[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 7/08 (2006.01)
G06F 7/38 (2006.01)

[12] 发明专利说明书

专利号 ZL 03106814.6

[45] 授权公告日 2008 年 6 月 25 日

[11] 授权公告号 CN 100397332C

[22] 申请日 2003.3.3 [21] 申请号 03106814.6

[30] 优先权

[32] 2002. 3. 1 [33] JP [31] 56238/02

[73] 专利权人 惠普开发有限公司

地址 美国德克萨斯州

[72] 发明人 T·卡瓦塔尼

[56] 参考文献

US5671333A 1997.9.23

US6192360B1 2001.2.20

CN1310825A 2001.8.29

JP2000 - 194723A 2000.7.14

CN1363899A 2002.8.14

WO97/33250A1 1997.9.12

JP11 - 53394A 1999.2.26

CN1158460A 1997.9.3

Data Compression and Local Metrics for NearestNerghbor Classification. Francesco Ricci and Paolo Avesani. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, Vol. 21 No. 4. 1999

Discriminant Adaptive Nearest Neighbor Classification. Trevor Hastie and Robert Tibshirani. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, Vol. 18 No. 6. 1996

审查员 冯丹琼

[74] 专利代理机构 中国专利代理(香港)有限公司

代理人 吴立明 陈 霁

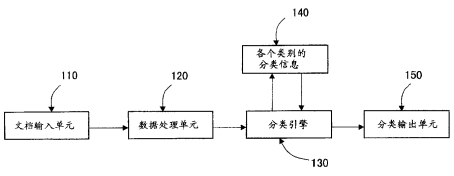
权利要求书 7 页 说明书 11 页 附图 7 页

[54] 发明名称

文档分类方法和设备

[57] 摘要

通过从文档里出现的项中选择分类中所用的项来把文档归类到至少一种文档类别。 用为每个文档类别保存的信息来计算输入文档和各个类别之间的相似度。 然后修正计算出来的对各个类别的相似度。 最的根据对各个类别的修正相似度确定输入文档所属的类别。



1.一种把给定的输入文档归类到至少一种文档类别的方法，该方法包括以下步骤：

(a)从输入文档中存在的项中选择用于分类的项；

(b)把输入文档分成预定单元的文档段；

(c)产生文档段向量，其分量是与在文档段中出现的选中的项的频率有关的数值，还产生文档向量，其中所有的文档段向量都被加在一起；

(d)用为每个文档类别保存的信息计算输入文档和每个类别之间的相似度；

(e)修正到每个类别的相似度；以及

(f)依照到每个类别的修正相似度确定输入文档所属的类别，

其中到每个类别的相似度是通过把为每个文档类别所保存的至少一个正主题差异因子向量和各自的文档段向量之间的点积平方的加权和加到输入文档到每个类别的相似度进行修正的；并且

通过从对每个类别的相似度减去为每个文档类别保存的至少一个负主题差异因子向量和各自的文档段向量之间的点积平方的加权和来对相似度进行进一步的修正。

2.如权利要求1限定的方法，其中用于修正相似度的每个类别的正负主题差异因子向量是通过以下步骤确定的：

(a)计算给定的练习文档集合中包括的练习文档和单个类别之间的相似度，并对练习文档进行分类；

(b)在对练习文档集合的分类结果的基础上找到一组竞争文档，其中每个文档的相似度都超过为各个类别所选的阈值，不管它还属于另一类别；

(c)找到每个类别的正主题差异因子向量作为最大化一个分数的投影轴，该分数的分子是在属于相干类别的所有或选中的文档的文档段向量被投影到该投影轴上时获得的投影值的平方和，分母是在相干类别的竞争文档的文档段向量被投影到该投影轴时获得的投影值的平方和；以及

(d)找到每个类别的负主题差异因子向量作为最大化一个分数的投影轴，该分数的分母是当属于相干类别的所有或选中的文档的文档

段向量被投影到该投影轴时获得的投影值的平方和，分子是当相干类别的竞争文档的文档段向量被投影到该投影轴时获得的投影值的平方和。

3.如权利要求1限定的方法，其中文档段向量和文档向量是通过由它们各自的标准对它们进行分割而实现规范化的。

4.如权利要求1限定的方法，其中每个正或负主题差异因子向量和文档段向量之间的点积平方的加权和是通过由文档段中包括的项的数量对它们进行分割而实现规范化的。

5.如权利要求1限定的方法，其中每个正或负主题差异因子向量和文档段向量之间的点积平方的加权和是通过由输入文档中包含的文档段的数量对它们进行分割而实现规范化的。

6.一种用来把给定的输入文档归类到至少一个预先定义的文档类别的设备，该设备拥有文档输入单元、数据处理单元、分类引擎、分类信息单元和分类输出单元，该设备包括:

(a)选择器，从输入到文档输入单元的输入文档中出现的项中选择用于分类的项;

(b)分割器，把输入文档分割成预定单元的文档段;

(c)向量发生器，产生文档段向量，它的分量是与在文档段中出现的所选的项的频率有关的数值，并产生文档向量，其中文档段向量被加在一起;

(d)第一计算器，用预先为每个文档类别存储的信息计算输入文档和各个类别之间的相似度;

(e)加法器，把预先为每个文档类别存储的至少一个正主题差异因子向量和各自的文档段向量之间的点积平方的加权和加到输入文档到各个类别的相似度上;

(f)减法器，从各个类别的相似度减去预先为每个文档类别存储的至少一个负主题差异因子向量和各自的文档段向量之间的点积平方的加权和;

(g)判断器，根据对各个类别的修正相似度确定并输出输入文档所属的类别。

7.如权利要求6限定的设备，其中用于修正相似度的每种类别的正或负主题差异因子向量是由以下部件确定的:

(a)第二计算器，用于计算给定的练习文档集合中所包括的练习文档与各个类别之间的相似度，并对练习文档归类；

(b)第一探测器，用于在练习文档集合的分类结果的基础上找到一组竞争文档，其中的每个文档具有到各个类别的超过预定的阈值的相似度，而不管它们还属于其它类别；

(c)第二探测器，用于找到各个类别的正主题差异因子向量作为最大化一个分数的投影轴，该分数的分子是在属于相干类别的所有或选中的文档的文档段向量被投影到该投影轴上时获得的投影值的平方和，其分母是在相干类别的竞争文档的文档段向量被投影到投影轴上时获得的投影值的平方和；

(d)第三探测器，用于找到各个类别的负主题差异因子向量作为最大化一个分数的投影轴，该分数的分母是在属于相干类别的所有或选中的文档的文档段向量被投影到该投影轴上时获得的投影值的平方和，其分子是在相干类别的竞争文档的文档段向量被投影到投影轴上时获得的投影值的平方和。

