

特征词的确定方法、装置和服务

技术领域

[01] 本说明书属于互联网技术领域，尤其涉及一种特征词的确定方法、装置和服务。

背景技术

5 [02] 在网络交易监管中，常常会通过检索与网络交易相关的数据所携带的文本数据中的特征词（例如能反映某一属性类型的网络交易共性的词组），来确定该交易的属性类型，从而可以有针对性地对某种属性类型的网络交易进行监控管理。

[03] 例如，可以通过检索识别与网络交易相关的文本数据中是否携带有“上分”、“下分”等黑词（一种表征违法交易的特征词），来判断该交易是否属于违法交易，从而可以
10 以及时地发现并处理这类违法交易。

[04] 基于上述方法来确定网络交易的属性类型时，所使用的特征词是否准确、涵盖的范围是否全面，会对网络交易的属性类型的判断是否准确产生较大的影响。而且随着人们语言习惯、用词方式的改变，对于同一属性类型的网络交易，能够有效表征该属性类型交易共性的特征词也会发生变化。例如，可能会出现一些新的词组成为表征该属性类型
15 交易的特征词。因此，亟需一种能够准确地确定出用于表征网络交易的属性类型的特征词的方法。

发明内容

[05] 本说明书目的在于提供一种特征词的确定方法、装置和服务，以避免提取第二特征词的过程中，由于分词错误导致的特征词提取不准确、不全面，达到能精确、高效地
20 从文本数据中挖掘出符合要求的新的特征词的技术效果。

[06] 本说明书提供的一种特征词的确定方法、装置和服务是这样实现的：

[07] 一种特征词的确定方法，包括：获取文本数据；从所述文本数据中提取出第一特征词；根据所述第一特征词，更新分词库，得到更新后的分词库，其中，所述分词库包含多个用于表征预设属性类型的预设特征词；根据所述更新后的分词库，从所述文本数据
25 中提取第二特征词。

[08] 一种特征词的确定装置，包括：获取模块，用于获取文本数据；第一提取模块，用

于从所述文本数据中提取出第一特征词；更新模块，用于根据所述第一特征词，更新分词库，得到更新后的分词库，其中，所述分词库包含多个用于表征预设属性类型的预设特征词；第二提取模块，用于根据所述更新后的分词库，从所述文本数据中提取第二特征词。

- 5 [09]一种服务器，包括处理器以及用于存储处理器可执行指令的存储器，所述处理器执行所述指令时实现获取文本数据；从所述文本数据中提取出第一特征词；根据所述第一特征词，更新分词库，得到更新后的分词库，其中，所述分词库包含多个用于表征预设属性类型的预设特征词；根据所述更新后的分词库，从所述文本数据中提取第二特征词。

- [10]一种计算机可读存储介质，其上存储有计算机指令，所述指令被执行时实现获取文本数据；从所述文本数据中提取出第一特征词；根据所述第一特征词，更新分词库，得到更新后的分词库，其中，所述分词库包含多个用于表征预设属性类型的预设特征词；根据所述更新后的分词库，从所述文本数据中提取第二特征词。

- [11]本说明书提供的一种特征词的确定方法、装置和服务器，通过先对文本数据进行新词提取得到第一特征词；再利用第一特征词对分词库进行更新；进而利用更新后的分词库和预设特征词从文本数据提取出新的特征词作为第二特征词，从而避免了提取特征词过程中，由于分词错误导致的特征词提取不准确、不全面，达到能精确地从文本数据中挖掘出符合要求的新的特征词的技术效果。

附图说明

- [12]为了更清楚地说明本说明书实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本说明书中记载的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

- [13]图 1 是在一个场景示例中，应用本说明书实施例提供的特征的确定方法的一种实施例的示意图；

- 25 [14]图 2 是在一个场景示例中，应用本说明书实施例提供的特征的确定方法的一种实施例的示意图；

- [15]图 3 是在一个场景示例中，应用本说明书实施例提供的特征的确定方法的一种实施例的示意图；

[16]图 4 是在一个场景示例中，应用本说明书实施例提供的特征的确定方法的一种实施例的示意图；

[17]图 5 是本说明书实施例提供的特征词的确定方法的流程的一种实施例的示意图；

[18]图 6 是本说明书实施例提供的服务器的结构的一种实施例的示意图；

5 [19]图 7 是本说明书实施例提供的特征词的确定装置的结构的一种实施例的示意图。

具体实施方式

[20]为了使本技术领域的人员更好地理解本说明书中的技术方案，下面将结合本说明书实施例中的附图，对本说明书实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本说明书一部分实施例，而不是全部的实施例。基于本说明书中的实
10 施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都应当属于本说明书保护的范围。

[21]考虑到现有的特征词的确定方法通常是先直接对所采集的文本数据进行分词处理；再通过深度学习算法得到分词的向量；继而通过语义向量距离和 PageRank 等方法计算分词的向量与已有特征词之间的向量距离；最后根据向量距离确定出新的特征词。

15 [22]然而，在互联网领域，使用的词组、短语的变化速度相对较快。例如，可能每隔一两天互联网上就会出现大量之前没有出现过的新的词语。但上述方法具体实施时，并没有考虑到类似新词的影响，导致在分词处理过程容易将新词组合错误地进行拆分，以致后续无法得到新词真正的向量，难以挖掘出新的特征词。

[23]针对上述情况，本说明书考虑可以将新词的发现与特征词的确定相结合。具体的，
20 可以先对所采集的文本数据进行新词挖掘，提取得到文本数据中的新词作为第一特征词；再根据第一特征词更新包含有已有的预设特征词的分词词库；通过更新后的分词词库对文本数据进行分词处理，再结合预设特征词，根据分词结果的向量确定出新的特征词，即挖掘出了第二特征词。从而避免了现有方法在提取特征词的过程中，由于分词错误导致的特征词提取不准确，达到能精确地从文本数据中挖掘出符合要求的新的特征词的技术效果。
25

[24]本说明书实施方式提供一种可以应用本说明书提供的特征词的确定方法的系统架构中。其中，该系统中具体可以包括服务器，且该服务器可以与网络平台的业务系统对接，进而可以采集得到网络平台上的文本数据；再从所述文本数据中提取出第一特征词；根

据所述第一特征词，更新分词库，得到更新后的分词库；通过所述更新后的分词库和所述预设特征词，从所述文本数据中提取第二特征词，即挖掘出新的特征词。

[25]在本实施方式中，所述服务器可以为一种应用于后台的，能够实现数据获取、数据处理等功能的服务器。具体的，所述服务器可以为一个具有数据运算、存储功能以及网络交互功能的电子设备；也可以为运行于该电子设备中，为数据处理、存储和网络交互提供支持的软件程序。在本实施方式中，并不具体限定所述服务器的数量。所述服务器具体可以为一个服务器，也可以为几个服务器，或者，由若干服务器形成的服务器集群。

[26]在一个场景示例中，可以参阅图 1 所示，可以应用本说明书实施例提供的特征词的确定方法对在某交易网络平台上采集得到的相关文本数据进行具体分析处理，从中提取得到能够用于表征非法交易的新的特征词，以便后续可以根据上述新的特征词，及时地发现交易网络平台上的非法交易，并及时进行监控处理。

[27]在本场景示例中，具体实施时，参阅图 2 所示，服务器可以先搜集该交易网络平台上最近一周的网络交易数据（例如转账数据）中所携带的交易附言等作为文本数据。其中，上述交易网络平台具体可以是一种以中文为主要语言的购物网站、理财网站、支付平台等，相应的，所涉及的文本数据通常会包含大量的中文汉字。

[28]上述文本数据具体可以理解作为一种交易网络平台上包含有汉字、英文字母、拼音符号或数字等信息的数据。具体的，上述文本数据可以包括交易附言，也可以包括文本标签等。当然，上述所列举的交易附言、文本标签只是为了更好地说明本说明书实施方式。具体实施时，根据具体的应用场景和使用需求，上述文本数据还可以包括有其他类型或形式的数据。对此，本说明书不作限定。

[29]其中，上述交易附言具体可以理解作为一种由交易数据发起者自定义编辑的用以描述交易数据的相关特征的信息数据。具体的，交易数据发起者在发送交易数据时，可以通过预设的文本数据输入界面，输入或设置用以描述交易数据的例如交易目的、交易时间或交易对象等特征信息，作为交易附言。例如，用户 A 在通过交易网络平台向用户 B 支付 XX 货款 10 元时，在确认发送该笔交易数据前，可以在预设的留言界面中输入“XX 货款”作为交易附言，以表征该笔交易数据的交易目的是用于支付 XX 货款，进而交易网络平台会将上述交易附言连同交易数据一同发送给用户 B，以便用户 B 可以知晓所收到的该笔交易数据的所对应的交易目的。

