(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110019829 A
(43)申请公布日 2019.07.16

(21)申请号 201710848242.X

(22)申请日 2017.09.19

(71)申请人 小草数语(北京)科技有限公司
地址 100089 北京市海淀区西小口路66号中关村东升科技园B-6号楼B座5层

(72)发明人 宋奇 王思睿 姜萌芽 钟磊 秦锋剑

(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
代理人 张润

(51)Int.Cl.
G06F 16/36(2019.01)

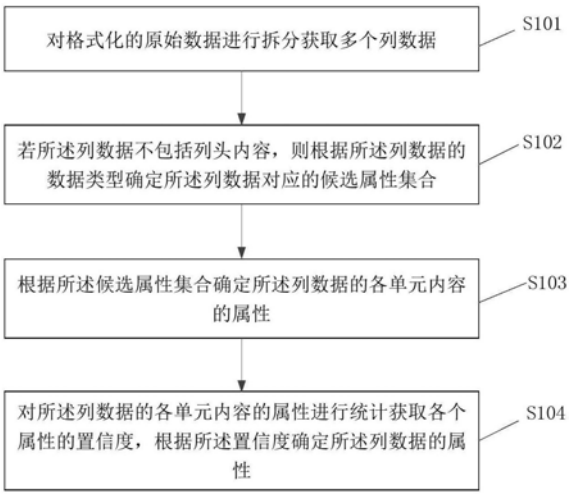
权利要求书2页 说明书13页 附图5页

(54)发明名称

数据属性确定方法、装置

(57)摘要

本发明公开了一种数据属性确定方法及装置,其中,该方法包括:对格式化的原始数据进行拆分获取多个列数据;若所述列数据不包括列头内容,则根据所述列数据的数据类型确定所述列数据对应的候选属性集合;根据所述候选属性集合确定所述列数据的各单元内容的属性;对所述列数据的各单元内容的属性进行统计获取各个属性的置信度,根据所述置信度确定所述列数据的属性。该方法通过分门别类地查找列数据的候选属性集合,以及通过统计各个单元内容的属性确定列数据的属性,实现了尽可能地减少数据属性识别过程中的运算量,提升了数据属性的识别效率和准确率。



CN 110019829 A

1.一种数据属性确定方法,其特征在于,包括:

对格式化的原始数据进行拆分获取多个列数据;

若所述列数据不包括列头内容,则根据所述列数据的数据类型确定所述列数据对应的候选属性集合;

根据所述候选属性集合确定所述列数据的各单元内容的属性;

对所述列数据的各单元内容的属性进行统计获取各个属性的置信度,根据所述置信度确定所述列数据的属性。

2.如权利要求1所述的方法,其特征在于,所述根据所述列数据的数据类型确定所述列数据对应的候选属性集合,包括:

在所述数据类型为数字字母类型时,获取与所述列数据对应的候选正则表达式集合,将候选正则表达式集合确定为所述列数据对应的候选属性集合,其中,所述候选正则表达式集合包括多个正则表达式,所述正则表达式关联数据属性;

所述根据所述候选属性集合确定所述列数据的各单元内容的属性,包括:

将所述单元内容与所述候选正则表达式集合中的各个正则表达式进行逐一匹配;

在所述正则表达式与所述单元内容匹配时,将所述正则表达式关联的数据属性确定为单元内容对应的属性。

3.如权利要求1所述的方法,其特征在于,所述根据所述列数据的数据类型确定所述列数据对应的候选属性集合,包括:

在所述数据类型为非数字字母类型时,获取与所述列数据对应的候选哈希词典集合,将所述候选哈希词典集合确定为所述列数据对应的候选属性集合,其中,所述候选哈希词典集合包括多个哈希词典,所述哈希词典关联数据属性;

所述根据所述候选属性集合确定所述列数据的各单元内容的属性,包括:

将单元内容输入到所述候选哈希词典集合中的各个哈希词典进行逐一查询;

当在所述哈希词典中查询到所述单元内容时,将所述哈希词典关联的数据属性确定为单元内容对应的属性。

4.如权利要求3所述的方法,其特征在于,

所述对所述列数据的各单元内容的属性进行统计获取各个属性的置信度,根据所述置信度确定所述列数据的属性,包括:

对所述列数据的各单元内容的关联于哈希词典的属性进行统计获取各个关联于哈希词典的属性的置信度;

根据关联于哈希词典的置信度确定关联于哈希词典的列数据的属性;

在所述根据关联于哈希词典的置信度确定关联于哈希词典的列数据的属性之后,还包括:

在所述非数字字母类型为短文本且关联于哈希词典的列数据的属性的置信度低于设定阈值时,确定关联于Trie树的列数据的属性;

将关联于哈希词典的列数据的属性与关联于Trie树的列数据的属性进行比较,将置信度大的列数据的属性确定为列数据的目标属性。

5.如权利要求1至4任一项所述的方法,其特征在于,还包括:

在确定所述列数据包括列头内容时,根据所述列头内容查询预设属性映射词典获取与

所述列头内容匹配的属性，将所述与所述列头内容匹配的属性确定为列数据的属性。

6.一种数据属性确定装置，其特征在于，包括：

拆分模块，用于对格式化的原始数据进行拆分获取多个列数据；

第一确定模块，用于若所述列数据不包括列头内容，则根据所述列数据的数据类型确定所述列数据对应的候选属性集合；

第二确定模块，用于根据所述候选属性集合确定所述列数据的各单元内容的属性；

第三确定模块，用于对所述列数据的各单元内容的属性进行统计获取各个属性的置信度，根据所述置信度确定所述列数据的属性。

7.如权利要求6所述的装置，其特征在于，所述第一确定模块包括第一单元；第二确定模块包括第二单元；

所述第一单元，用于在所述数据类型为数字字母类型时，获取与所述列数据对应的候选正则表达式集合，将候选正则表达式集合确定为所述列数据对应的候选属性集合，其中，所述候选正则表达式集合包括多个正则表达式，所述正则表达式关联数据属性；

所述第二单元，用于将所述单元内容与所述候选正则表达式集合中的各个正则表达式进行逐一匹配；在所述正则表达式与所述单元内容匹配时，将所述正则表达式关联的数据属性确定为单元内容对应的属性。

8.如权利要求6所述的装置，其特征在于，所述第一确定模块还包括第三单元；所述第二确定模块还包括第四单元；

所述第三单元，用于在所述数据类型为非数字字母类型时，获取与所述列数据对应的候选哈希词典集合，将所述候选哈希词典集合确定为所述列数据对应的候选属性集合，其中，所述候选哈希词典集合包括多个哈希词典，所述哈希词典关联数据属性；

所述第四单元，用于将单元内容输入到所述候选哈希词典集合中的各个哈希词典进行逐一查询；当在所述哈希词典中查询到所述单元内容时，将所述哈希词典关联的数据属性确定为单元内容对应的属性。

9.一种计算机设备，其特征在于，包括：处理器和存储器；

其中，所述处理器通过读取所述存储器中存储的可执行程序代码来运行与所述可执行程序代码对应的程序，以用于实现如权利要求1至5中任一项所述的数据属性确定方法。

10.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该计算机程序被处理器执行时实现如权利要求1至5中任一项数据属性确定方法。

数据属性确定方法、装置

技术领域

[0001] 本发明涉及数据挖掘技术领域，尤其涉及一种数据属性确定方法、装置。

背景技术

[0002] 在关系图谱相关的大数据分析中，对结构化的原始数据进行按项拆分和识别，以便把数据映射到实体－属性，便于实体建模和分析。

[0003] 典型的应用场景是：对于一个结构化完好的excel输入表格，通过依据表格的文件名、表头、每列内容，可以把整个表格内容对应的各列属性猜测出来，并对应到E-R(entity relationship model，实体联系模式)图的相关模型属性。这样可以实现原始输入信息到图谱模型的关系映射，便于后续图谱挖掘等更深操作。然而，如何提高数据属性的识别率一直是亟待解决的技术问题。

发明内容

[0004] 本发明的目的旨在至少在一定程度上解决上述的技术问题之一。

[0005] 为此，本发明的第一个目的在于提出一种数据属性确定方法，通过分门别类地查找列数据的候选属性集合，以及通过统计各个单元内容的属性确定列数据的属性，实现了尽可能地减少数据属性识别过程中的运算量，提升了数据属性的识别效率和准确率。