8.一种把给定输入文档归类到至少一种文档类别的文档分类方法，该方法包括以下步骤：

(a)从输入文档里出现的项中选择用于分类的项；

(b)用预先为每种文档类别保存的信息计算输入文档和各个类别之间的相似度；

(c)修正计算出来的相似度；

(d)根据对各个类别的修正相似度确定输入文档所属的类别，

其中计算出来的相似度是通过以下步骤进行修正的：

把预先为每种文档类别保存的至少一个正主题差异因子向量和各自的文档段向量之间的点积平方的加权和加到输入文档对各个类别的相似度；

从对各个类别的相似度减去预先为每个类别保存的至少一个负主题差异因子向量和各自的文档段向量之间的点积平方的加权和。

9.如权利要求8限定的方法，其中用于修正相似度的每种类别的正负主题差异因子向量是由以下步骤决定的：

(a)计算给定的练习文档集合中包括的练习文档和单个类别之间的相似度，并对练习文档进行分类；

(b)在对练习文档集合的分类结果的基础上找到一组竞争文档，其中每个文档具有超过到各个类别的所选的一个阈值的相似度，不管它还属于其它类别；

(c)找到每个类别的正主题差异因子向量作为最大化一个分数的投影轴，该分数的分子是在属于相干类别的所有或选中的文档的文档段向量被投影到该投影轴上时获得的投影值的平方和，分母是在相干类别的竞争文档的文档段向量被投影到该投影轴时获得的投影值的平方和；以及

(d)找到每个类别的负主题差异因子向量作为最大化一个分数的投影轴，该分数的分母是当属于相干类别的所有或选中的文档的文档段向量被投影到该投影轴时获得的投影值的平方和，分子是当相干类别的竞争文档的文档段向量被投影到该投影轴时获得的投影值的平方和。

10.如权利要求 8 限定的文档分类方法，其中文档段向量和文档向量是通过用它们各自的标准对其进行分割而实现规范化的。

11.如权利要求 9 限定的方法，其中文档段向量和文档向量是通过用它们各自的标准对其进行分割而实现规范化的。

12.如权利要求 8 限定的文档分类方法，其中每个正或负主题差异因子向量和文档段向量之间的点积平方的加权和是通过用文档段中所包括的项的个数对其进行分割而实现规范化的。

13.如权利要求 8 限定的文档分类方法，其中每个正或负主题差异因子向量和文档段向量之间的点积平方的加权和是通过用输入文档中包括的文档段的个数对其进行分割而实现规范化的。

14.一种用来把给定的输入文档归类到至少一个预先定义的文档类别的设备，该设备拥有文档输入单元、数据处理单元、分类引擎、分类信息单元和分类输出单元，该设备包括：

(a)选择器，从输入到文档输入单元中的输入文档中出现的项中选择用于分类的项；

(b)第一计算器，用预先为每种文档类别保存的信息计算输入文档和各个类别之间的相似度；

(c)修正器，用来修正相似度；

(d)判断器，根据对各个类别的修正相似度确定并输出输入文档

所属的类别，

其中修正器包括：

加法器，把预先为每种文档类别保存的至少一个正主题差异因子向量和各自的文档段向量之间的点积平方的加权和加到输入文档到各个类别的相似度；

减法器，从输入文档到各个类别的相似度减去预先为每个类别保存的至少一个负主题差异因子向量和各自的文档段向量之间的点积平方的加权和。

15.如权利要求14限定的设备，还包括第二计算器用来计算修正相似度中所用的各个类别的正负主题差异因子向量，该第二计算器包括：

(a)第三计算器，用来计算给定的练习文档集合中包括的练习文档和单个类别之间的相似度，并对练习文档进行分类；

(b)第一探测器，用来在对练习文档集合的分类结果的基础上找到一组竞争文档，其中每个文档具有超过到各个类别预定的一个阈值的相似度，不管它还属于其它类别；

(c)第二探测器，用来找到每个类别的正主题差异因子向量作为最大化一个分数的投影轴，该分数的分子是在属于相干类别的所有或选中的文档的文档段向量被投影到该投影轴上时获得的投影值的平方和，分母是在相干类别的竞争文档的文档段向量被投影到该投影轴时获得的投影值的平方和；以及

(d)第三探测器，找到每个类别的负主题差异因子向量作为最大化一个分数的投影轴，该分数的分母是当属于相干类别的所有或选中的文档的文档段向量被投影到该投影轴时获得的投影值的平方和，分子是当相干类别的竞争文档的文档段向量被投影到该投影轴时获得的投影值的平方和。

16.一种用于把给定文档归类到至少一种文档类别的设备，该设备包括下列步骤的处理装置：

从输入文档中存在的项中选择用于分类的项；

把输入文档分成预定单元中的文档段；

产生文档段向量，其分量是与在文档段中出现的选中的项的频率有关的数值，还产生文档向量，其中所有的文档段向量都被加在一

起;

用为每个文档类别保存的信息计算输入文档和每个类别之间的相似度;

修正到每个类别的相似度;以及

依照到每个类别的修正相似度确定输入文档所属的类别,

所述设备还包括通过把为每种文档类别保存的至少一个正主题差异因子向量和各自的文档段向量之间的点积平方的加权和加到输入文档到各个类别的相似度来修正对各个类别的相似度;

还包括通过从输入文档到各个类别的相似度减去为每个文档类别保存的至少一个负主题差异因子向量和各自的文档段向量之间的点积平方的加权和来进一步修正相似度。

17.如权利要求16限定的设备,还包括用以下步骤确定修正相似度中所用的各个类别的正负主题差异因子向量:

计算给定的练习文档集合中包括的练习文档和单个类别之间的相似度,并对练习文档进行分类;

在对练习文档集合的分类结果的基础上找到一组竞争文档,其中每个文档具有超过到各个类别所选的一个阈值的相似度,不管它还属于其它类别;

找到每个类别的正主题差异因子向量作为最大化一个分数的投影轴,该分数的分子是在属于相干类别的所有或选中的文档的文档段向量被投影到该投影轴上时获得的投影值的平方和,分母是在相干类别的竞争文档的文档段向量被投影到该投影轴时获得的投影值的平方和;以及

找到每个类别的负主题差异因子向量作为最大化一个分数的投影轴,该分数的分母是当属于相干类别的所有或选中的文档的文档段向量被投影到该投影轴时获得的投影值的平方和,分子是当相干类别的竞争文档的文档段向量被投影到该投影轴时获得的投影值的平方和。

18.如权利要求16限定的设备,还包括通过用文档段向量和文档向量各自的标准对它们进行分割来实现对它们的规范化。

19.如权利要求17限定的设备,还包括通过用文档段向量和文档向量各自的标准对它们进行分割来实现对它们的规范化。

20.如权利要求16限定的设备，还包括通过用文档段中包括的项的个数对每个正负主题差异因子向量和文档段向量之间的点积平方的加权和进行分割来实现对它们的规范化。

21.如权利要求16限定的设备，还包括通过用输入文档中包括的文档段的个数对每个正负主题差异因子向量和文档段向量之间的点积平方的加权和进行分割来实现对它们的规范化。

文档分类方法和设备

技术领域

本发明与包括文档分类的自然语言处理有关。更准确地说，本发明允许人们精确地提取文档集合之间的区别，由此提高处理性能。

背景技术

文档分类是把文档归类到预定组的一种技术，并且在信息流通中变得更重要了，且有上升趋势。关于文档分类，至今已经研究并开发了多种方法，例如向量空间模型、k 最近邻居方法（KNN 方法）、自然贝叶斯方法、决策树方法、支持向量机方法以及增压方法。M. Nagata 和 H. Hira 已经在“文本分类—识别理论的样本对”中详细描述了文档分类最近的趋势，该文被收集在日本信息处理会议录 42 卷 1 号（2001 年 1 月）中。在任何分类方法中，以任意形式描述文档类别上的信息并把它和输入文档进行比较。下面应该称其为“类别模型”。例如，类别模型在向量空间模型中由属于每个类别的文档的平均向量来表示，在 KNN 方法中由属于每种类别的文档的向量组表示，在增压方法中由一组简单的假定来表示。为了实现精确的分类，类别模型必须精确地描述每个类别。也可以说，在至今提出的高性能分类方法中，类别模型更精确地描述每个类别。

