇的聚类中所包括的非结构化数据均与该聚类对应的多个子正则表达式进行转化，得到与每一非结构化数据对应的多个初始转化结果；

初始转化结果串接单元，用于将同一非结构化数据对应的多个初始转化结果进行串接，得到该非结构化数据对应的结构化数据。

在一实施例中，在初始转化结果串接单元中，是将同一非结构化数据对应的多个初始转化结果尾部均增加一换行符后再进行串接，得到该非结构化数据对应的结构化数据。

在一实施例中，如图 7 所示，所述海量数据结构化装置 100，还包括：

分类存储单元 104，用于将每一簇的聚类中的每一结构化数据根据文本分类算法填充至该聚类对应的数据表格中进行存储。

在一实施例中，如图 10 所示，所述分类存储单元 104 包括：

表格建立单元 1041，用于根据每一簇的聚类的 ID 编号，一一对应建立数据表格；

字段建立单元 1042，用于根据每一簇的聚类中所选定结构化数据的字段，一一对应建立数据表格的字段；

分类填充单元 1043，用于将每一簇的聚类中所包括结构化数据根据 Fasttext

文本分类算法填充至该聚类对应的数据表格的对应字段中。

可见，该装置实现由聚类算法将海量的非结构化数据做聚类，每个簇对应生成一个正则表达式并应用到该簇的所有数据，这样海量非结构化数据都能快速转化为结构化数据，能快速的满足深度学习所需训练数据的需求。

上述海量数据结构化装置可以实现为一种计算机程序的形式，该计算机程序可以在如图 11 所示的计算机设备上运行。

请参阅图 11，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。该计算机设备 500 设备可以是终端。该终端可以是平板电脑、笔记本电脑、台式电脑、个人数字助理等电子设备。

参阅图 11，该计算机设备 500 包括通过系统总线 501 连接的处理器 502、存储器和网络接口 505，其中，存储器可以包括非易失性存储介质 503 和内存存储器 504。

该非易失性存储介质 503 可存储操作系统 5031 和计算机程序 5032。该计算机程序 5032 包括程序指令，该程序指令被执行时，可使得处理器 502 执行一种海量数据结构化方法。

该处理器 502 用于提供计算和控制能力，支撑整个计算机设备 500 的运行。

该内存存储器 504 为非易失性存储介质 503 中的计算机程序 5032 的运行提供环境，该计算机程序 5032 被处理器 502 执行时，可使得处理器 502 执行一种海量数据结构化方法。

该网络接口 505 用于进行网络通信，如发送分配的任务等。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计算机设备 500 的限定，具体的计算机设备 500 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 502 用于运行存储在存储器中的计算机程序 5032，以实现如下功能：将非结构化数据进行聚类，得到与预设的聚类簇数对应的聚类结果，并设置与每一簇的聚类一一对应的 ID 编号；在聚类结果中每一簇的聚类中均获取一非结构化数据，并将所获取的非结构化数据对应转化为正则表达式；将每一簇的聚类中所包括的非结构化数据均根据该聚类对应的正则表达式进行转化，得到结构化数据。

在一实施例中，处理器 502 还执行如下操作：将每一簇的聚类中的每一结构化数据根据文本分类算法填充至该聚类对应的数据表格中进行存储。

在一实施例中，处理器 502 还执行如下操作：在多个非结构化数据中选取与预设的聚类簇数相同个数的非结构化数据，将所选取的非结构化数据作为每一簇的初始聚类中心；根据多个非结构化数据与各初始聚类中心的相异值，将多个非结构化数据进行划分，得到初始聚类结果；根据初始聚类结果，获取每一簇的调整后聚类中心；根据调整后聚类中心，将多个非结构化数据进行划分，直至聚类结果保持相同的次数多于预设的次数，得到与预设的聚类簇数对应的聚类结果。

在一实施例中，处理器 502 还执行如下操作：将所获取的与每一簇的聚类一一对应的非结构化数据根据字符类型进行划分，得到与所获取的非结构化数据一一对应的划分结果；其中，每一划分结果均包括多个子字符串；根据每一划分结果所包括子字符串，生成与每一划分结果对应的多个子正则表达式；将同一划分结果内对应的多个子正则表达式串接，得到该划分结果对应的正则表达式。

