(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 105005589 A
(43)申请公布日 2015.10.28

(21)申请号 201510364152.4

(22)申请日 2015.06.26

(71)申请人 腾讯科技(深圳)有限公司
地址 518000 广东省深圳市福田区振兴路赛格科技园2栋东403室

(72)发明人 邹缘孙

(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 刘映东

(51)Int.Cl.
G06F 17/30(2006.01)

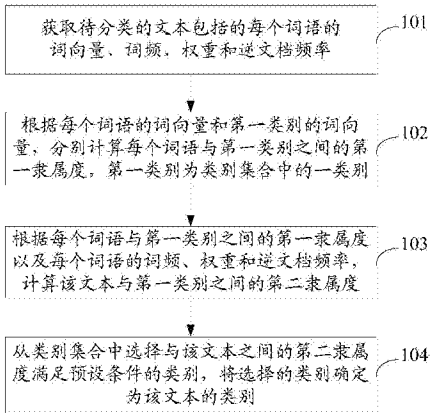
权利要求书2页 说明书10页 附图4页

(54)发明名称

一种文本分类的方法和装置

(57)摘要

本发明公开了一种文本分类的方法和装置，属于互联网技术领域。方法包括：获取待分类的文本包括的每个词语的词向量、词频、权重和逆文档频率；根据所述每个词语的词向量和第一类别的词向量，分别计算所述每个词语与所述第一类别之间的第一隶属度，所述第一类别为类别集合中的任一类别；根据所述每个词语与所述第一类别之间的第一隶属度以及所述每个词语的词频、权重和逆文档频率，计算所述文本与所述第一类别之间的第二隶属度；从所述类别集合中选择与所述文本之间的第二隶属度满足预设条件的类别，将所述选择的类别确定为所述文本的类别。装置包括：第一获取模块，第一计算模块，第二计算模块和分类模块。本发明提供了文本分类的准确性。



CN 105005589 A

1. 一种文本分类的方法，其特征在于，所述方法包括：

获取待分类的文本包括的每个词语的词向量、词频、权重和逆文档频率；

根据所述每个词语的词向量和第一类别的词向量，分别计算所述每个词语与所述第一类别之间的第一隶属度，所述第一类别为类别集合中的任一类别；

根据所述每个词语与所述第一类别之间的第一隶属度以及所述每个词语的词频、权重和逆文档频率，计算所述文本与所述第一类别之间的第二隶属度；

从所述类别集合中选择与所述文本之间的第二隶属度满足预设条件的类别，将所述选择的类别确定为所述文本的类别。

2. 如权利要求 1 所述的方法，其特征在于，所述根据所述每个词语的词向量和第一类别的词向量，分别计算所述每个词语与所述第一类别之间的第一隶属度，包括：

获取第一类别对应的词语集合中的各词语的词向量；

计算所述获取到词向量的平均词向量并将所述平均词向量作为所述第一类别的词向量；

分别计算所述每个词语的词向量与所述第一类别的词向量之间的距离，将所述每个词语的词向量与所述第一类别的词向量之间的距离分别作为所述每个词语与所述第一类别之间的第一隶属度。

3. 如权利要求 2 所述的方法，其特征在于，所述方法还包括：

获取多个文本样本；

将所述多个文本样本中的每个文本样本进行分词，将得到的词语组成训练集合；

对所述训练集合中的词语进行聚类，得到多个词语集合以及所述多个词语集合中的每个词语集合的类别。

4. 如权利要求 3 所述的方法，其特征在于，所述对所述训练集合中的词语进行聚类，得到多个词语集合以及所述多个词语集合中的每个词语集合的类别，包括：

获取所述训练集合中的各词语的词向量；

根据所述各词语的词语向量，计算所述各词语中的任意两词语之间的距离；

将距离小于预设距离的多个词语组成一个词语集合，以及获取用户标注的所述词语集合的类别。

5. 如权利要求 1 所述的方法，其特征在于，所述根据所述每个词语与所述第一类别之间的第一隶属度以及所述每个词语的词频、权重和逆文档频率，计算所述文本与所述第一类别之间的第二隶属度，包括：

分别计算所述每个词语的词频、权重、逆文档频率以及与所述第一类别之间的第一隶属度的乘积，得到所述每个词语与所述第一类别之间的第三隶属度；

将所述每个词语与所述第一类别之间的第三隶属度进行累加，得到所述文本与所述第一类别之间的第二隶属度。

6. 一种文本分类的装置，其特征在于，所述装置包括：

第一获取模块，用于获取待分类的文本包括的每个词语的词向量、词频、权重和逆文档频率；

第一计算模块，用于根据所述每个词语的词向量和第一类别的词向量，分别计算所述每个词语与所述第一类别之间的第一隶属度，所述第一类别为类别集合中的任一类别；

第二计算模块，用于根据所述每个词语与所述第一类别之间的第一隶属度以及所述每个词语的词频、权重和逆文档频率，计算所述文本与所述第一类别之间的第二隶属度；

分类模块，用于从所述类别集合中选择与所述文本之间的第二隶属度满足预设条件的类别，将所述选择的类别确定为所述文本的类别。

7. 如权利要求 6 所述的装置，其特征在于，所述第一计算模块，包括：

第一获取单元，用于获取第一类别对应的词语集合中的各词语的词向量；

第一计算单元，用于计算所述获取到词向量的平均词向量并将所述平均词向量作为所述第一类别的词向量；

第二计算单元，用于分别计算所述每个词语的词向量与所述第一类别的词向量之间的距离，将所述每个词语的词向量与所述第一类别的词向量之间的距离分别作为所述每个词语与所述第一类别之间的第一隶属度。