在这点上，虽然很多分类方法针对类别模型的描述精度，但它们没有考虑类别模型重叠。例如，在向量空间模型或 KNN 方法中，一个特定类别的模型还包括与另一类别匹配的信息。如果在类别模型间发生重叠，它就有可能存在于特定的输入文档和该输入文档并不相属的类别之间并且会导致错误分类。为了消除错误分类的起因，类型模型需要通过找到每种类别的与众不同的信息来描述以减少类别模型的重叠。

发明内容

鉴于上述原因，依照本发明提供一种方法用于提取在每个给定类别中出现但很少在任意其它类别中出现的特征，以及出现在任意其它类别中但很少出现在给定类别中的特征。分类方案包括两个阶段，构

造主分类器和子分类器以有效地使用这些特征。在主分类方案中，采用了一种现有的高性能分类方法，同时在子分类方案中使用这些特征。假定主分类方案是以输入文档和各个类别之间的相似度为基础来对输入文档进行分类。

如下所述，使用所有带有指示各个单独文档的类别的标记的练习文档来提取在子分类方案中所用的特征。首先，在主分类方案中，针对每个练习文档为各个类别获取相似度。与一个相干类别之间的相似度超出预设阈值的文档被确定为属于该相干类别。这些文档被分到两个集合中，在第一个集合中文档被正确地归类为它们的正确类别（以下称为“给定类别文档集合”），在第二个集合中文档被归类到给定的类别文档集合中而不管它们还属于其它类别（以下称为“竞争文档集合”）。每个文档由一组句子向量表示。句子向量的每个成分是在相干句子中出现的每个项的频率或者与频率对应的一个量，而其维度是在所有练习文档中出现的项的种类的数量或者所选的项的种类的数量。假定所有文档的所有句子向量都被投影到一个特定的投影轴上。优先采用来自给定类别文档集合的投影值的平方和和来自竞争文档集合的投影值平方和的比值作为指示集合间的差异程度的判别函数。使用使最大化判别函数的投影轴提取用在子分类方案中的特征。

多个这样的投影轴可以表示为普遍的特征值问题的特征向量。更准确地说，当判别函数由（来自给定类别文档集合的投影值的平方和）/（来自竞争文档集合的特征值的平方和）表示时，最大化判别函数的投影轴有一个较大的值作为来自给定类别文档集合的投影值的平方和，并且有一个较小的值作为来自竞争文档集合的投影值的平方和。因此，投影轴反映很少在任意竞争文档中出现但经常出现在给定类别中的信息。因此，这样的投影轴可以称为“正主题差异因子向量”。相反地，当判别函数由（来自竞争文档集合的特征值的平方和）/（来自给定类别文档集合的投影值的平方和）表示时，最大化判别函数的投影轴反映很少出现在给定类别中但经常出现在任意竞争文档中的信息。因此，这样的投影轴被称为“负主题差异因子微量”。

在子分类方案中，把输入文档的句子向量和每个类别的一定数量的正主题差异因子向量之间的点积平方的加权和加到在主分类方案中获得的相干类别的相似度中。从相干类别的相似度中减去输入文档的

句子向量和每个类别的一定数量的负主题差异因子向量之间的点积平方的加权和。把这样修正的相似度和每个类别的预定阈值进行比较。

如前所述，在本发明中，由子分类方案修正主分类方案计算出的相似度。如果由子分类方案在一个特定类别中计算输入文档的句子向量和一定数量的正主题差异因子向量之间的点积平方的加权和，正主题差异因子向量就规定该类别中存在的特征。因此，如果输入文档属于该相干类别，上面的加权和通常有大的值，而且相似度也被修正为一个大的值。另一方面，如果输入文档不属于该相干类别，上述加权和通常具有小的值，而且相似度变化也小。此外，如果在该特定类别中计算输入文档的句子向量和一定数量的负主题差异因子向量之间的点积平方的加权和，负主题差异因子向量就规定不应该存在于该类别中的特征。因此，如果输入文档属于该相干类别，上述加权和通常具有小的值而相似度被修正为一个小的值。然而，当输入文档不属于该相干类别时，上述加权和通常具有大的值而相似度被修正为一个小的值。既然用这种方式修正相似度，修正通常会导致对扩大输入文档所属的类别的相似度并减小对输入文档不相属的类别的相似度。因而提高了分类精度。

附图说明

如果结合附图读对示例实施方案的下列详细描述以及权利要求能够明白前面的描述并更好地理解本发明，所有这些都构成了本发明公开的一部分。尽管前面和后面的书面并带有插图的发明公开集中于公开本发明的示例实施方案，但应该清楚地理解只通过图解和示例也是一样的，而且并不是为了把本发明限制在那里。仅由所附权利要求的各项限制本发明的精神和范围。

附图的简短描述表示如下，其中：

图 1 是显示依照本发明的一种实施方案的文档分类设备的框图；

图 2 是本发明的一种实施方案的流程图；

图 3A-3C 是解释文档向量的图示；

图 4 是依照 KNN 方法计算输入文档的相似度（图 2 中的步骤 14）的步骤的流程图；

图 5 是获取正负主题差异因子向量以便修正相似度的步骤的流程图，这些步骤使用给定类别的文档集合以及被错误归类到该给定类别

的文档集合或者可能被错误归类到其中的文档集合;

图 6A-6C 属于类别 1 的文档的结构的图示;

图 7 是分类步骤的流程图（图 5 的步骤 22）。

具体实施方式

在开始对主题发明的详细描述之前，先顺序提及下列叙述。在适当时，在区分附图时使用类似的参考数字和字母来指示相同的、对应的或类似的成分。此外，在接下来的详细描述中，给出示例性的大小/模型/数值/范围，尽管本发明并不受限于此。以框图形式显示排列以避免模糊本发明，还由于关于这样的框图排列的实现的细节高度取决于实现本发明的平台，即这样的细节正好在本领域技术人员的范围内。陈述特定的细节，例如电路或流程图，以便描述本发明的示例实施方案，本领域的技术人员应该明白可以在有或没有这些特定细节的变体的情况下实践本发明。最后，应该明白区分硬布线电路和软件指令的组合可以用来实现本发明的实施方案，也就是说，本发明并不局限于硬件和软件的特定实施方案。

图 1 是依照本发明的一种实施方案的文档分类设备的框图。首先，把要分类的文档输入输入单元 110。在数据处理单元 120 中，对输入的文档进行数据处理，例如项提取和文档段提取。在分类引擎 130 中，参考包含各个类别的分类信息的单元 140，由主分类方案计算相似度，由子分类方案修正相似度。用修正后的相似度确定输入文档所属的类别并把它输出到分类输出单元 150。

图 2 是图 1 中设备的处理步骤的流程图，从文档输入执行到类别判定。在输入步骤 11 中向单元 110 提供文档。在步骤 12 中，单元 120 提取并选择项。在步骤 13 中，单元 120 提取文档段向量。在步骤 14 和 15 中，引擎 130 分别执行相似度计算和相似度修正。在步骤 16 中，单元 140 进行类别判定。步骤 11 到 14 对应于主分类方案，而步骤 15 和 16 对应于子分类方案。下面用英文文档描述了一个实例。

