(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 102799666 A
(43)申请公布日 2012.11.28

(21)申请号 201210244558.5

(22)申请日 2012.07.13

(71)申请人 北京航空航天大学
地址 100191 北京市海淀区学院路 37 号

(72)发明人 袁满 欧阳元新 赵海 熊璋

(74)专利代理机构 北京科迪生专利代理有限责任公司 11251
代理人 杨学明 顾炜

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

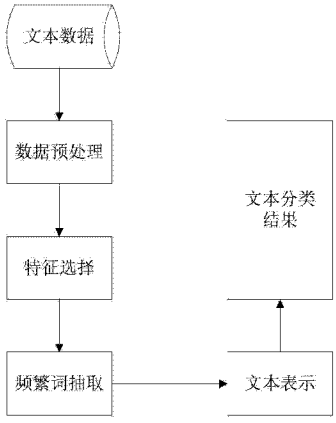
权利要求书 2 页 说明书 4 页 附图 2 页

(54)发明名称

一种基于频繁词集的网络新闻自动文本分类的方法

(57)摘要

本发明公开了一种基于频繁词集的网络新闻自动文本分类的方法，该方法具体为：步骤(1)数据预处理和特征选择；步骤(2)提取频繁词集；步骤(3)基于频繁词集的文本表示；步骤(4)训练分类器并对文本进行分类。本发明能够利用频繁词集的新的约束参数 AD-Sup，使频繁特征具有更好的类别区分能力。并且针对频繁特征上的数据稀疏性问题，提出了基于特征相似度的权重计算方法，有效的降低了文档在新增频繁词特征上的数据稀疏性。



CN 102799666 A

1. 一种基于频繁词集的网络新闻自动文本分类的方法，其特征在于：该方法具体步骤如下：

步骤 (1)、数据预处理和特征选取：利用词干提取和停等词去除来进行文本数据预处理，利用信息增益来对文本数据进行特征选取；

步骤 (2)、频繁词集抽取：从步骤 (1) 生成的候选词集中发现支持度高于最低支持度的词集；频繁词集抽取的过程利用 Apriori 算法，通过宽度优先的策略逐级生成各项频繁项目集；

步骤 (3)、文本表示：把频繁词集和初始单词作为一个整体，初始单词视为只包含一个频繁项的单元词集；当频繁词集数量为 0，特征空间就是由传统的 bag of words 组成，而当初始单词数量为 0，那么特征空间就仅包含有频繁词集；并且利用基于相似度的权重的计算方法解决数据稀释的问题；

步骤 (4)、文本分类：在训练集上构建分类器，并且在测试集上进行分类，利用的分类器是 SVM。

2. 根据权利要求 1 所述的一种基于频繁词集的网络新闻自动文本分类的方法，其特征在于：所述步骤（2）中对候选词集的频繁词抽取，具体如下：

步骤①、采用新的文本分类的频繁集生成约束参数：均偏差支持率；假定文档集合包含 n 个类别 $\{class_1, \cdots class_i, \cdots, class_n\}$，令 FS 表示词集，t 为词集 FS 中的词条，则 AD-Sup 的表达式为公式（1），而在公式（2）中，$Sup(FS)_i$ 指的是词集 FS 在类别 i 中的支持数，在公式（3）中，df(t) 指的是词条 t 的文档频率：

$$AD\text{-}Sup(FS) = \frac{\sqrt{\sum_{i=1}^{n}\{Sup(FS)_i - Ave(Sup(FS))\}^2}}{Ave(Sup(FS))} \qquad (1)$$

$$Ave(Sup(FS)) = \frac{\sum_{i=1}^{n} Sup(FS)_i}{n} \qquad (2)$$

$$Sup(FS)_i = \min\{df(t)_1 \ldots df(t)_m\} \qquad (3)$$

步骤②、利用步骤①产生的新的约束参数，对候选词集进行频繁词抽取，抽取的过程利用了 Apriori 算法，通过宽度优先的策略逐级生成各项频繁项目集；频繁集的提取是先用传统的支持度 min-sup 作提取，然后用提出的新的约束参数，对提取后的频繁集进行优化，并不是只用新约束参数提取一次。