8. 如权利要求 7 所述的装置，其特征在于，所述装置还包括：

第二获取模块，用于获取多个文本样本；

分词模块，用于将所述多个文本样本中的每个文本样本进行分词，将得到的词语组成训练集合；

聚类模块，用于对所述训练集合中的词语进行聚类，得到多个词语集合以及所述多个词语集合中的每个词语集合的类别。

9. 如权利要求 8 所述的装置，其特征在于，所述聚类模块，包括：

第二获取单元，用于获取所述训练集合中的各词语的词向量；

第三计算单元，用于根据所述各词语的词语向量，计算所述各词语中的任意两词语之间的距离；

聚类单元，用于将距离小于预设距离的多个词语组成一个词语集合；

第三获取单元，用于获取用户标注的所述词语集合的类别。

10. 如权利要求 6 所述的装置，其特征在于，所述第二计算模块，包括：

第四计算单元，用于分别计算所述每个词语的词频、权重、逆文档频率以及与所述第一类别之间的第一隶属度的乘积，得到所述每个词语与所述第一类别之间的第三隶属度；

累加单元，用于将所述每个词语与所述第一类别之间的第三隶属度进行累加，得到所述文本与所述第一类别之间的第二隶属度。

一种文本分类的方法和装置

技术领域

[0001] 本发明涉及互联网技术领域，特别涉及一种文本分类的方法和装置。

背景技术

[0002] 随着互联网技术的发展，互联网上的文本越来越多，大量的文本给用户提供方便的同时也给用户的查找带来了很大的不便，面对这个问题，文本分类被提出来了，文本分类能够按照预先定义的主题类别，为文本确定一个类别，将文本按照类别进行分类，从而方便用户查找。

[0003] 现有技术提供了一种文本分类的方法，可以为：服务器获取大量人工标注的文本样本，获取这些文本样本的特征，根据这些文本样本的特征对分类器进行训练；对分类器训练完成之后，服务器可以采用该分类器对需要分类的文本进行分类，具体过程为：服务器取待分类的文本的特征，根据待分类的文本的特征，通过训练后的分类器对待分类的文本进行分类。

[0004] 在实现本发明的过程中，发明人发现现有技术至少存在以下问题：

[0005] 待分类的文本的特征往往是待分类的文本中的一个关键的词语，仅仅根据待分类的文本中的一个关键的词语对待分类的文本进行分类显然不准确，例如，一个关于描述开发游戏资金消耗问题的文本，服务器获取的这个文本的特征可能是“游戏”，根据该特征“游戏”确定该文本的类别为“游戏”，然而该文本的重点主要是资金消耗问题，将该文本的类别确定为“财经”更合适，因此，通过该文本的特征对该文本进行分类的准确性低。

发明内容

[0006] 为了解决现有技术的问题，本发明提供了一种文本分类的方法和装置。技术方案如下：

[0007] 一种文本分类的方法，所述方法包括：

[0008] 获取待分类的文本包括的每个词语的词向量、词频、权重和逆文档频率；

[0009] 根据所述每个词语的词向量和第一类别的词向量，分别计算所述每个词语与所述第一类别之间的第一隶属度，所述第一类别为类别集合中的任一类别；

[0010] 根据所述每个词语与所述第一类别之间的第一隶属度以及所述每个词语的词频、权重和逆文档频率，计算所述文本与所述第一类别之间的第二隶属度；

[0011] 从所述类别集合中选择与所述文本之间的第二隶属度满足预设条件的类别，将所述选择的类别确定为所述文本的类别。

[0012] 一种文本分类的装置，所述装置包括：

[0013] 第一获取模块，用于获取待分类的文本包括的每个词语的词向量、词频、权重和逆文档频率；

[0014] 第一计算模块，用于根据所述每个词语的词向量和第一类别的词向量，分别计算所述每个词语与所述第一类别之间的第一隶属度，所述第一类别为类别集合中的任一类

别；

[0015] 第二计算模块，用于根据所述每个词语与所述第一类别之间的第一隶属度以及所述每个词语的词频、权重和逆文档频率，计算所述文本与所述第一类别之间的第二隶属度；

[0016] 分类模块，用于从所述类别集合中选择与所述文本之间的第二隶属度满足预设条件的类别，将所述选择的类别确定为所述文本的类别。

[0017] 在本发明实施例中，根据待分类的文本包括的每个词语的词向量、词频、权重和逆文档频率和第一类别的词向量，计算该文本与第一类别之间的第二隶属度，第一类别为类别集合中的任一类别，根据与该文本之间的第二隶属度，从类别集合中选择类别；由于本发明在对待分类的文本进行分类时，考虑了该文本包括的每个词语，因此提高了分类的准确性。

附图说明

[0018] 图 1 是本发明实施例 1 提供的一种文本分类的方法流程图；

[0019] 图 2-1 是本发明实施例 2 提供的一种文本分类的方法流程图；

[0020] 图 2-2 是本发明实施例 2 提供的一种生成每个类别的词语集合的示意图；

[0021] 图 3 是本发明实施例 3 提供的一种文本分类的装置结构示意图；

[0022] 图 4 是本发明实施例 4 提供的一种服务器的结构示意图。

具体实施方式

[0023] 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。

[0024] 实施例 1

[0025] 本发明实施例提供了一种文本分类的方法，参见图 1，其中，该方法包括：

[0026] 步骤 101：获取待分类的文本包括的每个词语的词向量、词频、权重和逆文档频率；

[0027] 步骤 102：根据每个词语的词向量和第一类别的词向量，分别计算每个词语与第一类别之间的第一隶属度，第一类别为类别集合中的任一类别；

[0028] 步骤 103：根据每个词语与第一类别之间的第一隶属度以及每个词语的词频、权重和逆文档频率，计算该文本与第一类别之间的第二隶属度；

[0029] 步骤 104：从类别集合中选择与该文本之间的第二隶属度满足预设条件的类别，将选择的类别确定为该文本的类别。

[0030] 在本发明实施例中，根据待分类的文本包括的每个词语的词向量、词频、权重和逆文档频率和第一类别的词向量，计算该文本与第一类别之间的第二隶属度，第一类别为类别集合中的任一类别，根据与该文本之间的第二隶属度，从类别集合中选择类别；由于本发明在对待分类的文本进行分类时，考虑了该文本包括的每个词语，因此提高了分类的准确性。

[0031] 实施例 2

[0032] 本发明实施例提供了一种文本分类的方法，当服务器对需要分类的文本进行分类

时，为了提高分类的准确性，服务器可以采用本发明实施例提供的文本分类的方法对待分类的文本进行分类，从而提高分类的准确性。该方法的执行主体为服务器；参见图 2-1，其中，该方法包括：

