(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107436875 B
(45)授权公告日 2020.12.04

(21)申请号 201610354930.6
(22)申请日 2016.05.25
(65)同一申请的已公布的文献号
申请公布号 CN 107436875 A
(43)申请公布日 2017.12.05
(73)专利权人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼
(72)发明人 刘炳源 张旭
(74)专利代理机构 北京三高永信知识产权代理有限责任公司 11138
代理人 罗振安
(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/36(2019.01)

(56)对比文件
US 2012254083 A1,2012.10.04
US 7692573 B1,2010.04.06
WO 2009026433 A1,2009.02.26
CN 103678483 A,2014.03.26

审查员 曾伟涛

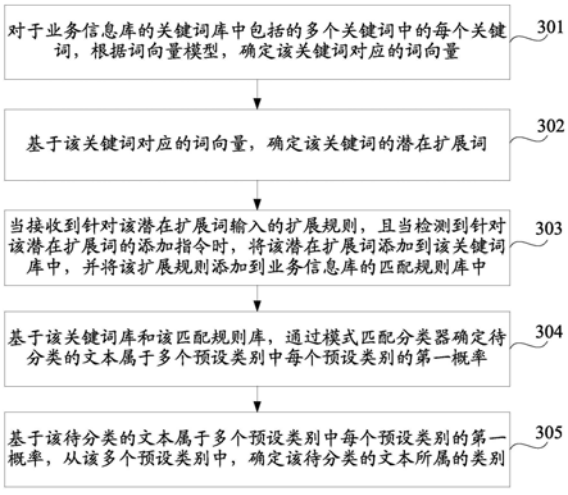
权利要求书4页 说明书19页 附图4页

(54)发明名称

文本分类方法及装置

(57)摘要

本发明公开了一种文本分类方法及装置，属于计算机技术领域。所述方法包括：对于业务信息库的关键词库中包括的多个关键词中的每个关键词，根据词向量模型，确定关键词对应的词向量；基于关键词对应的词向量，确定关键词的潜在扩展词；当接收到针对潜在扩展词输入的扩展规则，且当检测到针对潜在扩展词的添加指令时，将潜在扩展词添加到关键词库中，并将扩展规则添加到匹配规则库中；基于关键词库和匹配规则库，通过模式匹配分类器确定待分类的文本属于多个预设类别中每个预设类别的第一概率；基于第一概率，从多个预设类别中，确定待分类的文本所属的类别。本发明可以降低构建业务信息库的人工成本，且可以提高文本分类的覆盖率和准确率。



CN 107436875 B

1.一种文本分类方法,其特征在于,所述方法包括:

对于业务信息库的关键词库中包括的多个关键词中的每个关键词,根据词向量模型,确定所述关键词对应的词向量;

基于所述关键词对应的词向量,确定所述关键词的潜在扩展词;

当接收到针对所述潜在扩展词输入的扩展规则,且当检测到针对所述潜在扩展词的添加指令时,将所述潜在扩展词添加到所述关键词库中,并将所述扩展规则添加到所述业务信息库的匹配规则库中;

从待分类的文本中,通过模式匹配分类器获取与所述关键词库中的关键词相同的词语;将获取的词语确定为所述待分类的文本的关键词;

基于所述待分类的文本的关键词和所述匹配规则库,通过所述模式匹配分类器对所述待分类的文本进行分类,得到所述待分类的文本属于多个预设类别中每个预设类别的第一概率;其中,所述匹配规则库包括多个匹配规则,所述多个匹配规则中的每个匹配规则包含所述关键词库中的至少一个关键词,所述每个预设类别对应所述匹配规则库中的至少一个匹配规则;

基于所述待分类的文本属于所述多个预设类别中每个预设类别的第一概率,从所述多个预设类别中,确定所述待分类的文本所属的类别。

2.如权利要求1所述的方法,其特征在于,所述基于所述关键词对应的词向量,确定所述关键词的潜在扩展词,包括:

根据所述词向量模型,确定文本数据集包括的各个词语对应的词向量;

计算所述关键词对应的词向量与所述文本数据集包括的各个词语对应的词向量之间的相似度;

将所述文本数据集中的相似词语确定为所述关键词的潜在扩展词,所述相似词语对应的词向量与所述关键词对应的词向量之间的相似度大于指定相似度。

3.如权利要求1所述的方法,其特征在于,所述基于所述待分类的文本属于所述多个预设类别中每个预设类别的第一概率,从所述多个预设类别中,确定所述待分类的文本所属的类别,包括:

基于语义分类器中包括的多个预设语义特征向量,确定所述待分类的文本属于所述多个预设类别中每个预设类别的第二概率,所述多个预设语义特征向量与所述多个预设类别一一对应;

基于所述待分类的文本属于所述多个预设类别中每个预设类别的第一概率和所述待分类的文本属于所述多个预设类别中每个预设类别的第二概率,从所述多个预设类别中,确定所述待分类的文本所属的类别。

4.如权利要求3所述的方法,其特征在于,所述基于语义分类器中包括的多个预设语义特征向量,确定所述待分类的文本属于所述多个预设类别中每个预设类别的第二概率,包括:

通过所述语义分类器确定所述待分类的文本中每个词语的语义特征向量;

基于所述待分类的文本中每个词语的语义特征向量,通过所述语义分类器确定所述待分类的文本的语义特征向量;

对于所述多个预设语义特征向量中的每个预设语义特征向量,通过所述语义分类器计

算所述预设语义特征向量与所述待分类的文本的语义特征向量之间的相似度；

将计算得到的相似度确定为所述待分类的文本属于所述预设语义特征向量对应的预设类别的第二概率。

5.如权利要求3或4所述的方法，其特征在于，所述基于所述待分类的文本属于所述多个预设类别中每个预设类别的第一概率和所述待分类的文本属于所述多个预设类别中每个预设类别的第二概率，从所述多个预设类别中，确定所述待分类的文本所属的类别，包括：

从所述多个预设类别中，确定至少一个目标预设类别；

对于所述至少一个目标预设类别中的每个目标预设类别，确定所述目标预设类别对应的第一概率和第二概率；

基于所述模式匹配分类器对应的第一权重和所述语义分类器对应的第二权重，对所述目标预设类别对应的第一概率和第二概率进行加权平均，得到加权概率；

当所述加权概率大于指定概率时，确定所述目标预设类别为所述待分类的文本所属的类别。

6.如权利要求5所述的方法，其特征在于，所述从所述多个预设类别中，确定至少一个目标预设类别，包括：

将所述多个预设类别中的每个预设类别确定为所述目标预设类别。

7.如权利要求5所述的方法，其特征在于，所述从所述多个预设类别中，确定至少一个目标预设类别包括如下两种方式中的至少一种：

按照所述多个预设类别对应的多个第一概率由大到小的顺序，从所述多个第一概率中，获取N个第一概率，并将所述N个第一概率中每个第一概率对应的预设类别确定为所述目标预设类别，所述N为大于或等于1的自然数；或者，

按照所述多个预设类别对应的多个第二概率由大到小的顺序，从所述多个第二概率中，获取N个第二概率，并将所述N个第二概率中每个第二概率对应的预设类别确定为所述目标预设类别。

8.一种文本分类装置，其特征在于，所述装置包括：

第一确定单元，用于对于业务信息库的关键词库中包括的多个关键词中的每个关键词，根据词向量模型，确定所述关键词对应的词向量；

第二确定单元，用于基于所述关键词对应的词向量，确定所述关键词的潜在扩展词；

添加单元，用于当接收到针对所述潜在扩展词输入的扩展规则，且当检测到针对所述潜在扩展词的添加指令时，将所述潜在扩展词添加到所述关键词库中，并将所述扩展规则添加到所述业务信息库的匹配规则库中；

第三确定单元，用于从待分类的文本中，通过模式匹配分类器获取与所述关键词库中的关键词相同的词语；将获取的词语确定为所述待分类的文本的关键词；基于所述待分类的文本的关键词和所述匹配规则库，通过所述模式匹配分类器对所述待分类的文本进行分类，得到所述待分类的文本属于多个预设类别中每个预设类别的第一概率；其中，所述匹配规则库包括多个匹配规则，所述多个匹配规则中的每个匹配规则包含所述关键词库中的至少一个关键词，所述每个预设类别对应所述匹配规则库中的至少一个匹配规则；

第四确定单元，用于基于所述待分类的文本属于所述多个预设类别中每个预设类别的

第一概率，从所述多个预设类别中，确定所述待分类的文本所属的类别。

9.如权利要求8所述的装置，其特征在于，

所述第二确定单元，用于基于所述关键词对应的词向量，确定所述关键词的潜在扩展词，具体为：

根据所述词向量模型，确定文本数据集包括的各个词语对应的词向量；

计算所述关键词对应的词向量与所述文本数据集包括的各个词语对应的词向量之间的相似度；

将所述文本数据集中的相似词语确定为所述关键词的潜在扩展词，所述相似词语对应的词向量与所述关键词对应的词向量之间的相似度大于指定相似度。

10.如权利要求8所述的装置，其特征在于，

所述第四确定单元，用于基于所述待分类的文本属于所述多个预设类别中每个预设类别的第一概率，从所述多个预设类别中，确定所述待分类的文本所属的类别，具体为：

基于语义分类器中包括的多个预设语义特征向量，确定所述待分类的文本属于所述多个预设类别中每个预设类别的第二概率，所述多个预设语义特征向量与所述多个预设类别一一对应；

基于所述待分类的文本属于所述多个预设类别中每个预设类别的第一概率和所述待分类的文本属于所述多个预设类别中每个预设类别的第二概率，从所述多个预设类别中，确定所述待分类的文本所属的类别。

11.如权利要求10所述的装置，其特征在于，

所述第四确定单元，用于基于语义分类器中包括的多个预设语义特征向量，确定所述待分类的文本属于所述多个预设类别中每个预设类别的第二概率，具体为：

通过所述语义分类器确定所述待分类的文本中每个词语的语义特征向量；

基于所述待分类的文本中每个词语的语义特征向量，通过所述语义分类器确定所述待分类的文本的语义特征向量；

对于所述多个预设语义特征向量中的每个预设语义特征向量，通过所述语义分类器计算所述预设语义特征向量与所述待分类的文本的语义特征向量之间的相似度；

将计算得到的相似度确定为所述待分类的文本属于所述预设语义特征向量对应的预设类别的第二概率。

12.如权利要求10或11所述的装置，其特征在于，

所述第四确定单元，用于基于所述待分类的文本属于所述多个预设类别中每个预设类别的第一概率和所述待分类的文本属于所述多个预设类别中每个预设类别的第二概率，从所述多个预设类别中，确定所述待分类的文本所属的类别，具体为：