在一实施例中，处理器 502 还执行如下操作：根据每一簇的聚类的 ID 编号，一一对应建立数据表格；根据每一簇的聚类中所选定结构化数据的字段，一一对应建立数据表格的字段；将每一簇的聚类中所包括结构化数据根据 Fasttext 文本分类算法填充至该聚类对应的数据表格的对应字段中。

在一实施例中，处理器 502 还执行如下操作：获取每一簇的聚类所对应的正则表达式所包括的多个子正则表达式，将每一簇的聚类中所包括的非结构化数据均与该聚类对应的多个子正则表达式进行转化，得到与每一非结构化数据对应的多个初始转化结果；将同一非结构化数据对应的多个初始转化结果进行串接，得到该非结构化数据对应的结构化数据。

在一实施例中，处理器 502 还执行如下操作：将同一非结构化数据对应的多个初始转化结果尾部均增加一转行符后再进行串接，得到该非结构化数据对应的结构化数据。

本领域技术人员可以理解，图 11 中示出的计算机设备的实施例并不构成对计算机设备具体构成的限定，在其他实施例中，计算机设备可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。例如，在一些实

施例中，计算机设备可以仅包括存储器及处理器，在这样的实施例中，存储器及处理器的结构及功能与图 11 所示实施例一致，在此不再赘述。

应当理解，在本申请实施例中，处理器 502 可以是中央处理单元 (Central Processing Unit, CPU)，该处理器 502 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。其中，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一种存储介质。该存储介质可以为计算机可读存储介质。该存储介质存储有计算机程序，其中计算机程序包括程序指令。该程序指令被处理器执行时实现本申请实施例的海量数据结构化方法。

所述存储介质可以是前述设备的内部存储单元，例如设备的硬盘或内存。所述存储介质也可以是所述设备的外部存储设备，例如所述设备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，闪存卡 (Flash Card) 等。进一步地，所述存储介质还可以既包括所述设备的内部存储单元也包括外部存储设备。

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，上述描述的设备、装置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。

另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以是两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

以上所述，仅为本发明的具体实施方式，但本发明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发明揭露的技术范围内，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发明的保护范围之内。因此，本发明的保护范围应以权利要求的保护范围为准。

权 利 要 求 书

1. 一种海量数据结构化方法，其特征在于，包括：

将非结构化数据进行聚类，得到与预设的聚类簇数对应的聚类结果，并设置与每一簇的聚类一一对应的 ID 编号；

在聚类结果中每一簇的聚类中均获取一非结构化数据，并将所获取的非结构化数据对应转化为正则表达式；

将每一簇的聚类中所包括的非结构化数据均根据该聚类对应的正则表达式进行转化，得到结构化数据。

2. 根据权利要求 1 所述的海量数据结构化方法，其特征在于，所述将每一簇的聚类中所包括的非结构化数据均根据该聚类对应的正则表达式进行转化，得到结构化数据之后，还包括：