[0033] 步骤 201：获取多个文本样本；

[0034] 文本样本用于训练类别集合中的每个类别对应的词语集合；并且，多个文本样本中的每个文本样本对应的一个类别，在本发明实施例中多个文本样本可以为任一类别的文本样本，为了提高分类的准确性，多个文本样本可以包括类别集合中的每个类别对应的文本样本。例如，类别集合中包括：财经、娱乐、体育、时尚、汽车、房产、科技、教育等。在选择文本样本时，多个文本样本可以包括类别为财经的文本样本，类别为娱乐的文本样本，类别为体育的文本样本，类别为时尚的文本样本，类别为汽车的文本样本，类别为房产的文本样本，类别为房产的文本样本，类别为教育的文本样本。

[0035] 在本发明实施例中，用户可以选择多个文本样本，然后输入多个文本样本给服务器；服务器接收用户输入的多个文本样本。

[0036] 步骤 202：将多个文本样本中的每个文本样本进行分词，将得到的词语组成训练集合；

[0037] 利用现有的分词工具，将多个文本样本中的每个文本样本进行分词，得到每个文本样本包括的词语；将每个文本包括的词语组成训练集合。

[0038] 其中，利用分词工具对文本样本进行分词的过程为现有技术，在此不再详细说明。

[0039] 得到训练集合之后，执行步骤 203，采用现有的聚类方法对训练集合中的词语进行聚类。

[0040] 步骤 203：对训练集合中的词语进行聚类，得到多个词语集合以及多个词语集合中的每个词语集合的类别；

[0041] 其中，本步骤可以通过以下步骤 (1) 至 (3) 实现，包括：

[0042] (1)：获取训练集合中的各词语的词向量；

[0043] 其中，词语的词向量用于描述词语特性的向量表述，在本发明实施例中词语的词向量特指基于词嵌入技术构造的词语向量的表述。

[0044] 在本发明实施例中可以采用任一获取词向量的方法获取训练集合中的各词语的词向量，例如使用神经网络语言模型中词嵌入技术 word2vec 方法，获取该词语的词向量。并且使用神经网络语言模型中词嵌入技术 word2vec 方法，获取该词语的词向量具体过程为现有技术，在此不再详细说明。

[0045] 其中，训练集合中的各词语的词向量都为 n 维向量，可以表示为 $Wi = (w_1, w_2, \cdots\cdots, w_n)$。Wi 为第 i 个词语的词向量，$W_n$ 为第 n 维向量的向量值。

[0046] 由于“的”、“了”和“吗”之类的语气词对文本进行分类时不起关键作用，因此，为了减少运算量以及提高分类的准确性，在本步骤中可以将“的”、“了”“吗”之类的语气词去除，只获取训练集合中剩余词语的词向量，则本步骤可以为：

[0047] 从训练集合中获取预设类型的词语，从训练集合中去除该获取的词语，得到训练集合中剩余词语，获取剩余词语的词向量。

[0048] 其中，预设类型的词语可以为语气词或者助词等。并且获取剩余词语的词向量的过程和获取训练集合中的各词语的词向量的过程相同，在此不再赘述。

[0049] 进一步地，获取到训练集合中的各词语的词向量之后，将各词语和各词语的词向量存储在词语和词语的词向量的对应关系中，以便于对待分类的文本进行分类时，获取该文本中包括的词语的词向量时，直接从词语和词语的词向量的对应关系中获取词语的词向量，节省了获取词语的词向量的时间，提高了对文本分类的效率。

[0050] (2)：根据各词语的词语向量，计算各词语中的任意两词语之间的距离；

[0051] 对于各词语中的任意两词语，分别根据这两个词语的词向量，按照以下公式(1)计算这两个词语之间的距离。

[0052] $$dist(W_i, W_j) = \frac{\sum_{k=1}^{n} W_{i,k} \bullet W_{j,k}}{|W_i| * |W_j|} \qquad (1)$$

[0053] 其中，Wi 为第 i 个词语的词向量，$|W_i|$ 为第 i 个词语的向量的的绝对值；Wj 为第 j 个词语的词向量，$|W_j|$ 为第 j 个词语的向量的绝对值，dist(Wi，Wj) 为第 i 个词语与第 j 个词语之间的距离。

[0054] 其中，如果步骤(1)中只获取训练集合中的各描述词语的词向量，则步骤可以为：

[0055] 根据训练集合中的各描述词语的词向量，计算各描述词语中的任意两描述词语之间的距离。

[0056] (3)：将距离小于预设距离的多个词语组成一个词语集合，以及获取用户标注的该词语集合的类别。

[0057] 两个词语之间的距离用于表示两个词语之间的相似度，如果两个词语之间的距离小于预设距离，则确定这两个词语为相近的词语，将这两个词语放到一个词语集合中，并确定这两个词语属于同一类别。通过这种方法能够将训练集合中的各词语分词进行分类并组成多个词语集合；用户根据多个词语集合中的每个词语集合中包括的词语，确定每个词语集合的类别；对每个词语集合进行标注，得到每个词语集合的类别，然后向服务器输入每个词语集合的类别；服务器接收用户输入的每个词语集合的类别。

[0058] 预设距离可以根据需要进行设置并更改，在本发明实施例中对预设距离不作具体限定；例如，预设距离可以为 0.2 或者 0.5 等。

[0059] 需要说明的是，在本发明实施例中可以采用任一聚类方法对训练集合中的词语进行聚类得到多个词语集合；例如，采用分层聚类的方法，则可以获取多个词语集合以及多个词语集合的关系，如图 2-2 所示，每一个圆圈代表一个词语，不同层级表示聚类的层级结构，在聚类结果中，通过人工浏览每一个层级结构包含的词语来对该层对应的词语集合进行标注。

[0060] 在本发明实施例中采用聚类思想对多个文本样本中包括的词语进行聚类，通过变标注多个文本样本变为对多个词语集合进行标注，得到每个类别的词语集合，因此，本发明只需要进行少量标注即可，节省了人力资源，并缩短了标注时间，提高了分类效率。并且，在本发明实施例中获取每个类别的词语集合时，只需要获取少量的文本样本即可，也不需要对文本样本进行标注，从而节省了时间和人力资源，从而达到较快的分类效率，尤其是在互联网行业中，通常文本类别多，数量巨大，为了快速对文本进行分类，可以采用本发明实施例提供的方法，缩短了分类时间，提高了分类效率。

