(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 102141977 A
(43)申请公布日 2011.08.03

(21)申请号 201010104512.4

(22)申请日 2010.02.01

(71)申请人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼岛资本大厦一座四层 847 号邮箱

(72)发明人 孙翔

(74)专利代理机构 北京同达信恒知识产权代理有限公司 11291
代理人 郭润湘

(51)Int.Cl.
G06F 17/21(2006.01)
G06F 17/30(2006.01)

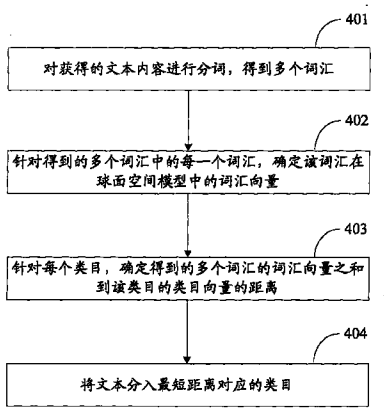
权利要求书 1 页 说明书 6 页 附图 4 页

(54)发明名称

一种文本分类的方法及装置

(57)摘要

本申请公开了一种文本分类的方法，用于实现文本分类，简化分类操作，并提高文本分类的准确度。所述方法包括：对获得的文本内容进行分词，得到多个词汇；针对得到的多个词汇中的每一个词汇，确定该词汇在球面空间模型中的词汇向量；词汇的词汇向量包括该词汇在各类目上的词频值进行归一化后得到的归一化词频值；球面空间模型是以单位长度为半径的多维球体模型，球面空间的维度等于类目的个数，类目对应球面空间中的一个类目向量；针对每个类目，确定得到的多个词汇的词汇向量之和到该类目的类目向量的距离；将文本分入最短距离对应的类目。本申请还公开了用于实现所述方法的装置。



CN 102141977 A

1. 一种文本分类的方法，其特征在于，包括以下步骤：

对获得的文本内容进行分词，得到多个词汇；

针对得到的多个词汇中的一个词汇，确定该词汇在球面空间模型中的词汇向量，其中球面空间的维度等于类目的个数，类目对应球面空间中的一个类目向量；

针对每个类目，确定得到的多个词汇的词汇向量之和到该类目的类目向量的距离；

将文本分入最短距离对应的类目。

2. 如权利要求 1 所述的方法，其特征在于，还包括步骤：在得到多个词汇后，对得到的多个词汇进行过滤，得到符合过滤条件的多个词汇。

3. 如权利要求 1 所述的方法，其特征在于，词汇向量到类目向量的距离为直线距离或球面距离。

4. 如权利要求 1、2 或 3 所述的方法，其特征在于，词汇的词汇向量包括该词汇在各类目上的词频值进行归一化后得到的归一化词频值；球面空间模型是以单位长度为半径的多维球体模型。

5. 如权利要求 4 所述的方法，其特征在于，所述单位长度为 1。

6. 如权利要求 1、2 或 3 所述的方法，其特征在于，确定得到的多个词汇的词汇向量之和到各类目的类目向量的距离的步骤包括：将得到的多个词汇的词汇向量在该类目上的归一化词频值进行累加，得到归一化词汇向量和；

将文本分入最短距离对应的类目的步骤包括：将文本分入归一化词汇向量和的最大分量对应的类目。

7. 一种用于文本分类的装置，其特征在于，包括：

分词模块，用于对获得的文本内容进行分词，得到多个词汇；

查询模块，用于针对得到的多个词汇中的一个词汇，确定该词汇在球面空间模型中的词汇向量，其中球面空间的维度等于类目的个数，类目对应球面空间中的一个类目向量；

计算模块，用于针对每个类目，确定得到的多个词汇的词汇向量之和到该类目的类目向量的距离；

分类模块，用于将文本分入最短距离对应的类目。

8. 如权利要求 7 所述的装置，其特征在于，还包括：过滤模块，用于对得到的多个词汇进行过滤，得到符合过滤条件的多个词汇。

9. 如权利要求 7 所述的装置，其特征在于，词汇向量到类目的距离为直线距离或球面距离。

10. 如权利要求 7、8 或 9 所述的装置，其特征在于，词汇的词汇向量包括该词汇在各类目上的词频值进行归一化后得到的归一化词频值；球面空间模型是以单位长度为半径的多维球体模型。