将每一簇的聚类中的每一结构化数据根据文本分类算法填充至该聚类对应的数据表格中进行存储。

3. 根据权利要求 1 所述的海量数据结构化方法，其特征在于，所述将非结构化数据进行聚类，得到与预设的聚类簇数对应的聚类结果，包括：

在多个非结构化数据中选取与预设的聚类簇数相同个数的非结构化数据，将所选取的非结构化数据作为每一簇的初始聚类中心；

根据多个非结构化数据与各初始聚类中心的相异值，将多个非结构化数据进行划分，得到初始聚类结果；

根据初始聚类结果，获取每一簇的调整后聚类中心；

根据调整后聚类中心，将多个非结构化数据进行划分，直至聚类结果保持相同的次数多于预设的次数，得到与预设的聚类簇数对应的聚类结果。

4. 根据权利要求 1 所述的海量数据结构化方法，其特征在于，所述将所获取的非结构化数据对应转化为正则表达式，包括：

将所获取的与每一簇的聚类一一对应的非结构化数据根据字符类型进行划分，得到与所获取的非结构化数据一一对应的划分结果；其中，每一划分结果均包括多个子字符串；

根据每一划分结果所包括子字符串，生成与每一划分结果对应的多个子正则表达式；

将同一划分结果内对应的多个子正则表达式串接，得到该划分结果对应的正则表达式。

5. 根据权利要求 2 所述海量数据结构化方法，其特征在于，所述将每一簇的聚类中的每一结构化数据根据文本分类算法填充至该聚类对应的数据表格中进行存储，包括：

根据每一簇的聚类的 ID 编号，一一对应建立数据表格；

根据每一簇的聚类中所选定结构化数据的字段，一一对应建立数据表格的字段；

将每一簇的聚类中所包括结构化数据根据 Fasttext 文本分类算法填充至该聚类对应的数据表格的对应字段中。

6. 根据权利要求 4 所述海量数据结构化方法，其特征在于，所述将每一簇的聚类中所包括的非结构化数据均根据该聚类对应的正则表达式进行转化，得到结构化数据，包括：

获取每一簇的聚类所对应的正则表达式所包括的多个子正则表达式，将每一簇的聚类中所包括的非结构化数据均与该聚类对应的多个子正则表达式进行转化，得到与每一非结构化数据对应的多个初始转化结果；

将同一非结构化数据对应的多个初始转化结果进行串接，得到该非结构化数据对应的结构化数据。

7. 根据权利要求 6 所述海量数据结构化方法，其特征在于，所述将同一非结构化数据对应的初始转化结果进行串接，得到该非结构化数据对应的结构化数据，包括：

将同一非结构化数据对应的多个初始转化结果尾部均增加一转行符后再进行串接，得到该非结构化数据对应的结构化数据。

8. 一种海量数据结构化装置，其特征在于，包括：

聚类单元，用于将非结构化数据进行聚类，得到与预设的聚类簇数对应的聚类结果，并设置与每一簇的聚类一一对应的 ID 编号；

正则表达式获取单元，用于在聚类结果中每一簇的聚类中均获取一非结构化数据，并将所获取的非结构化数据对应转化为正则表达式；

数据结构化单元，用于将每一簇的聚类中所包括的非结构化数据均根据该聚类对应的正则表达式进行转化，得到结构化数据。

9. 根据权利要求 8 所述海量数据结构化装置, 其特征在于, 还包括:

分类存储单元, 用于将每一簇的聚类中的每一结构化数据根据文本分类算法填充至该聚类对应的数据表格中进行存储。

10. 根据权利要求 8 所述海量数据结构化装置, 其特征在于, 所述聚类单元包括:

初始中心选择单元, 用于在多个非结构化数据中选取与预设的聚类簇数相同个数的非结构化数据, 将所选取的非结构化数据作为每一簇的初始聚类中心;

初始聚类单元, 用于根据多个非结构化数据与各初始聚类中心的相异值, 将多个非结构化数据进行划分, 得到初始聚类结果;

聚类中心调整单元, 用于根据初始聚类结果, 获取每一簇的调整后聚类中心;

聚类调整单元, 用于根据调整后聚类中心, 将多个非结构化数据进行划分, 直至聚类结果保持相同的次数多于预设的次数, 得到与预设的聚类簇数对应的聚类结果。

11. 根据权利要求 8 所述海量数据结构化装置, 其特征在于, 所述正则表达式获取单元包括:

字符串划分单元, 用于将所获取的与每一簇的聚类一一对应的非结构化数据根据字符类型进行划分, 得到与所获取的非结构化数据一一对应的划分结果; 其中, 每一划分结果均包括多个子字符串;

分段生成单元, 用于根据每一划分结果所包括子字符串, 生成与每一划分结果对应的多个子正则表达式;

串接单元, 用于将同一划分结果内对应的多个子正则表达式串接, 得到该划分结果对应的正则表达式。

12. 根据权利要求 9 所述海量数据结构化装置, 其特征在于, 所述分类存储单元包括:

表格建立单元, 用于根据每一簇的聚类的 ID 编号, 一一对应建立数据表格;

字段建立单元, 用于根据每一簇的聚类中所选定结构化数据的字段, 一一对应建立数据表格的字段;

分类填充单元, 用于将每一簇的聚类中所包括结构化数据根据 Fasttext 文本分类算法填充至该聚类对应的数据表格的对应字段中。

13. 根据权利要求 11 所述海量数据结构化装置，其特征在于，所述数据结构化单元包括：

初始转化结果获取单元，用于获取每一簇的聚类所对应的正则表达式所包括的多个子正则表达式，将每一簇的聚类中所包括的非结构化数据均与该聚类对应的多个子正则表达式进行转化，得到与每一非结构化数据对应的多个初始转化结果；