[0006] 本发明的第二个目的在于提出一种数据属性确定装置。

[0007] 本发明的第三个目的在于提出一种计算机设备。

[0008] 本发明的第四个目的在于提出一种计算机可读存储介质。

[0009] 为了实现上述目的，本发明第一方面实施例的数据属性确定方法，包括：对格式化的原始数据进行拆分获取多个列数据；

[0010] 若所述列数据不包括列头内容，则根据所述列数据的数据类型确定所述列数据对应的候选属性集合；

[0011] 根据所述候选属性集合确定所述列数据的各单元内容的属性；

[0012] 对所述列数据的各单元内容的属性进行统计获取各个属性的置信度，根据所述置信度确定所述列数据的属性。

[0013] 如上所述的方法，所述根据所述列数据的数据类型确定所述列数据对应的候选属性集合，包括：

[0014] 在所述数据类型为数字字母类型时，获取与所述列数据对应的候选正则表达式集合，将候选正则表达式集合确定为所述列数据对应的候选属性集合，其中，所述候选正则表达式集合包括多个正则表达式，所述正则表达式关联数据属性；

[0015] 所述根据所述候选属性集合确定所述列数据的各单元内容的属性，包括：

[0016] 将所述单元内容与所述候选正则表达式集合中的各个正则表达式进行逐一匹配；

[0017] 在所述正则表达式与所述单元内容匹配时，将所述正则表达式关联的数据属性确定为单元内容对应的属性。

[0018] 如上所述的方法，所述根据所述列数据的数据类型确定所述列数据对应的候选属性集合，包括：

[0019] 在所述数据类型为非数字字母类型时，获取与所述列数据对应的候选哈希词典集合，将所述候选哈希词典集合确定为所述列数据对应的候选属性集合，其中，所述候选哈希词典集合包括多个哈希词典，所述哈希词典关联数据属性；

[0020] 所述根据所述候选属性集合确定所述列数据的各单元内容的属性，包括：

[0021] 将单元内容输入到所述候选哈希词典集合中的各个哈希词典进行逐一查询；

[0022] 当在所述哈希词典中查询到所述单元内容时，将所述哈希词典关联的数据属性确定为单元内容对应的属性。

[0023] 如上所述的方法，所述对所述列数据的各单元内容的属性进行统计获取各个属性的置信度，根据所述置信度确定所述列数据的属性，包括：

[0024] 对所述列数据的各单元内容的关联于哈希词典的属性进行统计获取各个关联于哈希词典的属性的置信度；

[0025] 根据关联于哈希词典的置信度确定关联于哈希词典的列数据的属性；

[0026] 在所述根据关联于哈希词典的置信度确定关联于哈希词典的列数据的属性之后，还包括：

[0027] 在所述非数字字母类型为短文本且关联于哈希词典的列数据的属性的置信度低于设定阈值时，确定关联于Trie树的列数据的属性；

[0028] 将关联于哈希词典的列数据的属性与关联于Trie树的列数据的属性进行比较，将置信度大的列数据的属性确定为列数据的目标属性。

[0029] 如上所述的方法，还包括：

[0030] 在确定所述列数据包括列头内容时，根据所述列头内容查询预设属性映射词典获取与所述列头内容匹配的属性，将所述与所述列头内容匹配的属性确定为列数据的属性。

[0031] 为了实现上述目的，本发明第二方面实施例的数据属性确定装置，包括：拆分模块，用于对格式化的原始数据进行拆分获取多个列数据；

[0032] 第一确定模块，用于若所述列数据不包括列头内容，则根据所述列数据的数据类型确定所述列数据对应的候选属性集合；

[0033] 第二确定模块，用于根据所述候选属性集合确定所述列数据的各单元内容的属性；

[0034] 第三确定模块，用于对所述列数据的各单元内容的属性进行统计获取各个属性的置信度，根据所述置信度确定所述列数据的属性。

[0035] 为了实现上述目的，本发明第三方面实施例的计算机设备，包括：处理器和存储器；

[0036] 其中，所述处理器通过读取所述存储器中存储的可执行程序代码来运行与所述可执行程序代码对应的程序，以用于实现第一方面所述的数据属性确定方法。

[0037] 为了实现上述目的，本发明第四方面实施例的计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时实现第一方面所述的数据属性确定方法。

[0038] 本发明附加的方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

[0039] 本发明上述的和/或附加的方面和优点从下面结合附图对实施例的描述中将变得明显和容易理解,其中,

[0040] 图1是本发明一实施例的数据属性确定方法的流程图;

[0041] 图2是本发明又一实施例的数据属性确定方法的流程图;

[0042] 图3是本发明另一实施例的数据属性确定方法的流程图;

[0043] 图4是本发明再一实施例的数据属性确定方法的流程图;

[0044] 图5是本发明一实施例的数据属性确定方法的流程图;

[0045] 图6是本发明一实施例的数据属性确定装置的结构示意图;

[0046] 图7是本发明一实施例的计算机设备的结构示意图。

具体实施方式

[0047] 下面详细描述本发明的实施例,实施例的示例在附图中示出,其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的,旨在用于解释本发明,而不能理解为对本发明的限制。

[0048] 下面参考附图描述本发明实施例的数据属性确定方法及装置。

[0049] 图1是本发明一实施例的数据属性确定方法的流程图。本实施例的数据属性确定方法由数据属性确定装置执行,该装置可以集成在服务器中。

[0050] 如图1所示,本实施例的数据属性确定方法,包括:

[0051] 步骤S101、对格式化的原始数据进行拆分获取多个列数据。

[0052] 具体地,格式化的原始数据蕴藏海量的有价值的信息宝库,对格式化的原始数据进行数据挖掘能够帮助用户更为科学的决策。本实施例中的格式化的原始数据可以是各类报表,诸如客户名单、产品清单、物品清单、订单、发货单等数据的报表,报表的形式可以是PDF、Word、Excel和Power Point等。

[0053] 以表1为例,表1的文件名为物流管理报表,表1共有3列数据,各列数据的列头内容分别为:运单号、发件公司、收件公司。一般来说,列头内容即为对应列数据的属性,即表1中的运单号所在列的属性为运单号,表1中的发件公司所在列的属性为发件公司,以此类推。

[0054] 当集成了数据属性确定装置的服务器采集到大量的诸如表1所示的物流管理报表,首先,识别物流管理报表的结构形式,确定物流管理报表的列数;接着,按列拆分,即得到3列数据,分别为运单号、发件公司、收件公司;再接着,根据列内容或是列数据中的各个单元内容挖掘各列数据的属性。

[0055] 表1

[0056]

运单号	发件公司	收件公司
4506442377787	江西公司	山东公司
4523447706787	北京公司	山西公司
8744235077647	海南公司	河南公司
7643507474287	湖南公司	河北公司

3587442077647	上海公司	江西公司

[0057] 步骤S102、若所述列数据不包括列头内容，则根据所述列数据的数据类型确定所述列数据对应的候选属性集合。

[0058] 举例来说，服务器在采集到诸如表1所示的列数据即包括列头内容，又包括单元内容时，由于列头内容表征列数据的属性，这时服务器会优先识别列头内容来挖掘列数据的属性。然而，并不是采集到的格式化的原始数据都包括列头，这时，服务器根据各列数据的各个单元内容来挖掘列数据的属性。

[0059] 具体地，列数据的属性的数据类型可以分成数字字母类型和非数字字母类型。数字字母类型的属性对应的单元内容可以是由数字、26个英文字母、下划线、空格、制表符、换页符等组成的内容，诸如运单号、邮政编码、身份证、MAC地址等属性对应的单元内容。非数字字母类型可以是枚举类型、短文本类型。枚举类型的属性有民族、性别等；短文本类型的属性有公司、学校等。

[0060] 举例来说，当服务器采集到表1中的运单号对应的列数据时，首先，服务器会对3个单元内容进行识别、分类、聚类等操作，猜测关联于运单号的列数据的属性可能为运单号、邮政编码、身份证等数据字母类型属性，这时服务器调取预先存储的运单号属性识别模型、邮政编码的属性识别模型、身份证属性识别模型。如果服务器针对不同的物流公司建立了不同运单号属性识别模型，此时服务器会调取出所有运单号属性识别模型。所有运单号属性识别模型、邮政编码的属性识别模型、身份证属性识别模型即组成本实施例中的列数据对应的候选属性集合(即关联于运单号的列数据对应的候选属性集合)