首先，在文档输入步骤 11 输入要分类的文档。在项提取和选择步骤 12，从文档中提取单词、公式、一系列符号，等。所有单词和系列符号在后面都称为“项”。在书面英文的情况下，已经建立了一种单词分开书写的标记方法，并且因此使得对项的检测更加容易。在项提取和选择步骤 12，把项包括在一个项列表中用于分类，并且从输入文档

中存在的项之间提取项。用大量标记过的练习文档可以实现对分类中所用的项的选择，而且 tf-idf（项频率—反向文档频率）技术、采用 X^2 统计的方法、采用相互信息的方法等等都是提供有利结果的已知方法的实例。文档段向量提取步骤 13 把文档分成文档片段，并为每个文档段创建一个向量。把文档分割成文档段中最基本的处理是把文档分成句子单元。在书面英语中，句子以句号结束并且后跟一个空格，因此可以很容易地提取句子。其它把文档分割成文档段的方法包括把多个句子收集进文档段中以使文档段的项的数量实际相等的方法，以及从头开始分割文档而不考虑句子以使文档段中所包括的项的数量实际相等的方法，等等。还可以把整个文档作为一个文档段。随后，为每个文档段创建一个向量。向量的分量表示分类中所用的各个项在相干文档段中的频率。换句话说，频率被乘以权重。已经对如何设置权重进行了研究，本领域的技术人员已经知道一些设置权重的有效方法。通过把所有文档段向量加起来而产生的向量称为“文档向量”。下面的描述假定句子向量是文档段向量。当输入由 K 个句子（图 3A）组成的输入文档 X 时，第 k 个句子向量由 x_k（图 3B）表示，文档向量由 x 表示（图 3C）。在图 3B 底部的数字是为了举例说明句子向量的成分。也就是说，这些数字指示对应句子向量 x_k 的各个成分对应的项的频率。

相似度计算步骤 14（图 2）计算输入文档到各个类别 1 的相似度。已知有多种找到相似度的方法。在向量空间模型情况下，用练习主体找到各个类别的平均文档向量并把它们保存起来。假设类别 1 的平均向量为 m_1，输入文档到类别 1 的相似度 sim(X, 1)可以表示为:

$$sim(X,1) = x^T m_1 / (\|x\| \times \|m_1\| \qquad \cdots\cdots(1)$$

这里，$\|x\|$表示 x 的标准，上标 T 代表向量转置。

现在参考图 4 中所示的流程图描述由图 1 的设备执行的 KNN 方法。在 KNN 方法中，假设 Y_t 表示练习文档集合中的第 t 个文档，并且假设 y_t 表示第 t 个文档的文档向量，输入文档 X 到文档 Y_t 的相似度 sim(X, Y_t) 由下面的公式获得:

$$sim(X, Y_t) = x^T y_t / (\|x\| \times \|y_t\| \quad \cdots\cdots(2)$$

在已经获得输入文档 X 对所有练习文档的相似度之后（步骤 142），选择 k 个对输入文档 X 的相似度最大的文档（步骤 144）。此后，根据每个文档所附的标记为每个类别对 k 个选中的文档进行分类（步骤 146）。随后，计算输入文档到类别 1 的相似度 sim（X，1）（步骤 148）。相似度 sim（X，1）定义为输入文档 X 到分类到类别 1 的文档的相似度之和。也就是说，按照如下公下计算相似度 sim（X，1）：

$$sim(X,l) = \sum_{Y_t \in \Omega_l} sim(X, Y_t) \quad \cdots\cdots(3)$$

这里，Ω_l 表示在 k 个文档中属于类别 1 的练习文档的集合。

在相似度修正步骤 15（图 2），用已经为每个类别保存的正主题差异因子向量和负主题差异因子向量修正相似度。用在相似度修正中的类别 1 的正主题差异因子向量由 $\{\alpha_i\}$（i=1，…，L_{ci}）表示，负主题差异因子向量由 $\{\beta_i\}$（i=1，…，L_p）表示。然后，类别 1 的修正相似度由 sim_c（X，1）表示，由下列公式给出：

$$sim_c(X,l) = sim(X,l) + a\sum_{i=1}^{L_c}\sum_{k=1}^{K}(x_k^T\alpha_i)^2 - b\sum_{i=1}^{L_p}\sum_{k=1}^{K}(x_k^T\beta_i)^2 \quad (4)$$

注意 a 和 b 是正数的参数，已经和 L_p、L_c 一起被预先确定。可以确定参数 a、b、L_p、L_c 的值，以便在接连改变各个参数 a、b、L_p、L_c 的值的时候发现不用于向量 $\{\alpha_i\}$ 和 $\{\beta_i\}$ 的计算的文档集合的性能，并选择提供最大 F 度量的数值的组合。F 度量定义如下：

精度=（正确分配到各个文档作为分类结果的总类别数）/（分配到各个文档作为分类结果的总类别数）

重复度=（正确分配到各个文档作为分类结果的总类别数）/（每个文档应该所属的类别总数）

F 度量=精度×重复度×2/（精度+重复度）

修正后的相似度 sim_c（X，1）由下列公式计算

$$sim_c(X,1) = sim(X,1) + \sum_{i=1}^{L_C}\sum_{k=1}^{K} a_i(x_k^T\alpha_i)^2 - \sum_{i=1}^{L_P}\sum_{k=1}^{K} b_i(x_k^T\beta_i)^2 \quad (5)$$

这种情况下，a_i和b_i分别是第 i 个正主题差异因子和第 i 个负主题差异因子的权重。当给定L_P和L_C时，可以通过采用线性判别式分析可获得权重 a_i 和 b_i。更准确地说，为每个未用于向量$\{\alpha_i\}$和$\{\beta_i\}$的计算的文档准备一个 L_P+L_C+1 维的向量，而给定$(x_k^T\alpha_i)$ $(i=1, \ldots, L_C)$、$(x_k^T\beta_i)$ $(i=1, \ldots, L_P)$以及 sim(X,1)作为分量。随后，在类别 1 的文档集合和属于另一类别的文档集合之间进行线性判定式分析，并为各自的分量确定最优分离这两个文档集合的权重。“属于另一类别的文档集合”表示属于另一类别的文档，其中到类别 1 的相似度 sim(X,1)超过了特定阈值，在分类步骤 22（图 5）作为分类结果。通常用线性判别式分析可以发现最优分离两组向量集合的投影轴。计算投影轴以使各个组的平均向量的差额向量被乘以在其中已加上了各个组的协方差矩阵的矩阵的逆矩阵。此后，由 sim(X,1)的权重分割$(x_k^T\alpha_i)$ $(i=1, \ldots, L_C)$和$(x_k^T\beta_i)$ $(i=1, \ldots, L_P)$的权重，由此分别确定 a_i 和 b_i。为所有的L_P和L_C值的组合执行这样的处理，并可以采用提供最佳分类结果的权重 a_i 和 b_i 的值。

在分类判定步骤 16（图 2），通过比较各个类别的预定阈值和修正后的相似度确定输入文档所属的类别。如果类别 1 的修正相似度大于类别 1 的阈值，就确定输入文档属于类别 1。

图 5 是确定用来修在图 2 的步骤 15 修正相似度的正主题差异因子向量和负主题差异因子向量的步骤的流程图。在步骤 21，准备练习文档。在步骤 22，分类发生。在步骤 23，完成文档集合编辑。在步骤 24，实现主题差异向量分析。

在练习文档准备步骤 21，准备用于确定正负主题差异因子向量的练习文档集合，为每个这样的文档获取文档向量和文档段向量。在随后的分类步骤 22，选择每个练习文档作为输入文档以便计算它到所有其它练习文档的相似度并由此确定它所属的类别（图 2 的步骤 14 和 16）。通过执行这样的操作对所有练习文档分类。但在这种情况下不执行图 2 中步骤 15 的相似度修正。

下面参考图 7 的流程图描述图 5 中的分类步骤 22.