从所述多个预设类别中，确定至少一个目标预设类别；

对于所述至少一个目标预设类别中的每个目标预设类别，确定所述目标预设类别对应的第一概率和第二概率；

基于所述模式匹配分类器对应的第一权重和所述语义分类器对应的第二权重，对所述目标预设类别对应的第一概率和第二概率进行加权平均，得到加权概率；

当所述加权概率大于指定概率时，确定所述目标预设类别为所述待分类的文本所属的类别。

13.如权利要求12所述的装置,其特征在于,

所述第四确定单元,用于所述从所述多个预设类别中,确定至少一个目标预设类别,具体为:

将所述多个预设类别中的每个预设类别确定为所述目标预设类别。

14.如权利要求12所述的装置,其特征在于,

所述第四确定单元,用于所述从所述多个预设类别中,确定至少一个目标预设类别,具体为:

按照所述多个预设类别对应的多个第一概率由大到小的顺序,从所述多个第一概率中,获取N个第一概率,并将所述N个第一概率中每个第一概率对应的预设类别确定为所述目标预设类别,所述N为大于或等于1的自然数;或者,

按照所述多个预设类别对应的多个第二概率由大到小的顺序,从所述多个第二概率中,获取N个第二概率,并将所述N个第二概率中每个第二概率对应的预设类别确定为所述目标预设类别。

15.一种文本分类装置,其特征在于,所述装置包括:

处理器和存储器;

其中,所述存储器中存有计算机可读程序;

所述处理器通过运行所述存储器中的程序,以用于完成权利要求1-7任一所述的方法。

文本分类方法及装置

技术领域

[0001] 本发明涉及计算机技术领域，特别涉及一种文本分类方法及装置。

背景技术

[0002] 随着计算机技术的快速发展，海量的信息资源以文本的形式存在。由于海量的信息资源中往往蕴含有大量真实而有价值的信息，如当该信息资源为某一业务中用户与客服的对话数据时，该对话数据往往会体现出业务开展情况、业务问题反馈等方面的信息，因此，为了从海量的信息资源中快速而有效发掘出有价值的信息，需要对该信息资源对应的文本进行分类。

[0003] 目前，提供了一种文本分类方法，具体为：技术人员手动构建业务信息库的关键词库和匹配规则库，构建完成之后，基于该关键词库和该匹配规则库，确定待分类的文本属于多个预设类别中某个预设类别的概率，当该概率大于指定概率时，将该预设类别确定为该待分类的文本所属的类别。

[0004] 由于业务信息库的关键词库和匹配规则库完全由人工手动构建，因此，该关键词库和该匹配规则库的构建时间较长，构建效率较低，且建立和维护该关键词库和该匹配规则库的人工成本较大。另外，由于该关键词库和该匹配规则库完全依赖于技术人员的经验总结，因此，上述文本分类方法的准确率和覆盖率都较低。

发明内容

[0005] 为了解决现有技术的问题，本发明实施例提供了一种文本分类方法及装置。所述技术方案如下：

[0006] 第一方面，提供了一种文本分类方法，所述方法用于文本分类装置中，所述方法包括：

[0007] 对于业务信息库的关键词库中包括的多个关键词中的每个关键词，根据词向量模型，确定所述关键词对应的词向量。由于词向量模型可以将词语转换为词向量，因此，可以先根据该词向量模型，确定关键词对应的词向量，以便后续可以基于该关键词对应的词向量确定该关键词的潜在扩展词。其中，词向量模型是由词向量工具经过训练得到的，该词向量工具例如可以为Word2vec(Word to vector)工具等。

[0008] 基于所述关键词对应的词向量，确定所述关键词的潜在扩展词。

[0009] 当接收到针对所述潜在扩展词输入的扩展规则，且当检测到针对所述潜在扩展词的添加指令时，将所述潜在扩展词添加到所述关键词库中，并将所述扩展规则添加到所述业务信息库的匹配规则库中。其中，添加指令用于指示将该潜在扩展词添加到关键词库中，并将针对该潜在扩展词输入的扩展规则添加到匹配规则库中，该添加指令可以由技术人员触发，该技术人员可以通过指定操作触发，该指定操作可以为单击操作、双击操作、语音操作等。

[0010] 基于所述关键词库和所述匹配规则库，通过模式匹配分类器确定待分类的文本属

于多个预设类别中每个预设类别的第一概率。其中，模式匹配分类器用于根据业务信息库的关键词库和匹配规则库对待分类的文本进行分类。另外，待分类的文本可以是文本数据集中一个，也可以是服务器在与客户端进行通信的过程中实时获取的。

[0011] 基于所述待分类的文本属于所述多个预设类别中每个预设类别的第一概率，从所述多个预设类别中，确定所述待分类的文本所属的类别。其中，该多个预设类别可以预先设置，如该多个预设类别可以为积分兑换-积分查询、国际漫游-资费咨询、信用开机-停机原因。

[0012] 在本发明实施例中，关键词库和匹配规则库是采用半人工的方式构建的，从而不仅节省了构建时间，提高了构建效率，且降低了建立和维护该关键词库和该匹配规则库的人工成本。另外，由于可以根据潜在扩展词来对该关键词库和该匹配规则库进行扩展更新，因此，可以提高该关键词库和该匹配规则库的覆盖率，且由于该潜在扩展词和该扩展规则是基于添加指令添加到该关键词库和该匹配规则库中的，也即是该潜在扩展词和该扩展规则是在技术人员确认后添加到该关键词库和该匹配规则库中的，因此，可以提高该关键词库和该匹配规则库的准确率，进而可以提高基于该关键词库和该匹配规则库进行文本分类时的覆盖率和准确率。

[0013] 在一种可能的设计中，所述基于所述关键词对应的词向量，确定所述关键词的潜在扩展词，包括：

[0014] 根据所述词向量模型，确定文本数据集包括的各个词语对应的词向量。其中，该文本数据集中包括多个文本，该多个文本为与多个预设类别相关的文本，如该多个文本可以为该多个预设类别的说明文本、客服文本等，本发明实施例对此不做具体限定。另外，说明文本为记录有预设类别的说明数据的文本，客服文本为记录有与预设类别相关的用户和客服的对话数据的文本。

[0015] 计算所述关键词对应的词向量与所述文本数据集包括的各个词语对应的词向量之间的相似度。由于关键词的潜在扩展词的语义信息与该关键词的语义信息相近，因此，可以计算该关键词对应的词向量与该文本数据集包括的各个词语对应的词向量之间的相似度，以便根据该相似度确定某个词语是否为该关键词的潜在扩展词。可选地，可以计算两者之间的欧式距离，将计算得到的欧式距离确定为两者之间的相似度；或者，可以计算两者之间的余弦相似度，将计算得到的余弦相似度确定为两者之间的相似度，本发明实施例对此不做具体限定。

[0016] 将所述文本数据集中的相似词语确定为所述关键词的潜在扩展词，所述相似词语对应的词向量与所述关键词对应的词向量之间的相似度大于指定相似度。由于当某个词语对应的词向量与关键词对应的词向量之间的相似度大于指定相似度时，表明该词语的语义信息与该关键词的语义信息较为相近，因此，可以将该词语确定为该关键词的潜在扩展词。其中，指定相似度可以根据具体的业务需要预先设置，如该指定相似度可以为0.75、0.8等，本发明实施例对此不做具体限定。

[0017] 在本发明实施例中，可以直接根据该关键词对应的词向量与该文本数据集包括的各个词语对应的词向量之间的相似度，来从该文本数据集包括的多个词语中，确定该关键词的潜在扩展词，操作简单，便于实现，且由于某个词语的词向量可以体现该词语的语义信息，因此，通过本发明实施例中的方法确定的潜在扩展词的准确度较高。

[0018] 在一种可能的设计中，所述基于所述关键词库和所述匹配规则库，通过模式匹配分类器确定待分类的文本属于多个预设类别中每个预设类别的第一概率，包括：

[0019] 从所述待分类的文本中，通过所述模式匹配分类器获取与所述关键词库中的关键词相同的词语。例如，该模式匹配分类器可以通过Wu-Manber算法等多模式匹配算法，从待分类的文本中，获取与该关键词库中的关键词相同的词语。由于Wu-Manber算法可以通过跳转表来快速进行字符串匹配，因此，该模式匹配分类器通过Wu-Manber算法，从待分类的文本中，获取与该关键词库中的关键词相同的词语时，可以提高获取效率。

[0020] 将获取的词语确定为所述待分类的文本的关键词。由于关键词库中的多个关键词是与匹配规则库中的多个匹配规则相对应的，因此，当该待分类的文本中包含与该关键词库中的关键词相同的词语时，可以将该词语确定为该待分类的文本的关键词，以便后续可以基于该待分类的文本的关键词，从匹配规则库中，查找与该待分类的文本的关键词对应的匹配规则。

[0021] 基于所述待分类的文本的关键词和所述匹配规则库，通过所述模式匹配分类器对所述待分类的文本进行分类，得到所述待分类的文本属于所述多个预设类别中每个预设类别的第一概率。例如，对于该匹配规则库包括的多个匹配规则中的每个匹配规则，该模式匹配分类器可以基于该待分类的文本的关键词，判断该待分类的文本是否满足该匹配规则，当该待分类的文本满足该匹配规则时，确定该待分类的文本属于该匹配规则对应的预设类别的第一概率为1，而当该待分类的文本不满足该匹配规则时，可以基于该待分类的文本的关键词，确定该待分类的文本的关键词向量，并基于该匹配规则中包含的至少一个关键词，确定该匹配规则的关键词向量，之后，计算该待分类的文本的关键词向量与该匹配规则的关键词向量之间的相似度，将计算得到的相似度确定为该待分类的文本属于该匹配规则对应的预设类别的第一概率。通过模式匹配分类器对该待分类的文本中的关键词进行识别，该识别过程简单方便，且由于该关键词库是经过扩展更新的，因此，可以更为准确地识别出该待分类文本的关键词，之后，可以基于该待分类文本的关键词和该匹配规则库，通过该模式匹配分类器对该待分类的文本进行分类，得到该待分类的文本属于该多个预设类别中每个预设类别的第一概率，由于该待分类的关键词具有准确度，因此，可以保证得到的第一概率准确度。

[0022] 在一种可能的设计中，所述基于所述待分类的文本属于所述多个预设类别中每个预设类别的第一概率，从所述多个预设类别中，确定所述待分类的文本所属的类别，包括：

[0023] 基于语义分类器中包括的多个预设语义特征向量，确定所述待分类的文本属于所述多个预设类别中每个预设类别的第二概率，所述多个预设语义特征向量与所述多个预设类别一一对应。其中，语义分类器用于基于待分类的文本的语义信息对该待分类的文本进行分类，如该语义分类器可以为递归卷积神经网络等。

[0024] 基于所述待分类的文本属于所述多个预设类别中每个预设类别的第一概率和所述待分类的文本属于所述多个预设类别中每个预设类别的第二概率，从所述多个预设类别中，确定所述待分类的文本所属的类别。

[0025] 在本发明实施例中，可以将模式匹配分类器和语义分类器结合起来进行文本分类，从而避免了单一依赖某一方法来进行文本分类带来的不准确性，提高了文本分类的准确率。