[0061] 举例来说，当服务器采集到表1中的发件公司对应的列数据时，首先，服务器会对3个单元内容进行识别、分类、聚类等操作，猜测关联于发件公司的列数据的属性可能为发件公司、收件公司、中转公司等非数据字母类型属性，这时服务器调取预先存储的发件公司属性识别模型、收件公司属性识别模型、中转公司属性识别模型，如果服务器针对不同的发件公司、收件公司、中转公司分别建立了相应的属性识别模型，这时服务器会调取出所有发件公司属性识别模型、收件公司属性识别模型、中转公司属性识别模型。所有发件公司属性识别模型、收件公司属性识别模型、中转公司属性识别模型即组成本实施例中的列数据对应的候选属性集合(即关联于发件公司的列数据对应的候选属性集合)。

[0062] 需要指出的是，诸如运单号属性识别模型、邮政编码的属性识别模型、身份证属性识别模型、发件公司属性识别模型、收件公司属性识别模型、中转公司属性识别模型等示例性的属性识别模型，由数据挖掘开发公司根据实际需求建立。举例来说，数据挖掘开发公司通过采集海量的历史格式化的原始数据，通过机器学习等分析方法建立相应的属性识别模型，或者是，数据挖掘开发公司通过统计分析海量的历史格式化的原始数据，基于预设算法建立相应的属性识别模型。举例来说，属性识别模型可以是基于属性建立的哈希词典、基于属性建立的Trie树、基于属性建立的正则表达式等识别模型。

[0063] 本实施例在确定列数据不包括列头内容时，通过先确定列数据的数据类型进行分门别类地查找列数据的候选属性集合，尽可能地减少数据属性识别过程中的运算量，提升数据属性的识别效率和准确率。

[0064] 步骤S103、根据所述候选属性集合确定所述列数据的各单元内容的属性。

[0065] 步骤S104、对所述列数据的各单元内容的属性进行统计获取各个属性的置信度，

根据所述置信度确定所述列数据的属性。

[0066] 举例来说，当服务器对表1中的发件公司对应的5个单元内容中的属性进行识别时，将各单元内容输入到候选属性集合进行逐一匹配。其中，候选属性集合包括：发件公司属性识别模型、收件公司属性识别模型、中转公司属性识别模型。

[0067] 具体的，将江西公司依次输入到发件公司属性识别模型、收件公司属性识别模型、中转公司属性识别模型进行逐一匹配时，发现江西公司输入到收件公司属性模型进行匹配，匹配成功，则确定江西公司对应的单元内容的属性为收件公司。以此类推，确定北京公司对应的单元内容的属性为发件公司；确定海南公司对应的单元内容的属性为发件公司；确定湖南公司对应的单元内容的属性为发件公司；确定上海公司对应的单元内容的属性为发件公司。

[0068] 具体的，经过统计，关联于发件公司的列数据的属性存在两种，分别为发件公司、收件公司；其中，四个单元内容的属性识别为发件公司，一个单元内容的属性识别为收件公司。经过计算，列数据的属性为发件公司的概率为80％，

[0069] 列数据的属性为收件公司的概率为20％。本实施例中属性的置信度可以理解为属性的概率，比如，列数据的属性为发件公司的概率为80％即为列数据的属性为发件公司的置信度为80％，列数据的属性为收件公司的概率为20％即为列数据的属性为收件公司的置信度为20％。

[0070] 其中，根据置信度确定所述列数据的属性的具体实现方式为：

[0071] 第一种实现方式，对同一列数据对应的各个置信度进行比较，选取置信度最大对应的属性确定为列数据的属性。

[0072] 第二种实现方式，确定同一列数据对应的各个置信度是否满足设定条件，将符合条件的置信度对应的属性确定为列数据的属性。其中，设定条件可以是将置信度与设定阈值进行大小比较，也可以是确定置信度是否在设定范围内，但并不以此为限。需要指出的是，可能会出现多个符合条件的置信度，相应地，所确定的列数据的属性可能会有多个。

[0073] 第三种实现方式，将同一列数据对应的各个置信度均呈现给用户，由用户根据置信度自主选择列数据的属性。

[0074] 本实施例提供的数据属性确定方法，包括：对格式化的原始数据进行拆分获取多个列数据；若所述列数据不包括列头内容，则根据所述列数据的数据类型确定所述列数据对应的候选属性集合；根据所述候选属性集合确定所述列数据的各单元内容的属性；对所述列数据的各单元内容的属性进行统计获取各个属性的置信度，根据所述置信度确定所述列数据的属性。该方法通过分门别类地查找列数据的候选属性集合，以及通过统计各个单元内容的属性确定列数据的属性，实现了尽可能地减少数据属性识别过程中的运算量，提升了数据属性的识别效率和准确率。

[0075] 图2是本发明又一实施例的数据属性确定方法的流程图。在上述实施例的基础上，当列数据具有列头内容时，利用列头内容可以简单高效地确定更为准确的列数据的属性，也加快了列数据的属性的识别效率。

[0076] 如图2所示，本实施例的数据属性确定方法，包括：

[0077] 步骤S201、对格式化的原始数据进行拆分获取多个列数据，执行步骤S202或步骤S205；

[0078] 步骤S202、若所述列数据不包括列头内容，则根据所述列数据的数据类型确定所述列数据对应的候选属性集合，执行步骤S203；

[0079] 步骤S203、根据所述候选属性集合确定所述列数据的各单元内容的属性，执行步骤S204。

[0080] 步骤S204、对所述列数据的各单元内容的属性进行统计获取各个属性的置信度，根据所述置信度确定所述列数据的属性。

[0081] 本实施例中的步骤S201、S202、S203、S204的实现方式分别与上述实施例中的S101、S102、S103、S104的实现方式相同，在此不再赘述。

[0082] 步骤S205、在确定所述列数据包括列头内容时，根据所述列头内容查询预设属性映射词典获取与所述列头内容匹配的属性，将所述与所述列头内容匹配的属性确定为列数据的属性。

[0083] 举例来说，用户设计报表时，列头内容所表征的即为列数据的属性，诸如表1所示的列数据包括运单号、发件公司、收件公司等列头内容。

[0084] 具体的，预设属性映射词典事先存储在服务器中，预设属性映射词典由数据挖掘开发公司根据各行各业的特点进行设计，并不断更新。预设属性映射词典中包含的属性是经过专业人员认证的属性，权威性高。本实施例通过将列内容输入到预设属性映射词典中，将在预设属性映射词典查到属性确定为列数据的属性，方法简单高效，所确定的列数据属性更为精准。

[0085] 本实施例提供的数据属性确定方法，在确定所述列数据包括列头内容时，根据所述列头内容查询预设属性映射词典获取与所述列头内容匹配的属性，将所述与所述列头内容匹配的属性确定为列数据的属性，方法简单高效，所确定的列数据属性更为精准。

[0086] 图3是本发明另一实施例的数据属性确定方法的流程图。在上述实施例的基础上，在所述数据类型为数字字母类型时，通过将单元内容与正则表达式进行逻辑判断，确定单元内容的属性，进而确定列数据的属性。

[0087] 如图3所示，本实施例的数据属性确定方法，包括：

[0088] 步骤S301、对格式化的原始数据进行拆分获取多个列数据，执行步骤S302；

[0089] 步骤S302、若所述列数据不包括列头内容，则根据所述列数据的数据类型确定所述列数据对应的候选属性集合，执行步骤S303；

[0090] 步骤S303、在所述数据类型为数字字母类型时，获取与所述列数据对应的候选正则表达式集合，将候选正则表达式集合确定为所述列数据对应的候选属性集合，执行步骤S304。

[0091] 其中，所述候选正则表达式集合包括多个正则表达式，所述正则表达式关联数据属性。

[0092] 具体地，正则表达式是对字符串操作的一种逻辑公式，就是用事先定义好的一些特定字符、及这些特定字符的组合，组成一个“规则字符串”，这个“规则字符串”用来表达对字符串的一种过滤逻辑。正则表达式灵活性、逻辑性和功能性非常强；正则表达式用于字符串处理、表单验证等场合，实用高效。