初始转化结果串接单元，用于将同一非结构化数据对应的多个初始转化结果进行串接，得到该非结构化数据对应的结构化数据。

14. 根据权利要求 13 所述海量数据结构化装置，其特征在于，所述初始转化结果串接单元中，是将同一非结构化数据对应的多个初始转化结果尾部均增加一转行符后再进行串接，得到该非结构化数据对应的结构化数据。

15. 一种计算机设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现以下步骤：

将非结构化数据进行聚类，得到与预设的聚类簇数对应的聚类结果，并设置与每一簇的聚类一一对应的 ID 编号；

在聚类结果中每一簇的聚类中均获取一非结构化数据，并将所获取的非结构化数据对应转化为正则表达式；

将每一簇的聚类中所包括的非结构化数据均根据该聚类对应的正则表达式进行转化，得到结构化数据。

16. 根据权利要求 15 所述的计算机设备，其特征在于，所述将每一簇的聚类中所包括的非结构化数据均根据该聚类对应的正则表达式进行转化，得到结构化数据之后，还包括：

将每一簇的聚类中的每一结构化数据根据文本分类算法填充至该聚类对应的数据表格中进行存储。

17. 根据权利要求 15 所述的计算机设备，其特征在于，所述将非结构化数据进行聚类，得到与预设的聚类簇数对应的聚类结果，包括：

在多个非结构化数据中选取与预设的聚类簇数相同个数的非结构化数据，将所选取的非结构化数据作为每一簇的初始聚类中心；

根据多个非结构化数据与各初始聚类中心的相异值，将多个非结构化数据

进行划分，得到初始聚类结果；

根据初始聚类结果，获取每一簇的调整后聚类中心；

根据调整后聚类中心，将多个非结构化数据进行划分，直至聚类结果保持相同的次数多于预设的次数，得到与预设的聚类簇数对应的聚类结果。

18. 一种存储介质，其特征在于，所述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时使所述处理器执行以下操作：

将非结构化数据进行聚类，得到与预设的聚类簇数对应的聚类结果，并设置与每一簇的聚类一一对应的 ID 编号；

在聚类结果中每一簇的聚类中均获取一非结构化数据，并将所获取的非结构化数据对应转化为正则表达式；

将每一簇的聚类中所包括的非结构化数据均根据该聚类对应的正则表达式进行转化，得到结构化数据。

19. 根据权利要求 18 所述的存储介质，其特征在于，所述将每一簇的聚类中所包括的非结构化数据均根据该聚类对应的正则表达式进行转化，得到结构化数据之后，还包括：