[0061] 在本发明实施例中通过配置类别与词语集合的对应关系，从而实现分类模型的迁移，给定不同的业务场景文本可能是篇幅较长的新闻，也可能是较短的视频的标题或者用

户的微博等文本，不同的业务可能关注的类别不一样，基于文本的思想，只需要在类别集合中增加类别，并建立该增加的类别的词语集合即可，从而能够实现分类模型的迁移，解决模型应对新场景的分类问题，使得分类模型能够快速响应不同业务场景下的分类需求。

[0062] 进一步地，在本发明实施例中也可以不采用聚类思想获取每个类别对应的词语集合，采用用户直接标注的方式获取每个类别对应的词语集合，则步骤 201-203 可以替换为：用户获取多个词语组成训练集合，并根据训练集合中的词语，对训练集合中的词语进行分类，得到多个词语集合，并对多个词语集合中的每个词语集合进行标注，得到每个词语集合的类别，然后向服务器输入每个词语集合以及每个词语集合的类别；服务器接收用户输入的每个词语集合的类别以及每个词语集合的类别。

[0063] 进一步地，获取到每个词语集合的类别时，根据每个类别的词语集合，计算每个类别的词向量，将每个类别和每个类别的词向量存储在类别和词向量的对应关系中，以便于之后获取类别的词向量时，不需要进行重复计算，直接该类别，从类别和词向量的对应关系中获取该类别的词向量。

[0064] 其中，对于每个类别，计算该类别的词向量的过程可以为：

[0065] 获取该类别的词语集合中的各词语的词向量；计算获取的词向量的平均词向量并将该平均词向量作为该类别的词向量。

[0066] 需要说明的是，步骤 201-203 为训练每个类别的词语集合的过程，因此，步骤 201-203 只需要执行一次即可，之后根据每个类别的词语集合对需要分类的文本进行分类时，不需要执行步骤 201-203，只需要执行步骤 204 至 208 对需要分类的文本进行分类即可。

[0067] 步骤 204：根据第一类别的词语集合，获取第一类别的词向量，第一类别为类别集合中的任一类别；

[0068] 具体地，获取第一类别的词语集合中的各词语的词向量；计算获取的词向量的平均词向量并将该平均词向量作为第一类别的词向量。或者，根据第一类别，从类别和词向量的对应关系中获取第一类别的词向量。

[0069] 其中，使用神经网络语言模型中词嵌入技术 word2vec 方法获取第一类别的词语集合中的各词语的词向量；或者，根据第一类别的词语集合中的各词语，从词语和词语的词向量的对应关系中获取各词语的词向量；并且通过这种方法获取类别集合中的每个类别的词语集合中的各词语的词向量。

[0070] 步骤 205：获取待分类的文本包括的每个词语的词向量、词频、权重和逆文档频率；

[0071] 利用现有的分词工具，对待分类的文本进行分词，得到该文本包括的每个词语；使用神经网络语言模型中词嵌入技术 word2vec 方法，获得每个词语的词向量，或者根据每个词语，从词语和词语的词向量的对应关系中获取每个词语的词向量；对于该文本包括的每个词语，统计该词语在该文本中出现的次数作为该词语的词频；获取该词语在该文本中的位置，根据该词语在该文本中的位置，获取该词语的权重；以及获取该词语在训练集合中的逆文档频率。

[0072] 其中，逆文档频率又称反文档频率，是文档频率的倒数；服务器获取该词语在训练集合中的逆文档频率的过程可以为：

[0073] 获取该词语在训练集合中出现的次数，获取训练集合中包括的词语的个数；计算

该词语的词频与该次数的比值得到第一数值，计算第一数值与该个数得到该词语在训练集合中的逆文档频率。

[0074] 其中，服务器设置词语在文本中的位置和权重的对应关系，则根据该词语在该文本中的位置，获取该词语的权重的步骤可以为：

[0075] 根据该词语在该文本中的位置，从位置和权重的对应关系中获取该词语的权重。

[0076] 其中，服务器设置词语在文本中的位置和权重的对应关系时，可以为该文本的标题、摘要或者其他比较重要的位置的词语设置较高的权重，为该文本的正文中的词语设置较低的权重。

[0077] 步骤 206：根据每个词语的词向量和第一类别的词向量，分别计算每个词语与第一类别之间的第一隶属度；

[0078] 在本发明实施例中用每个词语的词向量与第一类别的词向量之间的距离来衡量每个词语与第一类别之间的第一隶属度，则本步骤可以为：

[0079] 分别计算每个词语的词向量与第一类别的词向量之间的距离，将每个词语的词向量与第一类别的词向量之间的距离分别作为每个词语与第一类别之间的第一隶属度。

[0080] 需要说明的是，对于类别集合中的每个类别都按以上方法计算每个词语与该类别之间的第一隶属度。并且，计算每个词语的词向量与第一类别的词向量之间的距离的过程和步骤 203 中计算两个向量之间的距离的过程相同，在此不再赘述。

[0081] 步骤 207：根据每个词语与第一类别之间的第一隶属度以及每个词语的词频、权重和逆文档频率，计算文本与第一类别之间的第二隶属度；

[0082] 其中，本步骤可以通过以下步骤 (1) 和 (2) 实现，包括：

[0083] (1)：分别计算每个词语的词频、权重、逆文档频率以及与第一类别之间的第一隶属度的乘积，得到每个词语与第一类别之间的第三隶属度；

[0084] 对于每个词语，根据每个词语的词频、权重、逆文档频率以及与第一类别之间的第一隶属度的乘积，按照以下公式 (2) 计算每个词语与第一类别之间的第三隶属度：

[0085] $f_{wi} = p_{wi} * tf_{wi} * idf_{wi} * b_{wi,c}$ (2)

[0086] 其中，f_{wi}为第 i 个词语与第一类别之间的第三隶属度，p_{wi}为第 i 个词语的权重，tf_{wi}为第 i 个词语的词频，idf_{wi}为第 i 个词语的逆文档频率，$b_{wi,c}$为第 i 个词语与第一类别之间的第一隶属度。