11. 如权利要求 10 所述的装置，其特征在于，所述单位长度为 1。

12. 如权利要求 7、8 或 9 所述的装置，其特征在于，计算模块将得到的多个词汇的词汇向量在该类目上的归一化词频值进行累加，得到归一化词汇向量和；

分类模块将文本分入归一化词汇向量和的最大分量对应的类目。

一种文本分类的方法及装置

技术领域

[0001] 本申请涉及计算机及通信领域，特别是涉及文本分类的方法及装置。

背景技术

[0002] 文本分类是文本挖掘的一个重要内容，是指按照预先定义的主题类别，为文档集合中的每个文档确定一个类别。通过自动文本分类系统把文档进行归类，可以帮助人们更好地寻找需要的信息和知识。在人们看来，分类是对信息的一种最基本的认知形式。传统的文献分类研究有着丰富的研究成果和相当的实用水平。但随着文本信息的快速增长，特别是互联网 (Internet) 上在线文本信息的激增，文本自动分类已经成为处理和组织大量文档数据的关键技术。现在，文本分类正在各个领域得到广泛的应用。但是，随着信息量日趋丰富，人们对于内容搜索的准确率，查全率等方面的要求会越来越高，因而对文本分类技术需求大为增加，如何构造一个有效的文本分类系统仍然是文本挖掘的一个主要研究方向。

[0003] 在自然语言处理领域，文本的表示主要采用向量空间模型 (Vector spacemodel，VSM)，这种方法认为每篇文本都包含一些用概念词表达的揭示其内容的独立属性，而每个属性都可以看成是概念空间的一个维数，这些独立属性称为文本特征项，文本就可以表示为这些特征项的集合。特征向量的相近程度常用夹角余弦来衡量。然后根据文本向量与候选类别的特征向量的相近程度来判定文本的类别。

[0004] 现有技术中需要计算每个文本向量与候选类别的所有特征向量相近程度，每次计算均需要采用夹角余弦来衡量，计算量非常大，并且现有技术对文本的语义没有任何约束，其分类的准确度不是很好。

发明内容

[0005] 本申请实施例提供一种文本分类的方法及装置，用于实现文本分类，简化分类操作，并提高文本分类的准确度。

[0006] 一种文本分类的方法，包括以下步骤：

[0007] 对获得的文本内容进行分词，得到多个词汇；

[0008] 针对得到的多个词汇中的每一个词汇，确定该词汇在球面空间模型中的词汇向量；词汇的词汇向量包括该词汇在各类目上的词频值进行归一化后得到的归一化词频值；球面空间模型是以单位长度为半径的多维球体模型，球面空间的维度等于类目的个数，类目对应球面空间中的一个类目向量；

[0009] 针对每个类目，确定得到的多个词汇的词汇向量之和到该类目的类目向量的距离；

[0010] 将文本分入最短距离对应的类目。

[0011] 一种用于文本分类的装置，包括：

[0012] 分词模块，用于对获得的文本内容进行分词，得到多个词汇；

[0013] 查询模块，用于针对得到的多个词汇中的每一个词汇，确定该词汇在球面空间模

型中的词汇向量；词汇的词汇向量包括该词汇在各类目上的词频值进行归一化后得到的归一化词频值；球面空间模型是以单位长度为半径的多维球体模型，球面空间的维度等于类目的个数，类目对应球面空间中的一个类目向量；

[0014] 计算模块，针对每个类目，确定得到的多个词汇的词汇向量之和到该类目的类目向量的距离；

[0015] 分类模块，用于将文本分入最短距离对应的类目。

[0016] 本申请实施例预先构造一球面空间模型，并基于该球面空间模型对文本进行分类，在分类过程中，计算文本中各词汇的向量和与各类目向量的距离，从而确定文本应分入的类目。本申请实施例实现了文本分类，并且相对于现有技术中的夹角余弦算法，计算量明显减少。以及本申请实施例中球面空间模型以单位长度为半径，则一个词汇在各类目上的归一化后的词汇向量的平方和也为单位长度，相当于将一个词汇的语义信息量等价为单位长度，对语义信息量进行了约束，因此相对于现有技术可提高文本分类的准确度。

附图说明

[0017] 图 1 为本申请实施例中装置的主要结构图；

[0018] 图 2 为本申请实施例中装置的详细结构图；

[0019] 图 3 为本申请实施例中球面空间的示意图；

[0020] 图 4 为本申请实施例中文本分类的主要方法流程图；

[0021] 图 5 为本申请实施例中通过距离和进行文本分类的方法流程图；

[0022] 图 6 为本申请实施例中通过词汇向量和进行文本分类的方法流程图。

具体实施方式

[0023] 本申请实施例预先构造一球面空间模型，并基于该球面空间模型对文本进行分类，在分类过程中，计算文本中各词汇的向量和与各类目向量的距离，从而确定文本应分入的类目。本申请实施例实现了文本分类，并且相对于现有技术中的夹角余弦算法，计算量明显减少。以及本申请实施例中球面空间模型以单位长度为半径，则一个词汇在各类目上的归一化后的词汇向量的平方和也为单位长度，相当于将一个词汇的语义信息量等价为单位长度，对语义信息量进行了约束，因此相对于现有技术可提高文本分类的准确度。