步骤 221：为所有练习文档执行项提取和文档段提取这样的数据处

理。

步骤 222：选择一个练习文档作为输入文档。

步骤 223：计算输入文档和其它练习文档之间的相似度以根据公式（3）获取到各自的类别的相似度。

步骤 224：判断是否已经为所有练习文档获得了到各自的类别的相似度。

步骤 225：相似度大于给定类别的阈值的文档被分为包括正确分类文档的文档集合和包括不正确分类文档的竞争文档集合。

现在详细描述图 5 的流程图。M 个文档的集合被正确分类为属于类别 1，由 D（图 6A）表示。假定集合 D 的第 m 个文档 Dm 由 $K_D(m)$ 个句子组成，第 k 个句子向量由 d_{mk} 表示（图 6B）。竞争文档集合编辑步骤 23（图 5）创建竞争文档的集合，每个竞争文档都被错误地归类为类别 1 或者可能被错误地归类到其中，每个类别在分类步骤 22 上以分类结果为基础。通过选择到类别 1 的相似度 sim(X,1) 超过特定阈值来提取类别 1 的任意竞争文档。该阈值可以根据要选择的竞争文档数任意确定。假定类别 1 的竞争文档集合 T 由 N 个文档组成。假定集合 T 的第 n 个文档 Tn 由 $K_T(n)$ 个句子组成，第 k 个句子向量由 t_{nk} 表示（图 6C）。主题差因子分析步骤 24（图 5）使用属于每个类别的文档集合和它的竞争文档集合计算正负主题差异因子向量。用作正主题差异因子向量的投影轴由 α 表示。假定 P_D 和 P_T 分别表示在文档集合 D 和 T 的所有句子向量被投影到坐标轴 α 的情况下投影值的平方和，获取正主题差异因子向量作为最大化判别函数 $J(\alpha)=P_D(\alpha)/P_T(\alpha)$ 的 α。最大化 J(α) 的 α 反映存在于文档集合 D 但很少存在于文档集合 T 中的特征，因为它应该有一个较大的值作为文档集合 D 的句子向量的投影值的平方和，以及有一个较小的值作为文档集合 T 的句子向量的投影值的平方和。这种情况下，$P_D(\alpha)$ 和 $P_T(\alpha)$ 分别表示如下：

$$P_D(\alpha) = \sum_{m=1}^{M} \sum_{k=1}^{Kd(m)} (d_{mk}^T \alpha)^2 = \alpha^T S_D \alpha \quad \cdots\cdots(6)$$

$$S_D = \sum_{m=1}^{M} \sum_{k=1}^{Kd(m)} d_{mk} d_{mk}^T \quad \cdots\cdots(7)$$

$$P_T(\alpha) = \sum_{n=1}^{N} \sum_{k=1}^{Kt(n)} (t_{nk}^T \alpha)^2 = \alpha^T S_T \alpha \quad \cdots\cdots(8)$$

$$S_T = \sum_{n=1}^{N} \sum_{k=1}^{Kt(n)} t_{nk} t_{nk}^T \quad \cdots\cdots(9)$$

因此，判别函数 J(α) 可以写成：

$$J(\alpha) = \frac{P_D(\alpha)}{P_T(\alpha)} = \frac{\alpha^T S_D \alpha}{\alpha^T S_T \alpha} \quad \cdots\cdots(10)$$

由方程（10）给出的最大化判别函数 J(α) 的 α 可以通过对 α 微分方程（10）然后设置结果等于 0 来求得。也就是说，把它当作下列普遍的特征值问题的特征向量：

$$S_D\alpha = \lambda S_T\alpha \quad \cdots\cdots(11)$$

通常可以从公式（11）获得多个特征向量，从它们中间选取的第 1 到第 L_G 个特征向量在图 2 的步骤 15 成为正主题差异因子向量 $\{\alpha_i\}$（$i=1, \ldots, L_{Gi}$）。如果 β 表示要找的其它投影轴，并且 $J(\beta)=P_T(\beta)/P_D(\beta)$ 表示判别函数，那么最大化判别函数 J(β) 的 β 就表示应该存在于文档集合 T 但很少存在于文档集合 D 中的特征。这种情况下，最大化判别函数 J(β) 的 β 被作为下列普遍的特征值问题的特征向量给出，同样对方程（11）：

$$S_T\beta = \lambda S_D\beta \quad \cdots\cdots(12)$$

在从方程（12）获得的多个特征向量之间选取的第 1 到第 L_P 个特征向量在图 2 的步骤 15 变成负主题差异因子向量 $\{\beta_i\}$（$i=1, \ldots, L_P$）。就方程（11）来说，矩阵 S_T 必须是要获取的特征向量的正则矩阵。但实际上在练习文档集合中的句子数小于项的数目时或者特定数量的项对总是一起出现时不可能获取矩阵 S_T 作为正则矩阵。这种情况下，允许通过根据下列方程正则化矩阵 S_T 获得特征向量：

$$\hat{S}_T = S_t + \sigma^2 I \qquad \cdots\cdots(13)$$

其中 σ^2 表示一个参数，I 表示恒等矩阵。在采用方程（13）的情况下，判别函数 J(α) 与如下方程对应：

$$J(\alpha) = Pb(\alpha) / (Pr(\alpha) + \sigma^2 \qquad \cdots\cdots(14)$$

在上述实施方案中，没有考虑文档和句子的长度。因此，即使在不考虑文档长度的情况下已经获得了输入文档到每个类别的相似度，也还存在对相似度的修正量级对较长的文档扩大的更多或者相似度的修正量级受长文档影响较大的问题。因此，在图 2 的步骤 15 可以替换方程（4）为：

$$sim_c(X,l) = sim(X,l) + a\sum_{i=1}^{I_b}\sum_{k=1}^{K}(x_k^T\alpha_i)^2/K - b\sum_{i=1}^{I_r}\sum_{k=1}^{K}(x_k^T\beta_i)^2/K \qquad \cdots\cdots(15)$$

如前所述，K 表示输入文档 X 中的句子的个数。因而，可以减少文档长度的影响。方程（5）同样如此。换句话说，假设 N_k 表示输入文档中第 k 个句子中出现的项的个数，可以替代方程（4）为：

$$sim_c(X,l) = sim(X,l) + a\sum_{i=1}^{I_b}\sum_{k=1}^{K}(x_k^T\alpha_i)^2/N_k - b\sum_{i=1}^{I_r}\sum_{k=1}^{K}(x_k^T\beta_i)^2/N_k \qquad \cdots\cdots(16)$$