[0026] 在一种可能的设计中，所述基于语义分类器中包括的多个预设语义特征向量，确定所述待分类的文本属于所述多个预设类别中每个预设类别的第二概率，包括：

[0027] 通过所述语义分类器确定所述待分类的文本中每个词语的语义特征向量；基于所述待分类的文本中每个词语的语义特征向量，通过所述语义分类器确定所述待分类的文本的语义特征向量。例如，可以基于该待分类的文本中每个词语的语义特征向量，通过语义分类器中的文本特征提取层确定该待分类的文本的语义特征向量。其中，待分类的文本的语义特征向量为可以体现该待分类的文本的语义信息的特征向量。

[0028] 对于所述多个预设语义特征向量中的每个预设语义特征向量，通过所述语义分类器计算所述预设语义特征向量与所述待分类的文本的语义特征向量之间的相似度。例如，可以通过语义分类器中的分类层计算该预设语义特征向量与该待分类的文本的语义特征向量之间的相似度。

[0029] 将计算得到的相似度确定为所述待分类的文本属于所述预设语义特征向量对应的预设类别的第二概率。

[0030] 需要说明的是，词特征提取层和文本特征提取层可以采用递归结构，以保证通过该词特征提取层确定的词语的语义特征向量不仅可以体现出该词语的语义信息，还可以体现出该词语在上下文中的依赖关系，进而保证基于该词语的语义特征向量，通过该文本特征提取层确定的待分类的文本的语义特征向量可以完整体现出该待分类的文本的语义信息。

[0031] 在本发明实施例中，语义分类器是基于待分类的文本的语义信息对该待分类的文本进行分类的，从而可以有效避免由于信息缺乏而导致的对待分类的文本的误分类，提高了文本分类的准确率。

[0032] 在一种可能的设计中，所述基于所述待分类的文本属于所述多个预设类别中每个预设类别的第一概率和所述待分类的文本属于所述多个预设类别中每个预设类别的第二概率，从所述多个预设类别中，确定所述待分类的文本所属的类别，包括：

[0033] 从所述多个预设类别中，确定至少一个目标预设类别；

[0034] 对于所述至少一个目标预设类别中的每个目标预设类别，确定所述目标预设类别对应的第一概率和第二概率；

[0035] 基于所述模式匹配分类器对应的第一权重和所述语义分类器对应的第二权重，对所述目标预设类别对应的第一概率和第二概率进行加权平均，得到加权概率。其中，第一权重和第二权重可以预先设置，且实际应用中，第一权重和第二权重可以根据模式匹配分类器和语义分类器的可靠性来进行设置和调整，且还可以根据不同的业务需求来进行设置和调整。

[0036] 当所述加权概率大于指定概率时，确定所述目标预设类别为所述待分类的文本所属的类别。

[0037] 在本发明实施例中，当根据模式匹配分类器和语义分类器的可靠性来设置和调整第一权重和第二权重时，可以保证得到的加权概率的准确度，进而保证基于该加权概率确定的该待分类的文本所属的类别的准确度。而当根据不同的业务需求来设置和调整第一权重和第二权重时，可以保证得到的加权概率符合技术人员的文本分类需求，进而保证基于该加权概率确定的该待分类的文本所属的类别的符合技术人员的文本分类需求。

[0038] 在一种可能的设计中，所述从所述多个预设类别中，确定至少一个目标预设类别，包括：

[0039] 将所述多个预设类别中的每个预设类别确定为所述目标预设类别。

[0040] 在本发明实施例中，可以直接将该多个预设类别中的每个预设类别确定为目标预设类别，此时，服务器无需再执行其它操作，从而可以提高目标预设类别的确定效率。

[0041] 在另一种可能的设计中，所述从所述多个预设类别中，确定至少一个目标预设类别包括如下两种方式中的至少一种：

[0042] 按照所述多个预设类别对应的多个第一概率由大到小的顺序，从所述多个第一概率中，获取N个第一概率，并将所述N个第一概率中每个第一概率对应的预设类别确定为所述目标预设类别，所述N为大于或等于1的自然数；或者，

[0043] 按照所述多个预设类别对应的多个第二概率由大到小的顺序，从所述多个第二概率中，获取N个第二概率，并将所述N个第二概率中每个第二概率对应的预设类别确定为所述目标预设类别。

[0044] 在本发明实施例中，可以基于第一概率和第二概率，确定至少一个目标预设类别，此时该至少一个目标预设类别为该多个预设类别中的部分预设类别，因此，服务器在后续步骤中不需对该多个预设类别都计算加权概率，而只需对该多个预设类别中的部分预设类别计算加权概率，从而可以节省服务器的处理资源，且可以提高文本分类效率。

[0045] 第二方面，本发明实施例提供了一种文本分类装置，该文本分类装置具有实现上述第一方面中文本分类装置行为的功能。所述功能可以通过硬件实现，也可以通过硬件执行相应的软件实现。所述硬件或软件包括一个或多个与上述功能相对应的模块。

[0046] 在一种可能的设计中，文本分类的装置的结构中包括处理器和存储器，所述存储器用于存储支持文本分类装置执行上述方法的程序，所述处理器被配置为用于执行所述存储器中存储的程序。所述文本分类装置还可以包括通信接口，用于所述文本分类装置与其他装置或通信网络通信。

[0047] 第三方面，本发明实施例提供了一种计算机存储介质，用于储存为上述文本分类装置所用的计算机软件指令，其包含用于执行上述方面为文本分类装置所设计的程序。

[0048] 相较于现有技术，在本发明实施例中，对于业务信息库的关键词库包括的多个关键词中的每个关键词，根据词向量模型，确定该关键词对应的词向量，并基于该关键词对应的词向量，确定该关键词的潜在扩展词，之后，当接收到针对该潜在扩展词输入的扩展规则，且当检测到针对该潜在扩展词的添加指令时，将该潜在扩展词添加到该关键词库中，并将该扩展规则添加到业务信息库的匹配规则库中，也即是，该关键词库和该匹配规则库是采用半人工的方式构建的，不仅可以节省构建时间，提高构建效率，且可以降低建立和维护该关键词库和该匹配规则库的人工成本。另外，由于可以根据服务器确定的潜在扩展词对该关键词库和该匹配规则库进行扩展更新，因此，可以提高该关键词库和该匹配规则库的覆盖率，且由于添加指令是由技术人员触发的，因此，该潜在扩展词和该扩展规则是在技术人员确认后添加到该关键词库和该匹配规则库中的，从而可以提高该关键词库和该匹配规则库的准确率。再者，由于提高了该关键词库和该匹配规则库的覆盖率和准确率，因此，基于该关键词库和该匹配规则库，确定待分类的文本属于多个预设类别中每个预设类别的第一概率，并基于该待分类的文本属于该多个预设类别中每个预设类别的第一概率，从该多

个预设类别中，确定该待分类的文本所属的类别时，可以提高文本分类的覆盖率和准确率。

[0049] 本发明的这些方面或其他方面在以下实施例的描述中会更加简明易懂。

附图说明

[0050] 为了更清楚地说明本发明实施例中的技术方案，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，下面描述中的附图仅仅是本发明的一些实施例，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它的附图。

[0051] 图1A是本发明实施例提供的一种客服文本的示意图；

[0052] 图1B是本发明实施例提供的一种文本分类方法所涉及的系统架构的示意图；

[0053] 图1C是本发明实施例提供的一种服务器的结构示意图；

[0054] 图2为本发明实施例提供的一种计算机设备的结构示意图；

[0055] 图3是本发明实施例提供的一种文本分类方法的流程图；

[0056] 图4是本发明实施例提供的一种文本分类装置的结构示意图。

具体实施方式

[0057] 为使本发明的目的、技术方案和优点更加清楚，下面将结合附图对本发明实施方式作进一步地详细描述。

[0058] 在对本发明实施例进行详细地解释说明之前，先对本发明实施例所涉及的应用场景予以说明。

[0059] 客服平台往往是电信运营商或者互联网运营商最重要的服务窗口，如移动10086平台、淘宝客服平台等。以移动10086平台为例，2015年上半年的日均客服接通量约为550万，该移动10086平台中每天都会有数百万条的客服数据被存储，一条客服数据就是一次用户与客服的对话记录。由于该客服数据往往是以录音形式进行存储的，因此，为了便于处理该客服数据，往往可以对该客服数据进行语音解析，以将该客服数据转换为客服文本，图1A所示为一条客服文本的具体示例。

[0060] 由于海量的客服数据中往往蕴含有大量真实而有价值的信息，如业务开展情况、业务问题反馈等方面的信息，这些信息可以为产品创新、营销完善、网络优化等措施的实施提供重要参考。因此，为了从海量的客服数据中快速而有效发掘出有价值的信息，可以对该客服数据对应的客服文本进行分类。

[0061] 相关技术中，技术人员可以手动构建业务信息库的关键词库和匹配规则库，构建完成之后，基于该关键词库和该匹配规则库，从多个预设类别中，确定待分类的文本所属的类别。由于当业务信息库的关键词库和匹配规则库完全由人工手动构建时，该业务信息库的构建时间较长，构建效率较低，且建立和维护该业务信息库的人工成本较大，又由于当业务信息库完全依赖于技术人员的经验总结时，基于该业务信息库进行的文本分类的准确率和覆盖率都将较低，因此，本发明实施例提供了一种文本分类方法，来在降低人工成本的前提下，提高文本分类的准确率和覆盖率。

[0062] 图1B是本发明实施例提供的一种文本分类方法所涉及的系统架构的示意图。参见图1B，该系统架构包括：服务器110，客户端120以及便于服务器110和客户端120建立通信连

接的网络130。其中，用户可以通过客户端120来与服务器110进行通信。需要说明的是，不同的用户可以通过不同类型的客户端120与服务器110进行通信，例如，用户A可以通过个人电脑120-1与服务器110进行通信，用户B可以通过移动终端120-2与服务器110进行通信，本发明实施例对此不做具体限定。用户通过客户端120与服务器110进行通信的过程中会产生通信数据，该通信数据可以以文本的形式进行表现，例如，该通信数据可以表现为图1A所示的客服文本。服务器110用于获取待分类的文本，并对该待分类的文本进行分类，其中，待分类的文本可以是服务器110在与客户端120进行通信的过程中实时获取的，也可以是事先存储在服务器110或其它存储设备上的文本，本发明实施例对此不做具体限定。

[0063] 图1C是本发明实施例提供的一种服务器110的结构示意图。参见图1C，服务器110包括：业务信息库构建模块1101，预处理模块1102，模式匹配分类模块1103，语义分类模块1104，加权融合模块1105和类别确定模块1106。