将每一簇的聚类中的每一结构化数据根据文本分类算法填充至该聚类对应的数据表格中进行存储。

20. 根据权利要求 18 所述的存储介质，其特征在于，所述将非结构化数据进行聚类，得到与预设的聚类簇数对应的聚类结果，包括：

在多个非结构化数据中选取与预设的聚类簇数相同个数的非结构化数据，将所选取的非结构化数据作为每一簇的初始聚类中心；

根据多个非结构化数据与各初始聚类中心的相异值，将多个非结构化数据进行划分，得到初始聚类结果；

根据初始聚类结果，获取每一簇的调整后聚类中心；

根据调整后聚类中心，将多个非结构化数据进行划分，直至聚类结果保持相同的次数多于预设的次数，得到与预设的聚类簇数对应的聚类结果。

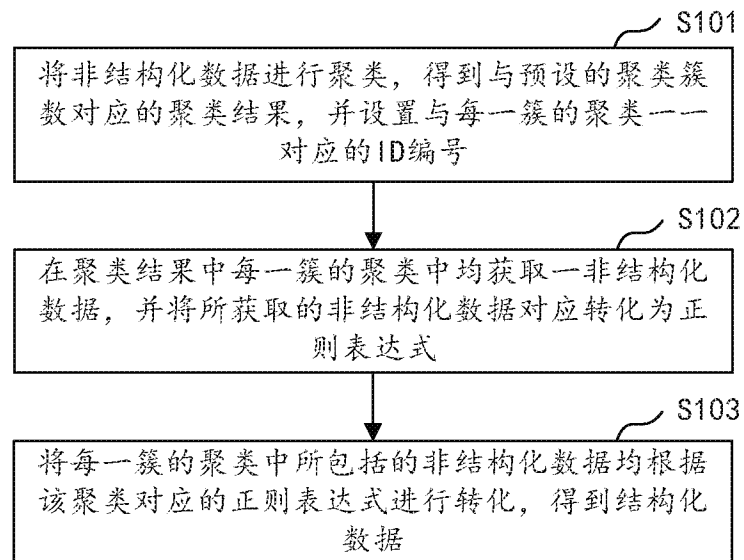


图 1

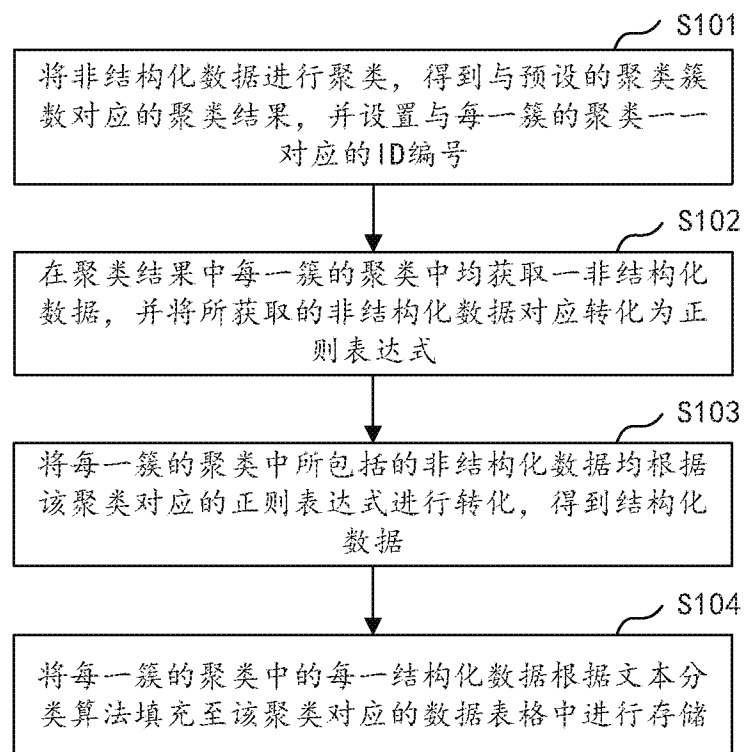


图 2

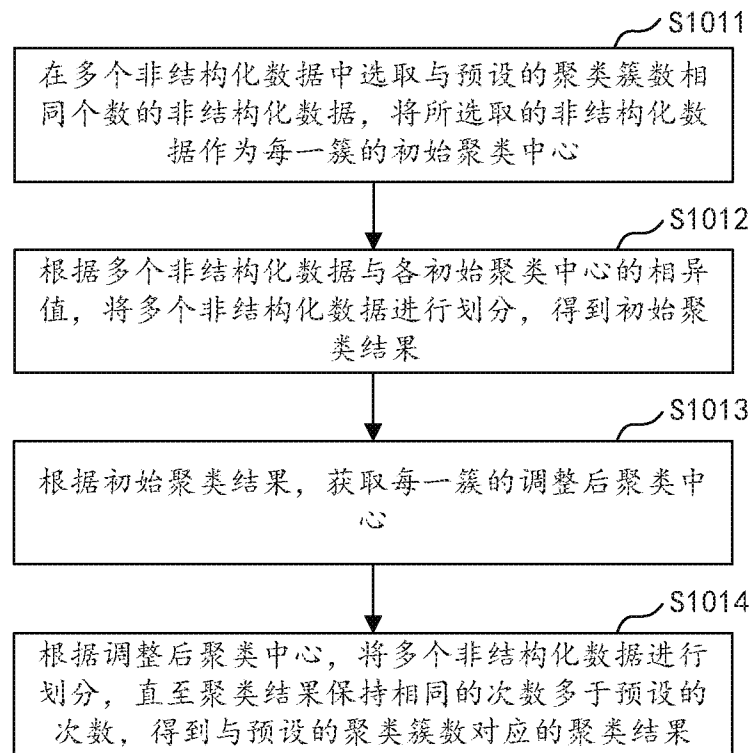


图 3

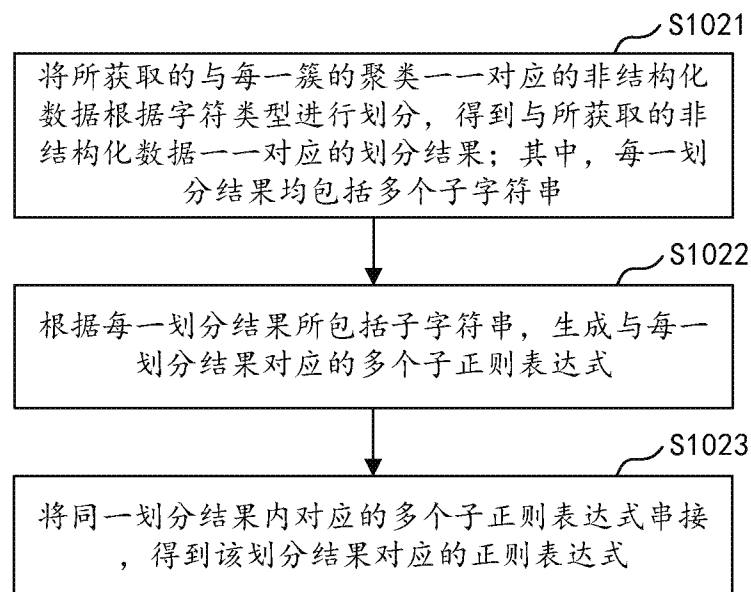


图 4

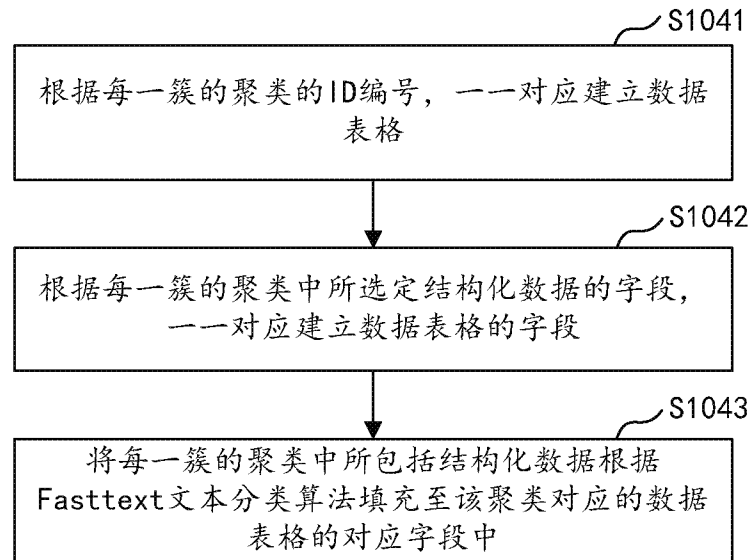


图 5

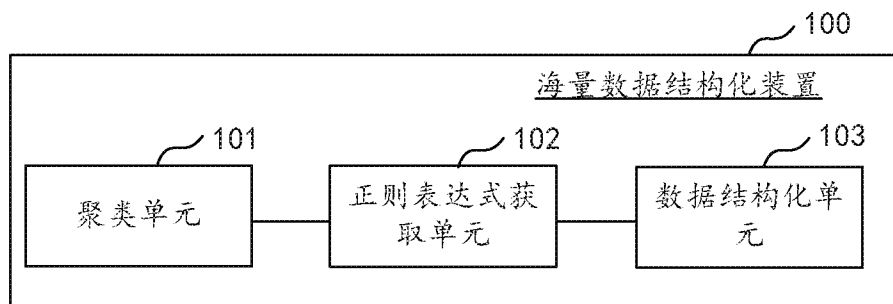


图 6

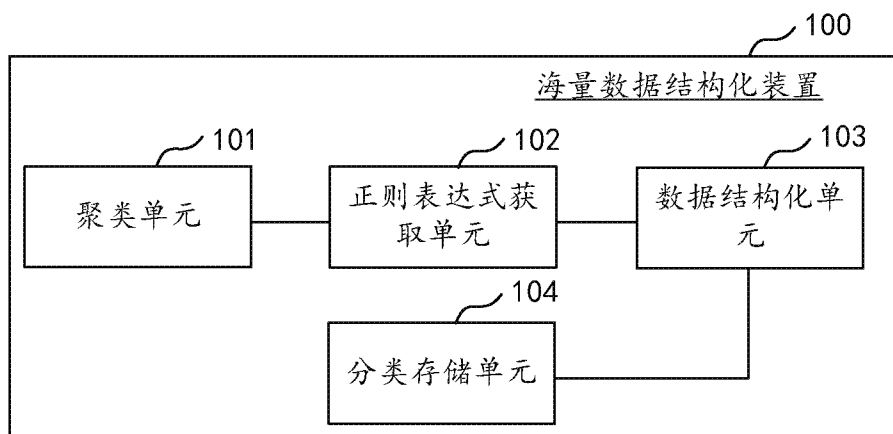


图 7

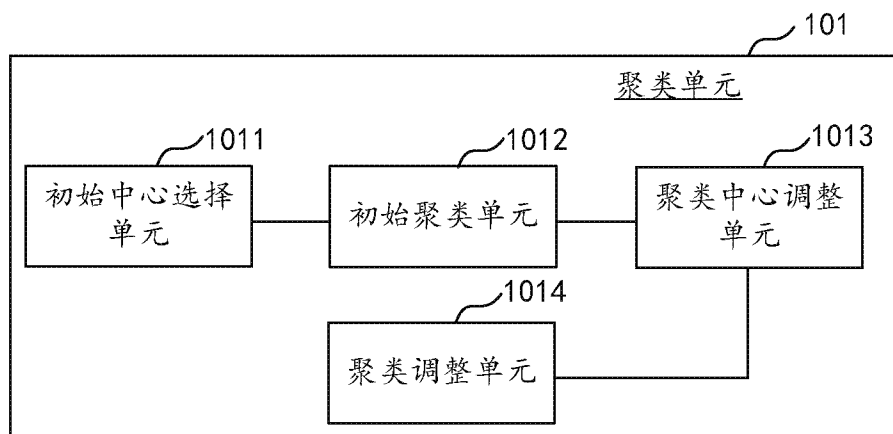


图 8

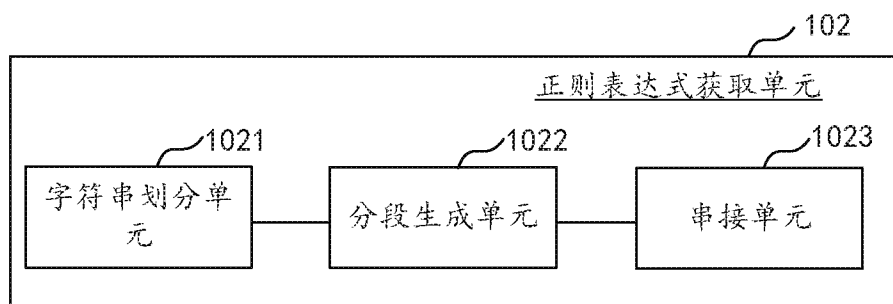


图 9

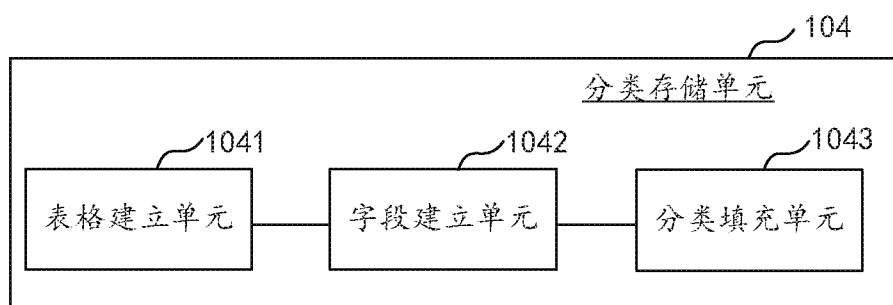


图 10

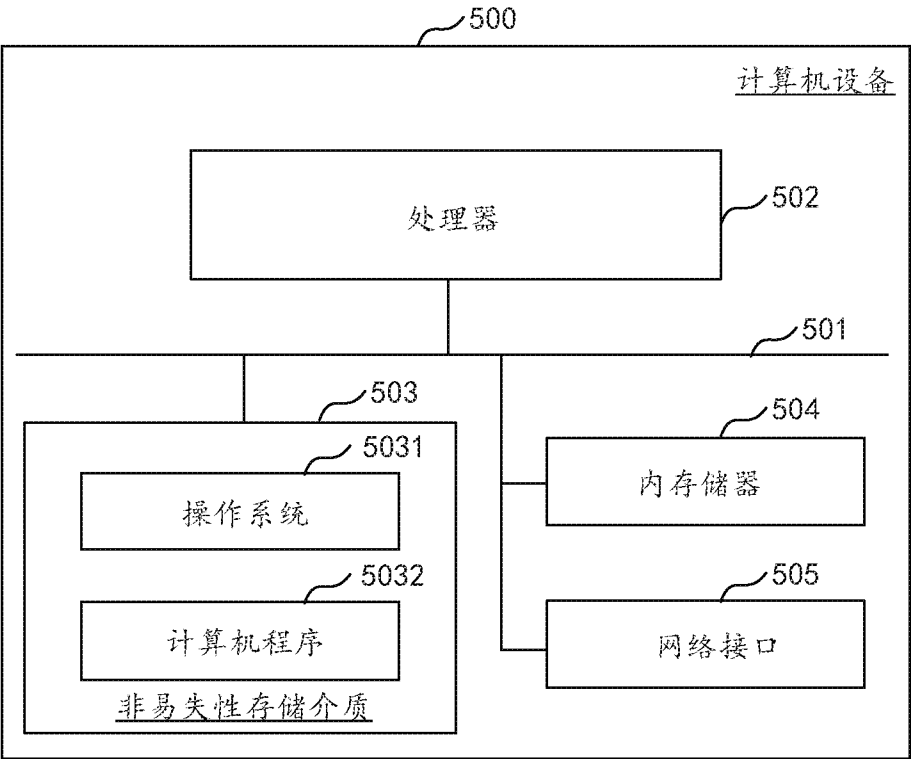


图 11

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/102468