[0093] 举例来说，服务器预先存储了各种正则表达式，一种正则表达式关联一种属性。比如，预存了运单号正则表达式、邮政编码正则表达式、身份证正则表达式。

[0094] 当服务器采集到表1中的运单号对应的列数据时，首先，服务器会对5个单元内容进行识别、分类、聚类等操作，猜测关联于运单号对应的列数据可能为运单号、邮政编码、身份证等数据字母类型属性。这时服务器调取预先存储的运单号正则表达式、邮政编码正则表达式、身份证正则表达式，即本实施例中的候选正则表达式集合包括运单号正则表达式、邮政编码正则表达式、身份证正则表达式。

[0095] 步骤S304、将所述单元内容与所述候选正则表达式集合中的各个正则表达式进行逐一匹配，执行步骤S305。

[0096] 步骤S305、在所述正则表达式与所述单元内容匹配时，将所述正则表达式关联的数据属性确定为单元内容对应的属性，执行步骤S306。

[0097] 步骤S306、对所述列数据的各单元内容的属性进行统计获取各个属性的置信度，根据所述置信度确定所述列数据的属性。

[0098] 举例来说，当服务器对表1中的运单号对应的5个单元内容中的属性进行识别时，将各单元内容输入到候选正则表达式集合进行逐一匹配。

[0099] 具体的，经过统计，关联于运单号的列数据的属性存在三种，分别为运单号、邮政编码、身份证；其中，3个单元内容的属性识别为运单号，1个单元内容的属性识别为邮政编码，1个单元内容的属性识别为身份证。

[0100] 经过计算，列数据的属性为运单号的概率(概率可以理解为置信度)为60％，列数据的属性为邮政编码的概率(概率可以理解为置信度)为20％；列数据的属性为身份证的概率(概率可以理解为置信度)为20％。

[0101] 举例来说，对同一列数据对应的各个置信度进行比较，选取置信度最大对应的属性确定为列数据的属性。那么，上述举例中置信度为60％为最大置信度，这时，所确定列数据的属性为运单号。

[0102] 本实施例的数据属性确定方法，在确定列数据的数据类型为数字字母类型时，先获取与所述列数据对应的候选正则表达式集合，接着，将所述单元内容与所述候选正则表达式集合中的各个正则表达式进行逐一匹配以确定为单元内容对应的属性，最后，对所述列数据的各单元内容的属性进行统计确定列数据的属性。该方法利用正则表达式来确定数字字母类型的列数据的属性，实用高效。由于正则表达式灵活性、逻辑性和功能性非常强，根据新增的属性编辑不同的正则表达式，具有很好的扩充性。

[0103] 图4是本发明再一实施例的数据属性确定方法的流程图。在上述实施例的基础上，在所述数据类型为非数字字母类型时，通过将单元内容与哈希词典进行逻辑判断，确定单元内容的属性，进而确定列数据的属性。

[0104] 如图4所示，本实施例的数据属性确定方法，包括：

[0105] 步骤S401、对格式化的原始数据进行拆分获取多个列数据，执行步骤S402；

[0106] 步骤S402、若所述列数据不包括列头内容，则根据所述列数据的数据类型确定所述列数据对应的候选属性集合，执行步骤S403；

[0107] 步骤S403、在所述数据类型为非数字字母类型时，获取与所述列数据对应的候选哈希词典集合，将所述候选哈希词典集合确定为所述列数据对应的候选属性集合，其中，所述候选哈希词典集合包括多个哈希词典，所述哈希词典关联数据属性，执行步骤S404。

[0108] 在此简要介绍哈希表：散列表(Hash table，也叫哈希表)是根据关键字(Key

value)而直接访问在内存存储位置的数据结构，即哈希表通过计算一个关于键值的函数，将所需查询的数据映射到表中一个位置来访问记录，加快了查找速度。其中，映射函数称做散列函数，存放记录的数组称做散列表，关键字和函数法则理论上可以任意确定。

[0109] 为了加快数据属性的查询速度，本实施例中的哈希词典可以建立根据行业的不同建立相应的首字母散列表。以表1为例，需要建立的哈希词典为发件公司哈希词典、收件公司哈希词典；在发件公司哈希词典中，分别建立了对应江西公司的首字母散列表、对应北京公司的首字母散列表、对应海南公司的首字母散列表、对应湖南公司的首字母散列表、对应上海公司的首字母散列表。当需要确定各个单元内容的属性时，只需将首字母在哈希词典中进行快速搜索。比如，确定单元内容为江西公司的属性时，将首字母j输入到发件公司哈希词典中，若在哈希词典中查询到江西公司，则确定单元内容为江西公司的发件公司。

[0110] 步骤S404、将单元内容输入到所述候选哈希词典集合中的各个哈希词典进行逐一查询，执行步骤S405；

[0111] 步骤S405、当在所述哈希词典中查询到所述单元内容时，将所述哈希词典关联的数据属性确定为单元内容对应的属性，执行步骤S406。

[0112] 步骤S406、对所述列数据的各单元内容的属性进行统计获取各个属性的置信度，根据所述置信度确定所述列数据的属性。

[0113] 举例来说，当服务器对表1中的发件公司对应的5个单元内容中的属性进行识别时，将各单元内容输入到候选哈希词典集合进行逐一查询。

[0114] 具体的，经过统计，关联于发件公司的列数据的属性存在二种，分别为发件公司、收件公司；其中，4个单元内容的属性识别为发件公司，1个单元内容的属性识别为收件公司。

[0115] 经过计算，列数据的属性为发件公司的概率(概率可以理解为置信度)为80％，列数据的属性为收件公司的概率(概率可以理解为置信度)为20％。

[0116] 举例来说，对同一列数据对应的各个置信度进行比较，选取置信度最大对应的属性确定为列数据的属性。那么，上述举例中置信度为80％为最大置信度，这时，所确定列数据的属性为发件公司。

[0117] 本实施例的数据属性确定方法，在所述数据类型为非数字字母类型时，获取与所述列数据对应的候选哈希词典集合；接着，将单元内容输入到所述候选哈希词典集合中的各个哈希词典进行逐一查询；当在所述哈希词典中查询到所述单元内容时，将所述哈希词典关联的数据属性确定为单元内容对应的属性；最后，对所述列数据的各单元内容的属性进行统计确定列数据的属性。该方法利用哈希词典快速查询非数字字母类型的列数据的属性；建立哈希词典时，其中的哈希表中的关键字和函数法则理论上可以任意确定，根据新增的属性编辑不同的哈希表，具有很好的扩充性。

[0118] 图5是本发明一实施例的数据属性确定方法的流程图。在上述实施例的基础上，对于诸如公司、学校等短文本的属性，由于各公司的名称、各学校的名称不断地更新变化，在哈希词典中可能不包括刚出现的公司名称、学校名称等，这时可能会出现选出置信度不高的列数据属性。针对上述情形，本实施例在利用哈希词典来确定列数据的属性之后，还通过Trie树来进一步确定列数据的属性，选出置信度较高的列数据属性作为列数据的目标属性。

[0119] 如图5所示，本实施例的数据属性确定方法，包括：

[0120] 步骤S501、对格式化的原始数据进行拆分获取多个列数据，执行步骤S502；

[0121] 步骤S502、若所述列数据不包括列头内容，则根据所述列数据的数据类型确定所述列数据对应的候选属性集合，执行步骤S503；

[0122] 步骤S503、在所述数据类型为非数字字母类型时，获取与所述列数据对应的候选哈希词典集合，将所述候选哈希词典集合确定为所述列数据对应的候选属性集合，其中，所述候选哈希词典集合包括多个哈希词典，所述哈希词典关联数据属性，执行步骤S504。

[0123] 步骤S504、将单元内容输入到所述候选哈希词典集合中的各个哈希词典进行逐一查询，执行步骤S505；

[0124] 步骤S505、当在所述哈希词典中查询到所述单元内容时，将所述哈希词典关联的数据属性确定为单元内容对应的属性，执行步骤S506。

[0125] 步骤S506、对所述列数据的各单元内容的关联于哈希词典的属性进行统计获取各个关联于哈希词典的属性的置信度；根据关联于哈希词典的置信度确定关联于哈希词典的列数据的属性，执行步骤S507。

[0126] 步骤S507、所述非数字字母类型为短文本且关联于哈希词典的列数据的属性的置信度低于设定阈值时，确定关联于Trie树的列数据的属性，执行步骤S508。

[0127] 具体地，本实施例中的非数字字母类型可以是枚举类型，也可以是短文本类型。诸如民族、性别等为枚举类型的属性，列数据的属性为性别为例，列数据的单元内容要么为女，要么对男。枚举类型的属性通过哈希词典就可以明确确定。