[0024] 参见图 1，本实施例中用于文本分类的装置包括：分词模块 101、查询模块 102、计算模块 103 和分类模块 104。

[0025] 分词模块 101 用于对获得的文本内容进行分词，得到多个词汇。

[0026] 查询模块 102 用于针对得到的多个词汇中的一个词汇，确定该词汇在球面空间模型中的词汇向量。词汇的词汇向量包括该词汇在各类目上的词频值进行归一化后得到的归一化词频值；球面空间模型是以单位长度为半径的多维球体模型，球面空间的维度等于类目的个数，类目对应球面空间中的一个类目向量。其中，单位长度可以为一常数，为了便于计算，本实施例中球面空间模型的半径为 1。文本中各词汇的向量和到各类目向量的距离为直线距离或球面距离。

[0027] 计算模块 103 用于针对每个类目，确定对文本分词后得到的多个词汇的词汇向量和到每个类目向量的距离。

[0028] 分类模块 104 用于将文本分入最短距离对应的类目。

[0029] 计算模块 103 在计算文本中词汇向量和到各类目向量的距离时，可将对文本分词后得到的多个词汇的词汇向量在相应类目上的归一化词频值进行累加，得到归一化词汇向量和。分类模块 104 将文本分入归一化词汇向量和的最大分量对应的类目。

[0030] 所述装置还包括：接口模块 105、过滤模块 106、构造模块 107 和存储模块 108，参见图 2 所示。

[0031] 接口模块 105 用于从装置外部获得待分类的文本。

[0032] 过滤模块 106 用于在对文本分词得到多个词汇后，对得到的多个词汇进行过滤，得到符合过滤条件的多个词汇。过滤条件有多种，如根据词汇在各类目上的词频值计算该词汇的变异系数，然后过滤出变异系数大于预设的变异系数阈值（如 0.5）的词汇。通过变异系数，可过滤掉在各类目中词频值变化不大的词（如你、我等在各类目的词频值基本一致），而保留在各类目中词频值变化较明显的词（如专业名词，在与其专业有关类目中的词频值明显高于其它类目下的词频值）。在各类目中词频值变化较明显的词，说明其主要出现在某一个或某几个类目中，这样的词对文本分类的准确性做出较多的贡献，本实施例认为这样的词属于优秀词，应通过过滤来筛选出优秀词。还可能有其它过滤条件，此处不一一列举。

[0033] 构造模块 107 用于构造球面空间模型。

[0034] 存储模块 108 用于存储构造的球面空间模型，以及分类存储各文本等。

[0035] 构造模块 107 构造球面空间模型的过程如下：

[0036] 设多维球面空间为 S，S 的维数与类目的总数相同。类目 Ci 是球面上的一个端点，同时对应球面空间中的一个类目向量，Ci = {0，.....，0，1，0，.....，0}，相当于从球心（相当于原点）指向球面端点，该类目向量的第 i 个维度值是 1，其余都是 0。本实施例中，假设任意一个词汇在任意两个类目 Ci 和 Cj 中出现的概率是概率独立的，则 Ci 和 Cj 在 S 中必然是相互垂直的，推广到一般，所有类目向量 {Ci} 是两两垂直的。

[0037] 本实施例中第 m 个词汇的词汇向量 Wm 为 S 中的一个向量，m = 1...M，M 为词汇的总数。Wm = {V1，V2，.....，VN}，Vi 是在类目 Ci 上的归一化词频值，i = 1...N，N 为类目的总数。该归一化词频值从球心指向球面端点，则可将归一化词频值表示为类目 Ci 上的坐标。词汇的词汇向量与类目向量的示意图参见图 3 所示，Ci、Cj 和 Ck 表示三个类目向量，O 表示球心，也是原点（坐标为 {0，0，.....，0}）。

[0038] 本实施例中设任一个词汇的语义信息量均为同一个常数，语义信息量是指认识主体所感知或所表述的事物的存在方式和运动状态的逻辑含义，是词汇内在含义因素的信息部分。定义该常数为单位长度，则词汇向量在 S 中的长度（即词汇向量的端点到原点 O 的距离）也为该常数，为了计算方便，设该常数为 1。词汇向量的端点到原点 O 的距离可表示为：|Wm-O| = 1（公式 1），进而根据 Wm = {V1，V2，.....，VN} 有 $\Sigma Vi^2 = 1$（公式 2）。由公式 1 可知，词汇向量 Wm 的端点均落在球面上。由于词汇向量 Wm 和类目向量 Ci 的端点都落在球面上，则任一个词汇的语义与类目的近似程度可以用 Wm 与 Ci 的距离来表示，距离越短则越接近。Wm 与 Ci 的距离可以通过直线距离或球面距离来计算。

[0039] 由于定义了任一个词汇的语义信息量均为同一个常数，则归一化词频值为词频值经过归一化后得到的，$\sqrt{\sum (Fi \times k)^2} = 1$，进而有 $\Sigma (Fi \times k)^2 = 1$，其中 Fi 为该词汇在类目 Ci