因而，可以减少句子长度中偏差的影响。这对方程（5）来说同样正确。

此外，图 3B 中输入文档的句子向量 x_k 可以很好的规范化如下，以便对方程（4）、（5）、（15）和（16）应用规范化的向量：

$$\hat{x}_k = x_k / \|x_k\| \qquad \cdots\cdots(17)$$

通过类似地规范化图 6B 和 6C 中的句子向量 d_{mk} 和 t_{mk} 来获取正负主题差异因子向量。

如上所述，依照本发明，每种类别的与众不同的信息可以用于分类，并因此显著地提高分类的精度。在一个采用路透社-21578 实验中（练习文档的数量是 7770，类别的数量是 87，测试文档的数量是 3019），现有 KNN 方法（其中没有进行证实对本发明的修正）的数据证明精度为 85.93%、重复度为 81.57%，F 度量为 83.69%。相反，根据依照本发明的方程（16）对相似度进行修正可以把精度、重复度和 F 度量分别增加到 90.03%、84.40%和 87.14%。

	精度	重复度	F 度量
现有 KNN 方法	85.93%	81.57%	83.69%
依照本发明的方法	90.03%	84.40%	87.14%

对精度、重复度和 F 度量的定义如前所述，在路透社-21578 中一个文档可以属于多个类别。

这包括对示例实施方案的描述。虽然至此已经参考多个示例性实施方案对本发明进行了描述，但应该理解本领域的技术人员也可以设计多种符合本发明的原理的精神和范围的其它改进和实施方案。更准确地说，在不偏离本发明的精神的前提下在前述发明公开、附图以及所附权利要求的范围中提供的组合配置的部件和/或配置也可以有合理的变体和改进。除了部件和/或配置的变体和改进外，本领域的技术人员还将明白替代的用途。