[30]上述文本标签具体可以理解作为一种由交易网络平台采集并根据交易数据发起者的所

实施的与该笔交易数据相关操作，自动生成的描述该笔交易数据的类型、目的或内容等的数据标签。例如，用户 C 在操作界面中，点击还贷选项，在所展示的下一级界面中输入“5000”元，发起一笔交易数据。交易网络平台可以根据用户 C 的上述操作确定该笔数据的目的是还贷，因此，可以自动生成“还贷”作为文本标签添加在该笔交易数据上，

5 以便进行记录。

[31]服务器在得到了上述文本数据后，可以先对上述文本数据进行新词挖掘。具体的，可以通过新词发现算法，对上述文本数据进行分析、处理，从中提取得到新词，作为第一特征词。

[32]其中，上述第一特征词具体可以理解为一种新词，即之前未被服务器发现，也没有被记录在服务器的数据库或分词库中的特定的汉字组合。

10

[33]需要说明的是，互联网领域区别于其他技术领域，互联网领域中的知识信息更新相对较为快速、频繁。例如，可能每隔一两天，网络中可能会出现大量之前没有出现过的，能够表征一些特定语义的网络词汇。根据现有的特征词的确定方法，往往忽略上述新词的影响，即没有及时地利用新词对分词库进行更新，导致后续进行分词处理时，容易

15 易出现错误，例如将一些新出现的汉字组合错误地进行拆，影响后续语义地确定，造成最终确定的特征词存在误差或者不够全面。

[34]例如，汉字组合“狗带”是一种网络新词。之前的分词库如果没有存储该词，后续在分词处理时，基于原有的分词库很有可能将该词错误拆分为“狗”和“带”，导致对该词的语义的确定产生误差，进而影响后续特征词确定。

[35]本说明书正是考虑到互联网领域中存在的上述特点，在具体确定特征词前，先对文本数据进行新词确定，得到第一特征词；进而后续可以结合第一特征词，更新分词库，再根据更新后的分词库，进行更加准确的分词处理。

20

[36]具体实施时，服务器可以从所述文本数据中筛选出字符长度小于预设长度阈值（例如 5 个字符长度）的字符串作为候选字符串；分别计算所述候选字符串中的各个字符串的指标参数；根据所述指标参数，从所述候选字符串中提取符合第一预设要求的候选字符串，作为新词，即第一特征词。

25

[37]具体的，考虑到汉语的组词习惯，通常一个词组的字符长度相对较短，因此，可以先根据字符长度，从文本数据中筛选掉明显不符合组词习惯的，具有较大概率不能构成词组的字符串，得到候选字符串。在本场景示例中，可以将预设长度阈值设为 5 个字符

长度,这样就可以先过滤掉字符长度大于5个字符的字符串,以减少后续的工作量,提高处理效率。需要说明的是,上述所列举的预设长度阈值只是一种示意性说明。具体实施时,根据具体情况还可以设置其他数值作为上述预设长度阈值。对此,本说明书不作限定。

- 5 [38]在本场景示例中,在从所述文本数据中筛选出字符长度小于预设长度阈值的字符串后,进一步可以根据已有的数据库(例如现有的分词库),从筛选出的字符串中过滤掉与已有的数据库中的已有字符串相同的字符串,将过滤后得到的字符串作为候选字符串。这样可以使得得到的候选字符串不会包含已有词词组所对应的字符串。

- 10 [39]上述指标参数具体可以理解作为一种用于表征字符串能否构造词组的特征参数。通常一个候选字符串的指标参数满足一定的要求,则可以认为该字符串可以构成一个词组,例如构成新词。

- [40]在本场景示例中,上述指标参数具体可以包括以下至少之一:凝固度、信息熵和频数等。当然,需要说明的是,上述所列举的指标参数只是为了更好地说明本说明书实施方式。具体实施时,根据具体情况还可以引入其他类型的特征参数作为上述指标参数。
- 15 对此,本说明书不作限定。

- [41]其中,上述凝固度具体可以理解为本场景中字符串中所包含的字符结合在一起的概率。通常一个字符串的凝固度数值越高,该字符串具有越高的概率成为一个词组。例如,通过计算发现字符串“电影院”的凝固度比字符串“的电影”的凝固度更高,则可以判断字符串“电影院”相对于“的电影”具有更高的概率、更容易成为一个词。
- 20 [42]具体计算字符串的凝固度时,可以先分别计算该字符串整体在文本数据中出现的概率、该字符串的各种拆分组合在文本数据中出现的概率;再计算该字符串整体在文本数据中出现的概率与该字符串的各种拆分组合在文本数据中出现的概率的比值;最后将数值最小的比值确定为该字符串的凝固度。

- [43]例如,在计算字符串“电影院”的凝固度时,可以先计算字符串“电影院”在文本
- 25 数据中出现的概率为 $p_0 = p(\text{电影院})$ 。字符串“电影院”可以拆分成两种拆分组合,即第一拆分组合:“电”和“影院”,以及第二拆分组合:“电影”和“院”。分别计算两种拆分组合在文本数据中出现的概率。以确定第一拆分组合的概率为例,可以先分别计算字符“电”的在文本数据中的出现概率和字符“影院”在文本数据中的出现概率;再将上述两个概率的乘积作为第一拆分组合在文本数据中出现的概率,可以表示为:

$p^1 = p(\text{电}) * p(\text{影院})$ 。按照类似的方式，可以计算出第二拆分组合在文本数据中出现的概率 $p^2 = p(\text{电影}) * p(\text{院})$ 。再分别计算 p^0 与 p^1 、 p^2 的比值为 p^0/p^1 、 p^0/p^2 。比较 p^0/p^1 、 p^0/p^2 的数值大小，将数值较小的 p^0/p^2 确定为字符串“电影院”的凝固度。

[44]上述信息熵，也可称为自由度，具体可以理解为字符串表征的某种特定信息的出现概率。通常当字符串所表征的某种特定信息的出现概率较高的时候，表明它被传播得较为广泛，或者说，被引用的程度相对较高。在说明书中，具体可以通过字符串的信息熵来衡量该字符段的左邻字集合和右邻字集合的随机性，从而可以利用该指标参数反映出该字符串作为词组所携带的信息量和使用的频繁程度。

[45]具体计算信息熵时，可以先分别搜索文本数据确定字符串的左邻字和右邻字，再分别统计各个左邻字的出现概率和各个右邻字的出现概率；根据各个左邻字的出现概率按照预设公式计算左邻字的信息熵，根据各个右邻字的出现概率计算右邻字的信息熵。

[46]例如，文本数据为“吃葡萄不吐葡萄皮不吃葡萄倒吐葡萄皮”，在计算字符串“葡萄”的信息熵时，可以先搜索该文本数据找到字符串“葡萄”的左邻字分别为：吃、吐、吃、吐，右邻字分别为：不、皮、倒、皮。再分别计算各个左邻字的出现概率为 $p(\text{吃})$ 为 $1/2$ ， $p(\text{吐})$ 为 $1/2$ ；各个右邻字的出现概率为 $p(\text{不})$ 为 $1/4$ ， $p(\text{倒})$ 为 $1/4$ ， $p(\text{皮})$ 为 $1/2$ 。再根据各个左邻字的出现概率按照以下预设公式 $H(x) = -\sum p(x_i) \log(p(x_i))$ 计算左邻字的信息熵。具体可以表示为 $-(1/2) \cdot \log(1/2) - (1/2) \cdot \log(1/2) \approx 0.693$ 。类似的，根据各个右邻字的出现概率可以得到右邻字的信息熵为 $-(1/2) \cdot \log(1/2) - (1/4) \cdot \log(1/4) - (1/4) \cdot \log(1/4) \approx 1.04$ 。通过比较左邻字的信息熵和右邻字的信息熵可以发现，对于字符串“葡萄”，右邻字的信息熵数值更大，因此，该字符串所对接的右邻字相对更丰富一些。

[47]上述频数具体可以理解为字符串在文本数据中出现的次数。通常一个字符串的频数越高，相对的具有较大概率成为一个词组。具体的，例如统计文本数据，字符串“机器”出现了 180 次，则可以将字符串“机器”的频数确定为 180。