[0128] 而对于诸如公司、学校等短文本的属性，由于各公司的名称、各学校的名称不断地更新变化，在哈希词典中可能不包括刚出现的公司名称、学校名称等，这时可能会出现选出置信度不高的列数据属性。

[0129] 针对上述情形，本实施例在利用哈希词典来确定列数据的属性之后，还通过Trie树来进一步确定列数据的属性，选出置信度较高的列数据属性作为列数据的目标属性。

[0130] Trie树，又称单词查找树，是一种树形结构，是一种哈希树的变种。典型应用是用于统计，排序和保存大量的字符串(但不仅限于字符串)，所以经常被搜索引擎系统用于文本词频统计。它的优点是：利用字符串的公共前缀来减少查询时间，最大限度地减少无谓的字符串比较，查询效率比哈希树高。

[0131] 在本实施例中，确定关联于Trie树的列数据的属性的具体实现方式为：

[0132] S1、获取与所述列数据对应的候选Trie树集合，将所述候选Trie树集合确定为所述列数据对应的候选属性集合，其中，所述候选Trie树集合包括多个Trie树，所述Trie树关联数据属性。

[0133] 以表1为例，需要建立的Trie树为发件公司Trie树、收件公司Trie树；在发件公司Trie树中，根据大数据挖掘的诸如江西公司、北京公司、海南公司、湖南公司、上海公司等等众多发件公司建立Trie树。

[0134] S2、将单元内容输入到所述候选Trie树集合中的各个Trie树进行逐一查询。

[0135] S3、当在所述Trie树中查询到所述单元内容时，将所述Trie树关联的数据属性确定为单元内容对应的属性。

[0136] 步骤S508、将关联于哈希词典的列数据的属性与关联于Trie树的列数据的属性进

行比较，将置信度大的列数据的属性确定为列数据的目标属性。

[0137] 举例来说，对发件公司所在列的属性确定，若根据哈希词典确定的列数据属性为收件公司，对应的置信度为40％(设定阈值为50％)；根据Trie树确定的列数据属性为发件公司，对应的置信度为30％，这时，根据哈希词典确定的列数据属性确定为列数据的目标属性(属性为收件公司)。若根据哈希词典确定的列数据属性为收件公司，对应的置信度为40％(设定阈值为50％)；根据Trie树确定的列数据属性为发件公司，对应的置信度为50％，这时，根据Trie树确定的列数据属性确定为列数据的目标属性(属性为发件公司)。

[0138] 本实施例的数据属性确定方法，非数字字母类型为短文本且关联于哈希词典的列数据的属性的置信度低于设定阈值时，确定关联于Trie树的列数据的属性，将关联于哈希词典的列数据的属性与关联于Trie树的列数据的属性进行比较，将置信度大的列数据的属性确定为列数据的目标属性。针对利用前期利用哈希词表选出的列数据属性不高的情形，还通过Trie树来进一步确定列数据的属性，选出置信度较高的列数据属性作为列数据的目标属性，提升数据属性的识别准确度。

[0139] 图6是本发明一实施例的数据属性确定装置的结构示意图。如图6所示，本实施例提供的数据属性确定装置，包括：

[0140] 拆分模块01，用于对格式化的原始数据进行拆分获取多个列数据；

[0141] 第一确定模块02，用于若所述列数据不包括列头内容，则根据所述列数据的数据类型确定所述列数据对应的候选属性集合；

[0142] 第二确定模块03，用于根据所述候选属性集合确定所述列数据的各单元内容的属性；

[0143] 第三确定模块04，用于对所述列数据的各单元内容的属性进行统计获取各个属性的置信度，根据所述置信度确定所述列数据的属性。

[0144] 进一步地，所述第一确定模块包括第一单元；第二确定模块包括第二单元；

[0145] 所述第一单元，用于在所述数据类型为数字字母类型时，获取与所述列数据对应的候选正则表达式集合，将候选正则表达式集合确定为所述列数据对应的候选属性集合，其中，所述候选正则表达式集合包括多个正则表达式，所述正则表达式关联数据属性；

[0146] 所述第二单元，用于将所述单元内容与所述候选正则表达式集合中的各个正则表达式进行逐一匹配；在所述正则表达式与所述单元内容匹配时，将所述正则表达式关联的数据属性确定为单元内容对应的属性。

[0147] 进一步地，所述第一确定模块还包括第三单元；所述第二确定模块还包括第四单元；

[0148] 所述第三单元，用于在所述数据类型为非数字字母类型时，获取与所述列数据对应的候选哈希词典集合，将所述候选哈希词典集合确定为所述列数据对应的候选属性集合，其中，所述候选哈希词典集合包括多个哈希词典，所述哈希词典关联数据属性；

[0149] 所述第四单元，用于将单元内容输入到所述候选哈希词典集合中的各个哈希词典进行逐一查询；当在所述哈希词典中查询到所述单元内容时，将所述哈希词典关联的数据属性确定为单元内容对应的属性。

[0150] 进一步地，所述第三确定模块，还用于对所述列数据的各单元内容的关联于哈希词典的属性进行统计获取各个关联于哈希词典的属性的置信度；

[0151] 根据关联于哈希词典的置信度确定关联于哈希词典的列数据的属性；

[0152] 在所述根据关联于哈希词典的置信度确定关联于哈希词典的列数据的属性之后，还包括：

[0153] 在所述非数字字母类型为短文本且关联于哈希词典的列数据的属性的置信度低于设定阈值时，确定关联于Trie树的列数据的属性；

[0154] 将关联于哈希词典的列数据的属性与关联于Trie树的列数据的属性进行比较，将置信度大的列数据的属性确定为列数据的目标属性。

[0155] 进一步地，第一确定模块，还用于在确定所述列数据包括列头内容时，根据所述列头内容查询预设属性映射词典获取与所述列头内容匹配的属性，将所述与所述列头内容匹配的属性确定为列数据的属性。

[0156] 关于本实施例中的装置，其中各个模块执行操作的具体方式已经在有关该方法的实施例中进行了详细描述，此处将不做详细阐述说明。

[0157] 本实施例提供的数据属性确定装置，包括：对格式化的原始数据进行拆分获取多个列数据；若所述列数据不包括列头内容，则根据所述列数据的数据类型确定所述列数据对应的候选属性集合；根据所述候选属性集合确定所述列数据的各单元内容的属性；对所述列数据的各单元内容的属性进行统计获取各个属性的置信度，根据所述置信度确定所述列数据的属性。该装置通过分门别类地查找列数据的候选属性集合，以及通过统计各个单元内容的属性确定列数据的属性，实现了尽可能地减少数据属性识别过程中的运算量，提升了数据属性的识别效率和准确率。

[0158] 为达上述目的，本发明实施例还提出了一种计算机设备。

[0159] 图7是本发明一实施例的计算机设备的结构示意图。

[0160] 如图7所示，该计算机设备包括：存储器11、处理器12及存储在存储器11上并可在处理器12上运行的计算机程序。

[0161] 处理器12执行所述程序时实现图1至图5任一所示的实施例中提供的数据属性确定方法。

[0162] 进一步地，计算机设备还包括：

[0163] 通信接口13，用于存储器11和处理器12之间的通信。

[0164] 存储器11，用于存放可在处理器12上运行的计算机程序。

[0165] 存储器11可能包含高速RAM存储器，也可能还包括非易失性存储器(non-volatile memory)，例如至少一个磁盘存储器。

[0166] 处理器12，用于执行所述程序时实现图1至图5所示的实施例中提供的数据属性确定方法。

[0167] 如果存储器11、处理器12和通信接口13独立实现，则通信接口13、存储器11和处理器12可以通过总线相互连接并完成相互间的通信。所述总线可以是工业标准体系结构(Industry Standard Architecture，简称ISA)总线、外部设备互连(Peripheral Component Interconnect，简称PCI)总线或扩展工业标准体系结构(Extended Industry Standard Architecture，简称EISA)总线等。所述总线可以分为地址总线、数据总线、控制总线等。为便于表示，图7中仅以一条粗线表示，但并不表示仅有一根总线或一种类型的总线。

[0168] 可选地，在具体实现时，如果存储器11、处理器12及通信接口13，集成在一块芯片上实现，则存储器11、处理器12及通信接口13可以通过内部接口完成相互间的通信。