上的词频值，k 为预设的归一化系数。由 $\Sigma (Fi \times k)^2 = 1$ 可以导出 $k = \sqrt{\frac{1}{\sum Fi^2}}$（公式 3）。再由 $Vi = Fi \times k$（公式 4）可得出词汇向量与词频值的转换函数（或称量化函数）$Wm = \delta(Fi) = \{Fi\} \times k$（公式 5）。

[0040] 通过以上描述，构造模块 107 构造出以原点为球心，以单位长度 1 为半径的球面空间，词汇向量 Wm 和类目向量 Ci 的端点都落在球面上。经过训练样本对该球面空间模型进行训练和学习，得到可直接应用的球面空间模型。样本训练过程与文本分类过程类似，也可通过其它的模式识别或人工等方式实现。

[0041] 对于一个文本 D 来说，$D = \Sigma Wm$，Wm 为文本中第 m 个词汇的词汇向量。计算模块 103 计算 ΣWm 与类目向量 Ci 的距离，则最短距离对应的类目即为文本应分入的类目。由于 ΣWm 不一定落在球面上，为了便于计算，计算模块 103 也可对 D 进行归一化，乘以归一化系数 k，再计算与类目向量 Ci 的距离。

[0042] 由于词汇向量 Wm 和与类目向量 Ci 的距离越短表示词汇向量 Wm 和与类目向量 Ci 的近似程度越大，为了简化计算过程，则可设 $P = \{Pi\} = \{\Sigma Vmi\}$（公式 6），Pi 表示第 i 个类目的权重分量，Pi 值越大对应到类目向量 Ci 的距离越短，相当于 ΣVmi 越大对应到类目向量 Ci 的距离越短。所以计算模块 103 在计算距离和时，可将得到的多个词汇在该类目上的归一化分量值进行累加，得到该类目的权重值。分类模块 104 将文本分入最大权重值对应的类目。

[0043] Pi 值越大对应到类目向量 Ci 的距离越短的实现原理如下：

[0044] 由于 $D = \Sigma Wm$，$Wm = \{V1, V2, \ldots\ldots, VN\}$，则 $D = \{\Sigma Vm1, \Sigma Vm2, \ldots\ldots, \Sigma Vmi, \ldots\ldots, \Sigma Vmn\}$，$\Sigma Vmi$ 是文档的所有词汇在第 i 个类目上的归一化词频值之和。不妨设 $Pi = \Sigma Vmi$，那么，$D = \{Pi\}$。D 到 Ci 的距离可表示为：

[0045] $|D-Ci| = |\{P1, P2, \ldots Pi, \ldots, Pn\} \times k - \{0,0,\ldots,0,1,0,\ldots,0\}|$

[0046] $= k \times |\{P1, P2, \ldots Pi, \ldots, Pn\} - \{0,0,\ldots,0,1/k,0,\ldots,0\}|$

[0047] $= k \times sqrt((P1-0)^2 + (P2-0)^2 + \ldots + (Pi-1/k)^2 + \ldots + (Pn-0)^2)$

[0048] $= k \times sqrt(P1^2 + P2^2 + \ldots + (Pi^2 - 2Pi/k + 1/k^2) + \ldots + Pn^2)$

[0049] $= k \times sqrt(\Sigma (Pi^2) - 2Pi/k + 1/k^2)$

[0050] $= sqrt(\Sigma ((Pi \times k)^2) - 2K \times Pi + 1)$

[0051] 由于 $\Sigma ((Pi \times k)^2) = 1$，所以：$sqrt(\Sigma ((Pi \times k)^2) - 2K \times Pi + 1) = sqrt(1 - 2K \times Pi + 1) = sqrt(2*(1 - K \times Pi))$。由此可知，D 到 Ci 的距离与 Pi 成反比关系，取 Pi 最大的类目，就是最接近的类目。其中 sqrt 表示 $\sqrt{\ }$。

[0052] 所述装置可以位于一个计算机设备内，或者所述装置中的各模块由不同的计算机设备实现，由多个计算机设备协作完成所述装置的功能。所述装置中的各模块可以是软件、硬件，也可以是软件和硬件相结合的形式实现。

[0053] 以上描述了文本分类装置的内部结构和功能，下面对文本分类的实现过程进行介绍。

[0054] 参见图 4，本实施例中文本分类的主要方法流程如下：

[0055] 步骤 401：对获得的文本内容进行分词，得到多个词汇。

[0056] 步骤 402：针对得到的多个词汇中的每一个词汇，确定该词汇在球面空间模型中

的词汇向量。词汇的词汇向量包括该词汇在各类目上的词频值进行归一化后得到的归一化词频值；球面空间模型是以单位长度为半径的多维球体模型，球面空间的维度等于类目的个数，类目对应球面空间中的一个类目向量。