[48]在按照上述方式计算得到了各个候选字符串的凝固度、信息熵和频数等指标参数，进一步可以将各个候选字符串的凝固度、信息熵、频数分别与预设的凝固度阈值、预设的信息熵阈值、预设的频数阈值进行比较，将凝固度小于等于预设的凝固度阈值、信息熵小于等于预设的信息熵阈值、频数小于等于预设的频数阈值的候选字符串确定为符合

第一预设要求的第一特征字符串，即新词。

[49]在本场景示例中，上述预设的凝固度阈值、预设的信息熵阈值、预设的频数阈值的具体数值可以根据具体情况和处理要求灵活设置。对此，本说明书不作限定。

[50]服务器在根据文本数据确定出了上述第一特征词后，进一步，可以根据上述第一特征词对分词库进行更新，得到更新后的分词库。

[51]上述分词库具体可以理解作为一种基于历史数据建立的，包含有多个用于表征预设属性类型的预设特征词的词库。其中，上述预设属性类型具体可以包括：合法和非法等。其中，合法类型又可以进一步细化为：还款、消费、借出等类型。类似的，非法类型又可以进一步细化为多种不同的非法类型等。上述预设特征词具体可以理解作为一种能够用于表征预设属性类型的已有词组（或已有词）。

[52]在本场景示例中，考虑到所应用的交易网络平台为使用中文的交易网络平台，主要分词处理的文本数据为中文文本数据，因此，具体可以使用 jieba（一种基于隐马尔可夫模型（HMM）的分词库）作为上述分词库。由于 jieba 针对中文汉字词组具有较好的匹配性，所以在本场景示例中利用 jieba 作为分词库可以使得后续的分词处理更加准确。

当然，需要说明的是，上述所列举的 jieba 只是一种示意性说明。具体实施时，根据具体的应用场景和所涉及的语言类型还可以选择使用其他类型的分词库。对此，本说明书不作限定。

[53]在本场景示例中，目的是为找到针对非法交易的新的特征词，因此，预设属性类型为非法类型。相应的预设特征词为能够表征非法类型的已有词组。具体的，例如，“上分”、“下分”是一种较为常见的与非法交易行为关联的关键词，因此，上述两个词组可以是分词库所包含的一种预设特征词。

[54]具体实施时，可以将新确定的第一特征词，即新词添加至已有分词库中，从而达到对已有分词库的更新、扩展，得到更加完整、准确的，包含有当前最新发现的新词的更新后的分词库。这样得到的更新后的分词库考虑到了当前刚出现的、已有分词库中没有记载的新词。因此，后续利用上述更新后的分词库取代已有分词库进行分词处理时，可以避免由于使用已有分词库，导致将一些已有分词库中未记载的新词进行错误拆分。

[55]在得到了上述更加准确、完整的分词库后，服务器可以根据上述更新后的分词库，从所述文本数据中提取第二特征词，即用于表征预设属性类型的新的特征词。

[56]上述第二特征词具体可以理解作为一种区别于预设特征词，服务器之前没有发现、确

定的，能表征预设属性类型的新的词组。

[57]具体实施时，可以根据所述更新后的分词库，对所述文本数据进行分词处理，得到多个分词单元；再对所述多个分词单元分别进行词向量化处理，得到分词单元的词向量；根据所述分词单元的词向量和预设特征词的词向量，从所述多个分词单元中确定符合第二预设要求的分词单元，作为所述第二特征词。

[58]在本场景示例中，由于是利用基于第一特征词更新后的分词库对文本数据进行分词处理，从而可以有效避免了对一些新词的错误拆分，使得的分词处理得到的分词单元更加准确。

[59]在得到分词单元后，具体实施时，可以对分词单元分别进行向量化处理，得到能够表征分词单元语义的词向量，进而可以后续可以根据上述分词单元的词向量与预设特征词的词向量，从所述分词单元中寻找出符合第二预设要求（即语义与预设特征词相对较接近，所表征的属性类型较为相似）的分词单元，作为第二特征词。

[60]在本场景示例中，考虑到所应用的是使用的中文的交易网络平台。而中文中汉字不同于英文中字母。英文字母是一种表音的文字，即单个的英文字母只表征发音，而不表征语义。而中文汉字是一种既表音又表音的文字（例如形声字、象形字等），即单个汉字除了表征发音外，汉字本身的结构（例如汉字内部的笔画组成）也能在一定程度上反映出语义。

[61]现有的对分词单元进行向量化处理大多是通过 PageRank 方法（一种词向量化算法）对汉字字符单元进行向量化处理。而 PageRank 方法本身是针对英文字母的特点设计，通过该方法得到的词向量只能反映出词的上下文外部信息，而无法反应出汉字内部结构所携带的语义信息。因此，根据现有的特征词的确定方法利用 PageRank 方法对汉字组成的分词单元进行向量处理得到的词向量所包含的信息往往会遗漏掉分词单元中汉字内部结构所表征的语义信息，即信息不够完整、准确，会影响后续第二特征词的提取。

[62]正是考虑到上述问题，为了使得所获取的分词单元的词向量所携带的信息更加丰富、完整，以进一步提高确定第二特征词的准确度，在本场景示例中，可以按照以下方式对多个分词单元中的各个分词单元分别进行词向量化处理。具体的，服务器可以先将分词单元中的汉字拆分为多个笔画；将不同的笔画映射为不同的数值，根据所述分词单元的多个笔画，建立分词单元的笔画向量；同时，搜索并获取文本数据中与所述分词单元相连的词语（例如左邻词和右邻词），作为上下文词语；并按照常用方法获取所述上下文

词语的词向量；再根据所述分词单元的笔画向量和所述上下文词语的词向量，确定所述分词单元的词向量。这样得到的分词单元的词向量即携带有分词单元的上下文外部信息，又包含了分词单元中汉字的内部结构所反映的语义信息，从而使得获取的分词单元的词向量更加的准确、完整。

- 5 [63]当然，上述所列举的获取分词单元的词向量的方式只是一种示意性说明。具体实施时，还可以通过 Cw2vec 算法（Learning Chinese Word Embeddings with Stroke n-gram Information，一种词向量化算法）对所述多个分词单元分别进行词向量化处理，以得到分词单元的词向量。对此，本说明书不作限定。

- 10 [64]在获取得到了上述分词单元的词向量后，进一步，服务器可以根据所述分词单元的词向量和预设特征词的词向量，从所述多个分词单元中确定符合第二预设要求（即与预设特征词语义或者所表征的属性类型相近）的分词单元，作为所述第二特征词，即得到了能够表征预设属性类型的新的特征词。

- 15 [65]在本场景示例中，具体实施时，可以先从多个预设特征词中提取预设数量的预设特征词作为测试词（也可称为 spy）；根据所述多个预设特征词中除测试词以外的预设特征词的词向量，建立标记样本集（也可称为黑样本，positive，记为 P），并将该样本集中的分词单元的词向量标记为 1；根据所述测试词的词向量和所述分词单元的词向量，建立非标记样本集（也可以称为白样本，定义为 unlabeled，记为 U），并将该样本集中的分词单元的词向量标记为 0。需要说明的是，上述标记样本集中的样本数量小于非标记样本集中的样本数量。

- 20 [66]在建立了上述标记样本集和非标记样本集后，可以根据所述标记样本集和非标记样本集，通过多次迭代拟合，确定拟合分数阈值。

- 25 [67]具体的，可以按照以下步骤处理确定出拟合分数阈值：对标记样本集和非标记样本集中的分词单元的词向量分别进行 GBDT（Gradient Boosting Decision Tree，梯度提升决策树）拟合，根据拟合结果对每个分词单元的词向量进行打分（记为 score）；对拟合结果的分值，进一步作如下处理：将属于 P 的分词单元的词向量的 score 置为 1，其余保持原来具体 score，然后进行归一化处理。按照上述步骤进行多次处理（例如重复两次），直到找出可以使得阈值比例（例如 90%）的被归入 U 中的预设特征词的词向量（即 spy）被识别出来阈值（记为 t）作为上述拟合分数阈值。