[0087] (2)：将每个词语与第一类别之间的第三隶属度进行累加，得到该文本与第一类别之间的第二隶属度。

[0088] 根据每个词语与第一类别之间的第三隶属度，按照以下公式 (3) 计算该文本与第一类别之间的第二隶属度：

[0089] $$F = f_1 + f_2 + \cdots\cdots + f_n = \sum_{wi \in c} p_{wi} * tf_{wi} * idf_{wi} * b_{wi,c} \quad (3)$$

[0090] 需要说明的是，对应类别集合中的每个类别都按以上方法计算该文本与该类别之间的第二隶属度。

[0091] 步骤 208：从类别集合中选择与该文本之间的第二隶属度满足预设条件的类别，将选择的类别确定为该文本的类别。

[0092] 第二隶属度用于表示该文本与该类别之间的相似度，预设条件可以为最大的第二

隶属度，也可以为大于第一预设数值的第二隶属度；当预设条件为最大的第二隶属度时，本步骤可以为：从类别集合中选择与该文本之间的第二隶属度中最大的第二隶属度的类别，将选择的类别确定为该文本的类别。

[0093] 当预设条件为大于第一预设数值的第二隶属度，则本步骤可以为：从类别集合中获取与该文本之间的第二隶属度大于第一预设数值的类别，从获取的类别中随机选择一个类别，将选择的类别确定为该文本的类别。

[0094] 进一步地，获取到该文本与类别集合中的每个类别之间的第二隶属度时，还采用如下公式（4）对第二隶属度进行归一化，得到归一化后的与该文本之间的第二隶属度；

[0095] $$F=\frac{\sum_{Wiec,\ c\in m(T)} wi*tfwi*idfwi*bwi,\ c}{\sum_{c\in C}\sum_{Wiec,\ c\in m(T)} wi*tfwi*idfwi*bwi,\ c} \qquad (4)$$

[0096] 则本步骤可以为：从类别集合中选择与该文本之间的归一化后的第二隶属度满足预设条件的类别，将选择的类别确定为该文本的类别。

[0097] 此时，预设条件可以为最大的归一化后的第二隶属度，也可以为大于第二预设数值的归一化后的第二隶属度。

[0098] 当预设条件为最大的归一化后的第二隶属度，则本步骤可以为：从类别集合中选择与该文本之间的归一化后的第二隶属度最大的归一化后的隶属度的类别，将选择的类别确定为该文本的类别。

[0099] 当预设条件为大于第二预设数值的归一化后的第二隶属度，则本步骤可以为：从类别集合中获取与该文本之间的归一化后的第二隶属度大于第二预设数值的类别，从获取的类别中随机选择一个类别，将选择的类别确定为该文本的类别。

[0100] 第一预设数值和第二预设数值都可以根据需要进行设置并更改，在本发明实施例中对第一预设数值和第二预设数值都不作具体限定。

[0101] 在本发明实施例中，根据待分类的文本包括的每个词语的词向量、词频、权重和逆文档频率和第一类别的词向量，计算该文本与第一类别之间的第二隶属度，第一类别为类别集合中的任一类别，根据与该文本之间的第二隶属度，从类别集合中选择类别；由于本发明在对待分类的文本进行分类时，考虑了该文本包括的每个词语，因此提高了分类的准确性。

[0102] 实施例 3

[0103] 本发明实施例提供了一种文本分类的装置，参见图 3，其中，该装置包括：

[0104] 第一获取模块 301，用于获取待分类的文本包括的每个词语的词向量、词频、权重和逆文档频率；

[0105] 第一计算模块 302，用于根据每个词语的词向量和第一类别的词向量，分别计算每个词语与第一类别之间的第一隶属度，第一类别为类别集合中的任一类别；

[0106] 第二计算模块 303，用于根据每个词语与第一类别之间的第一隶属度以及每个词语的词频、权重和逆文档频率，计算该文本与第一类别之间的第二隶属度；

[0107] 分类模块 303，用于从类别集合中选择与该文本之间的第二隶属度满足预设条件的类别，将选择的类别确定为该文本的类别。

[0108] 进一步地，第一计算模块 302，包括：

[0109] 第一获取单元，用于获取第一类别对应的词语集合中的各词语的词向量；

[0110] 第一计算单元，用于计算获取到词向量的平均词向量并将该平均词向量作为第一类别的词向量；

[0111] 第二计算单元，用于分别计算每个词语的词向量与第一类别的词向量之间的距离，将每个词语的词向量与第一类别的词向量之间的距离分别作为每个词语与第一类别之间的第一隶属度。

[0112] 进一步地，该装置还包括：

[0113] 第二获取模块，用于获取多个文本样本；

[0114] 分词模块，用于将多个文本样本中的每个文本样本进行分词，将得到的词语组成训练集合；

[0115] 聚类模块，用于对训练集合中的词语进行聚类，得到多个词语集合以及多个词语集合中的每个词语集合的类别。

[0116] 进一步地，聚类模块，包括：

[0117] 第二获取单元，用于获取训练集合中的各词语的词向量；

[0118] 第三计算单元，用于根据各词语的词语向量，计算各词语中的任意两词语之间的距离；

[0119] 聚类单元，用于将距离小于预设距离的多个词语组成一个词语集合；

[0120] 第三获取单元，用于获取用户标注的该词语集合的类别。

[0121] 进一步地，第二计算模块 303，包括：

[0122] 第四计算单元，用于分别计算每个词语的词频、权重、逆文档频率以及与第一类别之间的第一隶属度的乘积，得到每个词语与第一类别之间的第三隶属度；

[0123] 累加单元，用于将每个词语与第一类别之间的第三隶属度进行累加，得到该文本与第一类别之间的第二隶属度。

[0124] 在本发明实施例中，根据待分类的文本包括的每个词语的词向量、词频、权重和逆文档频率和第一类别的词向量，计算该文本与第一类别之间的第二隶属度，第一类别为类别集合中的任一类别，根据与该文本之间的第二隶属度，从类别集合中选择类别；由于本发明在对待分类的文本进行分类时，考虑了该文本包括的每个词语，因此提高了分类的准确性。