[0057] 步骤 403：针对每个类目，确定得到的多个词汇的词汇向量之和到该类目的类目向量的距离。

[0058] 步骤 404：将文本分入最短距离对应的类目。

[0059] 本实施例中可以通过距离进行文本分类，也可以根据文本中的词汇向量之和进行文本分类，下面针对这两种情况详细介绍文本分类的过程。

[0060] 参见图 5，本实施例中通过距离和进行文本分类的方法流程如下：

[0061] 步骤 501：对获得的文本内容进行分词，得到多个词汇。

[0062] 步骤 502：对得到的多个词汇进行过滤，得到符合过滤条件的多个词汇。过滤模块 106 可以根据词频值对词汇进行过滤。过滤条件有多种，如保留在所有类目上的词频均值大于预设数量；如词汇的归一化词汇向量中最大分量（即最大归一化词频值）大于预设的词频阈值的词汇等，此处不一一列举。

[0063] 步骤 503：针对符合过滤条件的每一个词汇，查询该词汇在各类目上的归一化词频值。其中，预先存有所有词汇在各类目上的归一化词频值，如果存储的词汇不够全，无法查询到该词汇，则该词汇在各类目上的归一化词频值均为 0。如果未存有词汇在各类目上的归一化词频值，而是存有词汇在各类目上的词频值，则可由查询模块 102 查询词频值，并对词频值进行归一化，得到归一化词频值，具体实现可参见公式 4。此步骤可过滤掉干扰词汇（如生僻词汇和常见词汇等），尽量过滤出专业词汇，以提高文本分类的准确度。

[0064] 步骤 504：针对每个类目，确定符合过滤条件的多个词汇的词汇向量和到各类目向量的距离。词汇向量和到类目向量的距离为直线距离或球面距离。

[0065] 在步骤 504 之前，还可以对得到的词汇向量和进行归一化，使归一化后的词汇向量和落入球面空间内。然后在步骤 504 中确定归一化后的词汇向量和到类目的距离。

[0066] 步骤 505：将文本分入最短距离对应的类目。

[0067] 可进一步在数据库中依据类目分类保存文本。

[0068] 参见图 6，本实施例中通过归一化词汇向量和进行文本分类的方法流程如下：

[0069] 步骤 601：对获得的文本内容进行分词，得到多个词汇。

[0070] 步骤 602：对得到的多个词汇进行过滤，得到符合过滤条件的多个词汇。

[0071] 步骤 603：针对符合过滤条件的每一个词汇，查询该词汇在各类目上的归一化词频值。其中，预先存有所有词汇在各类目上的归一化词频值。

[0072] 步骤 604：针对每个类目，将得到的多个词汇在该类目上的归一化词频值进行累加，得到归一化词汇向量和。具体实现可参见公式 6。

[0073] 步骤 605：将文本分入归一化词汇向量和的最大分量对应的类目。

[0074] 用于实现本申请实施例的软件可以存储于软盘、硬盘、光盘和闪存等存储介质。

[0075] 本申请实施例对 VSM 进行改进，预先构造一球面空间模型，并基于该球面空间模型对文本进行分类，在分类过程中，计算文本中词汇向量之和与类目向量的距离，从而确定文本应分入的类目。本申请实施例实现了文本分类，并且相对于现有技术中的夹角余弦算法，计算量明显减少。以及本申请实施例中球面空间模型以单位长度为半径，则一个词汇在

各类目上的归一化后的归一化词汇向量的平方和也为单位长度，相当于将一个词汇的语义信息量等价为单位长度，对语义信息量进行了约束，因此相对于现有技术可提高文本分类的准确度。

[0076] 有了较准确的文本分类，有利于提高文本分类存储和文本分类检索（或搜索）的准确度。

[0077] 显然，本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的精神和范围。这样，倘若对本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内，则本申请也意图包含这些改动和变型在内。