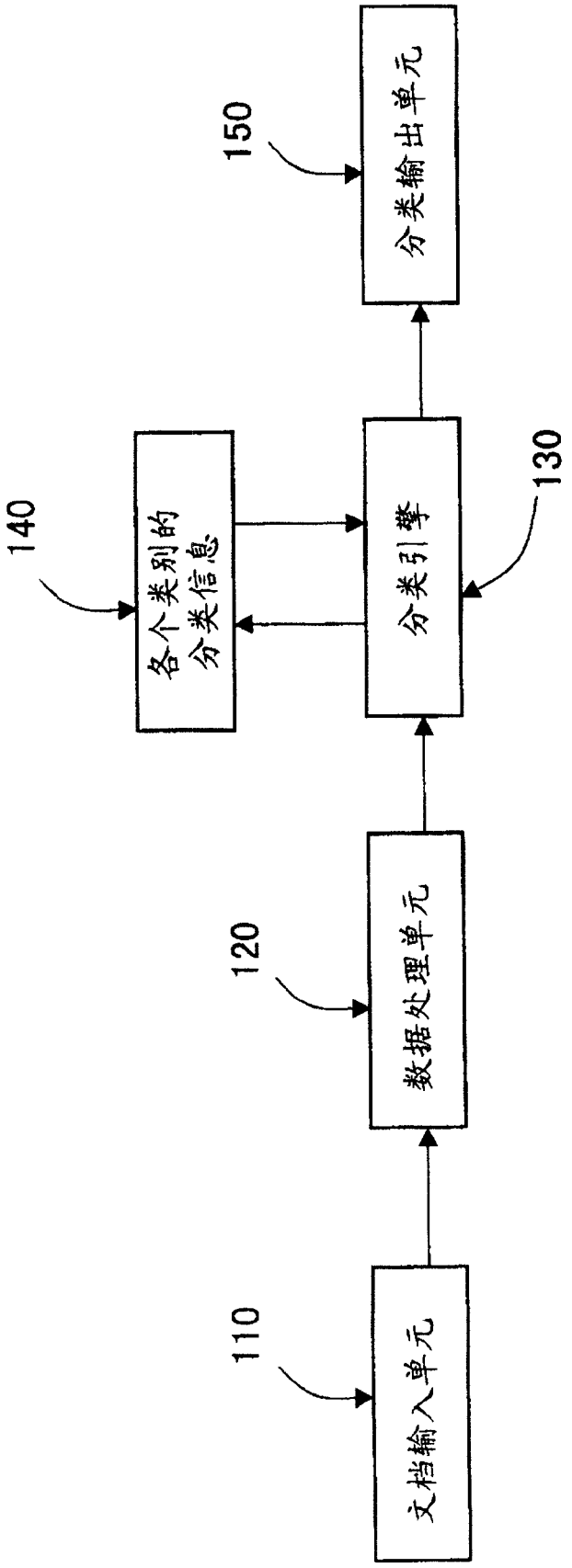


图 1

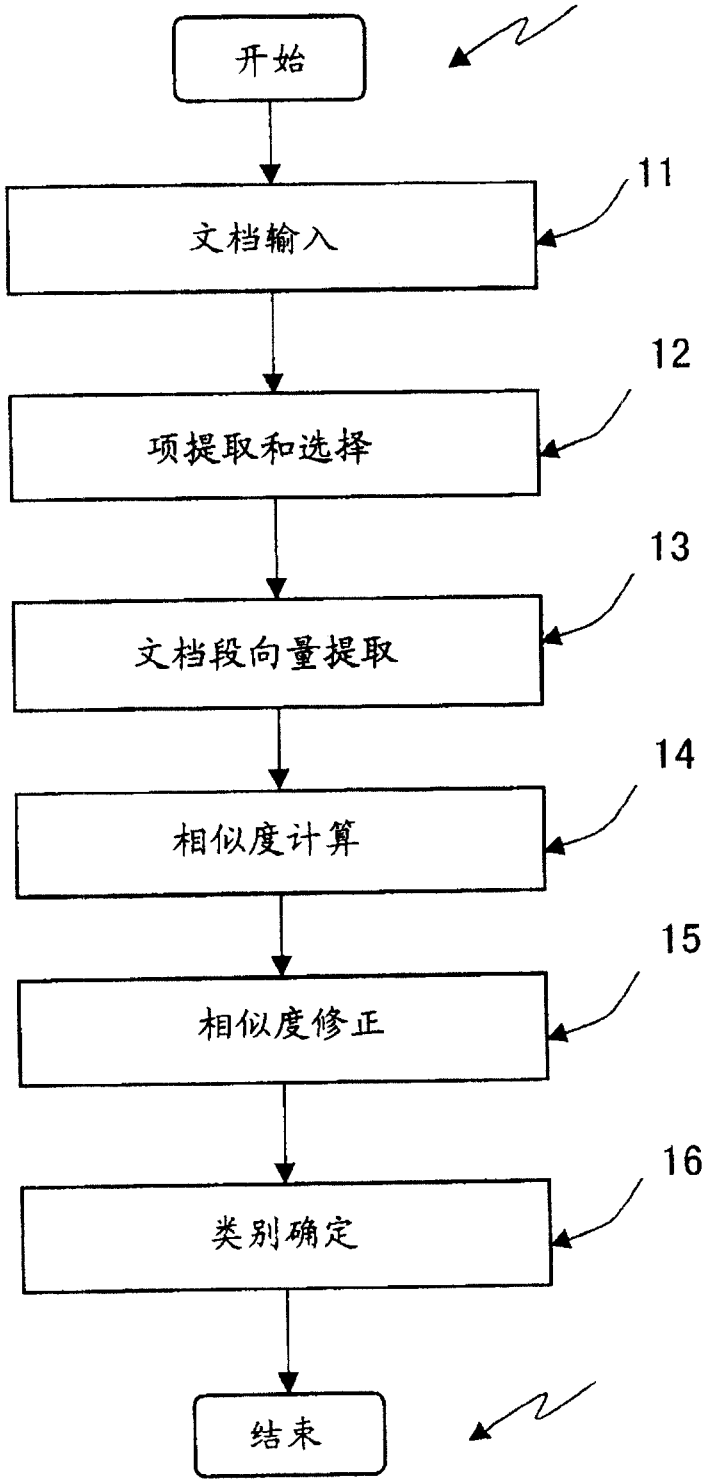


图 2

图 3A

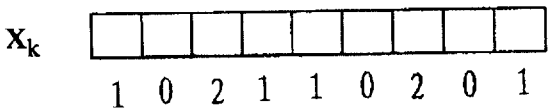


图 3B

$$x = \sum_{k=1}^{K} x_k$$

文档向量

图 3C

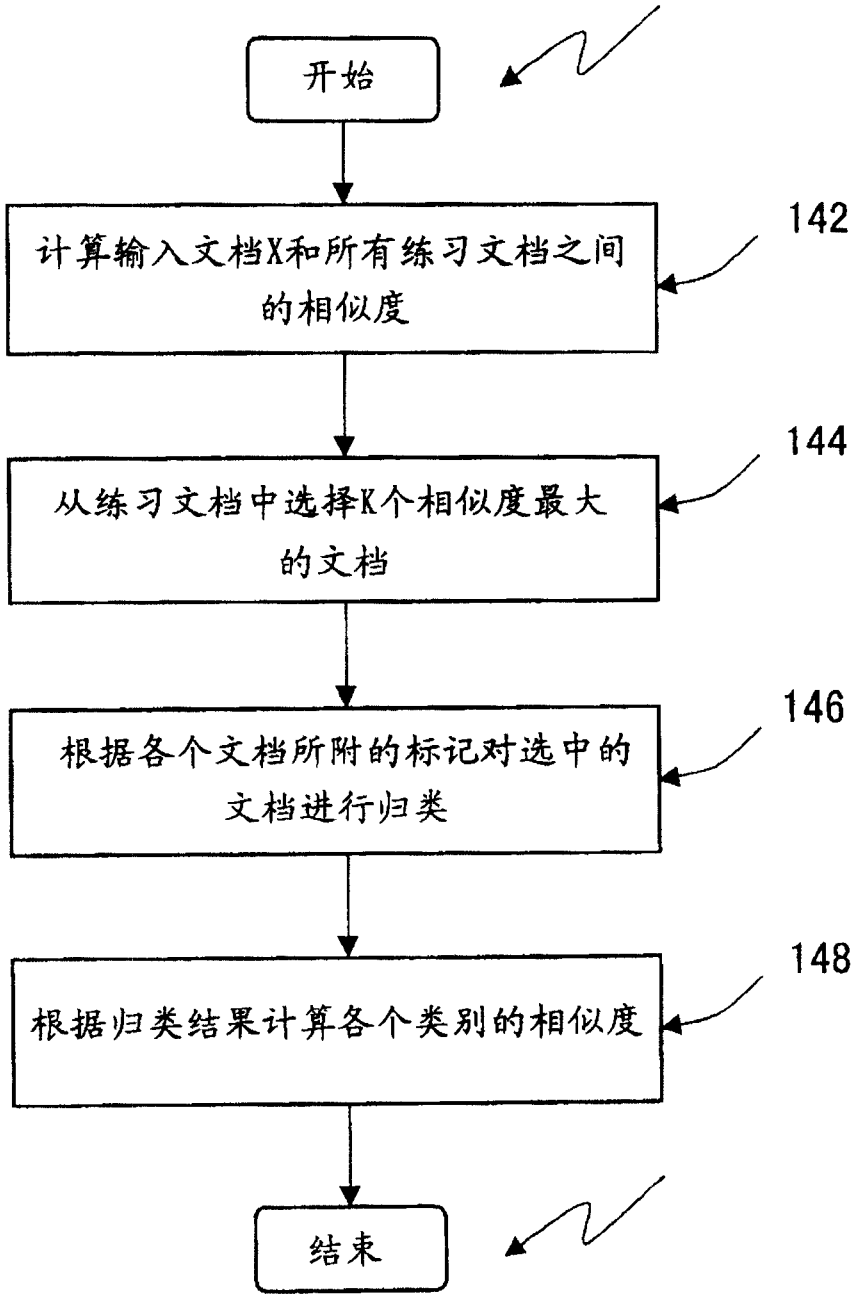

开始
计算输入文档X和所有练习文档之间的相似度
142
从练习文档中选择K个相似度最大的文档
144
根据各个文档所附的标记对选中的文档进行归类
146
根据归类结果计算各个类别的相似度
148
结束