[68]在确定出拟合分数阈值后，进一步可以根据上述拟合分数阈值，从分词单元向量中

确定出符合第二预设要求的分词单元向量所对应的分词单元作为第二特征词。

[69]具体的，可以按照以下步骤处理确定出第二特征词：将被归入 U 中的预设特征词的词向量（即 spy）重新归属回 P 中；再将剩下的 U 中所有拟合分数值小于拟合分数阈值（即 t）的分词单元的词向量的 score 赋为 0，P 中所有预设特征词的词向量的 score 赋为 1，其余保持当前的 score 然后进行归一化处理；再对所有的词向量进行 GBDT 拟合，并根据拟合结果对每个词的词向量进行重新打分得到拟合分数，记为 score'；根据拟合分数，将属于 P 的词向量的拟合分数 score' 置为 1，其余保持原有的 score'，然后进行归一化处理。按照上述步骤进行多次处理（例如重复 5 次），得到每个词向量最终的拟合分数，记为 score''，将 score'' 大于特征词分数阈值（记为 v）的词向量所对应的分词单元确定为第二特征词，即符合第二预设要求的新的能够表征该属性类型的词组。其中，上述特征词分数阈值具体可以根据具体情况和精度要求设置。对于特征词分数阈值的具体取值，本说明书不作限定。

[70]当然，上述所列举的确定第二特征词的方式只是一种示意性说明。具体实施时，还可以通过 PU_learning 算法（Learning from Positive and Unlabeled Example，正例和无标记样本学习算法）对所述多个分词单元的词向量进行分析处理，以确定出第二特征词。对此，本说明书不作限定。

[71]在本场景示例中，具体实施时，可以将特征词分数阈值为 0.5，进而可以按照上述处理步骤逐步从多个分词单元中确定出符合第二预设要求的能够表征违法的分词单元，作为第二特征词，即新的特征词。

[72]在按照上述方式确定出能够表征非法交易的新的特征词，即第二特征词后，服务器进一步可以将第二特征词与预设特征词合并；根据合并后特征词对交易网络平台上交易附言、文本标签等文本数据进行检测、识别，以便及时发现交易网络平台上的非法交易，并对非法交易进行有针对性的监控和处理。

[73]由上述场景示例可见，本说明书提供的特征词的确定方法，由于通过先从文本数据进行新词提取得到第一特征词；再利用第一特征词对分词库进行更新；进而利用更新后的分词库和预设特征词从文本数据提取出新的特征词作为第二特征词，从而避免了提取特征词过程中，由于分词错误导致的特征词提取不准确，达到能精确地从文本数据中挖掘出符合要求的新的特征词的技术效果。

[74]在另一个场景示例中，为了避免一些无意义的字符串的干扰，减少后续工作量，提

高处理效率，具体实施时，可以参阅图 3 所示，在从所述文本数据中筛选出字符长度小于预设长度阈值的字符串作为候选字符串之前，所述方法还可以包括以下内容：过滤所述文本数据中的无效字符串，得到过滤后的文本数据，再从过滤后的文本数据中筛选出字符长度小于预设长度阈值的字符串作为候选字符。

- 5 [75]在本场景示例中，上述无效字符串具体可以理解为不含有汉字，或明显不会构成词组的字符串。具体的，上述无效字符串可以是全由字母、数字或表情符号组成的字符串，也可以是网页链接，还可以是用于表征繁体转简体的文本数据等等。当然，需要说明的是，上述所列举的无效字符串只是一种示意性说明。具体实施时，根据具体的应用场景，上述无效字符串具体还可以包括其他类型的字符串。对此，本说明书不作限定。
- 10 [76]在另一个场景示例中，为了避免无意义的分词单元的干扰，进一步减少后续工作量，提高处理效率，具体实施时，可以参阅图 4 所示，在根据所述分词单元的词向量和预设特征词的词向量，从所述多个分词单元中确定符合第二预设要求的分词单元，作为所述第二特征词前，所述方法还可以包括以下内容：过滤所述分词单元的词向量中的停用词
- 15 的词向量，得到过滤后的分词单元的词向量；再从过滤后的分词单元的词向量中确定出符合第二预设要求的分词单元。

- [77]在本场景示例中，上述停用词具体可以理解为所表征的内容没有实际意义或与交易数据的属性类型无关的词组。具体的，上述停用词可以是一些连接词或助词，例如“的”、“是”、“了”等，也可以是一些与交易数据的无关，宽泛的代词，例如“我”、“这”、“那”，还可以是数字、字母或单个字的词等等。当然，需要说明的是上述所列举的停
- 20 用词只是一种示意性说明。具体实施时，根据具体的应用场景，上述停用词还可以包括其他的词，例如“在”、“有”、“人”、“一”等。对于上述停用词的具体内容，本说明书不作限定。

- [78]由上述场景示例可见，本说明书提供的特征词的确定方法，由于通过先对文本数据进行新词提取得到第一特征词；再利用第一特征词对分词库进行更新；进而利用更新后
- 25 的分词库和预设特征词从文本数据提取出新的特征词作为第二特征词，从而避免了提取特征词过程中，由于分词错误导致的特征词提取不准确、不全面，达到能精确地从文本数据中挖掘出符合要求的新的特征词的技术效果；又通过先将分词单元中的汉字拆分为多个笔画，得到分词单元的笔画向量；再根据分词单元的笔画向量和上下文词语的词向量，确定分词单元的词向量，从而使得得到的分词单元的词向量同时包含上下文外部信
- 30 息和汉字内部构造信息，能够反映出更加丰富、准确的语义信息，再基于上述分词单元

的词向量进行第二特征词的提取，提高了确定特征词的准确度；还通过先根据从多个预设特征词中提取出的预设个数的预设特征词和分词单元建立非标记样本集，根据剩余的预设特征词建立标记样本集；再基于上述非标记样本集和标记样本通过多次迭代拟合，确定出较为准确的拟合分数阈值，以便根据上述拟合分数阈值从分词单元中确定出第二特征词，进一步提高了确定特征词的准确度。

[79]参阅图 5 所示，本说明书实施例提供了一种特征词的确定方法，其中，该方法具体应用于后台服务器一侧。具体实施时，该方法可以包括以下内容：

[80]S51：获取文本数据。

[81]在本实施方式中，上述文本数据具体可以理解作为一种包含有汉字、英文字母、拼音符号或数字等信息的数据。具体的，上述文本数据可以包括交易附言，也可以包括文本标签等。当然，上述所列举的交易附言、文本标签只是为了更好地说明本说明书实施方式。具体实施时，根据具体的应用场景和使用需求，上述文本数据还可以包括有其他类型或形式的数据。对此，本说明书不作限定。

[82]其中，上述交易附言具体可以理解作为一种由交易数据发起者自定义编辑的用以描述交易数据的相关特征的信息数据。具体的，交易数据发起者在发送交易数据时，可以通过预设的文本数据输入界面，输入或设置用以描述交易数据的例如交易目的、交易时间或交易对象等特征信息，作为交易附言。上述文本标签具体可以理解作为一种由交易网络平台采集并根据交易数据发起者的所实施的与该笔交易数据相关操作，自动生成的用以描述该笔交易数据的类型、目的或内容等的的数据标签。

[83]在本实施方式中，上述获取文本数据具体可以包括：服务器搜集得到预设时间段（例如近一个月）内交易网络平台上的交易附言和/或文本标签等数据作为文本数据。

[84]S53：从所述文本数据中提取出第一特征词。

[85]在本实施方式中，上述第一特征词具体可以理解作为一种新词，即之前未被服务器发现，也没有被记录在服务器的数据库或分词库中的特定的汉字组合。

[86]在本实施方式中，具体实施时，服务器可以通过新词发现算法对上述文本数据进行分析处理，从上述文本数据中提取得到第一特征词。

[87]在本实施方式中，从所述文本数据中提取出第一特征词，具体可以包括以下内容：从所述文本数据中筛选出字符长度小于预设长度阈值的字符串作为候选字符串；计算所述候选字符串的指标参数；根据所述指标参数，从所述候选字符串中提取符合第一预设

要求的候选字符串，作为第一特征词。

[88]在本实施方式中，通过从所述文本数据中筛选出字符长度小于预设长度阈值的字符串作为候选字符串，可以先将明显不符合组词习惯的，具有较大概率不能构成词组的字符串事先过滤掉，从而可以避免上述类型字符串的干扰，减少工作量，提高处理效率。

- 5 [89]在本实施方式中，上述预设长度阈值具体可以设置为 5 个字符长度。当然，需要说明的是，上述所列举的预设长度阈值只是一种示意性说明。具体实施时，根据具体情况还可以设置其他数值作为上述预设长度阈值。对此，本说明书不作限定。

- [90]在本实施方式中，上述指标参数具体可以理解作为一种用于表征字符串能否构造词组的特征参数。通常一个候选字符串的指标参数满足一定的要求，则可以认为该字符串可以构成一个词组，例如构成新词。
- 10