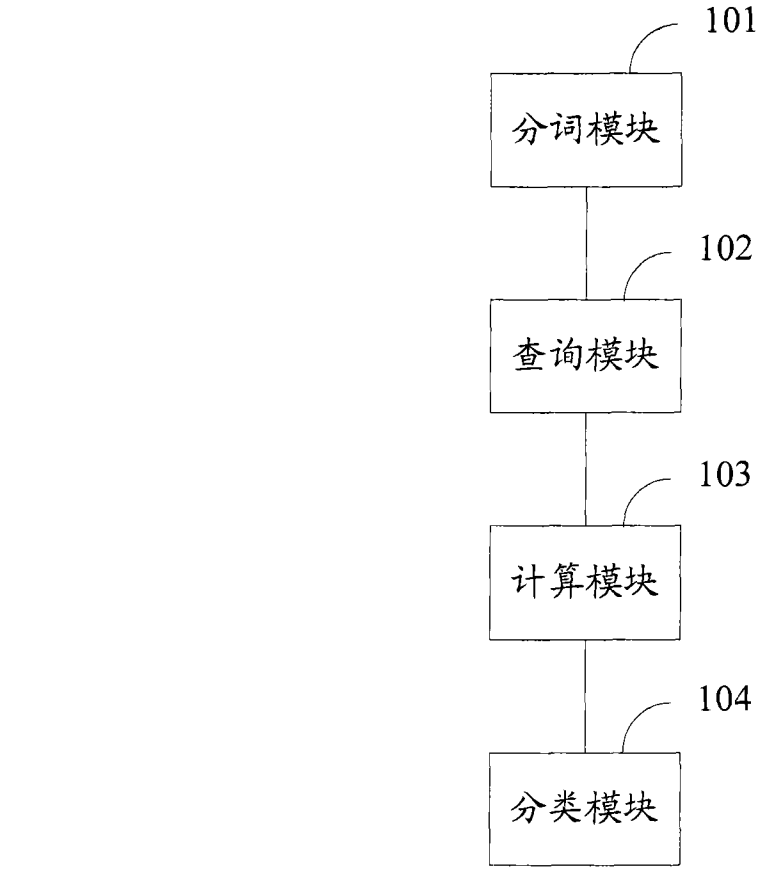

101
分词模块
102
查询模块
103
计算模块
104
分类模块

图 1

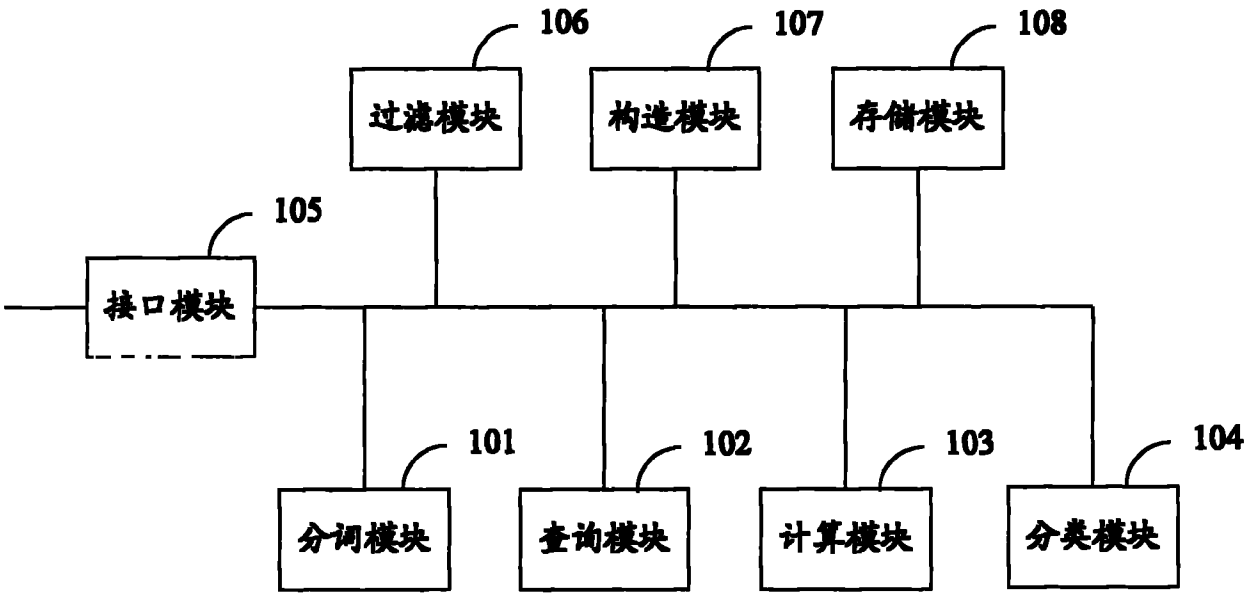

106
过滤模块
107
构造模块
108
存储模块
105
接口模块
101
分词模块
102
查询模块
103
计算模块
104
分类模块

图 2