[0169] 处理器12可以是一个中央处理器(Central Processing Unit，简称CPU)，或者是特定集成电路(Application Specific Integrated Circuit，简称ASIC)，或者是被配置成实施本发明实施例的一个或多个集成电路。

[0170] 为达上述目的，本申请实施例还提出了一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现如实现图1至图5任一所示的实施例中提供的数据属性确定方法。

[0171] 在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不必须针对的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任一个或多个实施例或示例中以合适的方式结合。此外，在不相互矛盾的情况下，本领域的技术人员可以将本说明书中描述的不同实施例或示例以及不同实施例或示例的特征进行结合和组合。

[0172] 此外，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括至少一个该特征。在本发明的描述中，“多个”的含义是至少两个，例如两个，三个等，除非另有明确具体的限定。

[0173] 流程图中或在此以其他方式描述的任何过程或方法描述可以被理解为，表示包括一个或更多个用于实现定制逻辑功能或过程的步骤的可执行指令的代码的模块、片段或部分，并且本发明的优选实施方式的范围包括另外的实现，其中可以不按所示出或讨论的顺序，包括根据所涉及的功能按基本同时的方式或按相反的顺序，来执行功能，这应被本发明的实施例所属技术领域的技术人员所理解。

[0174] 在流程图中表示或在此以其他方式描述的逻辑和/或步骤，例如，可以被认为是用于实现逻辑功能的可执行指令的定序列表，可以具体实现在任何计算机可读介质中，以供指令执行系统、装置或设备(如基于计算机的系统、包括处理器的系统或其他可以从指令执行系统、装置或设备取指令并执行指令的系统)使用，或结合这些指令执行系统、装置或设备而使用。就本说明书而言，"计算机可读介质"可以是任何可以包含、存储、通信、传播或传输程序以供指令执行系统、装置或设备或结合这些指令执行系统、装置或设备而使用的装置。计算机可读介质的更具体的示例(非穷尽性列表)包括以下：具有一个或多个布线的电连接部(电子装置)，便携式计算机盘盒(磁装置)，随机存取存储器(RAM)，只读存储器(ROM)，可擦除可编辑只读存储器(EPROM或闪速存储器)，光纤装置，以及便携式光盘只读存储器(CDROM)。另外，计算机可读介质甚至可以是可在其上打印所述程序的纸或其他合适的介质，因为可以例如通过对纸或其他介质进行光学扫描，接着进行编辑、解译或必要时以其他合适方式进行处理来以电子方式获得所述程序，然后将其存储在计算机存储器中。

[0175] 应当理解，本发明的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。如，如果用硬件来实现和在另一实施方式中一样，可用本领域公知的下列技

术中的任一项或他们的组合来实现:具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路,具有合适的组合逻辑门电路的专用集成电路,可编程门阵列(PGA),现场可编程门阵列(FPGA)等。

[0176] 本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通过程序来指令相关的硬件完成,所述的程序可以存储于一种计算机可读存储介质中,该程序在执行时,包括方法实施例的步骤之一或其组合。

[0177] 此外,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模块中,也可以是各个单元单独物理存在,也可以两个或两个以上单元集成在一个模块中。上述集成的模块既可以采用硬件的形式实现,也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时,也可以存储在一个计算机可读取存储介质中。

[0178] 上述提到的存储介质可以是只读存储器,磁盘或光盘等。尽管上面已经示出和描述了本发明的实施例,可以理解的是,上述实施例是示例性的,不能理解为对本发明的限制,本领域的普通技术人员在本发明的范围内可以对上述实施例进行变化、修改、替换和变型。

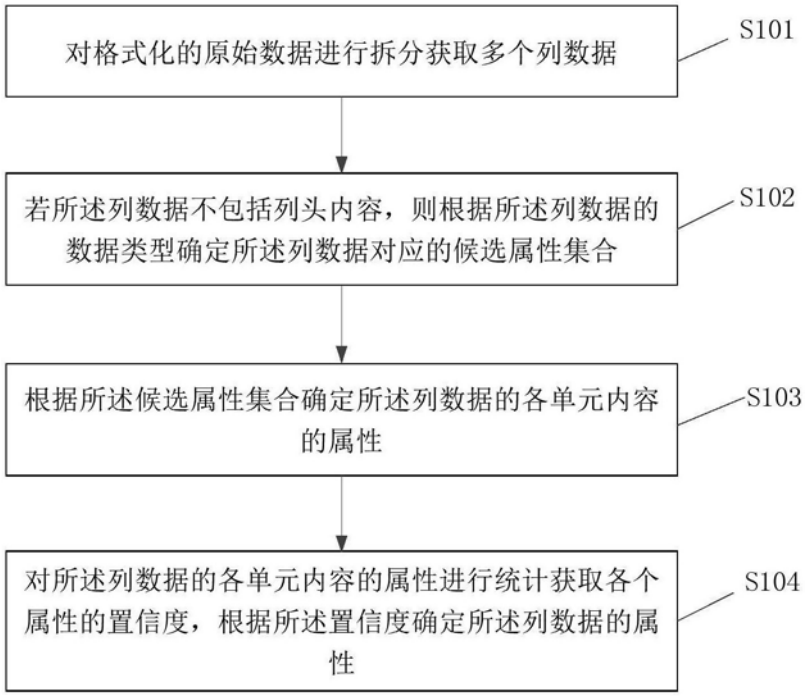

对格式化的原始数据进行拆分获取多个列数据
S101
若所述列数据不包括列头内容，则根据所述列数据的数据类型确定所述列数据对应的候选属性集合
S102
根据所述候选属性集合确定所述列数据的各单元内容的属性
S103
对所述列数据的各单元内容的属性进行统计获取各个属性的置信度，根据所述置信度确定所述列数据的属性
S104

图1

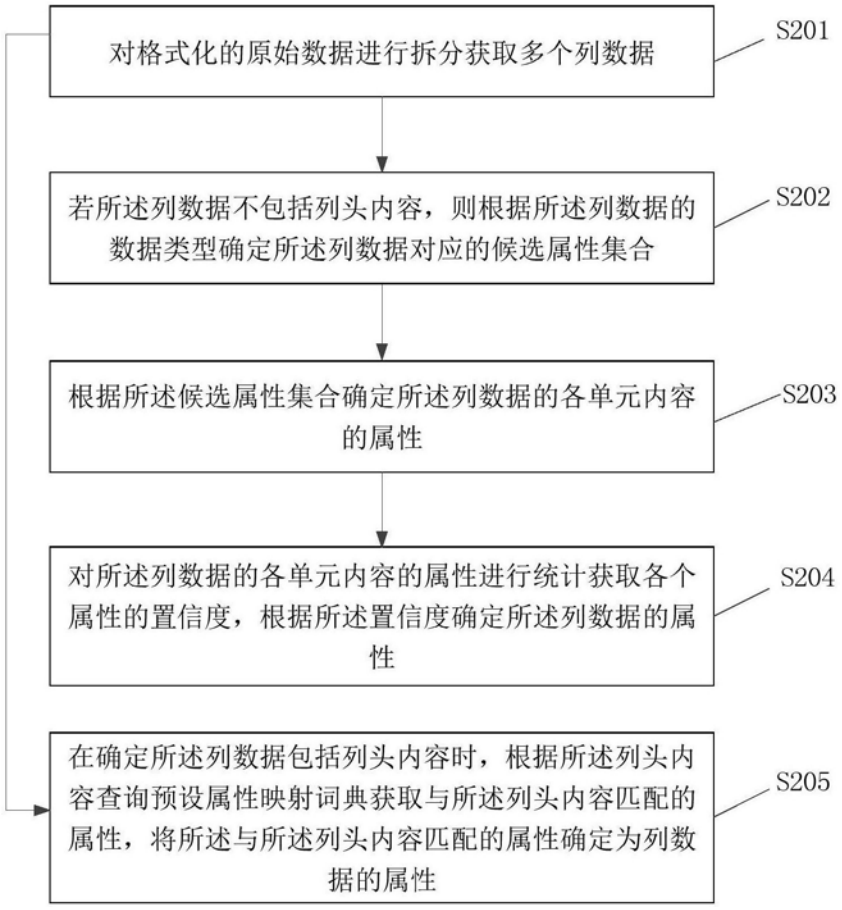

对格式化的原始数据进行拆分获取多个列数据
S201
若所述列数据不包括列头内容，则根据所述列数据的数据类型确定所述列数据对应的候选属性集合
S202
根据所述候选属性集合确定所述列数据的各单元内容的属性
S203
对所述列数据的各单元内容的属性进行统计获取各个属性的置信度，根据所述置信度确定所述列数据的属性
S204
在确定所述列数据包括列头内容时，根据所述列头内容查询预设属性映射词典获取与所述列头内容匹配的属性，将所述与所述列头内容匹配的属性确定为列数据的属性
S205