- [91]在本实施方式中，在从所述文本数据中筛选出字符长度小于预设长度阈值的字符串作为候选字符串后，在计算所述候选字符串的指标参数前，所述方法还包括：根据已有的数据库，过滤上述候选字符串中与已有的数据库中的字符串相同的字符串，保留已有的数据库中没有的候选字符串，以便进行指标参数的计算。这样可以先过滤掉已有的数据库中所记录的已有词组所对应的字符串，减少了后续处理的工作量，提高了处理效率。
- 15

[92]在本实施方式中，上述指标参数具体可以包括以下至少之一：凝固度、信息熵和频数等。当然，需要说明的是，上述所列举的指标参数只是为了更好地说明本说明书实施方式。具体实施时，根据具体情况还可以引入其他类型的特征参数作为上述指标参数。例如，互信息、tf-idf 等词频信息。对此，本说明书不作限定。

- 20 [93]其中，上述凝固度具体可以理解为文本数据中字符串中所包含的字符结合在一起的概率。通常一个字符串的凝固度数值越高，该字符串具有越高的概率成为一个词组。上述信息熵，也可称为自由度，具体可以理解为字符串表征的某种特定信息的出现概率。通常当字符串所表征的某种特定信息的出现概率较高的时候，表明它被传播得较为广泛，或者说，被引用的程度相对较高。上述频数具体可以理解为字符串在文本数据中出现的次数。通常一个字符串的频数越高，相对的具有较大概率成为一个词组。
- 25

[94]在本实施方式中，上述根据所述指标参数，从所述候选字符串中提取符合第一预设要求的候选字符串，作为第一特征词，具体实施时，可以包括以下内容：将各个候选字符串的凝固度、信息熵、频数分别与预设的凝固度阈值、预设的信息熵阈值、预设的频数阈值进行比较；根据比较结果，将凝固度小于等于预设的凝固度阈值、信息熵小于等

于预设的信息熵阈值、频数小于等于预设的频数阈值的候选字符串确定为符合第一预设要求的第一特征字符串，即新词。

[95]在本实施方式中，上述预设的凝固度阈值、预设的信息熵阈值、预设的频数阈值的具体数值可以根据具体情况和处理要求灵活设置。对此，本说明书不作限定。

- 5 [96]S55: 根据所述第一特征词，更新分词库，得到更新后的分词库，其中，所述分词库包含多个用于表征预设属性类型的预设特征词。

[97]在本实施方式中，上述分词库具体可以理解为一种基于历史数据建立的，包含有多个用于表征预设属性类型的预设特征词的词库。

- 10 [98]在本实施方式中，上述预设属性类型具体可以包括：合法和非法等。其中，合法类型又可以进一步细化为：还款、消费、借出等类型。类似的，非法类型又可以进一步细化为多种不同的非法类型等。

[99]在本实施方式中，上述预设特征词具体可以理解为一种能够用于表征预设属性类型的已有词组（或已有词）。其中，预设特征词与所表征的预设属性类型对应。例如，用于表征违法类型的预设特征词也可以称为黑词等。

- 15 [100] 在本实施方式中，进一步考虑到所应用的交易网络平台为使用中文的交易网络平台，主要分词处理的文本数据为中文文本数据，因此，具体可以使用 jieba（一种基于隐马尔可夫模型（HMM）的分词库）作为上述分词库。由于 jieba 针对中文汉字词组具有较好的匹配性，所以在本实施方式中利用 jieba 作为分词库可以使得后续的分词处理更加准确。当然，需要说明的是，上述所列举的 jieba 只是一种示意性说明。具体实施时，根据具体的应用场景和所涉及的语言类型还可以选择使用其他类型的分词库。对
20 此，本说明书不作限定。

[101] 在本实施方式中，上述更新后的分词库是在基于历史数据得到的已有分词库的基础上，增添了新确定的新词（即第一特征词），扩展后的分词库。

- 25 [102] 在本实施方式中，根据所述第一特征词，更新分词库，得到更新后的分词库，具体实施时，可以包括：可以将所确定的第一特征词，添加至已有分词库，从而达到对已有分词库的更新、扩展，得到更加完整、准确的，包含有当前最新发现的新词的更新后的分词库。这样后续利用上述更新后的分词库取代已有分词库进行分词处理时，可以避免由于使用已有分词库，导致将一些已有分词库中未记载的新词进行错误拆分。

[103] S57: 根据更新后的分词库，从所述文本数据中提取第二特征词根据所述更新后

的分词库，从所述文本数据中提取第二特征词。

[104] 在本实施方式中，上述第二特征词具体可以理解作为一种区别于预设特征词，之前没有发现、确定的，能表征预设属性类型的新的词组。

[105] 在本实施方式中，上述根据更新后的分词库，从所述文本数据中提取第二特征词根据所述更新后的分词库，从所述文本数据中提取第二特征词，具体实施时，可以包括以下内容：根据所述更新后的分词库，对所述文本数据进行分词处理，得到多个分词单元；对所述多个分词单元分别进行词向量化处理，得到分词单元的词向量；根据所述分词单元的词向量和预设特征词的词向量，从所述多个分词单元中确定符合第二预设要求的分词单元，作为所述第二特征词。

[106] 由上可见，本说明书实施例提供的特征词的确定方法，由于通过先对文本数据进行新词提取得到第一特征词；再利用第一特征词对分词库进行更新；进而利用更新后的分词库和预设特征词从文本数据提取出新的特征词作为第二特征词，从而避免了提取特征词过程中，由于分词错误导致的特征词提取不准确、不全面，达到能精确地从文本数据中挖掘出符合要求的新的特征词的技术效果。

[107] 在一个实施方式中，所述文本数据具体可以包括：交易附言，和/或，文本标签等。当然，需要说明的是上述所列举的交易附言、文本标签只是为了更好地说明本说明书实施方式。具体实施时，根据具体的应用场景和使用需求，上述文本数据还可以包括有其他类型或形式的数据。对此，本说明书不作限定。

[108] 在一个实施方式中，从所述文本数据中提取出第一特征词，具体实施时，可以包括以下内容：从所述文本数据中筛选出字符长度小于预设长度阈值的字符串作为候选字符串；计算所述候选字符串的指标参数；根据所述指标参数，从所述候选字符串中提取符合第一预设要求的候选字符串，作为第一特征词。

[109] 在一个实施方式中，所述指标参数具体可以包括以下至少之一：凝固度、信息熵和频数等。当然，需要说明的是，上述所列举的指标参数只是为了更好地说明本说明书实施方式。具体实施时，根据具体情况还可以引入其他类型的特征参数作为上述指标参数。对此，本说明书不作限定。

[110] 在一个实施方式中，在从所述文本数据中筛选出字符长度小于预设长度阈值的字符串作为候选字符串之前，所述方法具体实施时还可以包括以下内容：过滤所述文本数据中的无效字符串。

[111] 在本实施方式中，考虑到所获取的文本数据中可能包含有大量无意义的字符串，这类字符串通常不含有汉字，或明显不会构成词组，在本实施方式中，这类字符串可以称为无效字符串。具体的，可以是全由字母、数字或表情符号组成的字符串，也可以是网页链接，还可以是用于表征繁体转简体的文本数据等等。

- 5 [112] 在本实施方式中，为了避免后续的分析处理受到上述无效字符串的干扰，提高处理效率，具体实施时，可以在从所述文本数据中筛选出字符长度小于预设长度阈值的字符串作为候选字符串之前，先将文本数据中无效字符串过滤掉（或称清洗掉），得到过滤后的字符串，后续再基于过滤后的字符串进行候选字符串的确定。

- 10 [113] 在一个实施方式中，上述根据所述更新后的分词库，从所述文本数据中提取第二特征词，具体实施时，可以包括以下内容：根据所述更新后的分词库，对所述文本数据进行分词处理，得到多个分词单元；对所述多个分词单元分别进行词向量化处理，得到分词单元的词向量；根据所述分词单元的词向量和预设特征词的词向量，从所述多个分词单元中确定符合第二预设要求的分词单元，作为所述第二特征词。

- 15 [114] 在一个实施方式中，上述对所述多个分词单元分别进行词向量化处理，得到分词单元的词向量，具体可以包括以下内容：将分词单元中的汉字拆分为多个笔画；根据所述分词单元的多个笔画，建立分词单元的笔画向量；获取文本数据中与所述分词单元相连的词语，作为上下文词语；获取所述上下文词语的词向量；根据所述分词单元的笔画向量和所述上下文词语的词向量，确定所述分词单元的词向量。