3. 根据权利要求 1 或 2 所述的一种基于频繁词集的网络新闻自动文本分类的方法，其特征在于：所述步骤（3）中的文本的表示方法，具体如下：

步骤 A、把在步骤（2）中得到的频繁词集和初始单词作为一个整体考虑，初始单词视为只包含一个频繁项的单元词集；因此全局特征空间就包含了项目数从 1 到 n 的各级词集；

步骤 B、采用一种基于特征相似度的权重计算方法来解决数据稀疏性的问题；当一个文档包含某词集的一部分时，该部分词集即为原词集特征的一个相似特征；相似特征的权重可以通过原词集和部分词集之间的相似度来获得；若 FS′ 为相似特征，FS 为原特征，W 为特征权重，则二者的相似度可以通过公式（4）来计算：

$$FeatureSimilarity(FS', FS) = \frac{\sum_{i=1}^{n} W_i' \times W_i}{\sqrt{\sum_{i=1}^{n} (W_i')^2} \times \sqrt{\sum_{i=1}^{n} (W_i)^2}} \quad (4)$$

步骤 C、最终，FS’的权重为 FS’中的词频均值乘以 FS ′ 和 FS 的相似度：

$$W(FS') = \frac{(\sum_{i=1}^{n} TF_i')}{n} \times FeatureSimilarity(FS', FS) \quad (5)。$$

一种基于频繁词集的网络新闻自动文本分类的方法

技术领域

[0001] 本发明涉及信息检索、信息过滤、搜索引擎、推荐系统等技术领域，特别涉及一种基于频繁词集的网络新闻自动文本分类的方法。

背景技术

[0002] 随着互联网的发展，海量的网络数据使得有效的检索和管理变得越来越重要。由于绝大多数信息仍以文本形式呈现，基于文本内容的信息检索和数据挖掘成为备受关注的研究领域。文本分类(Text categorization)是文本挖掘中的重要内容之一，是在预先标注的类别集合上，对未标注的文本(文档)根据内容判定其类别。作为一种有效的信息组织和管理方法，文本分类便于用户和信息系统准确定位所需信息，在信息检索、信息过滤、搜索引擎、推荐系统等领域有着广泛的应用。目前文本分类的常用方法主要是基于机器学习的，典型的包括朴素贝叶斯、决策树、k-NN、Rocchio 和 SVM 等。在这些方法中，文本的表示是基于向量空间模型(VSM)的。在 VSM 中，文本内容被视为“Bag of words”(BOW)，BOW 的处理方法忽略了词条之间的关联性，不能保留文本的上下文和语法信息，而这种关联性却在自然语言中对文本内容所包含的具体含义有着重要的影响。

[0003] 频繁项目集是数据挖掘中的基本概念，指共同出现次数即支持度高于一定阈值的一组项目集合。频繁项目集隐含了其中各项之间的关联性，当其中的项目是文本中的词条，频繁项目集也就包含了更多的上下文信息。

发明内容

[0004] 本发明要解决的技术问题为：克服现有技术的不足，提供一种基于频繁词集的文本分类方法，该方法考虑文本上下文和语法信息，提出了一种新的文本表示策略，通过初始单词和频繁词集共同构建特征空间，并提高了文本分类的准确性。

[0005] 本发明解决上述技术问题的技术方案为：一种基于频繁词集的网络新闻自动文本分类的方法，该方法具体步骤如下：

[0006] 步骤 (1)、数据预处理和特征选取：利用词干提取和停等词去除来进行文本数据预处理，利用信息增益来对文本数据进行特征选取，利用基于特征相似度的权重计算方法，对包含部分频繁词集的特征进行权重预测，有效的降低了文档在新增频繁词特征上的数据稀疏性。

[0007] 步骤 (2)、频繁词集抽取：从步骤 (1) 生成的候选词集中发现支持度高于最低支持度的词集。频繁词集抽取的过程利用 Apriori 算法，通过宽度优先的策略逐级生成各项频繁项目集；通过 AD-Sup 对提起的频繁集进行了优化

[0008] 步骤 (3)、文本表示：把频繁词集和初始单词作为一个整体，特征空间同时包含单词和频繁词集，并且利用基于相似度的权重的计算方法解决频繁集特征的数据稀释问题；

[0009] 步骤 (4)、文本分类：在训练集上构建分类器，并且在测试集上进行分类。利用的分类器是 SVM。

[0010] 所述步骤(2)中对候选词集的频繁词抽取,具体如下:

[0011] 步骤①、采用新的文本分类的频繁集生成约束参数:均偏差支持率。假定文档集合包含 n 个类别 $\{class_1, \cdots class_i, \cdots class_n\}$,令 FS 表示词集,t 为词集 FS 中的词条,则 AD-Sup 的表达式为公式(1),而在公式(2)中,$Sup(FS)_i$ 指的是词集 FS 在类别 i 中的支持数,在公式(3)中,df(t) 指的是词条 t 的文档频率:

[0012] $$AD\text{-}Sup(FS) = \frac{\sqrt{\sum_{i=1}^{n}\{Sup(FS)_i - Ave(Sup(FS))\}^2}}{Ave(Sup(FS))} \quad (1)$$

[0013] $$Ave(Sup(FS)) = \frac{\sum_{i=1}^{n} Sup(FS)_i}{n} \quad (2)$$

[0014] $$Sup(FS)_i = \min\{df(t)_1 \ldots df(t)_m\} \quad (3)$$

[0015] 步骤②、利用步骤①产生的新的约束参数,对候选词集进行频繁词抽取,抽取的过程利用了 Apriori 算法,通过宽度优先的策略逐级生成各项频繁项目集,通过 AD-Sup 对提起的频繁集进行了优化;

[0016] 所述步骤(3)中的文本的表示方法,具体如下:

[0017] 步骤 A、把在步骤(2)中得到的频繁词集和初始单词作为一个整体考虑,初始单词视为只包含一个频繁项的单元词集。因此全局特征空间就包含了项目数从 1 到 n 的各级词集;

[0018] 步骤B、采用一种基于特征相似度的权重计算方法来解决数据稀疏性的问题。当一个文档包含某词集的一部分时,该部分词集即为原词集特征的一个相似特征。相似特征的权重可以通过原词集和部分词集之间的相似度来获得。若 FS' 为相似特征,FS 为原特征,W 为特征权重,则二者的相似度可以通过公式(4)来计算:

[0019] $$FeatureSimilarity(FS', FS) = \frac{\sum_{i=1}^{n} W_i' \times W_i}{\sqrt{\sum_{i=1}^{n}(W_i')^2} \times \sqrt{\sum_{i=1}^{n}(W_i)^2}} \quad (4)$$

[0020] 步骤 C、最终,FS' 的权重为 FS' 中的词频均值乘以 FS′ 和 FS 的相似度:

[0021] $$W(FS') = \frac{(\sum_{i=1}^{n} TF_i')}{n} \times FeatureSimilarity(FS', FS) \quad (5)$$

[0022] 本发明与现有技术相比的优点在于:

[0023] 本发明提出了新的约束参数,提取适用于分类的频繁词集特征,用频繁词集作为补充特征来表示文本,更多的保留了单词的上下文信息;针对数据稀疏性问题,提出了基于特征相似度的权重计算方法,对包含部分频繁词集的特征进行权重预测,有效的降低了文档在新增频繁词特征上的数据稀疏性,提高了分类效果。

附图说明

[0024] 图 1 为本发明的概要工作流程图；

[0025] 图 2 为本发明的详细工作流程图；

[0026] 图 3 为 Reuters-21578 数据集上的分类结果；

[0027] 图 4 为 WebKB 数据集上的分类结果。

具体实施方式

[0028] 现结合附图说明本发明的实施例。

[0029] 如图 2 所示，本发明包括四个主要步骤：

[0030] 步骤 (1)、数据预处理和特征选取：利用词干提取和停等词去除来进行文本数据预处理，利用信息增益来对文本数据进行特征选取，利用基于特征相似度的权重计算方法，对包含部分频繁词集的特征进行权重预测，有效的降低了文档在新增频繁词特征上的数据稀疏性。

[0031] 步骤 (2)、频繁词集抽取：从步骤 (1) 生成的候选词集中发现支持度高于最低支持度的词集。频繁词集抽取的过程利用 Apriori 算法，通过宽度优先的策略逐级生成各项频繁项目集；