[0125] 实施例 4

[0126] 图 4 是本发明实施例提供的服务器的结构示意图。该服务器 1900 可因配置或性能不同而产生比较大的差异，可以包括一个或一个以上中央处理器 (central processing units，CPU) 1922（例如，一个或一个以上处理器）和存储器 1932，一个或一个以上存储应用程序 1942 或数据 1944 的存储介质 1930（例如一个或一个以上海量存储设备）。其中，存储器 1932 和存储介质 1930 可以是短暂存储或持久存储。存储在存储介质 1930 的程序可以包括一个或一个以上模块（图示没标出），每个模块可以包括对服务器中的一系列指令操作。更进一步地，中央处理器 1922 可以设置为与存储介质 1930 通信，在服务器 1900 上执行存储介质 1930 中的一系列指令操作。

[0127] 服务器 1900 还可以包括一个或一个以上电源 1926，一个或一个以上有线或无线网络接口 1950，一个或一个以上输入输出接口 1958，一个或一个以上键盘 1956，和 / 或，

一个或一个以上操作系统 1941，例如 Windows ServerTM，Mac OS XTM，UnixTM，LinuxTM，FreeBSDTM 等等。

[0128] 服务器 1900 可以包括有存储器，以及一个或者一个以上的程序，其中一个或者一个以上程序存储于存储器中，且经配置以由一个或者一个以上处理器执行所述一个或者一个以上程序包含用于进行以下操作的指令：

[0129] 获取待分类的文本包括的每个词语的词向量、词频、权重和逆文档频率；

[0130] 根据所述每个词语的词向量和第一类别的词向量，分别计算所述每个词语与所述第一类别之间的第一隶属度，所述第一类别为类别集合中的任一类别；

[0131] 根据所述每个词语与所述第一类别之间的第一隶属度以及所述每个词语的词频、权重和逆文档频率，计算所述文本与所述第一类别之间的第二隶属度；

[0132] 从所述类别集合中选择与所述文本之间的第二隶属度满足预设条件的类别，将所述选择的类别确定为所述文本的类别。

[0133] 进一步地，所述根据所述每个词语的词向量和第一类别的词向量，分别计算所述每个词语与所述第一类别之间的第一隶属度，包括：

[0134] 获取第一类别对应的词语集合中的各词语的词向量；

[0135] 计算所述获取到词向量的平均词向量并将所述平均词向量作为所述第一类别的词向量；

[0136] 分别计算所述每个词语的词向量与所述第一类别的词向量之间的距离，将所述每个词语的词向量与所述第一类别的词向量之间的距离分别作为所述每个词语与所述第一类别之间的第一隶属度。

[0137] 进一步地，所述方法还包括：

[0138] 获取多个文本样本；

[0139] 将所述多个文本样本中的每个文本样本进行分词，将得到的词语组成训练集合；

[0140] 对所述训练集合中的词语进行聚类，得到多个词语集合以及所述多个词语集合中的每个词语集合的类别。

[0141] 进一步地，所述对所述训练集合中的词语进行聚类，得到多个词语集合以及所述多个词语集合中的每个词语集合的类别，包括：

[0142] 获取所述训练集合中的各词语的词向量；

[0143] 根据所述各词语的词语向量，计算所述各词语中的任意两词语之间的距离；

[0144] 将距离小于预设距离的多个词语组成一个词语集合，以及获取用户标注的所述词语集合的类别。

[0145] 进一步地，所述根据所述每个词语与所述第一类别之间的第一隶属度以及所述每个词语的词频、权重和逆文档频率，计算所述文本与所述第一类别之间的第二隶属度，包括：

[0146] 分别计算所述每个词语的词频、权重、逆文档频率以及与所述第一类别之间的第一隶属度的乘积，得到所述每个词语与所述第一类别之间的第三隶属度；

[0147] 将所述每个词语与所述第一类别之间的第三隶属度进行累加，得到所述文本与所述第一类别之间的第二隶属度。

[0148] 在本发明实施例中，根据待分类的文本包括的每个词语的词向量、词频、权重和逆

文档频率和第一类别的词向量，计算该文本与第一类别之间的第二隶属度，第一类别为类别集合中的任一类别，根据与该文本之间的第二隶属度，从类别集合中选择类别；由于本发明在对待分类的文本进行分类时，考虑了该文本包括的每个词语，因此提高了分类的准确性。

[0149] 需要说明的是：上述实施例提供的文本分类的装置在文本分类时，仅以上述各功能模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能模块完成，即将装置的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。另外，上述实施例提供的文本分类的装置与文本分类的方法实施例属于同一构思，其具体实现过程详见方法实施例，这里不再赘述。

[0150] 本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成，也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，上述提到的存储介质可以是只读存储器，磁盘或光盘等。

[0151] 以上所述仅为本发明的较佳实施例，并不用以限制本发明，凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