图 4

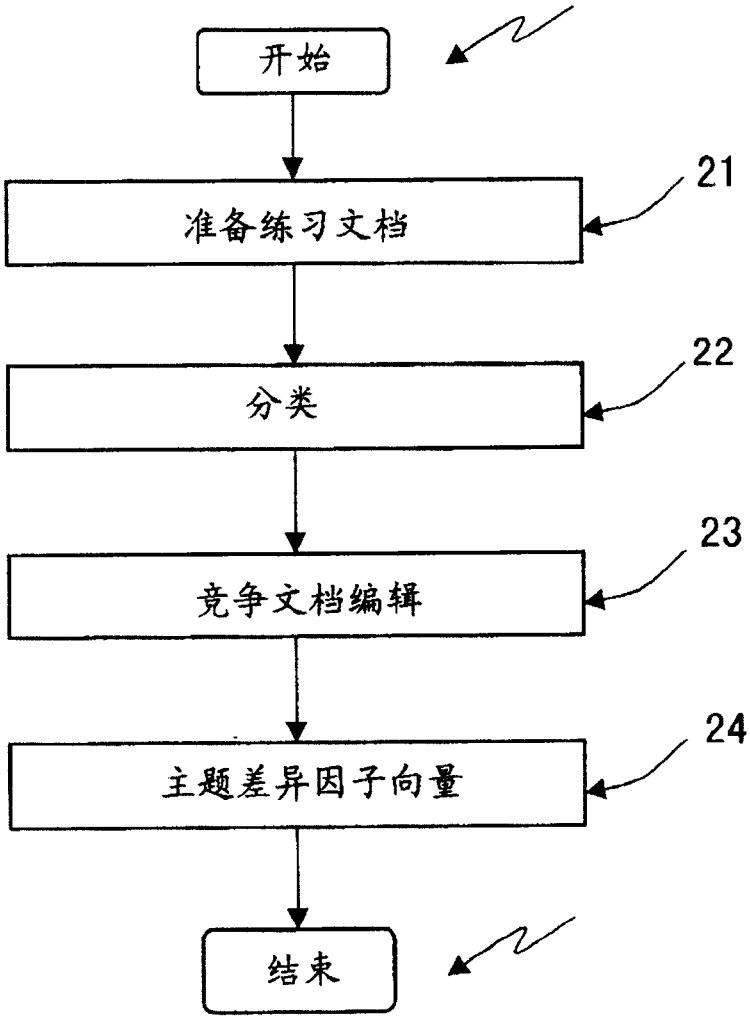


图 5

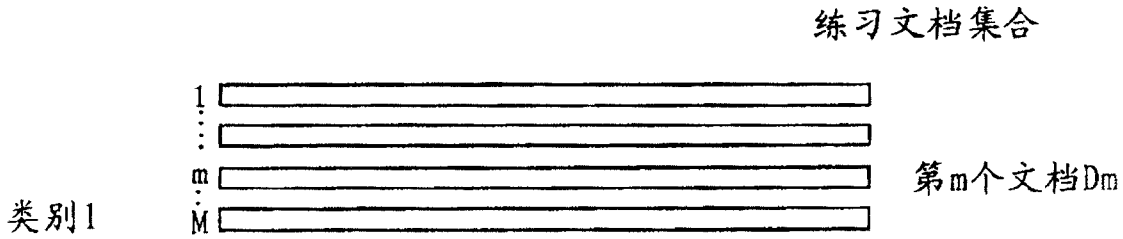

练习文档集合
1
m
M
类别l
第m个文档Dm

图 6A

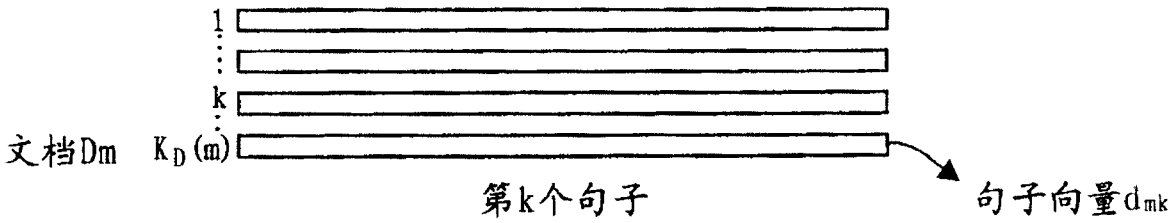

1
k
文档Dm
$K_D(m)$
第k个句子
句子向量d_{mk}

图 6B

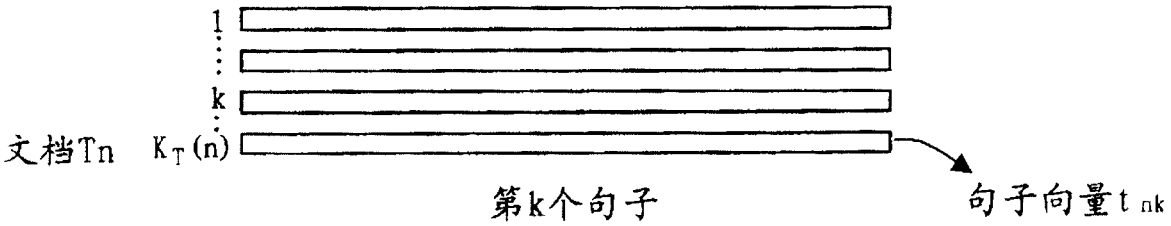

1
k
文档Tn
$K_T(n)$
第k个句子
句子向量t_{nk}

图 6C

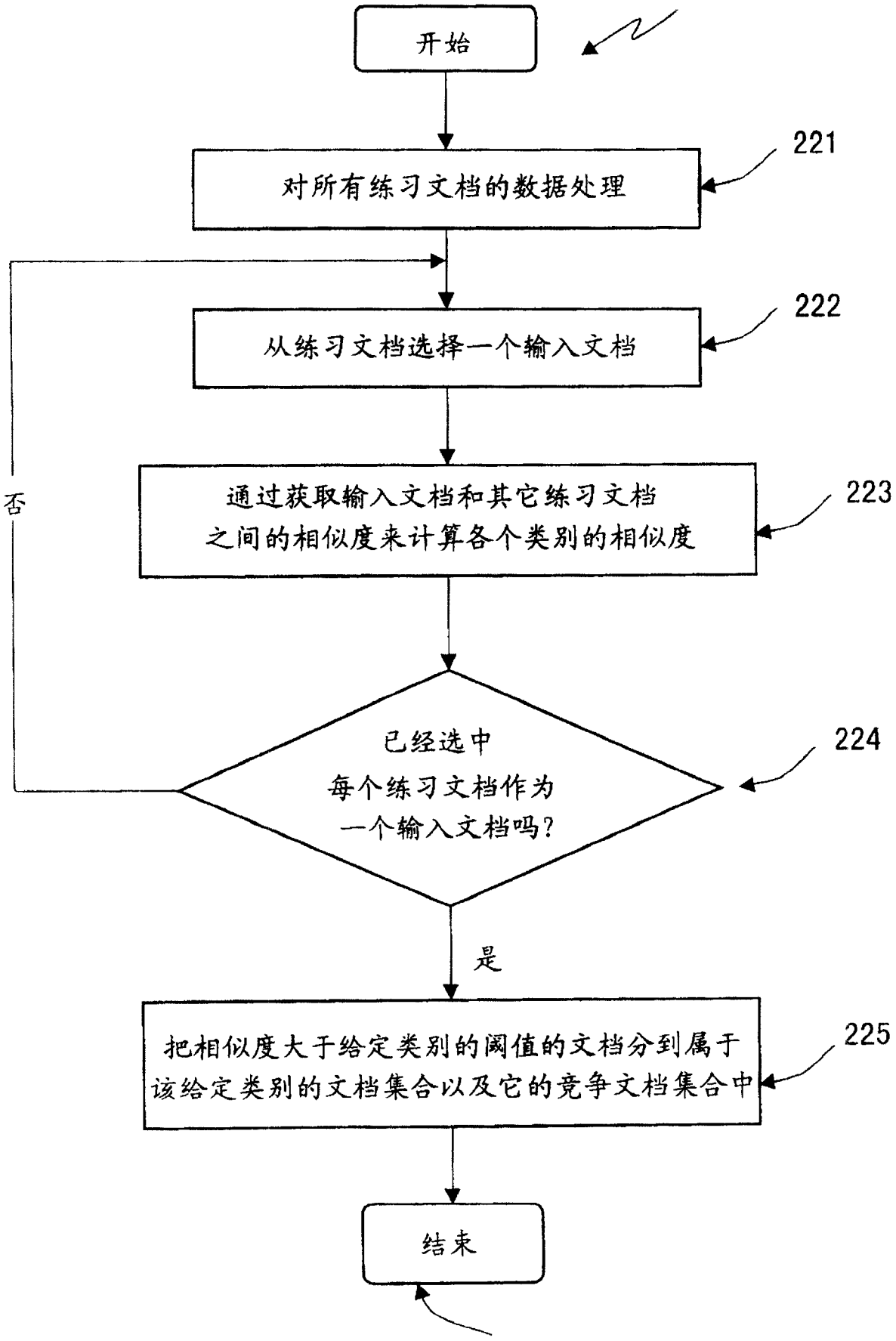

开始
221
对所有练习文档的数据处理
222
从练习文档选择一个输入文档
通过获取输入文档和其它练习文档
之间的相似度来计算各个类别的相似度
223
否
已经选中
每个练习文档作为
一个输入文档吗?
224
是
把相似度大于给定类别的阈值的文档分到属于
该给定类别的文档集合以及它的竞争文档集合中
225
结束

图 7