图2

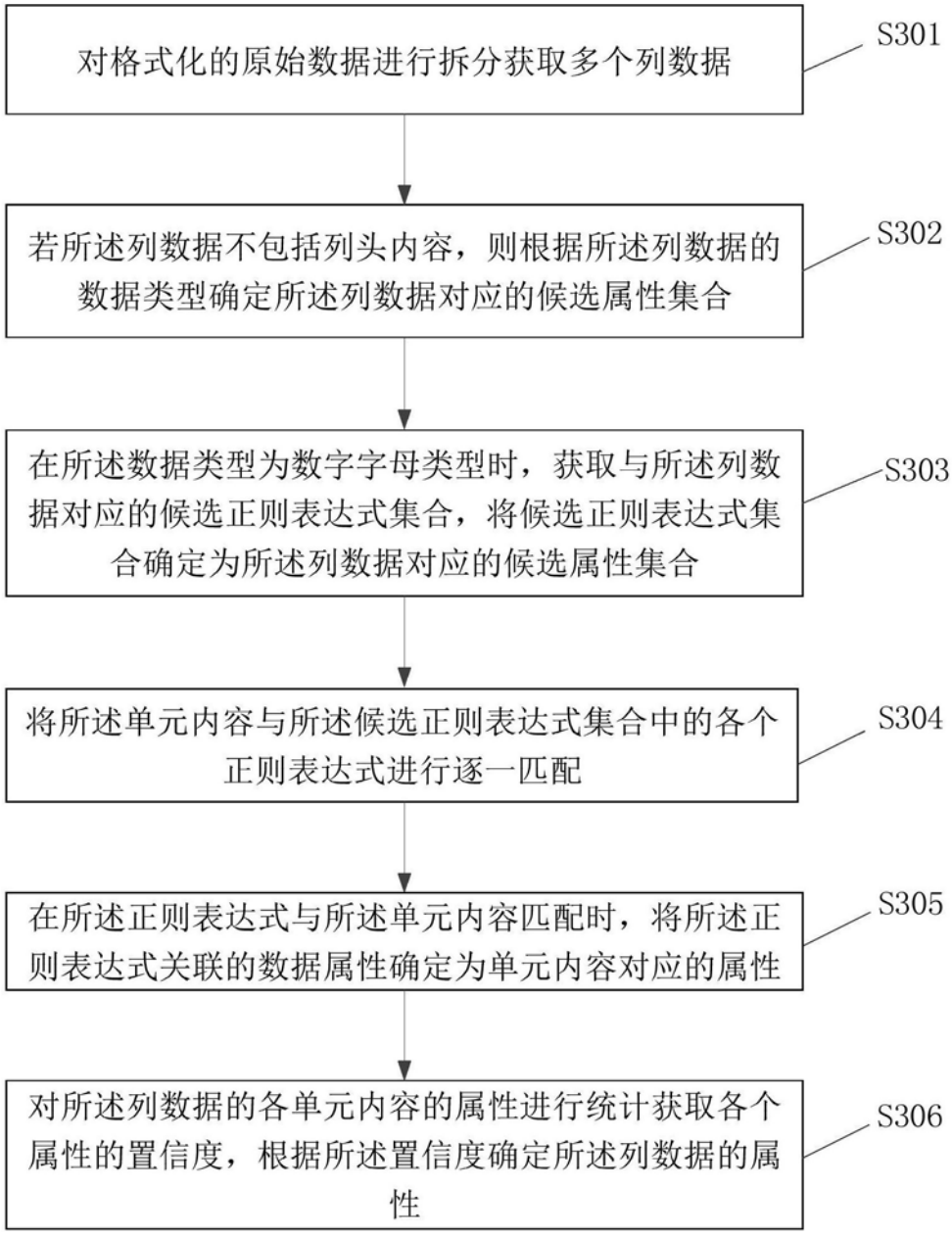

对格式化的原始数据进行拆分获取多个列数据
S301
若所述列数据不包括列头内容，则根据所述列数据的数据类型确定所述列数据对应的候选属性集合
S302
在所述数据类型为数字字母类型时，获取与所述列数据对应的候选正则表达式集合，将候选正则表达式集合确定为所述列数据对应的候选属性集合
S303
将所述单元内容与所述候选正则表达式集合中的各个正则表达式进行逐一匹配
S304
在所述正则表达式与所述单元内容匹配时，将所述正则表达式关联的数据属性确定为单元内容对应的属性
S305
对所述列数据的各单元内容的属性进行统计获取各个属性的置信度，根据所述置信度确定所述列数据的属性
S306

图3

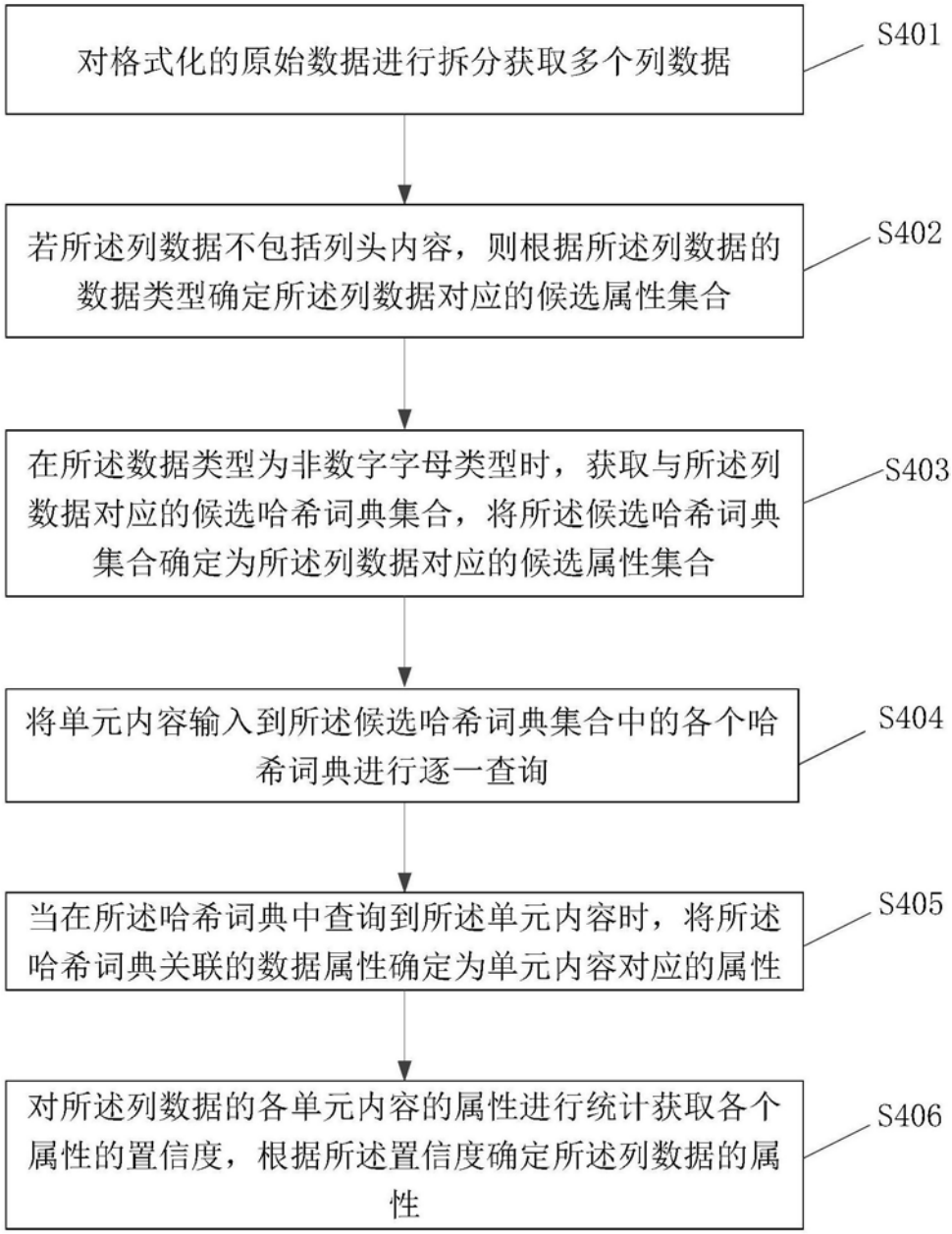

对格式化的原始数据进行拆分获取多个列数据
S401
若所述列数据不包括列头内容，则根据所述列数据的数据类型确定所述列数据对应的候选属性集合
S402
在所述数据类型为非数字字母类型时，获取与所述列数据对应的候选哈希词典集合，将所述候选哈希词典集合确定为所述列数据对应的候选属性集合
S403
将单元内容输入到所述候选哈希词典集合中的各个哈希词典进行逐一查询
S404
当在所述哈希词典中查询到所述单元内容时，将所述哈希词典关联的数据属性确定为单元内容对应的属性
S405
对所述列数据的各单元内容的属性进行统计获取各个属性的置信度，根据所述置信度确定所述列数据的属性
S406