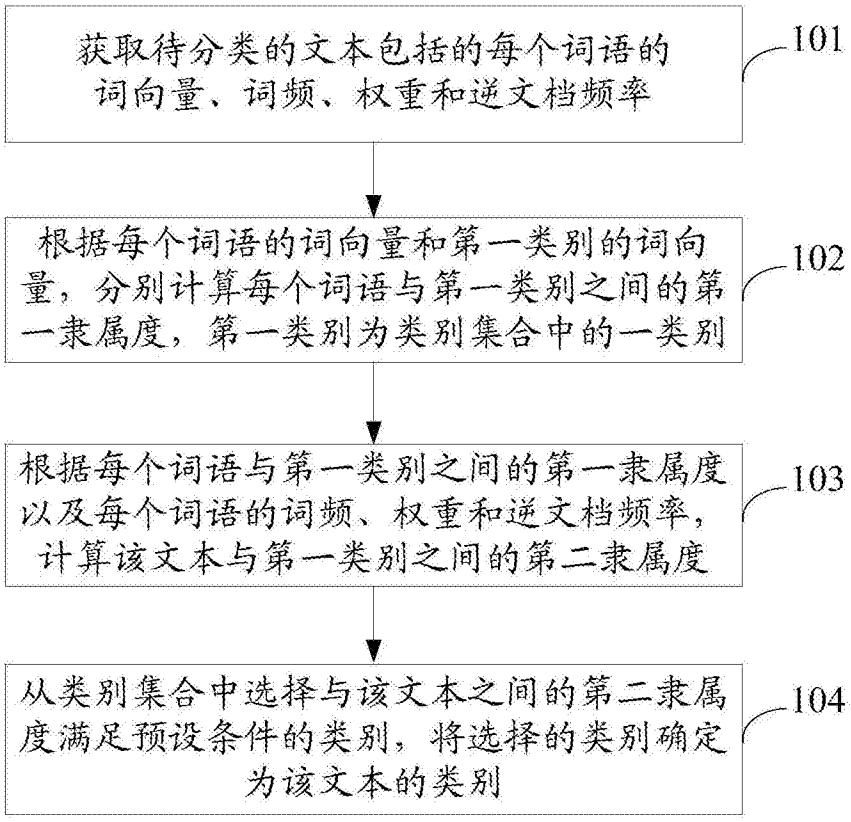

获取待分类的文本包括的每个词语的词向量、词频、权重和逆文档频率
101
根据每个词语的词向量和第一类别的词向量，分别计算每个词语与第一类别之间的第一隶属度，第一类别为类别集合中的一类别
102
根据每个词语与第一类别之间的第一隶属度以及每个词语的词频、权重和逆文档频率，计算该文本与第一类别之间的第二隶属度
103
从类别集合中选择与该文本之间的第二隶属度满足预设条件的类别，将选择的类别确定为该文本的类别
104

图 1

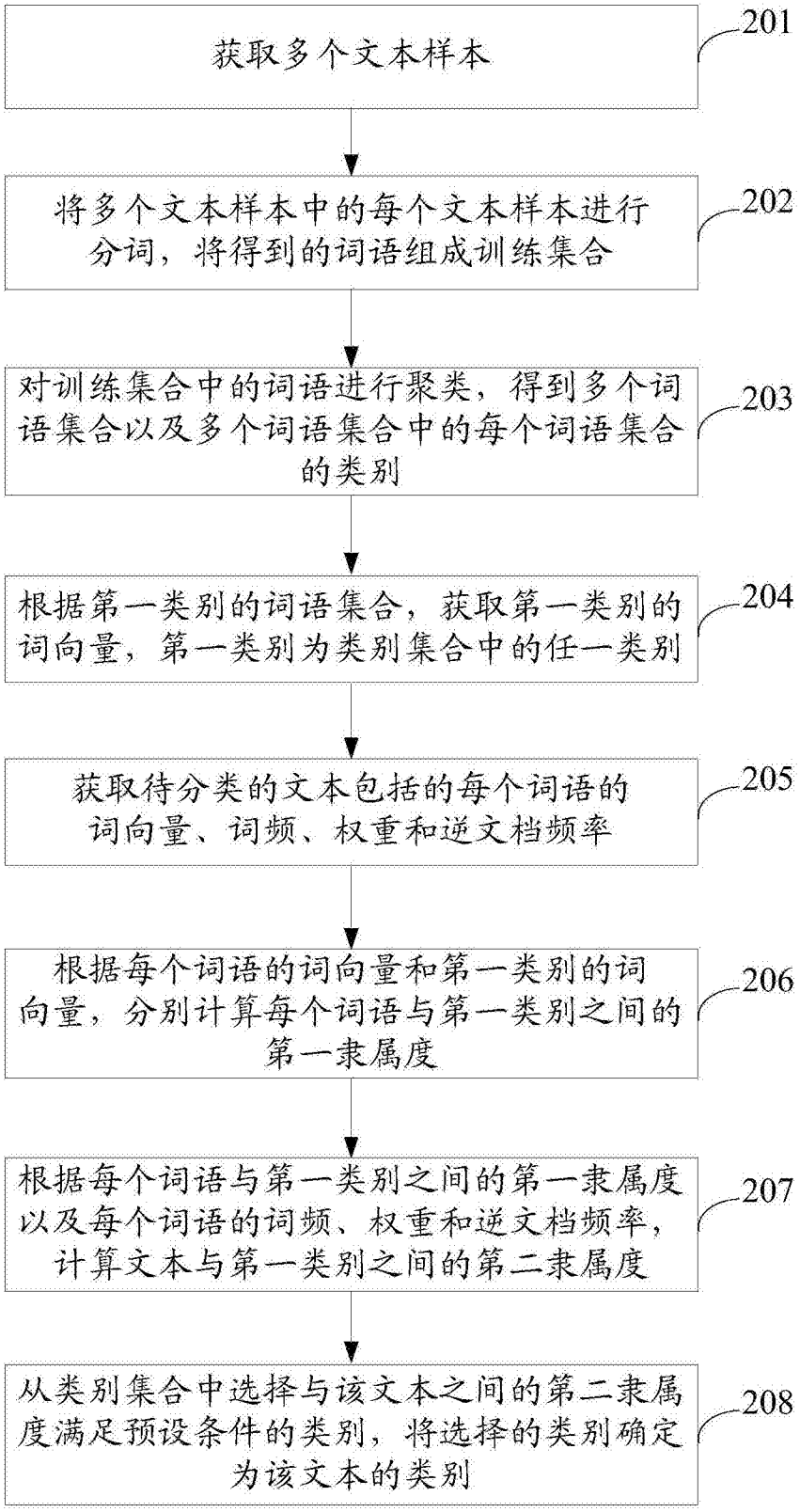

获取多个文本样本
201
将多个文本样本中的每个文本样本进行分词，将得到的词语组成训练集合
202
对训练集合中的词语进行聚类，得到多个词语集合以及多个词语集合中的每个词语集合的类别
203
根据第一类别的词语集合，获取第一类别的词向量，第一类别为类别集合中的任一类别
204
获取待分类的文本包括的每个词语的词向量、词频、权重和逆文档频率
205
根据每个词语的词向量和第一类别的词向量，分别计算每个词语与第一类别之间的第一隶属度
206
根据每个词语与第一类别之间的第一隶属度以及每个词语的词频、权重和逆文档频率，计算文本与第一类别之间的第二隶属度
207
从类别集合中选择与该文本之间的第二隶属度满足预设条件的类别，将选择的类别确定为该文本的类别
208