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/35(2019.01)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)

CNABS; CNTXT; CNKI; DWPI; SIPOABS: 数据, 非结构化, 结构化, 聚类, 分类, 相应, 对应, 正则, 表达式, data, unstructured, structured, cluster+, classify, correspond+, regular, expression

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 102346766 A (BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 08 February 2012 (2012-02-08) claims 1 and 2	1-20
Y	CN 105574156 A (HUAWEI TECHNOLOGIES CO., LTD.) 11 May 2016 (2016-05-11) claims 1 and 3	1-20
A	CN 106055608 A (BEIJING BAIDU NETCOM SCIENCE & TECHNOLOGY CO., LTD.) 26 October 2016 (2016-10-26) entire document	1-20
A	CN 102567405 A (BEIJING SAFE-CODE TECHNOLOGY CO., LTD.) 11 July 2012 (2012-07-11) entire document	1-20
A	CN 104881424 A (STATE GRID CORPORATION OF CHINA ET AL.) 02 September 2015 (2015-09-02) 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

12 February 2019

Date of mailing of the international search report

28 February 2019

Name and mailing address of the ISA/CN

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

Authorized officer

Facsimile No. (86-10)62019451

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/102468

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	102346766	A	08 February 2012	None			
CN	105574156	A	11 May 2016	WO	2017101541	A1	22 June 2017
CN	106055608	A	26 October 2016	None			
CN	102567405	A	11 July 2012	None			
CN	104881424	A	02 September 2015	CN	104881424	B	23 November 2018

国际检索报告

国际申请号

PCT/CN2018/102468

A. 主题的分类

G06F 16/35 (2019.01) 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))

CNABS;CNTXT;CNKI;DWPI;SIPOABS:数据, 非结构化, 结构化, 聚类, 分类, 相应, 对应, 正则, 表达式, data, un-structured, structured, cluster+, classify, correspond+, regular, expression

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
Y	CN 102346766 A (北京邮电大学) 2012年 2月 8日 (2012 - 02 - 08) 权利要求1、2	1-20
Y	CN 105574156 A (华为技术有限公司) 2016年 5月 11日 (2016 - 05 - 11) 权利要求1、3	1-20
A	CN 106055608 A (北京百度网讯科技有限公司) 2016年 10月 26日 (2016 - 10 - 26) 全文	1-20
A	CN 102567405 A (北京安码科技有限公司) 2012年 7月 11日 (2012 - 07 - 11) 全文	1-20
A	CN 104881424 A (国家电网公司 等) 2015年 9月 2日 (2015 - 09 - 02) 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

2019年 2月 12日

国际检索报告邮寄日期

2019年 2月 28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

郑嘉青

电话号码 86-(010)-62411684

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/102468

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	102346766	A	2012年 2月 8日	无			
CN	105574156	A	2016年 5月 11日	WO	2017101541	A1	2017年 6月 22日
CN	106055608	A	2016年 10月 26日	无			
CN	102567405	A	2012年 7月 11日	无			
CN	104881424	A	2015年 9月 2日	CN	104881424	B	2018年 11月 23日