[0064] 当服务器110需要对某一待分类的文本进行分类时，可以先将该待分类的文本输入到预处理模块1102中；该预处理模块1102可以对该待分类的文本进行预处理，以去除该待分类的文本中一些显著的噪声干扰，之后，预处理模块1102可以将该预处理后的待分类的文本分别输入到模式匹配分类模块1103和语义分类模块1104中；模式匹配模块1103可以基于业务信息库构建模块1101构建的业务信息库，通过模式匹配分类器对该待分类的文本进行分类，得到该待分类的文本属于多个预设类别中每个预设类别的第一概率，并将该第一概率输入到加权融合模块1105中；语义分类模块1104可以通过语义分类器对该待分类的文本进行分类，得到该待分类的文本属于多个预设类别中每个预设类别的第二概率，并将该第二概率输入到加权融合模块1105中；加权融合模块1105可以根据该待分类的文本属于多个预设类别中每个预设类别的第一概率和该待分类的文本属于多个预设类别中每个预设类别的第二概率，确定该待分类的文本属于该多个预设类别中某些预设类别的加权概率，并将该加权概率输入到类别确定模块1106中；类别确定模块1106可以根据该待分类的文本属于该多个预设类别中某些预设类别的加权概率，从该多个预设类别中，确定该待分类的文本所属的类别。

[0065] 具体地，将服务器110包括的多个模块各自的功能阐述如下：

[0066] 业务信息库构建模块1101，用于根据词向量模型构建业务信息库。业务信息库构建模块1101可以包括种子库构建子模块11011和关键词扩展子模块11012。其中，种子库构建子模块11011中包括技术人员事先构建的种子信息库，该种子信息库中包括关键词库和匹配规则库，该关键词库中包括多个关键词，该匹配规则库中包括多个匹配规则，该多个匹配规则中的每个匹配规则包含该关键词库中的至少一个关键词。关键词扩展子模块11012用于根据词向量模型，确定该多个关键词中每个关键词的潜在扩展词。之后，技术人员可以基于该多个关键词的潜在扩展词，对该种子信息库的关键词库和匹配规则库进行扩展更新，并将扩展更新后的种子信息库确定为业务信息库。

[0067] 预处理模块1102，用于对待分类的文本进行预处理。由于待分类的文本中包括了大量无用信息，如语气词、助词、连词等，因此，可以通过对待分类的文本进行预处理来去除该待分类的文本中一些显著的噪声干扰。一种可能的设计中，该预处理可以包括中文分词、词性过滤和停用词过滤中的至少一种。其中，中文分词是指将文本转换为中文单词的集合，词性过滤是指去掉文本中的语气词、助词、连词等，停用词过滤是指去掉文本中没有实际含

义的词，如去掉文本中的“的”、“那么”等。

[0068] 模式匹配分类模块1103，用于基于业务信息库构建模块1101构建的业务信息库，通过模式匹配分类器对预处理模块1102预处理后的待分类的文本进行分类，得到该待分类的文本属于多个预设类别中每个预设类别的第一概率。模式匹配分类模块1103可以包括关键词确定子模块11031和规则分类子模块11032。其中，关键词确定子模块11031用于基于业务信息库的关键词库，通过模式匹配分类器确定待分类的文本的关键词，一种可能的设计中，模式匹配分类器可以基于该关键词库，通过多模式匹配算法确定该待分类的文本的关键词。其中，规则分类子模块11032用于基于该待分类的文本的关键词和该业务信息库的匹配规则库，通过模式匹配分类器对该待分类的文本进行分类，得到该待分类的文本属于多个预设类别中每个预设类别的第一概率。

[0069] 语义分类模块1104，用于通过语义分类器对预处理模块1102预处理后的待分类的文本进行分类，得到该待分类的文本属于多个预设类别中每个预设类别的第二概率。语义分类模块1104对该待分类的文本进行分类时，可以基于语义分类器中包括的多个预设语义特征向量，确定该待分类的文本属于多个预设类别中每个预设类别的第二概率。具体地，可以通过语义分类器确定该待分类的文本中每个词语的语义特征向量；基于该待分类的文本中每个词语的语义特征向量，通过语义分类器确定该待分类的文本的语义特征向量；对于该多个预设语义特征向量中的每个预设语义特征向量，通过语义分类器计算该预设语义特征向量与该待分类的文本的语义特征向量之间的相似度；将计算得到的相似度确定为该待分类的文本属于该预设语义特征向量对应的预设类别的第二概率。

[0070] 加权融合模块1105，用于根据模式匹配分类模块1103得到的该待分类的文本属于多个预设类别中每个预设类别的第一概率和语义分类模块1104得到的该待分类的文本属于多个预设类别中每个预设类别的第二概率，确定该待分类的文本属于该多个预设类别中某些预设类别的加权概率。一种可能的设计中，可以基于模式匹配分类器对应的第一权重和语义分类器对应的第二权重，对某一预设类别对应的第一概率和第二概率进行加权平均，得到该待分类的文本属于该预设类别的加权概率。

[0071] 类别确定模块1106，用于根据加权融合模块1105得到的该待分类的文本属于该多个预设类别中某些预设类别的加权概率，从该多个预设类别中，确定该待分类的文本所属的类别。一种可能的设计中，当该待分类的文本属于该多个预设类别中某个预设类别的加权概率大于指定概率时，可以将该预设类别确定为该待分类的文本所属的类别。其中，指定概率可以根据具体的业务需求预先设置，本发明实施例对此不做具体限定。

[0072] 图2为本发明实施例提供的一种计算机设备的结构示意图，图1B或者图1C中的服务器110可以以图2中所示的计算机设备来实现。参见图2，该计算机设备包括至少一个处理器201，通信总线202，存储器203以及至少一个通信接口204。

[0073] 处理器201可以是一个通用中央处理器(CPU)，微处理器，特定应用集成电路(application-specific integrated circuit,ASIC)，或一个或多个用于控制本发明方案程序执行的集成电路。

[0074] 通信总线202可包括一通路，在上述组件之间传送信息。

[0075] 存储器203可以是只读存储器(read-only memory,ROM)或可存储静态信息和指令的其它类型的静态存储设备，随机存取存储器(random access memory,RAM))或者可存储

信息和指令的其它类型的动态存储设备，也可以是电可擦可编程只读存储器(Electrically Erasable Programmable Read-Only Memory，EEPROM)、只读光盘(Compact Disc Read-Only Memory，CD-ROM)或其它光盘存储、光碟存储(包括压缩光碟、激光碟、光碟、数字通用光碟、蓝光光碟等)、磁盘存储介质或者其它磁存储设备、或者能够用于携带或存储具有指令或数据结构形式的期望的程序代码并能够由计算机存取的任何其它介质，但不限于此。存储器203可以是独立存在，通过通信总线202与处理器201相连接。存储器203也可以和处理器201集成在一起。

[0076] 通信接口204，使用任何收发器一类的装置，用于与其它设备或通信网络通信，如以太网，无线接入网(RAN)，无线局域网(Wireless Local Area Networks，WLAN)等。

[0077] 在具体实现中，作为一种实施例，处理器201可以包括一个或多个CPU，例如图2中所示的CPU0和CPU1。

[0078] 在具体实现中，作为一种实施例，计算机设备可以包括多个处理器，例如图2中所示的处理器201和处理器208。这些处理器中的每一个可以是一个单核(single-CPU)处理器，也可以是一个多核(multi-CPU)处理器。这里的处理器可以指一个或多个设备、电路、和/或用于处理数据(例如计算机程序指令)的处理核。

[0079] 在具体实现中，作为一种实施例，计算机设备还可以包括输出设备205和输入设备206。输出设备205和处理器201通信，可以以多种方式来显示信息。例如，输出设备205可以是液晶显示器(liquid crystal display，LCD)，发光二级管(light emitting diode，LED)显示设备，阴极射线管(cathode ray tube，CRT)显示设备，或投影仪(projector)等。输入设备206和处理器201通信，可以以多种方式接收用户的输入。例如，输入设备206可以是鼠标、键盘、触摸屏设备或传感设备等。

[0080] 上述的计算机设备可以是一个通用计算机设备或者是一个专用计算机设备。在具体实现中，计算机设备可以是台式机、便携式电脑、网络服务器、掌上电脑(Personal Digital Assistant，PDA)、移动手机、平板电脑、无线终端设备、通信设备或者嵌入式设备。本发明实施例不限定计算机设备的类型。

[0081] 其中，存储器203用于存储执行本发明方案的程序代码，并由处理器201来控制执行。处理器201用于执行存储器203中存储的程序代码210。程序代码210中可以包括一个或多个软件模块(例如：业务信息库构建模块、预处理模块、模式匹配分类模块、语义分类模块、加权融合模块、类别确定模块等)。图1B或者图1C中所示的服务器110可以通过处理器201以及存储器203中的程序代码210中的一个或多个软件模块，对待分类的文本进行分类。

[0082] 图3是本发明实施例提供的一种文本分类方法的流程图，该方法用于服务器中。参见图3，该方法包括：

[0083] 步骤301：对于业务信息库的关键词库中包括的多个关键词中的每个关键词，根据词向量模型，确定该关键词对应的词向量。

[0084] 需要说明的是，业务信息库用于辅助模式匹配分类器对待分类的文本进行分类，且该业务信息库中可以包括关键词库，该关键词库中可以包括多个关键词。

[0085] 另外，词向量模型用于将词语转换为词向量，词向量模型是由词向量工具经过训练得到的，该词向量工具可以为Word2vec(Word to vector)工具等，本发明实施例对此不做具体限定。

[0086] 其中，根据词向量模型，确定该关键词对应的词向量的操作可以参考相关技术，本发明实施例对此不进行详细阐述。

[0087] 进一步地，根据词向量模型，确定该关键词对应的词向量之前，还可以获取文本数据集，并对该文本数据集中包括的多个文本进行预处理，得到第一训练语料库，之后，使用该第一训练语料库作为词向量工具的输入，对该词向量工具进行训练，得到词向量模型。

[0088] 需要说明的是，文本数据集可以存储在服务器中，也可以存储在其它存储设备中，且服务器与该其它存储设备可以通过有线网络或者无线网络进行通信，本发明实施例对此不做具体限定。该文本数据集中包括多个文本，该多个文本为与多个预设类别相关的文本，如该多个文本可以为该多个预设类别的说明文本、客服文本等，本发明实施例对此不做具体限定。其中，说明文本为记录有预设类别的说明数据的文本，客服文本为记录有与预设类别相关的用户和客服的对话数据的文本。

[0089] 另外，该多个预设类别可以预先设置，如该多个预设类别可以为积分兑换-积分查询、国际漫游-资费咨询、信用开机-停机原因等，本发明实施例对此不做具体限定。

[0090] 再者，预处理可以包括中文分词、词性过滤和停用词过滤中的至少一种，本发明实施例对此不做具体限定。中文分词是指将文本转换为中文单词的集合，词性过滤是指去掉文本中的语气词、助词、连词等，停用词过滤是指去掉文本中没有实际含义的词，如去掉文本中的“的”、“那么”等。

[0091] 需要说明的是，对该文本数据集中包括的多个文本进行预处理的操作可以参考相关技术，本发明实施例对此不进行详细阐述。

[0092] 另外，使用该第一训练语料库作为词向量工具的输入，对该词向量工具进行训练，得到词向量模型的操作可以参考相关技术，本发明实施例对此不进行详细阐述。

[0093] 再者，由词向量工具经过训练后获得的词向量模型可以将文本数据集包括的每个词语转换为一个K维的词向量，其中，K为大于或等于1的自然数，且该K的值可以由技术人员预先设置，本发明实施例对此不做具体限定。