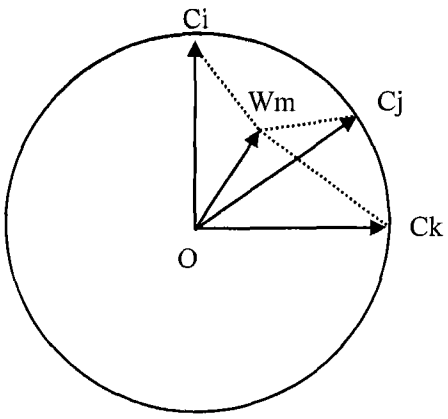

Ci
Wm
Cj
Ck
O

图 3

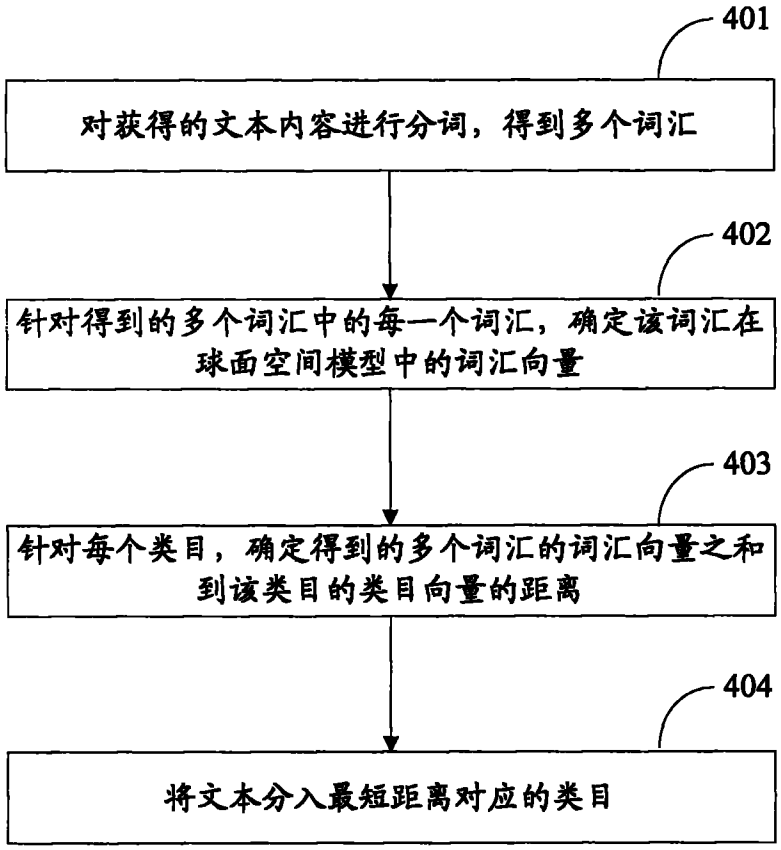

401
对获得的文本内容进行分词，得到多个词汇
402
针对得到的多个词汇中的每一个词汇，确定该词汇在球面空间模型中的词汇向量
403
针对每个类目，确定得到的多个词汇的词汇向量之和到该类目的类目向量的距离
404
将文本分入最短距离对应的类目