[0032] 步骤①、采用新的文本分类的频繁集生成约束参数：均偏差支持率。假定文档集合包含 n 个类别 $\{class_1, \cdots class_i, \cdots class_n\}$，令 FS 表示词集，t 为词集 FS 中的词条，则 AD-Sup 的表达式为公式 (1)，而在公式 (2) 中，$Sup(FS)_i$ 指的是词集 FS 在类别 i 中的支持数，在公式 (3) 中，df(t) 指的是词条 t 的文档频率：

[0033] $$AD\text{-}Sup(FS) = \frac{\sqrt{\sum_{i=1}^{n}\{Sup(FS)_i - Ave(Sup(FS))\}^2}}{Ave(Sup(FS))} \quad (1)$$

[0034] $$Ave(Sup(FS)) = \frac{\sum_{i=1}^{n} Sup(FS)_i}{n} \quad (2)$$

[0035] $$Sup(FS)_i = \min\{df(t)_1 \ldots df(t)_m\} \quad (3)$$

[0036] 步骤②、利用步骤①产生的新的约束参数，对候选词集进行频繁词抽取，抽取的过程利用了 Apriori 算法，通过宽度优先的策略逐级生成各项频繁项目集；

[0037] 步骤 (3)、文本表示：把频繁词集和初始单词作为一个整体，特征空间同时包含单词和频繁词集并且利用基于相似度的权重的计算方法解决数据稀释的问题；

[0038] 步骤 A、我们把在步骤 (2) 中得到的频繁词集和初始单词作为一个整体考虑，初始单词视为只包含一个频繁项的单元词集。因此全局特征空间就包含了项目数从 1 到 n 的各级词集；

[0039] 步骤 B、采用一种基于特征相似度的权重计算方法来解决数据稀疏性的问题。当一个文档包含某词集的一部分时，该部分词集即为原词集特征的一个相似特征。相似特征的权重可以通过原词集和部分词集之间的相似度来获得。若 FS′ 为相似特征，FS 为原特征，W 为特征权重，则二者的相似度可以通过公式 (4) 来计算：

[0040] $$FeatureSimilarity(FS',FS)=\frac{\sum_{i=1}^{n}W_i'\times W_i}{\sqrt{\sum_{i=1}^{n}(W_i')^2}\times\sqrt{\sum_{i=1}^{n}(W_i)^2}} \quad (4)$$

[0041] 步骤 C、最终，FS’的权重为 FS’中的词频均值乘以 FS ′ 和 FS 的相似度：

[0042] $$W(FS')=\frac{(\sum_{i=1}^{n}TF_i')}{n}\times FeatureSimilarity(FS',FS) \quad (5)$$

[0043] 步骤 (4)、文本分类：在训练集上构建分类器，并且在测试集上进行分类。利用的分类器是 SVM；

[0044] 本发明提出了一种基于频繁词集的网络新闻自动文本分类的方法，还可以应用于其他领域，如邮件过滤，文本检索，信息管理等，在频繁词集的选取中，我们引入了一个新的约束参数 AD-Sup，充分考虑了频繁词集在各类别中的分布差异性，使所选取的频繁词集特征具有更好的类别区分能力。针对数据稀疏性问题，我们提出了基于特征相似度的权重计算方法，对包含部分频繁词集的特征进行权重预测，有效的降低了文档在新增频繁词特征上的数据稀疏性。在 Reuters-21578 和 WebKB 数据集上，训练 SVM 进行文本分类，通过与单特征训练下的 SVM 对比分类结果验证了特征组合策略的有效性，并对比了不同权重计算方法下的分类结果。结果表明，通过 AD-Sup 选取的频繁词集和特征组合策略可以有效提高 SVM 的分类结果。结果如图 3，图 4 所示。

[0045] 本发明未详细阐述的部分属于本领域公知技术。

[0046] 以上实施例仅用以说明本发明的技术方案而非限制在具体实施方式的范围内，对本技术领域的普通技术人员来讲，只要各种变化在权利要求限定和确定的本发明的精神和范围内，这些变化是显而易见的，一切利用本发明构思的发明创造均在保护之列。