[0094] 需要说明的是，业务信息库的关键词库中的每个关键词可以由技术人员根据经验和对该文本数据集的观察后进行定义，本发明实施例对此不做具体限定。

[0095] 步骤302：基于该关键词对应的词向量，确定该关键词的潜在扩展词。

[0096] 具体地，根据词向量模型，确定文本数据集包括的各个词语对应的词向量；计算该关键词对应的词向量与该文本数据集包括的各个词语对应的词向量之间的相似度；将该文本数据集中的相似词语确定为该关键词的潜在扩展词，该相似词语对应的词向量与该关键词对应的词向量之间的相似度大于指定相似度。

[0097] 需要说明的是，指定相似度可以根据具体的业务需要预先设置，如该指定相似度可以为0.75、0.8等，本发明实施例对此不做具体限定。

[0098] 其中，计算该关键词对应的词向量与该文本数据集包括的各个词语对应的词向量之间的相似度时，可以计算两者之间的欧式距离，将计算得到的欧式距离确定为两者之间的相似度；或者，可以计算两者之间的余弦相似度，将计算得到的余弦相似度确定为两者之间的相似度，当然，实际应用中，也可以通过其它方法计算该关键词对应的词向量与该文本数据集包括的各个词语对应的词向量之间的相似度，本发明实施例对此不做具体限定。

[0099] 步骤303：当接收到针对该潜在扩展词输入的扩展规则，且当检测到针对该潜在扩

展词的添加指令时，将该潜在扩展词添加到该关键词库中，并将该扩展规则添加到业务信息库的匹配规则库中。

[0100] 需要说明的是，添加指令用于指示将该潜在扩展词添加到关键词库中，并将针对该潜在扩展词输入的扩展规则添加到匹配规则库中，该添加指令可以由技术人员触发，该技术人员可以通过指定操作触发，该指定操作可以为单击操作、双击操作、语音操作等，本发明实施例对此不做具体限定。

[0101] 另外，匹配规则库中可以包括多个匹配规则，该多个匹配规则中的每个匹配规则可以包含关键词库中的至少一个关键词，例如，匹配规则可以采用逻辑“与”符号(&)将多个关键词关联起来，表示当一个文本中同时包含该匹配规则中所有的关键词时，确定该文本满足该匹配规则。

[0102] 再者，该多个预设类别中的每个预设类别可以对应该匹配规则库中的至少一个匹配规则，本发明实施例对此不做具体限定。

[0103] 本发明实施例中，技术人员可以事先在业务信息库的关键词库中存储多个关键词，并事先在匹配规则库中存储多个匹配规则，该多个匹配规则中的每个匹配规则包含该多个关键词中的至少一个关键词，服务器可以根据词向量模型，确定该多个关键词中每个关键词的潜在扩展词，之后，技术人员可以对该潜在扩展词进行辨认分析，当技术人员确认该潜在扩展词可以用于构建某个预设类别的匹配规则时，该技术人员可以针对该潜在扩展词输入扩展规则，并将该潜在扩展词添加到关键词库中，将该扩展规则添加到匹配规则库中，从而完成对该关键词库和该匹配规则库的扩展更新。

[0104] 易于发现的是，本发明实施例中的关键词库和匹配规则库是采用半人工的方式构建的，从而不仅节省了构建时间，提高了构建效率，且降低了建立和维护该关键词库和该匹配规则库的人工成本。另外，由于可以根据服务器确定的潜在扩展词来对该关键词库和该匹配规则库进行扩展更新，因此，可以提高该关键词库和该匹配规则库的覆盖率，且由于该潜在扩展词和该扩展规则是基于添加指令添加到该关键词库和该匹配规则库中的，也即是该潜在扩展词和该扩展规则是在技术人员确认后添加到该关键词库和该匹配规则库中的，因此，可以提高该关键词库和该匹配规则库的准确率，进而可以提高后续基于该关键词库和该匹配规则库进行文本分类时的覆盖率和准确率。

[0105] 步骤304：基于该关键词库和该匹配规则库，通过模式匹配分类器确定待分类的文本属于多个预设类别中每个预设类别的第一概率。

[0106] 需要说明的是，待分类的文本可以是上述文本数据集中一个，也可以是服务器在与客户端进行通信的过程中实时获取的，本发明实施例对此不做具体限定。

[0107] 另外，基于该关键词库和该匹配规则库，通过模式匹配分类器确定待分类的文本属于多个预设类别中每个预设类别的第一概率之前，还可以对该待分类的文本进行预处理，以便后续可以基于该预处理后的待分类的文本，通过模式匹配分类器快速确定该待分类的文本属于多个预设类别中每个预设类别的第一概率，提高确定效率。

[0108] 需要说明的是，对待分类的文本进行预处理的操作与步骤301中的相关操作类似，本发明实施例对此不再赘述。

[0109] 具体地，基于该关键词库和该匹配规则库，通过模式匹配分类器确定待分类的文本属于多个预设类别中每个预设类别的第一概率时，服务器可以从待分类的文本中，通过

模式匹配分类器获取与该关键词库中的关键词相同的词语;将获取的词语确定为该待分类的文本的关键词;基于该待分类的文本的关键词和该匹配规则库,通过该模式匹配分类器对该待分类的文本进行分类,得到该待分类的文本属于该多个预设类别中每个预设类别的第一概率。

[0110] 其中,从待分类的文本中,通过模式匹配分类器获取与该关键词库中的关键词相同的词语时,该模式匹配分类器可以通过指定匹配算法,从待分类的文本中,获取与该关键词库中的关键词相同的词语,当然,实际应用中,该模式匹配分类器也可以通过其它方式,从待分类的文本中,获取与该关键词库中的关键词相同的词语,本发明实施例对此不做具体限定。

[0111] 需要说明的是,指定匹配算法可以预先设置,如该指定匹配算法可以为多模式匹配算法,该多模式匹配算法可以为Wu-Manber算法等,本发明实施例对此不做具体限定。另外,由于Wu-Manber算法可以通过跳转表来快速进行字符串匹配,因此,该模式匹配分类器通过Wu-Manber算法,从待分类的文本中,获取与该关键词库中的关键词相同的词语时,可以提高获取效率。

[0112] 另外,从待分类的文本中,通过模式匹配分类器获取与该关键词库中的关键词相同的词语的操作可以参考相关技术,本发明实施例对此不进行详细阐述。

[0113] 其中,基于该待分类的文本的关键词和该匹配规则库,通过该模式匹配分类器对该待分类的文本进行分类,得到该待分类的文本属于该多个预设类别中每个预设类别的第一概率时,该模式匹配分类器可以对于该匹配规则库包括的多个匹配规则中的每个匹配规则,基于该待分类的文本的关键词,判断该待分类的文本是否满足该匹配规则;当该待分类的文本满足该匹配规则时,确定该待分类的文本属于该匹配规则对应的预设类别的第一概率为1;当该待分类的文本不满足该匹配规则时,基于该待分类的文本的关键词,确定该待分类的文本的关键词向量,并基于该匹配规则中包含的至少一个关键词,确定该匹配规则的关键词向量,之后,计算该待分类的文本的关键词向量与该匹配规则的关键词向量之间的相似度,将计算得到的相似度确定为该待分类的文本属于该匹配规则对应的预设类别的第一概率。

[0114] 其中,计算该待分类的文本的关键词向量与该匹配规则的关键词向量之间的相似度的操作与步骤302中的相关操作类似,本发明实施例对此不再赘述。

[0115] 其中,基于该待分类的文本的关键词,判断该待分类的文本是否满足该匹配规则时,可以基于该待分类的文本的关键词,判断该待分类的文本中是否包含有该匹配规则中所有的关键词;当该待分类的文本中包含有该匹配规则中所有的关键词时,确定该待分类的文本满足该匹配规则;当该待分类的文本中未包含该匹配规则中所有的关键词时,确定该待分类的文本不满足该匹配规则。

[0116] 例如,该待分类的文本的关键词为国漫、取消,该匹配规则为国漫&取消,也即是该匹配规则中包含的至少一个关键词为国漫、取消,则可以确定该待分类的文本中包含有该匹配规则中所有的关键词,确定该待分类的文本满足该匹配规则。

[0117] 再例如,该待分类的文本的关键词为国漫、取消,该匹配规则为漫游&取消&国际,也即是该匹配规则中包含的至少一个关键词为漫游、取消、国际,则可以确定该待分类的文本中未包含该匹配规则中所有的关键词,确定该待分类的文本不满足该匹配规则。

[0118] 其中，当该待分类的文本不满足该匹配规则时，基于该待分类的文本的关键词，确定该待分类的文本的关键词向量时，可以确定该关键词库包括的多个关键词的个数，之后，将该待分类的文本所有的关键词转换为一个维数等于该个数的向量，并将该向量确定为该待分类的文本的关键词向量。

[0119] 例如，该关键词库包括的多个关键词的个数为8，该待分类的文本所有的关键词为国漫、取消，则可以将该待分类的文本所有的关键词转换为一个维数等于8的向量，如可以将该待分类的文本所有的关键词转换为向量(0,1,1,0,0,0,0,0)。

[0120] 其中，基于该匹配规则中包含的至少一个关键词，确定该匹配规则的关键词向量时，可以确定该关键词库包括的多个关键词的个数，之后，将该匹配规则中包含的至少一个关键词转换为一个维数等于该个数的向量，并将该向量确定为该匹配规则的关键词向量。

[0121] 例如，该关键词库包括的多个关键词的个数为8，该匹配规则为漫游&取消&国际，也即是该匹配规则中包含的至少一个关键词为漫游、取消、国际，则可以将该匹配规则中包含的至少一个关键词转换为一个维数等于8的向量，如可以将该匹配规则中包含的至少一个关键词转换为向量(1,0,1,1,0,0,0,0)。

[0122] 需要说明的是，基于该关键词库和该匹配规则库，通过模式匹配分类器确定待分类的文本属于多个预设类别中每个预设类别的第一概率的操作还可以参考相关技术，本发明实施例对此不再进行详细阐述。

[0123] 步骤305：基于该待分类的文本属于多个预设类别中每个预设类别的第一概率，从该多个预设类别中，确定该待分类的文本所属的类别。

[0124] 一种可能的设计中，通过步骤304可以确定待分类的文本属于多个预设类别中每个预设类别的第一概率，当该待分类的文本属于该多个预设类别中某个预设类别的第一概率大于指定概率时，可以将该预设类别确定为该待分类的文本所属的类别。

[0125] 需要说明的是，指定概率可以根据具体的业务需要预先设置，如该指定概率可以为0.8、0.85等，本发明实施例对此不做具体限定。

[0126] 例如，指定概率为0.8，该待分类的文本属于多个预设类别中某个预设类别的第一概率为0.85，则可以确定该待分类的文本属于该预设类别第一概率大于指定概率，确定该预设类别为该待分类的文本所属的类别。