图 2-1

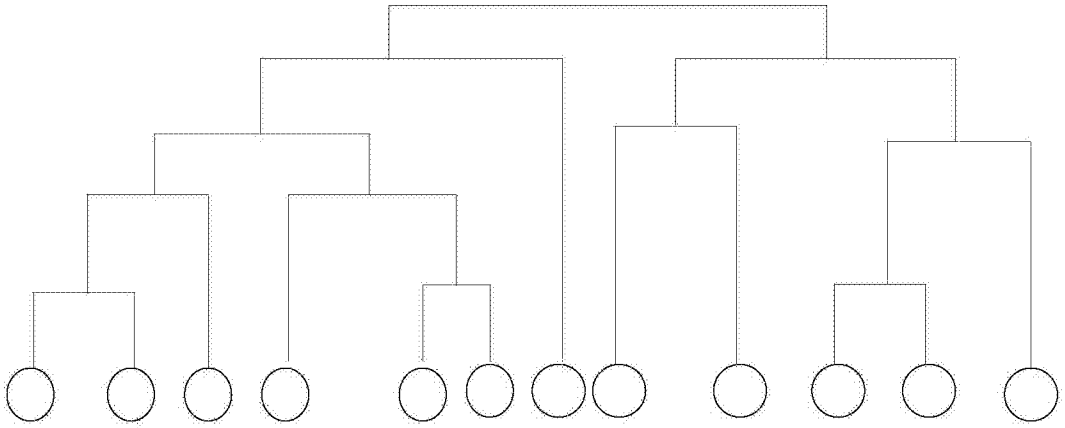


图 2-2

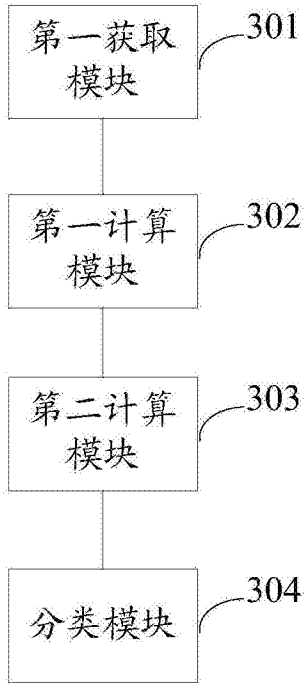


图 3

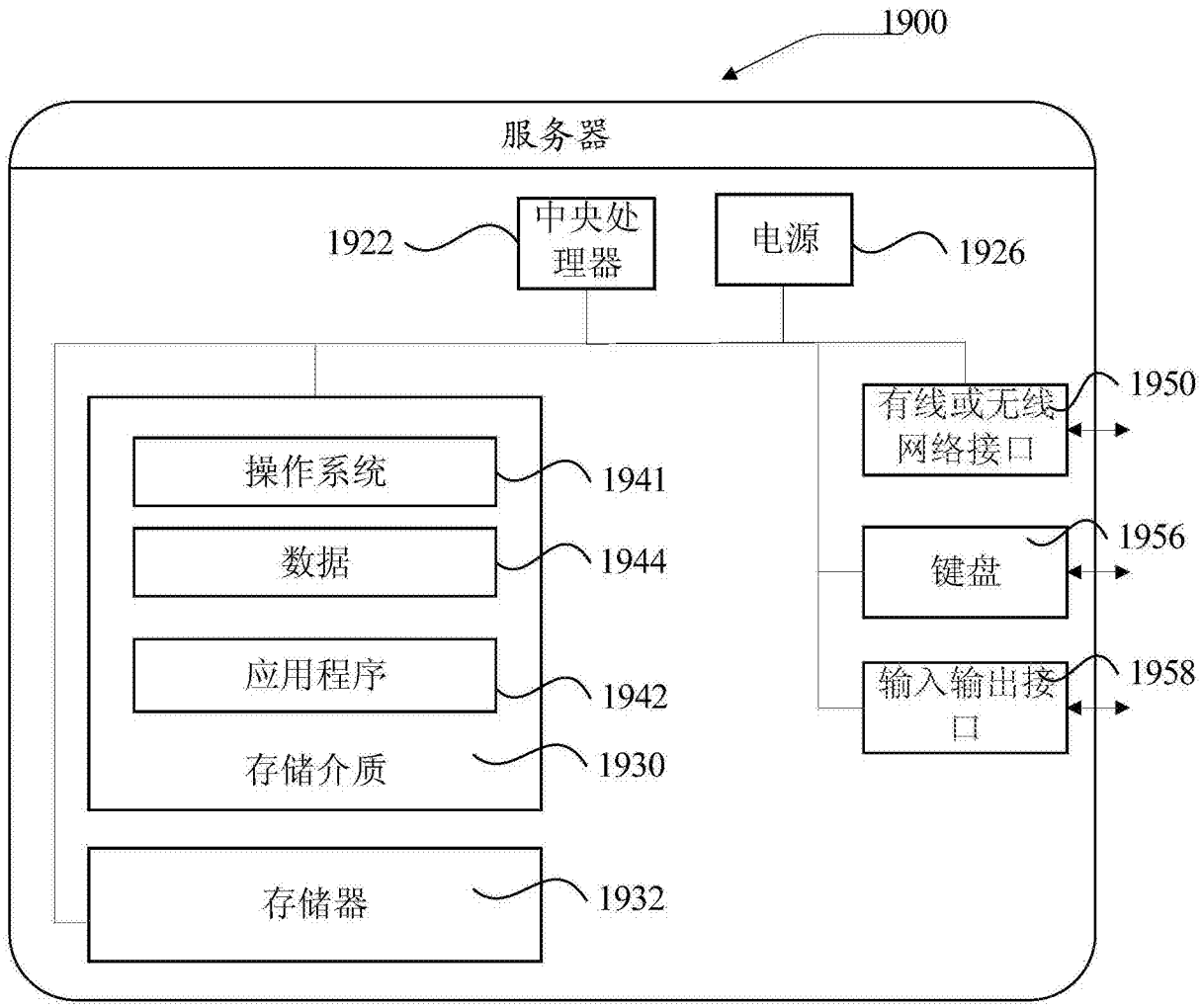

1900
服务器
1922
中央处理器
电源
1926
有线或无线网络接口
1950
操作系统
1941
数据
1944
键盘
1956
应用程序
1942
输入输出接口
1958
存储介质
1930
存储器
1932

图 4