- 20 [115] 在本实施方式中，考虑到现有的词向量化算法（例如 PageRank 方法）大多是针对英文字母的特点设计的，而单独的英文字母本身只能表音而无法表意。而在本实施方式中，所针对的分词单元主要是由汉字组成的。汉字不同于英文字母除了表音还能表意。即汉字本身的内部结构也能反映相应的语义信息。如果直接利用例如 PageRank 方法对由汉字构成的分词单元进行向量化处理得到的词向量所包含的信息往往会遗漏掉分词单元中汉字内部结构所表征的语义信息，即信息不够完整、准确，会影响后续第二特征
25 词的提取。

- [116] 正是考虑到上述问题，为了能够得到信息更为完整、准确的词向量，在本实施方式不是直接利用已有的 PageRank 等算法进行向量化处理，而是针对汉字的结构特点，根据分词单元中汉字的笔画，建立分词单元的笔画向量，以表征汉字的内部结构所携带的语义信息；并根据基于文本数据中与分词单元相连的词语（例如左邻词和右邻词），
30 得到上下文词语的词向量，以表征文本数据中分词单元的上下文外部信息；综合上述两

种不同的词向量，得到更为准确、完整的分词单元的词向量，有助于后续更准确地确定出第二特征词。

[117] 在一个实施方式中，具体实施时，还可以通过 Cw2vec 算法对所述多个分词单元分别进行词向量化处理，以得到分词单元的词向量。

- 5 [118] 在一个实施方式中，根据所述分词单元的词向量和预设特征词的词向量，从所述多个分词单元中确定符合第二预设要求的分词单元，作为所述第二特征词，具体实施时，可以包括以下内容：从多个预设特征词中提取预设数量的预设特征词作为测试词；根据所述多个预设特征词中除测试词以外的预设特征词的词向量，建立标记样本集；根据所述测试词的词向量和所述分词单元的词向量，建立非标记样本集；根据所述标记样本集和非标记样本集，通过迭代拟合，确定拟合分数阈值；根据所述拟合分数阈值，从
10 所述分词单元中确定出符合第二预设要求的分词单元，作为所述第二特征词。

- [119] 在本实施方式中，具体实施时，可以先从多个预设特征词中提取预设数量的预设特征词作为测试词（也可称为 spy）；根据所述多个预设特征词中除测试词以外的预设特征词的词向量，建立标记样本集（也可称为黑样本，positive，记为 P），并将该样
15 本集中的分词单元的词向量标记为 1；根据所述测试词的词向量和所述分词单元的词向量，建立非标记样本集（也可以称为白样本，定义为 unlabeled，记为 U），并将该样本集中的分词单元的词向量标记为 0。需要说明的是，上述标记样本集中的样本数量小于非标记样本集中的样本数量。

- [120] 在建立了上述标记样本集和非标记样本集后，可以根据所述标记样本集和非标
20 记样本集，通过多次迭代拟合，确定拟合分数阈值。

- [121] 具体的，可以按照以下步骤处理确定出拟合分数阈值：对标记样本集和非标记样本集中的分词单元的词向量分别进行 GBDT（Gradient Boosting Decision Tree，梯度提升决策树）拟合，根据拟合结果对每个分词单元的词向量进行打分（记为 score）；对拟合结果的分值，进一步作如下处理：将属于 P 的分词单元的词向量的 score 置为 1，
25 其余保持原来具体 score，然后进行归一化处理。按照上述步骤进行多次处理（例如重复两次），直到找出可以使得阈值比例（例如 90%）的被归入 U 中的预设特征词的词向量（即 spy）被识别出来阈值（记为 t）作为上述拟合分数阈值。

[122] 在确定出拟合分数阈值后，进一步可以根据上述拟合分数阈值，从分词单元向量中确定出符合第二预设要求的分词单元向量所对应的分词单元作为第二特征词。

[123] 具体的，可以按照以下步骤处理确定出第二特征词：将被归入 U 中的预设特征词的词向量（即 spy）重新归属回 P 中；再将剩下的 U 中所有拟合分数值小于拟合分数阈值（即 t）的分词单元的词向量的 score 赋为 0，P 中所有预设特征词的词向量的 score 赋为 1，其余保持当前的 score 然后进行归一化处理；再对所有的词向量进行 GBDT 拟合，并根据拟合结果对每个词的词向量进行重新打分得到拟合分数，记为 score'；根据拟合分数，将属于 P 的词向量的拟合分数 score' 置为 1，其余保持原有的 score'，然后进行归一化处理。按照上述步骤进行多次处理（例如重复 5 次），得到每个词向量最终的拟合分数，记为 score''，将 score'' 大于特征词分数阈值（记为 v）的词向量所对应的分词单元确定为第二特征词，即符合第二预设要求的新的能够表征该属性类型的词组。其中，上述特征词分数阈值具体可以根据具体情况和精度要求设置。对于特征词分数阈值的具体取值，本说明书不作限定。

[124] 在一个实施方式中，具体实施时，还可以通过 PU_learning 算法对所述多个分词单元的词向量进行分析处理，以确定出第二特征词。对此，本说明书不作限定。

[125] 在一个实施方式中，在根据所述分词单元的词向量和预设特征词的词向量，从所述多个分词单元中确定符合第二预设要求的分词单元，作为所述第二特征词前，所述方法还可以包括以下内容：过滤所述分词单元的词向量中的停用词的词向量。

[126] 在本实施方式中，上述停用词具体可以理解为所表征的内容没有实际意义或与交易数据的属性类型无关的词组。具体的，上述停用词可以是一些连接词或助词，例如“的”、“是”、“了”等，也可以是一些与交易数据的无关，宽泛的代词，例如“我”、“这”、“那”，还可以是数字、字母或单个字的词等等。当然，需要说明的是上述所列举的停用词只是一种示意性说明。具体实施时，根据具体的应用场景，上述停用词还可以包括其他的词，例如“在”、“有”、“人”、“一”等。对于上述停用词的具体内容，本说明书不作限定。

[127] 在本实施方式中，通过在根据所述分词单元的词向量和预设特征词的词向量，从所述多个分词单元中确定符合第二预设要求的分词单元，作为所述第二特征词前，过滤分词单元的词向量中的停用词的词向量，可以避免后续在对停用词的词向量进行分析处理时对时间、资源的占用，从而可以减少工作量，提高处理效率。

[128] 由上可见，本说明书实施例提供的特征词的确定方法，由于通过先对文本数据进行新词提取得到第一特征词；再利用第一特征词对分词库进行更新；进而利用更新后的分词库和预设特征词从文本数据提取出新的特征词作为第二特征词，从而避免了提取

特征词过程中，由于分词错误导致的特征词提取不准确、不全面，达到能精确地从文本数据中挖掘出符合要求的新的特征词的技术效果；又通过先将分词单元中的汉字拆分为多个笔画，得到分词单元的笔画向量；再根据分词单元的笔画向量和上下文词语的词向量，确定分词单元的词向量，从而使得得到的分词单元的词向量同时包含上下文外部信息

5 信息和汉字内部构造信息，能够反映出更加丰富、准确的语义信息；再基于上述分词单元的词向量进行第二特征词的提取，提高了确定特征词的准确度；还通过先根据从多个预设特征词中提取出的预设个数的预设特征词和分词单元建立非标记样本集，根据剩余的预设特征词建立标记样本集；再基于上述非标记样本集和标记样本通过多次迭代拟合，确定出较为准确的拟合分数阈值，以便根据上述拟合分数阈值从分词单元中确定出第二

10 特征词，进一步提高了确定特征词的准确度。

[129] 本说明书实施例还提供了一种服务器，包括处理器以及用于存储处理器可执行指令的存储器，所述处理器具体实施时可以根据指令执行以下步骤：获取文本数据；从所述文本数据中提取出第一特征词；根据所述第一特征词，更新分词库，得到更新后的分词库，其中，所述分词库包含多个用于表征预设属性类型的预设特征词；根据所述更

15 新后的分词库，从所述文本数据中提取第二特征词。

[130] 为了能够更加准确地完成上述指令，参阅图 6 所示，本说明书还提供了另一种具体的服务器，其中，所述服务器包括网络通信端口 601、处理器 602 以及存储器 603，上述结构通过内部线缆相连，以便各个结构可以进行具体的数据交互。

[131] 其中，所述网络通信端口 601，具体可以用于获取文本数据。

20 [132] 所述处理器 602，具体可以用于从所述文本数据中提取出第一特征词；根据所述第一特征词，更新分词库，得到更新后的分词库，其中，所述分词库包含多个用于表征预设属性类型的预设特征词；根据所述更新后的分词库，从所述文本数据中提取第二特征词。