[0127] 另一种可能的设计中，可以基于语义分类器中包括的多个预设语义特征向量，确定该待分类的文本属于该多个预设类别中每个预设类别的第二概率，该多个预设语义特征向量与该多个预设类别一一对应；基于该待分类的文本属于该多个预设类别中每个预设类别的第一概率和该待分类的文本属于该多个预设类别中每个预设类别的第二概率，从该多个预设类别中，确定该待分类的文本所属的类别。

[0128] 其中，基于语义分类器中包括的多个预设语义特征向量，确定该待分类的文本属于该多个预设类别中每个预设类别的第二概率时，可以通过语义分类器确定该待分类的文本中每个词语的语义特征向量；基于该待分类的文本中每个词语的语义特征向量，通过语义分类器确定该待分类的文本的语义特征向量；对于该多个预设语义特征向量中的每个预设语义特征向量，通过语义分类器计算该预设语义特征向量与该待分类的文本的语义特征向量之间的相似度；将计算得到的相似度确定为该待分类的文本属于该预设语义特征向量对应的预设类别的第二概率。

[0129] 需要说明的是，语义分类器用于基于待分类的文本的语义信息对该待分类的文本进行分类，如该语义分类器可以为递归卷积神经网络等，本发明实施例对此不做具体限定。

[0130] 另外，词语的语义特征向量为可以体现该词语的语义信息以及该词语在上下文中的依赖关系的特征向量，待分类的文本的语义特征向量为可以体现该待分类的文本的语义信息的特征向量，本发明实施例对此不做具体限定。

[0131] 再者，多个预设语义特征向量可以预先设置，本发明实施例对此不做具体限定。

[0132] 其中，通过语义分类器确定该待分类的文本中每个词语的语义特征向量时，可以通过语义分类器中的词特征提取层确定该待分类的文本中每个词语的语义特征向量，本发明实施例对此不做具体限定。另外，通过语义分类器确定该待分类的文本中每个词语的语义特征向量的操作可以参考相关技术，本发明实施例对此不进行详细阐述。

[0133] 其中，基于该待分类的文本中每个词语的语义特征向量，通过语义分类器确定该待分类的文本的语义特征向量时，可以基于该待分类的文本中每个词语的语义特征向量，通过语义分类器中的文本特征提取层确定该待分类的文本的语义特征向量，本发明实施例对此不做具体限定。另外，基于该待分类的文本中每个词语的语义特征向量，通过语义分类器确定该待分类的文本的语义特征向量的操作可以参考相关技术，本发明实施例对此不进行详细阐述。

[0134] 需要说明的是，词特征提取层和文本特征提取层可以采用递归结构，以保证通过该词特征提取层确定的词语的语义特征向量不仅可以体现出该词语的语义信息，还可以体现出该词语在上下文中的依赖关系，进而保证基于该词语的语义特征向量，通过该文本特征提取层确定的待分类的文本的语义特征向量可以完整体现出该待分类的文本的语义信息。

[0135] 其中，对于该多个预设语义特征向量中的每个预设语义特征向量，通过语义分类器计算该预设语义特征向量与该待分类的文本的语义特征向量之间的相似度时，可以通过语义分类器中的分类层计算该预设语义特征向量与该待分类的文本的语义特征向量之间的相似度，本发明实施例对此不做具体限定。另外，通过语义分类器计算该预设语义特征向量与该待分类的文本的语义特征向量之间的相似度的操作与上述步骤302中的相关操作类似，本发明实施例对此不再赘述。

[0136] 需要说明的是，相关技术中进行文本分类时，往往是先提取待分类的文本的词频-逆向文件频率(Term Frequency-Inverse Document Frequency，TF-IDF)特征，之后，通过支持向量机(英文：Support Vector Machine，简称：SVM)分类器确定该待分类的文本的TF-IDF特征所属的类别，并将该待分类的文本的TF-IDF特征所属的类别确定为该待分类的文本所属的类别。而由于TF-IDF特征只是对重要词语的词频的统计，缺乏对文本语义理解的高层信息，因此，属于不同类别的文本很有可能具有非常相似的TF-IDF特征，从而导致相关技术中的文本分类的准确率较低。而本发明实施例中，语义分类器是基于待分类的文本的语义信息对该待分类的文本进行分类的，从而可以有效避免由于信息缺乏而导致的对待分类的文本的误分类，提高了文本分类的准确率。

[0137] 进一步地，在基于语义分类器中包括的多个预设语义特征向量，确定该待分类的文本属于多个预设类别中每个预设类别的第二概率之前，还可以获取预设文本集，并对该预设文本集中包括的多个预设文本进行预处理，得到第二训练语料库，之后，使用该第二训

练语料库对待训练的语义分类器进行训练，得到该语义分类器，预设文本集中包括多个预设文本，该多个预设类别中的每个预设类别对应至少一个预设文本。

[0138] 需要说明的是，预设文本集可以预先设置，且该预设文本集可以存储在服务器中，也可以存储在其它存储设备中，且服务器与该其它存储设备可以通过有线网络或者无线网络进行通信，本发明实施例对此不做具体限定。

[0139] 另外，该多个预设文本也可以预先设置，且该多个预设文本可以为该多个预设类别的客服文本等，本发明实施例对此不做具体限定。该多个预设类别中的每个预设类别对应该预设文本集中的至少一个预设文本，也即是，该多个预设文本均为具有类别标识的文本。

[0140] 再者，本发明实施例在通过预设文本集对待训练的语义分类器进行训练时，可以采用监督学习的方式对待训练的语义分类器进行训练，该监督学习是指在给定该语义分类器输入与输出的情况下，通过指定调整算法不断调整该语义分类器中的参数，使该语义分类器达到所要求性能的过程。

[0141] 需要说明的是，指定调整算法可以预先设置，如该指定调整算法可以为随机梯度下降算法等，本发明实施例对此不做具体限定。

[0142] 另外，使用该第二训练语料库对待训练的语义分类器进行训练，得到该语义分类器的操作可以参考相关技术，本发明实施例对此不进行详细阐述。

[0143] 再者，对该预设文本集中包括的多个预设文本进行预处理的操作与上述步骤301中的相关操作类似，本发明实施例对此不再赘述。

[0144] 其中，基于该待分类的文本属于该多个预设类别中每个预设类别的第一概率和该待分类的文本属于该多个预设类别中每个预设类别的第二概率，从该多个预设类别中，确定该待分类的文本所属的类别时，可以从该多个预设类别中，确定至少一个目标预设类别；对于该至少一个目标预设类别中的每个目标预设类别，确定该目标预设类别对应的第一概率和第二概率；基于模式匹配分类器对应的第一权重和语义分类器对应的第二权重，对目标预设类别对应的第一概率和第二概率进行加权平均，得到加权概率；当该加权概率大于指定概率时，确定目标预设类别为该待分类的文本所属的类别。

[0145] 其中，从该多个预设类别中，确定至少一个目标预设类别的操作可以包括如下三种方式：

[0146] 第一种方式：将该多个预设类别中的每个预设类别确定为目标预设类别。

[0147] 第二种方式：按照该多个预设类别对应的多个第一概率由大到小的顺序，从该多个第一概率中，获取N个第一概率，并将该N个第一概率中每个第一概率对应的预设类别确定为目标预设类别，该N为大于或等于1的自然数。

[0148] 第三种方式：按照该多个预设类别对应的多个第二概率由大到小的顺序，从该多个第二概率中，获取N个第二概率，并将该N个第二概率中每个第二概率对应的预设类别确定为目标预设类别。

[0149] 需要说明的是，上述第二种方式和上述第三种方式不仅可以单独执行来从多个预设类别中，确定至少一个目标预设类别，当然，也可以将上述第二种方式和上述第三种方式结合起来从多个预设类别中，确定至少一个目标预设类别，本发明实施例对此不做具体限定。

[0150] 另外，本发明实施例可以直接将该多个预设类别中的每个预设类别确定为目标预设类别，此时，服务器无需再执行其它操作，从而可以提高目标预设类别的确定效率。且本发明实施例也可以基于第一概率和第二概率，确定至少一个目标预设类别，此时该至少一个目标预设类别为该多个预设类别中的部分预设类别，因此，服务器在后续步骤中不需对该多个预设类别都计算加权概率，而只需对该多个预设类别中的部分预设类别计算加权概率，从而可以节省服务器的处理资源，且可以提高文本分类效率。

[0151] 其中，基于模式匹配分类器对应的第一权重和语义分类器对应的第二权重，对目标预设类别对应的第一概率和第二概率进行加权平均，得到加权概率时，可以将第一权重与该目标预设类别对应的第一概率相乘，得到第一数值；将第二权重与该目标预设类别对应的第二概率相乘，得到第二数值；将第一数值与第二数值相加，得到加权概率。

[0152] 需要说明的是，第一权重和第二权重可以预先设置，且实际应用中，第一权重和第二权重可以根据模式匹配分类器和语义分类器的可靠性来进行设置和调整，且还可以根据不同的业务需求来进行设置和调整，本发明实施例对此不做具体限定。

[0153] 需要说明的是，本发明实施例中，可以将模式匹配分类器和语义分类器结合起来进行文本分类，从而避免了单一依赖某一方法来进行文本分类带来的不准确性，且由于可以根据不同的业务需求来设置和调整第一权重和第二权重，因此，可以保证得到的加权概率符合技术人员的文本分类需求。

[0154] 另外，本发明实施例中提供的文本分类方法，可以用于电信运营商或互联网运营商的客服平台中产生的客服文本的分类，由于本发明实施例中可以高效率地构建关键词库和业务规则库，且可以将模式匹配分类器和语义分类器结合起来进行文本分类，因此，在该客服文本的匹配规则较为繁琐、数据量较大的情况下，也可以较好地对该客服文本进行分类，文本分类的准确率可以得到保证。当然，本发明实施例中提供的文本分类方法也可以应用于其它领域，进行其它类型的文本分类，本发明实施例对此不做具体限定。

[0155] 需要说明的是，本发明实施例中，当服务器为由多个节点服务器组成的服务器集群时，该服务器的底层架构可以采用Hadoop分布计算式平台，本发明实施例对此不做具体限定。另外，当该服务器的底层架构采用Hadoop分布计算式平台时，该Hadoop分布计算式平台中可以包括HDFS、Map-Reduce和Hive等组件，且HDFS组件中用于存储文本、关键词库、匹配规则库等，Map-Reduce组件和Hive组件用于支撑文本分类的核心操作，该核心操作可以包括确定第一概率、第二概率、加权概率等操作。