图 4

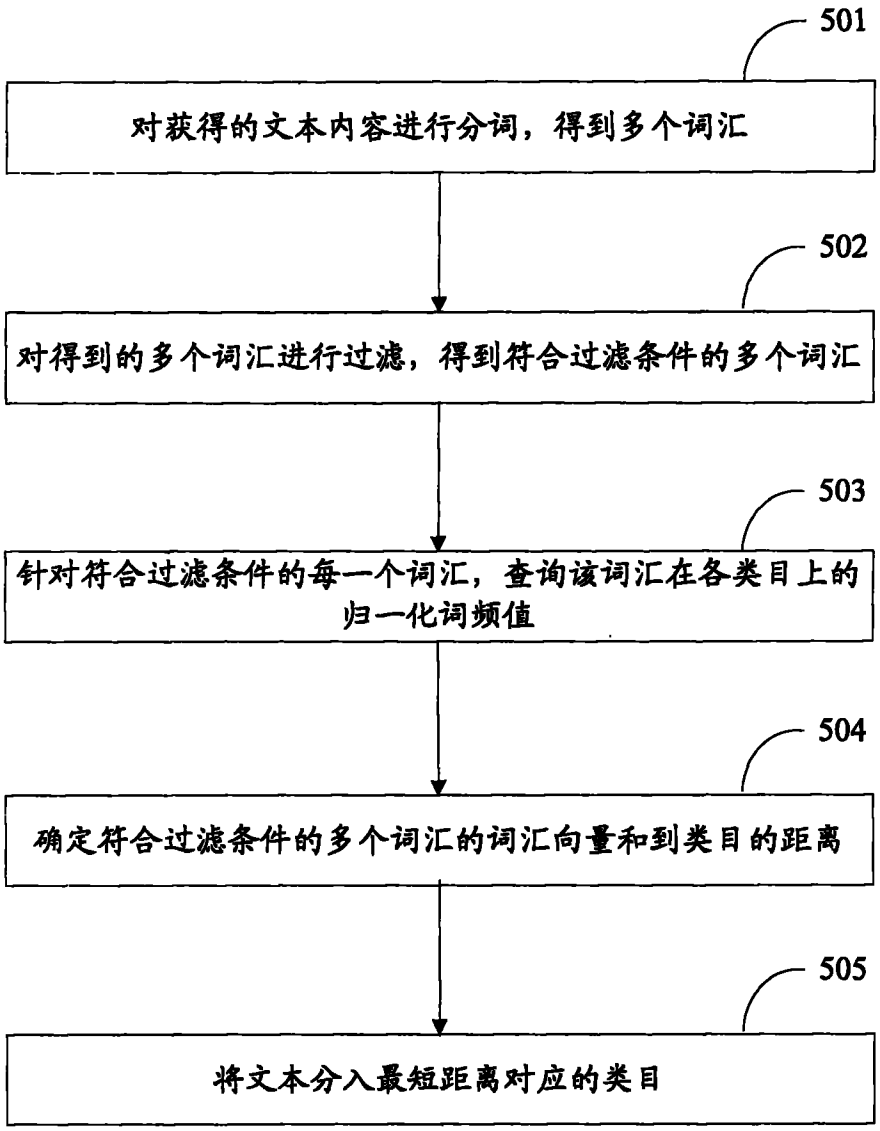

501
对获得的文本内容进行分词，得到多个词汇
502
对得到的多个词汇进行过滤，得到符合过滤条件的多个词汇
503
针对符合过滤条件的每一个词汇，查询该词汇在各类目上的归一化词频值
504
确定符合过滤条件的多个词汇的词汇向量和到类目的距离
505
将文本分入最短距离对应的类目

图 5

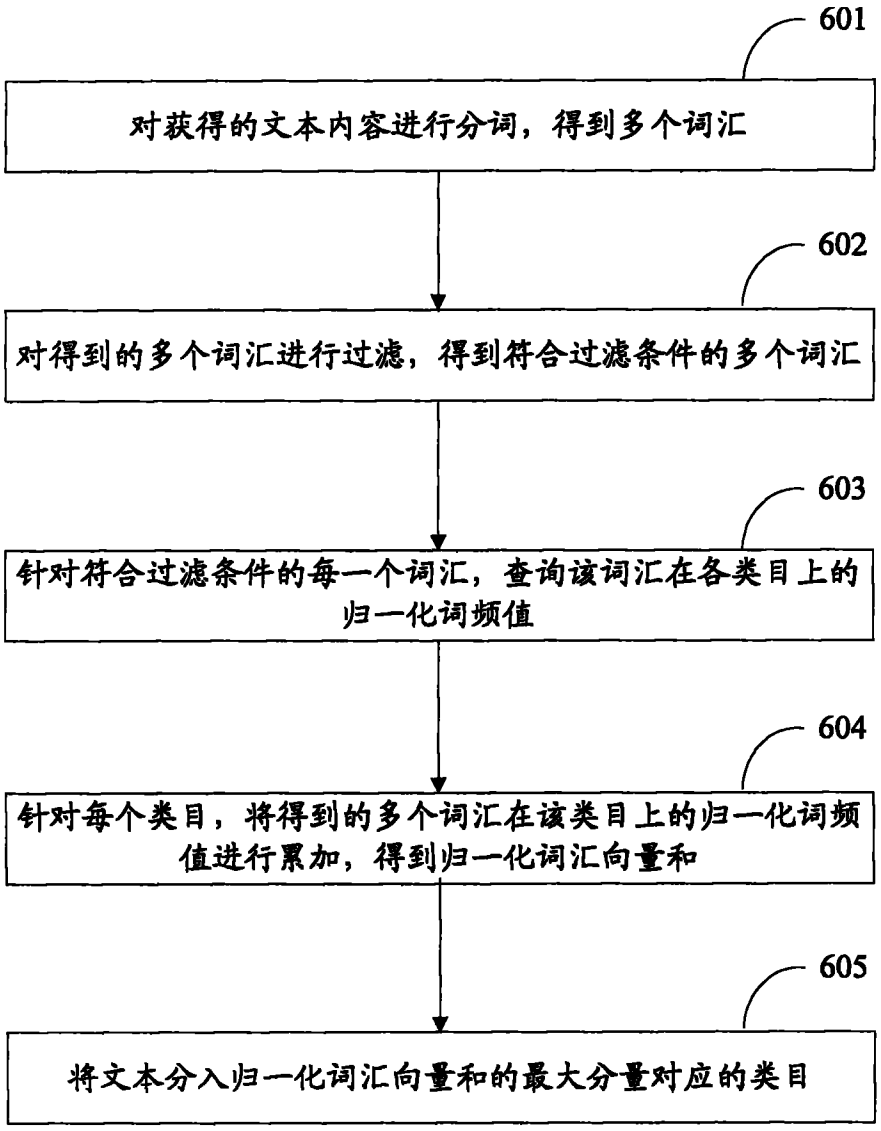

601
对获得的文本内容进行分词，得到多个词汇
602
对得到的多个词汇进行过滤，得到符合过滤条件的多个词汇
603
针对符合过滤条件的每一个词汇，查询该词汇在各类目上的归一化词频值
604
针对每个类目，将得到的多个词汇在该类目上的归一化词频值进行累加，得到归一化词汇向量和
605
将文本分入归一化词汇向量和的最大分量对应的类目

图 6