[133] 所述存储器 603，具体可以用于存储经网络通信端口 601 输入的文本数据，以及

25 相应的指令程序。

[134] 在本实施方式中，所述网络通信端口 601 可以是与不同的通信协议进行绑定，从而可以发送或接收不同数据的虚拟端口。例如，所述网络通信端口可以是负责进行 web 数据通信的 80 号端口，也可以是负责进行 FTP 数据通信的 21 号端口，还可以是负责进行邮件数据通信的 25 号端口。此外，所述网络通信端口还可以是实体的通信接口

或者通信芯片。例如，其可以为无线移动通信芯片，如 GSM、CDMA 等；其还可以为 Wifi 芯片；其还可以为蓝牙芯片。

5 [135] 在本实施方式中，所述处理器 602 可以按任何适当的方式实现。例如，处理器可以采取例如微处理器或处理器以及存储可由该（微）处理器执行的计算机可读程序代码（例如软件或固件）的计算机可读介质、逻辑门、开关、专用集成电路（Application Specific Integrated Circuit, ASIC）、可编程逻辑控制器和嵌入微控制器的形式等等。本说明书并不作限定。

10 [136] 在本实施方式中，所述存储器 603 可以包括多个层次，在数字系统中，只要能保存二进制数据的都可以是存储器；在集成电路中，一个没有实物形式的具有存储功能的电路也叫存储器，如 RAM、FIFO 等；在系统中，具有实物形式的存储设备也叫存储器，如内存条、TF 卡等。

[137] 本说明书实施例还提供了一种基于上述特征词的确定方法的计算机存储介质，所述计算机存储介质存储有计算机程序指令，在所述计算机程序指令被执行时实现：获取文本数据；从所述文本数据中提取出第一特征词；根据所述第一特征词，更新分词库，
15 得到更新后的分词库，其中，所述分词库包含多个用于表征预设属性类型的预设特征词；根据所述更新后的分词库，从所述文本数据中提取第二特征词。

[138] 在本实施方式中，上述存储介质包括但不限于随机存取存储器（Random Access Memory, RAM）、只读存储器（Read-Only Memory, ROM）、缓存（Cache）、硬盘（Hard Disk Drive, HDD）或者存储卡（Memory Card）。所述存储器可以用于存储计算机程序
20 指令。网络通信单元可以是依照通信协议规定的标准设置的，用于进行网络连接通信的接口。

[139] 在本实施方式中，该计算机存储介质存储的程序指令具体实现的功能和效果，可以与其它实施方式对照解释，在此不再赘述。

25 [140] 参阅图 7 所示，在软件层面上，本说明书实施例还提供了一种特征词的确定装置，该装置具体可以包括以下的结构模块：

[141] 获取模块 71，具体可以用于获取文本数据；

[142] 第一提取模块 72，具体可以用于从所述文本数据中提取出第一特征词；

[143] 更新模块 73，具体可以用于根据所述第一特征词，更新分词库，得到更新后的分词库，其中，所述分词库包含多个用于表征预设属性类型的预设特征词；

- [144] 第二提取模块 74, 具体可以用于根据所述更新后的分词库, 从所述文本数据中提取第二特征词。
- [145] 在一个实施方式中, 所述文本数据具体可以包括: 交易附言, 和/或, 文本标签等。当然, 需要说明的是, 上述所列举的文本数据只是一种示意性说明。对于文本数据的具体类型、形式和内容, 本说明书不作限定。
- [146] 在一个实施方式中, 所述第一提取模块 72 具体可以包括以下结构单元:
- [147] 筛选单元, 具体可以用于从所述文本数据中筛选出字符长度小于预设长度阈值的字符串作为候选字符串;
- [148] 计算单元, 具体可以用于计算所述候选字符串的指标参数;
- 10 [149] 提取单元, 具体可以用于根据所述指标参数, 从所述候选字符串中提取符合第一预设要求的候选字符串, 作为第一特征词。
- [150] 在一个实施方式中, 所述指标参数具体可以包括以下至少之一: 凝固度、信息熵和频数等。当然, 需要说明的是, 上述所列举的指标参数只是一种示意性说明。具体实施时, 根据具体情况和要求, 还可以引入其他类型的特征参数作为上述指标参数。对此, 本说明书不作限定。
- 15 [151] 在一个实施方式中, 所述第一提取模块 72 具体还可以包括: 第一过滤单元, 具体可以用于过滤所述文本数据中的无效字符串。
- [152] 在一个实施方式中, 所述第二提取模块 74 具体可以包括以下结构单元:
- [153] 第一处理单元, 具体可以用于根据所述更新后的分词库, 对所述文本数据进行分词处理, 得到多个分词单元;
- 20 [154] 第二处理单元, 具体可以用于对所述多个分词单元分别进行词向量化处理, 得到分词单元的词向量;
- [155] 确定单元, 具体可以用于根据所述分词单元的词向量和预设特征词的词向量, 从所述多个分词单元中确定符合第二预设要求的分词单元, 作为所述第二特征词。
- 25 [156] 在一个实施方式中, 所述第二处理单元具体可以包括以下结构子单元:
- [157] 拆分子单元, 具体可以用于将分词单元中的汉字拆分为多个笔画;
- [158] 第一建立子单元, 具体可以用于根据所述分词单元的多个笔画, 建立分词单元

的笔画向量;

[159] 第一获取子单元, 具体可以用于获取文本数据中与所述分词单元相连的词语, 作为上下文词语;

[160] 第二获取子单元, 具体可以用于获取所述上下文词语的词向量;

5 [161] 第三获取子单元, 具体可以用于根据所述分词单元的笔画向量和所述上下文词语的词向量, 确定所述分词单元的词向量。

[162] 在一个实施方式中, 所述第二处理单元具体可以通过 Cw2vec 算法对所述多个分词单元分别进行词向量化处理, 以得到分词单元的词向量。

[163] 在一个实施方式中, 所述确定单元具体可以包括以下结构子单元:

10 [164] 提取子单元, 具体可以用于从多个预设特征词中提取预设数量的预设特征词作为测试词;

[165] 第二建立子单元, 具体可以用于根据所述多个预设特征词中除测试词以外的预设特征词的词向量, 建立标记样本集;

15 [166] 第三建立子单元, 具体可以用于根据所述测试词的词向量和所述分词单元的词向量, 建立非标记样本集;

[167] 第一确定子单元, 具体可以用于根据所述标记样本集和非标记样本集, 通过迭代拟合, 确定拟合分数阈值;

[168] 第二确定子单元, 具体可以用于根据所述拟合分数阈值, 从所述分词单元中确定出符合第二预设要求的分词单元, 作为所述第二特征词。

20 [169] 在一个实施方式中, 所述第二提取模块具体还可以包括第二过滤单元, 具体可以用于过滤所述分词单元的词向量中的停用词的词向量。

[170] 在一个实施方式中, 所述确定单元具体可以通过 PU_learning 算法对所述多个分词单元的词向量进行分析处理, 以确定出第二特征词。

25 [171] 需要说明的是, 上述实施例阐明的单元、装置或模块等, 具体可以由计算机芯片或实体实现, 或者由具有某种功能的产品来实现。为了描述的方便, 描述以上装置时以功能分为各种模块分别描述。当然, 在实施本说明书时可以把各模块的功能在同一个或多个软件和/或硬件中实现, 也可以将实现同一功能的模块由多个子模块或子单元的组合实现等。以上所描述的装置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,例如多个单元或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另一点,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口,装置或单元的间接耦合或通信连接,可以是电性,机械或其它的形式。

5 [172] 由上可见,本说明书实施例提供的特征词的确定装置,由于通过第一提取模块先对文本数据进行新词提取得到第一特征词;再通过更新模块利用第一特征词对分词库进行更新;进而通过第二提取模块利用更新后的分词库和预设特征词从文本数据提取出新的特征词作为第二特征词,从而避免了提取特征词过程中,由于分词错误导致的特征词提取不准确、不全面,达到能精确地从文本数据中挖掘出符合要求的新的特征词的技术效果。

10 [173] 虽然本说明书提供了如实施例或流程图所述的方法操作步骤,但基于常规或者无创造性的手段可以包括更多或者更少的操作步骤。实施例中列举的步骤顺序仅仅为众多步骤执行顺序中的一种方式,不代表唯一的执行顺序。在实际中的装置或客户端产品执行时,可以按照实施例或者附图所示的方法顺序执行或者并行执行(例如并行处理器或者多线程处理的环境,甚至为分布式数据处理环境)。术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、产品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、产品或者设备所固有的要素。在没有更多限制的情况下,并不排除在包括所述要素的过程、方法、产品或者设备中还存在另外的相同或等同要素。第一,第二等词语用来表示名称,而并不表示任何特定的顺序。