[0156] 在本发明实施例中，对于业务信息库的关键词库包括的多个关键词中的每个关键词，根据词向量模型，确定该关键词对应的词向量，并基于该关键词对应的词向量，确定该关键词的潜在扩展词，之后，当接收到针对该潜在扩展词输入的扩展规则，且当检测到针对该潜在扩展词的添加指令时，将该潜在扩展词添加到该关键词库中，并将该扩展规则添加到业务信息库的匹配规则库中，也即是，该关键词库和该匹配规则库是采用半人工的方式构建的，不仅可以节省构建时间，提高构建效率，且可以降低建立和维护该关键词库和该匹配规则库的人工成本。另外，由于可以根据服务器确定的潜在扩展词对该关键词库和该匹配规则库进行扩展更新，因此，可以提高该关键词库和该匹配规则库的覆盖率，且由于添加指令是由技术人员触发的，因此，该潜在扩展词和该扩展规则是在技术人员确认后添加到该关键词库和该匹配规则库中的，从而可以提高该关键词库和该匹配规则库的准确率。再

者，由于提高了该关键词库和该匹配规则库的覆盖率和准确率，因此，基于该关键词库和该匹配规则库，确定待分类的文本属于多个预设类别中每个预设类别的第一概率，并基于该待分类的文本属于该多个预设类别中每个预设类别的第一概率，从该多个预设类别中，确定该待分类的文本所属的类别时，可以提高文本分类的覆盖率和准确率。

[0157] 图4是本发明实施例提供的一种与上述方法实施例属于同一发明构思下的文本分类装置的结构示意图。如图4所示，该文本分类装置的结构用于执行上述图3所示的方法实施例中服务器的功能，包括：第一确定单元401，第二确定单元402，添加单元403，第三确定单元404和第四确定单元405。

[0158] 第一确定单元401，用于对于业务信息库的关键词库中包括的多个关键词中的每个关键词，根据词向量模型，确定关键词对应的词向量；

[0159] 第二确定单元402，用于基于关键词对应的词向量，确定关键词的潜在扩展词；

[0160] 添加单元403，用于当接收到针对潜在扩展词输入的扩展规则，且当检测到针对潜在扩展词的添加指令时，将潜在扩展词添加到关键词库中，并将扩展规则添加到业务信息库的匹配规则库中；

[0161] 第三确定单元404，用于基于关键词库和匹配规则库，通过模式匹配分类器确定待分类的文本属于多个预设类别中每个预设类别的第一概率；

[0162] 第四确定单元405，用于基于待分类的文本属于多个预设类别中每个预设类别的第一概率，从多个预设类别中，确定待分类的文本所属的类别。

[0163] 可选地，该第二确定单元402，用于：

[0164] 根据词向量模型，确定文本数据集包括的各个词语对应的词向量；

[0165] 计算关键词对应的词向量与文本数据集包括的各个词语对应的词向量之间的相似度；

[0166] 将文本数据集中的相似词语确定为关键词的潜在扩展词，相似词语对应的词向量与关键词对应的词向量之间的相似度大于指定相似度。

[0167] 可选地，该第三确定单元404，用于：

[0168] 从待分类的文本中，通过模式匹配分类器获取与关键词库中的关键词相同的词语；

[0169] 将获取的词语确定为待分类的文本的关键词；

[0170] 基于待分类的文本的关键词和匹配规则库，通过模式匹配分类器对待分类的文本进行分类，得到待分类的文本属于多个预设类别中每个预设类别的第一概率；其中，匹配规则库中包括多个匹配规则，多个匹配规则中的每个匹配规则包含关键词库中的至少一个关键词，每个预设类别对应匹配规则库中的至少一个匹配规则。

[0171] 可选地，该第四确定单元405，用于：

[0172] 基于语义分类器中包括的多个预设语义特征向量，确定待分类的文本属于多个预设类别中每个预设类别的第二概率，多个预设语义特征向量与多个预设类别一一对应；

[0173] 基于待分类的文本属于多个预设类别中每个预设类别的第一概率和待分类的文本属于多个预设类别中每个预设类别的第二概率，从多个预设类别中，确定待分类的文本所属的类别。

[0174] 可选地，该第四确定单元405，还用于：

[0175] 通过语义分类器确定待分类的文本中每个词语的语义特征向量；

[0176] 基于待分类的文本中每个词语的语义特征向量，通过语义分类器确定待分类的文本的语义特征向量；

[0177] 对于多个预设语义特征向量中的每个预设语义特征向量，通过语义分类器计算预设语义特征向量与待分类的文本的语义特征向量之间的相似度；

[0178] 将计算得到的相似度确定为待分类的文本属于预设语义特征向量对应的预设类别的第二概率。

[0179] 可选地，该第四确定单元405，还用于：

[0180] 从多个预设类别中，确定至少一个目标预设类别；

[0181] 对于至少一个目标预设类别中的每个目标预设类别，确定目标预设类别对应的第一概率和第二概率；

[0182] 基于模式匹配分类器对应的第一权重和语义分类器对应的第二权重，对目标预设类别对应的第一概率和第二概率进行加权平均，得到加权概率；

[0183] 当加权概率大于指定概率时，确定目标预设类别为待分类的文本所属的类别。

[0184] 可选地，该第四确定单元405，还用于：

[0185] 将多个预设类别中的每个预设类别确定为目标预设类别。

[0186] 可选地，该第四确定单元，还用于如下两种方式中的至少一种：

[0187] 按照多个预设类别对应的多个第一概率由大到小的顺序，从多个第一概率中，获取N个第一概率，并将N个第一概率中每个第一概率对应的预设类别确定为目标预设类别，N为大于或等于1的自然数；或者，

[0188] 按照多个预设类别对应的多个第二概率由大到小的顺序，从多个第二概率中，获取N个第二概率，并将N个第二概率中每个第二概率对应的预设类别确定为目标预设类别。

[0189] 在本发明实施例中，文本分类的装置是以功能单元的形式来呈现。这里的“单元”可以指ASIC，执行一个或多个软件或固件程序的处理器和存储器，集成逻辑电路，和/或其他可以提供上述功能的器件。在一个简单的实施例中，本领域的技术人员可以想到文本分类装置可以采用图2所示的形式。第一确定单元401，第二确定单元402，添加单元403，第三确定单元404和第四确定单元405可以通过图2的处理器和存储器来实现，具体地，第一确定单元401、第二确定单元402和添加单元403可以通过由处理器执行业务信息库构建模块来实现，第三确定单元404可以通过由处理器执行模式匹配分类模块来实现，第四确定单元405可以通过由处理器执行语义分类模块、加权融合模块和类别确定模块来实现。

[0190] 本发明实施例还提供了一种计算机存储介质，用于储存实现上述图4所示的文本分类装置的计算机软件指令，其包含用于执行上述方法实施例所设计的程序。通过执行存储的程序，可以实现对待分类的文本的分类。

[0191] 需要说明的是，对于前述的各方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本发明并不受所描述的动作顺序的限制，因为依据本发明，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定是本发明所必须的。

[0192] 尽管在此结合各实施例对本发明进行了描述，然而，在实施所要求保护的本发明

过程中，本领域技术人员通过查看所述附图、公开内容、以及所附权利要求书，可理解并实现所述公开实施例的其他变化。在权利要求中，“包括”(comprising)一词不排除其他组成部分或步骤，“一”或“一个”不排除多个的情况。单个处理器或其他单元可以实现权利要求中列举的若干项功能。相互不同的从属权利要求中记载了某些措施，但这并不表示这些措施不能组合起来产生良好的效果。

[0193] 本领域技术人员应明白，本发明的实施例可提供为方法、装置(设备)、或计算机程序产品。因此，本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。计算机程序存储/分布在合适的介质中，与其它硬件一起提供或作为硬件的一部分，也可以采用其他分布形式，如通过Internet或其它有线或无线电信系统。

[0194] 本发明是参照本发明实施例的方法、装置(设备)和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0195] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0196] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0197] 尽管结合具体特征及其实施例对本发明进行了描述，显而易见的，在不脱离本发明的精神和范围的情况下，可对其进行各种修改和组合。相应地，本说明书和附图仅仅是所附权利要求所界定的本发明的示例性说明，且视为已覆盖本发明范围内的任意和所有修改、变化、组合或等同物。显然，本领域的技术人员可以对本发明进行各种改动和变型而不脱离本发明的精神和范围。这样，倘若本发明的这些修改和变型属于本发明权利要求及其等同技术的范围之内，则本发明也意图包含这些改动和变型在内。

客服：您好，很高兴为您服务。
用户：啊，我这手机怎么看积分啊。
客服：那您想要查询，我这边看不到我给您转自动台，您听一下多少积分呢？
用户：啊，行行行。
客服：嗯，现在给您转接了。
用户：嗯。