图4

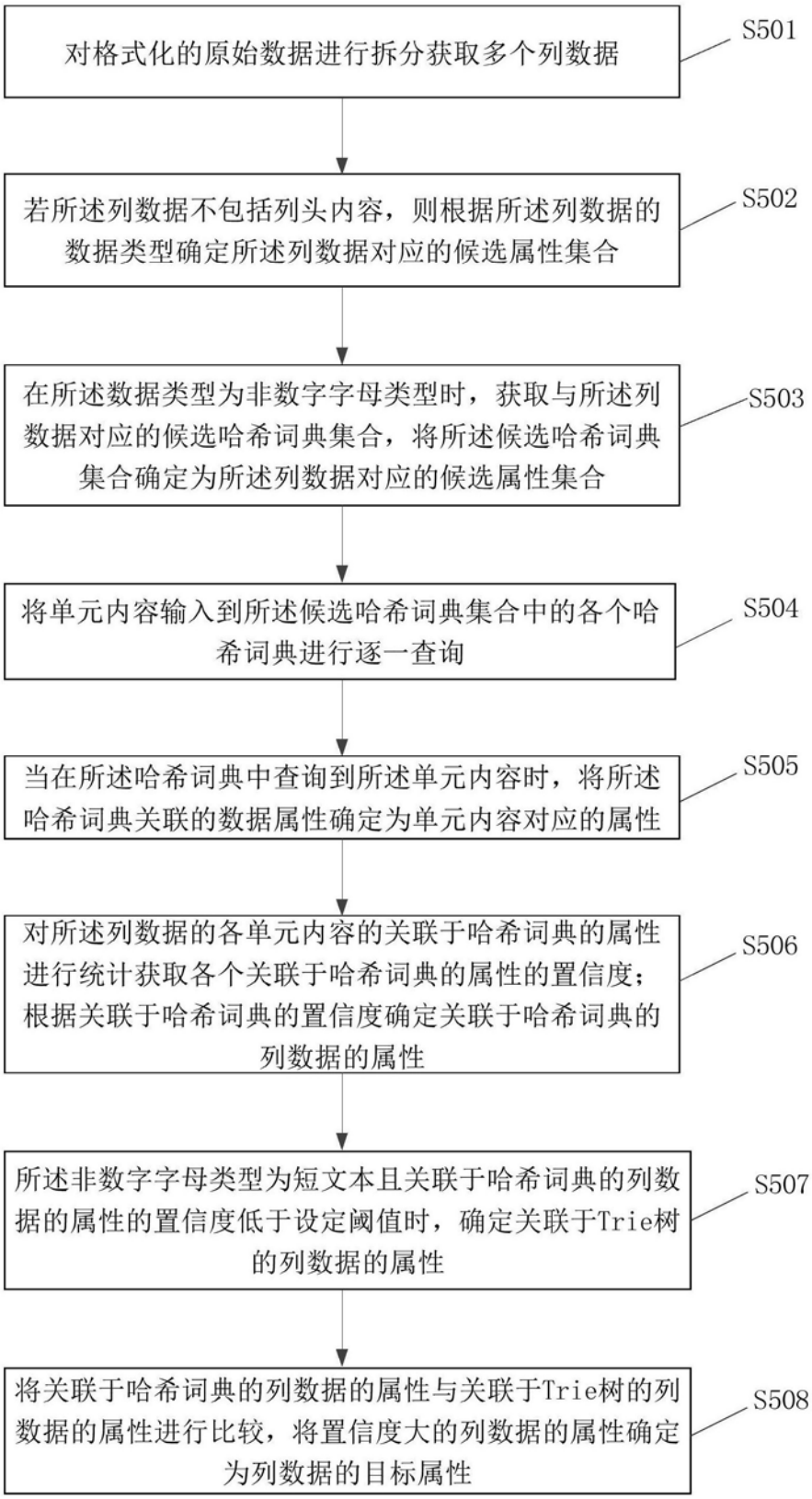

对格式化的原始数据进行拆分获取多个列数据
S501
若所述列数据不包括列头内容，则根据所述列数据的数据类型确定所述列数据对应的候选属性集合
S502
在所述数据类型为非数字字母类型时，获取与所述列数据对应的候选哈希词典集合，将所述候选哈希词典集合确定为所述列数据对应的候选属性集合
S503
将单元内容输入到所述候选哈希词典集合中的各个哈希词典进行逐一查询
S504
当在所述哈希词典中查询到所述单元内容时，将所述哈希词典关联的数据属性确定为单元内容对应的属性
S505
对所述列数据的各单元内容的关联于哈希词典的属性进行统计获取各个关联于哈希词典的属性的置信度；根据关联于哈希词典的置信度确定关联于哈希词典的列数据的属性
S506
所述非数字字母类型为短文本且关联于哈希词典的列数据的属性的置信度低于设定阈值时，确定关联于Trie树的列数据的属性
S507
将关联于哈希词典的列数据的属性与关联于Trie树的列数据的属性进行比较，将置信度大的列数据的属性确定为列数据的目标属性
S508

图5

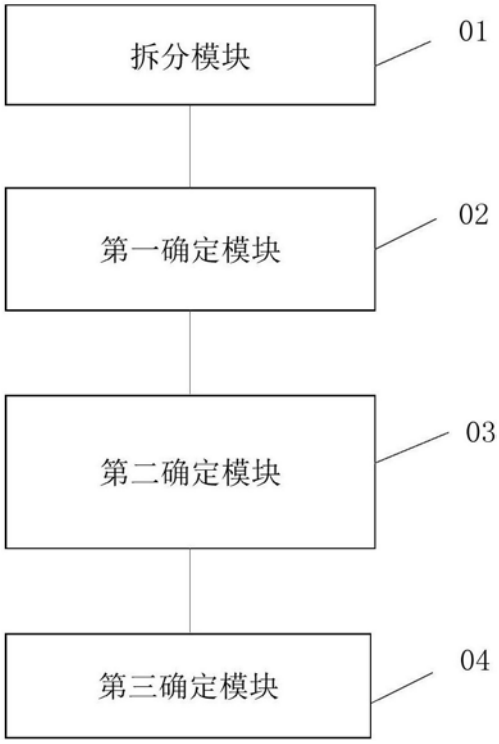


图6

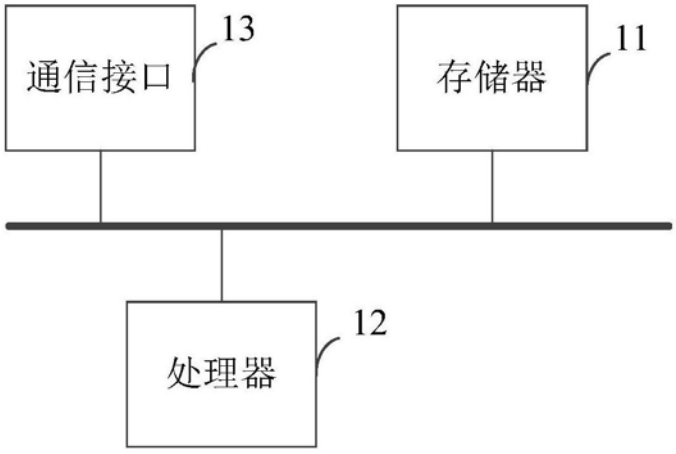


图7