15 [174] 本领域技术人员也知道,除了以纯计算机可读程序代码方式实现控制器以外,完全可以通过将方法步骤进行逻辑编程来使得控制器以逻辑门、开关、专用集成电路、可编程逻辑控制器和嵌入微控制器等的形式来实现相同功能。因此这种控制器可以被认为是一种硬件部件,而对其内部包括的用于实现各种功能的装置也可以视为硬件部件内的结构。或者甚至,可以将用于实现各种功能的装置视为既可以是实现方法的软件模块又可以是硬件部件内的结构。

20 [175] 本说明书可以在由计算机执行的计算机可执行指令的一般上下文中描述,例如程序模块。一般地,程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构、类等等。也可以在分布式计算环境中实践本说明书,在这些分布式计算环境中,由通过通信网络而被连接的远程处理设备来执行任务。在分布式计算

环境中，程序模块可以位于包括存储设备在内的本地和远程计算机存储介质中。

[176] 通过以上的实施方式的描述可知，本领域的技术人员可以清楚地了解到本说明书可借助软件加必需的通用硬件平台的方式来实现。基于这样的理解，本说明书的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机
5 机软件产品可以存储在存储介质中，如 ROM/RAM、磁碟、光盘等，包括若干指令用以使得一台计算机设备（可以是个人计算机，移动终端，服务器，或者网络设备等）执行本说明书各个实施例或者实施例的某些部分所述的方法。

[177] 本说明书中的各个实施例采用递进的方式描述，各个实施例之间相同或相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。本说明书可
10 用于众多通用或专用的计算机系统环境或配置中。例如：个人计算机、服务器计算机、手持设备或便携式设备、平板型设备、多处理器系统、基于微处理器的系统、置顶盒、可编程的电子设备、网络 PC、小型计算机、大型计算机、包括以上任何系统或设备的分布式计算环境等等。

[178] 虽然通过实施例描绘了本说明书，本领域普通技术人员知道，本说明书有许多
15 变形和变化而不脱离本说明书的精神，希望所附的权利要求包括这些变形和变化而不脱离本说明书的精神。

权利要求书

- 1.一种特征词的确定方法, 包括:
获取文本数据;
从所述文本数据中提取出第一特征词;
- 5 根据所述第一特征词, 更新分词库, 得到更新后的分词库, 其中, 所述分词库包含多个用于表征预设属性类型的预设特征词;
根据所述更新后的分词库, 从所述文本数据中提取第二特征词。
- 2.根据权利要求1所述的方法, 所述文本数据包括: 交易附言, 和/或, 文本标签。
- 3.根据权利要求1所述的方法, 从所述文本数据中提取出第一特征词, 包括:
- 10 从所述文本数据中筛选出字符长度小于预设长度阈值的字符串作为候选字符串;
计算所述候选字符串的指标参数;
根据所述指标参数, 从所述候选字符串中提取符合第一预设要求的候选字符串, 作为第一特征词。
- 4.根据权利要求3所述的方法, 所述指标参数包括以下至少之一: 凝固度、信息熵
- 15 和频数。
- 5.根据权利要求3所述的方法, 在从所述文本数据中筛选出字符长度小于预设长度阈值的字符串作为候选字符串之前, 所述方法还包括:
过滤所述文本数据中的无效字符串。
- 6.根据权利要求1所述的方法, 根据所述更新后的分词库, 从所述文本数据中提取
- 20 第二特征词, 包括:
根据所述更新后的分词库, 对所述文本数据进行分词处理, 得到多个分词单元;
对所述多个分词单元分别进行词向量化处理, 得到分词单元的词向量;
根据所述分词单元的词向量和预设特征词的词向量, 从所述多个分词单元中确定符合第二预设要求的分词单元, 作为所述第二特征词。
- 25 7.根据权利要求6所述的方法, 对所述多个分词单元分别进行词向量化处理, 得到分词单元的词向量, 包括:
将分词单元中的汉字拆分为多个笔画;
根据所述分词单元的多个笔画, 建立分词单元的笔画向量;
获取文本数据中与所述分词单元相连的词语, 作为上下文词语;
- 30 获取所述上下文词语的词向量;
根据所述分词单元的笔画向量和所述上下文词语的词向量, 确定所述分词单元的词

向量。

8.根据权利要求6所述的方法,根据所述分词单元的词向量和预设特征词的词向量,从所述多个分词单元中确定符合第二预设要求的分词单元,作为所述第二特征词,包括:

从多个预设特征词中提取预设数量的预设特征词作为测试词;

5 根据所述多个预设特征词中除测试词以外的预设特征词的词向量,建立标记样本集;

根据所述测试词的词向量和所述分词单元的词向量,建立非标记样本集;

根据所述标记样本集和非标记样本集,通过迭代拟合,确定拟合分数阈值;

根据所述拟合分数阈值,从所述分词单元中确定出符合第二预设要求的分词单元,作为所述第二特征词。

10 9.根据权利要求6所述的方法,在根据所述分词单元的词向量和预设特征词的词向量,从所述多个分词单元中确定符合第二预设要求的分词单元,作为所述第二特征词前,所述方法还包括:

过滤所述分词单元的词向量中的停用词的词向量。

10.一种特征词的确定装置,包括:

15 获取模块,用于获取文本数据;

第一提取模块,用于从所述文本数据中提取出第一特征词;

更新模块,用于根据所述第一特征词,更新分词库,得到更新后的分词库,其中,所述分词库包含多个用于表征预设属性类型的预设特征词;

第二提取模块,用于根据所述更新后的分词库,从所述文本数据中提取第二特征词。

20 11.根据权利要求10所述的装置,所述文本数据包括:交易附言,和/或,文本标签。

12.根据权利要求10所述的装置,所述第一提取模块包括:

筛选单元,用于从所述文本数据中筛选出字符长度小于预设长度阈值的字符串作为候选字符串;

计算单元,用于计算所述候选字符串的指标参数;

25 提取单元,用于根据所述指标参数,从所述候选字符串中提取符合第一预设要求的候选字符串,作为第一特征词。

13.根据权利要求12所述的装置,所述指标参数包括以下至少之一:凝固度、信息熵和频数。

30 14.根据权利要求12所述的装置,所述第一提取模块还包括:第一过滤单元,用于过滤所述文本数据中的无效字符串。

15.根据权利要求10所述的装置,所述第二提取模块包括:

第一处理单元,用于根据所述更新后的分词库,对所述文本数据进行分词处理,得到多个分词单元;

第二处理单元,用于对所述多个分词单元分别进行词向量化处理,得到分词单元的词向量;

- 5 确定单元,用于根据所述分词单元的词向量和预设特征词的词向量,从所述多个分词单元中确定符合第二预设要求的分词单元,作为所述第二特征词。

16.根据权利要求 15 所述的装置,所述第二处理单元包括:

拆分子单元,用于将分词单元中的汉字拆分为多个笔画;

第一建立子单元,用于根据所述分词单元的多个笔画,建立分词单元的笔画向量;

- 10 第一获取子单元,用于获取文本数据中与所述分词单元相连的词语,作为上下文词语;

第二获取子单元,用于获取所述上下文词语的词向量;

第三获取子单元,用于根据所述分词单元的笔画向量和所述上下文词语的词向量,确定所述分词单元的词向量。

- 15 17.根据权利要求 15 所述的装置,所述确定单元包括:

提取子单元,用于从多个预设特征词中提取预设数量的预设特征词作为测试词;

第二建立子单元,用于根据所述多个预设特征词中除测试词以外的预设特征词的词向量,建立标记样本集;

- 20 第三建立子单元,用于根据所述测试词的词向量和所述分词单元的词向量,建立非标记样本集;

第一确定子单元,用于根据所述标记样本集和非标记样本集,通过迭代拟合,确定拟合分数阈值;

第二确定子单元,用于根据所述拟合分数阈值,从所述分词单元中确定出符合第二预设要求的分词单元,作为所述第二特征词。

- 25 18.根据权利要求 15 所述的装置,所述第二提取模块还包括第二过滤单元,用于过滤所述分词单元的词向量中的停用词的词向量。

19.一种服务器,包括处理器以及用于存储处理器可执行指令的存储器,所述处理器执行所述指令时实现权利要求 1 至 9 中任一项所述方法的步骤。

- 30 20.一种计算机可读存储介质,其上存储有计算机指令,所述指令被执行时实现权利要求 1 至 9 中任一项所述方法的步骤。