图1A

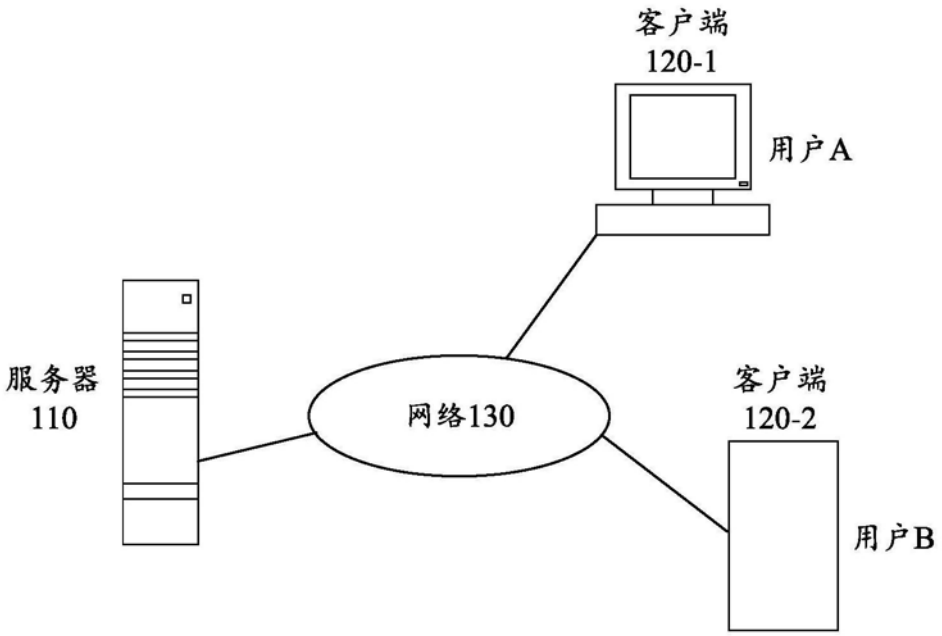


图1B

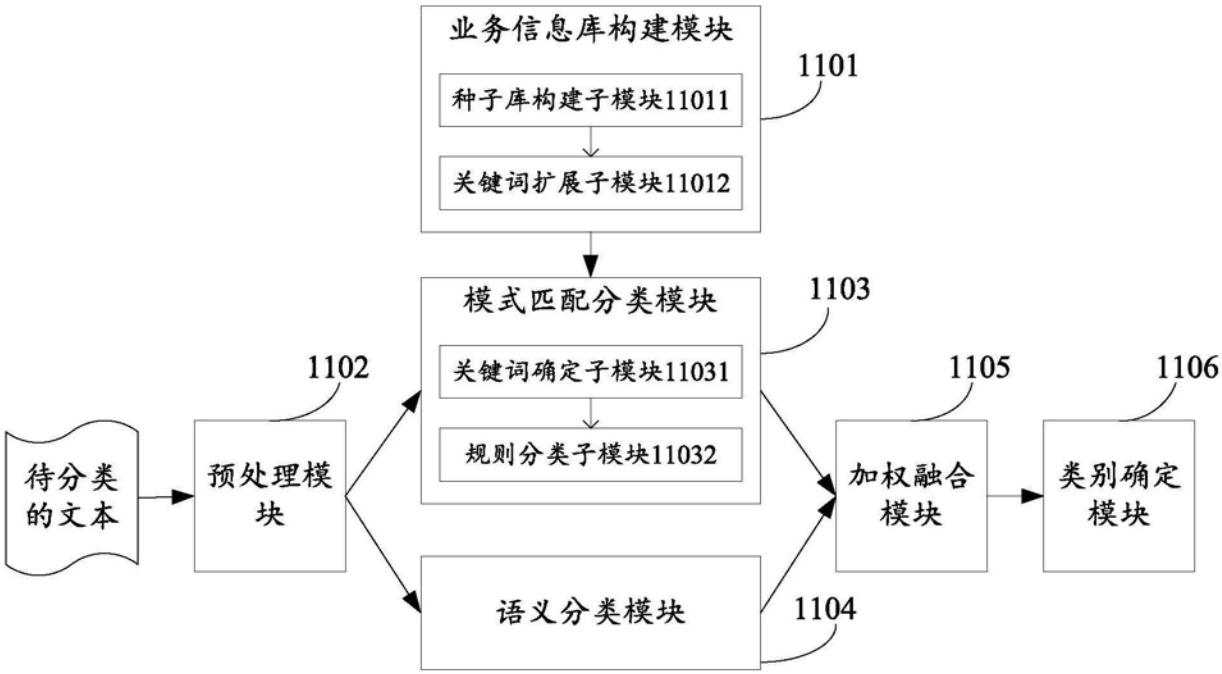

业务信息库构建模块
种子库构建子模块11011
关键词扩展子模块11012
1101
模式匹配分类模块
关键词确定子模块11031
规则分类子模块11032
1103
1102
1105
1106
待分类的文本
预处理模块
加权融合模块
类别确定模块
语义分类模块
1104

图1C

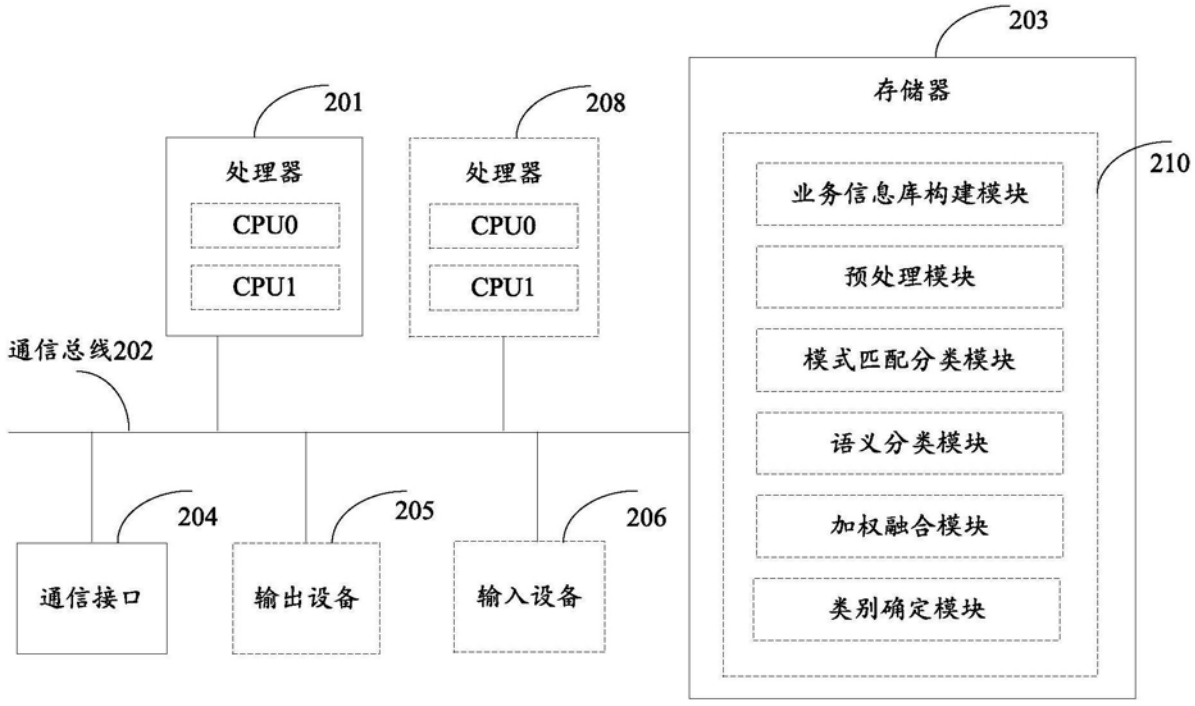

203
存储器
201
208
处理器
CPU0
CPU1
处理器
CPU0
CPU1
210
业务信息库构建模块
预处理模块
模式匹配分类模块
语义分类模块
加权融合模块
类别确定模块
通信总线202
204
205
206
通信接口
输出设备
输入设备

图2

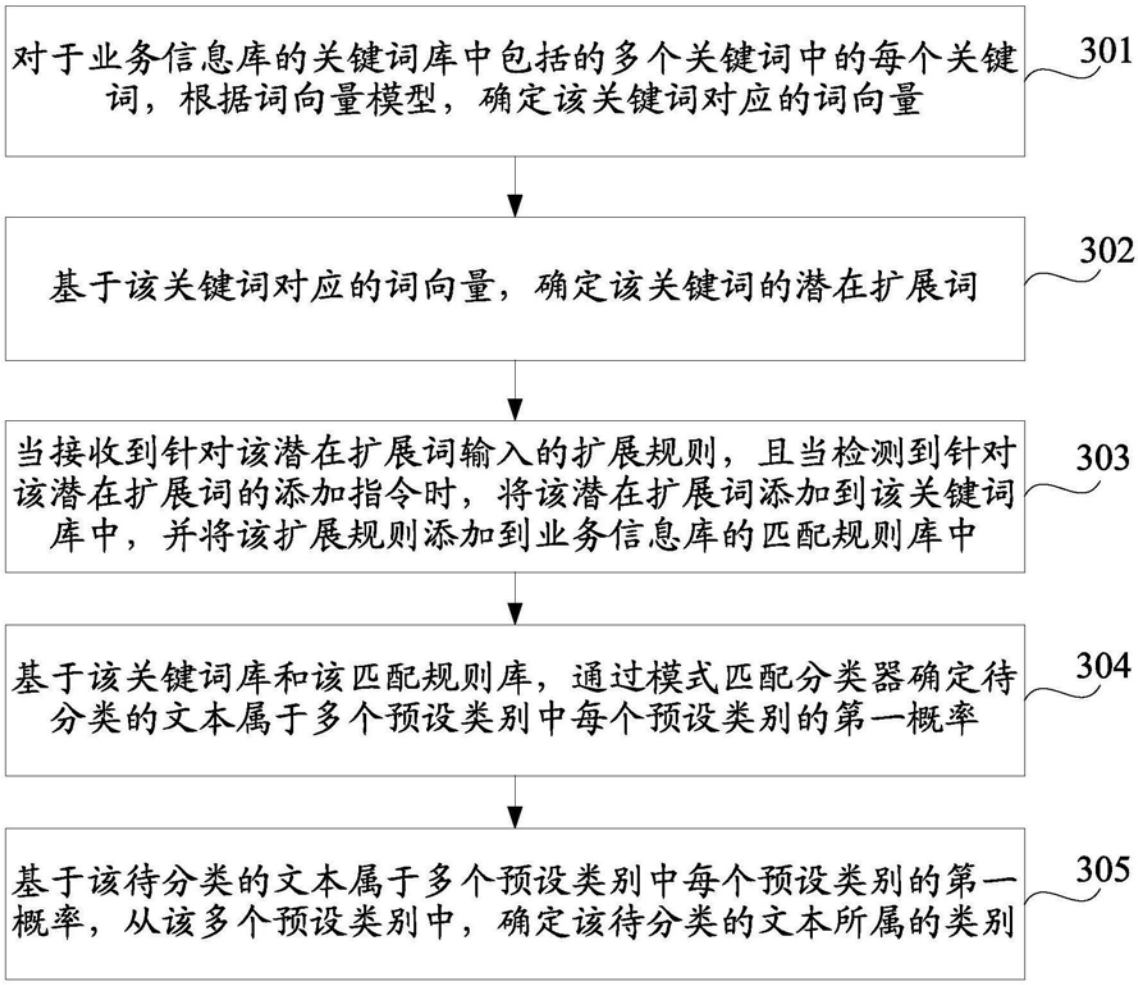

对于业务信息库的关键词库中包括的多个关键词中的每个关键词，根据词向量模型，确定该关键词对应的词向量
301
基于该关键词对应的词向量，确定该关键词的潜在扩展词
302
当接收到针对该潜在扩展词输入的扩展规则，且当检测到针对该潜在扩展词的添加指令时，将该潜在扩展词添加到该关键词库中，并将该扩展规则添加到业务信息库的匹配规则库中
303
基于该关键词库和该匹配规则库，通过模式匹配分类器确定待分类的文本属于多个预设类别中每个预设类别的第一概率
304
基于该待分类的文本属于多个预设类别中每个预设类别的第一概率，从该多个预设类别中，确定该待分类的文本所属的类别
305

图3

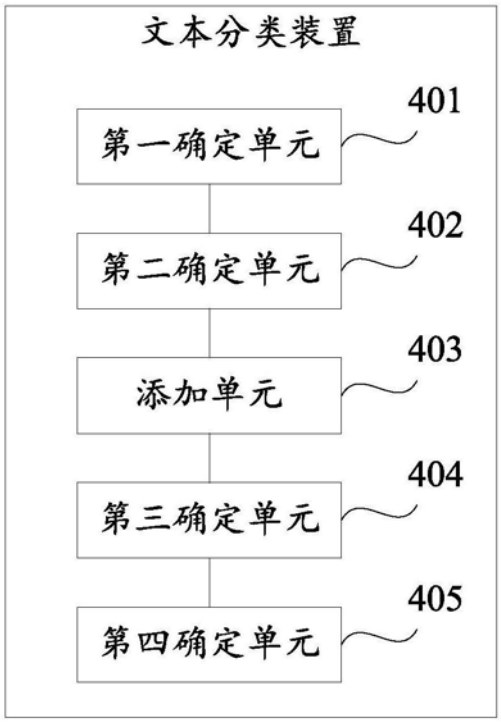


图4