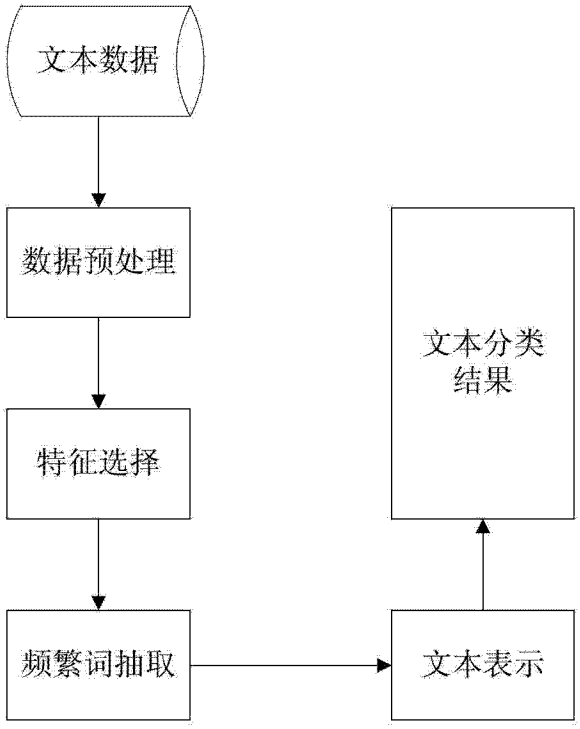

文本数据
数据预处理
特征选择
频繁词抽取
文本表示
文本分类结果

图 1

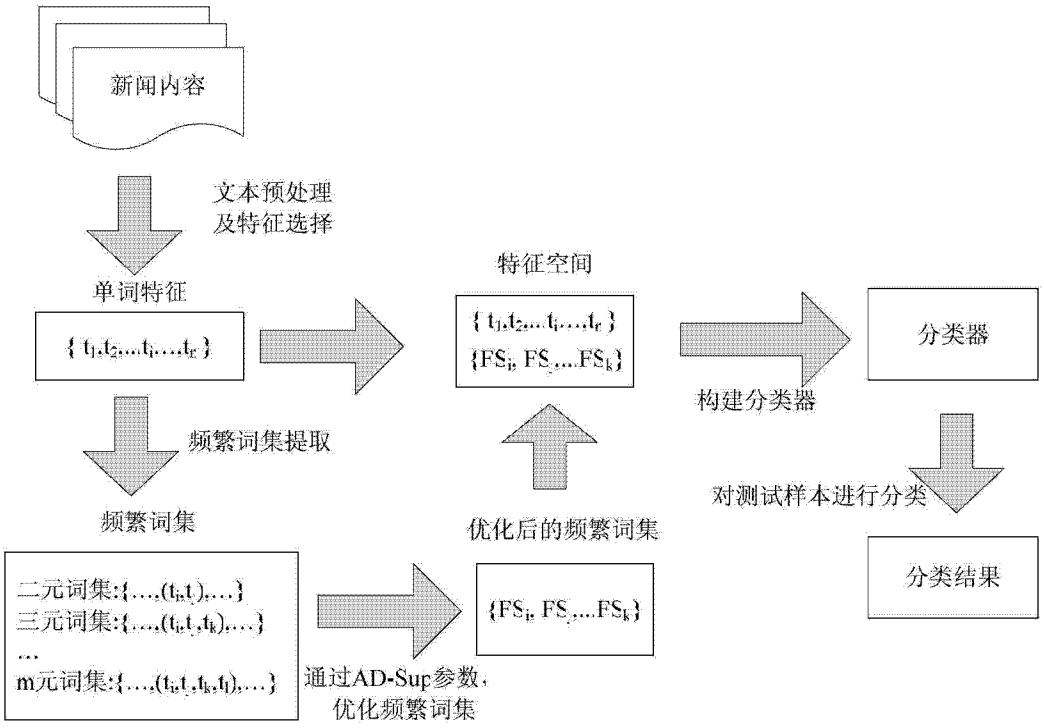

新闻内容
文本预处理及特征选择
单词特征
{ t1,t2,...ti,...tE }
频繁词集提取
频繁词集
二元词集:{...,(ti,tj),...}
三元词集:{...,(ti,tj,tk),...}
...
m元词集:{...,(ti,tj,tk,tl),...}
通过AD-Sup参数，优化频繁词集
{FSi, FSj,...FSk}
特征空间
{ t1,t2,...ti,...tE }
{FSi, FSj,...FSk}
优化后的频繁词集
构建分类器
分类器
对测试样本进行分类
分类结果

图 2

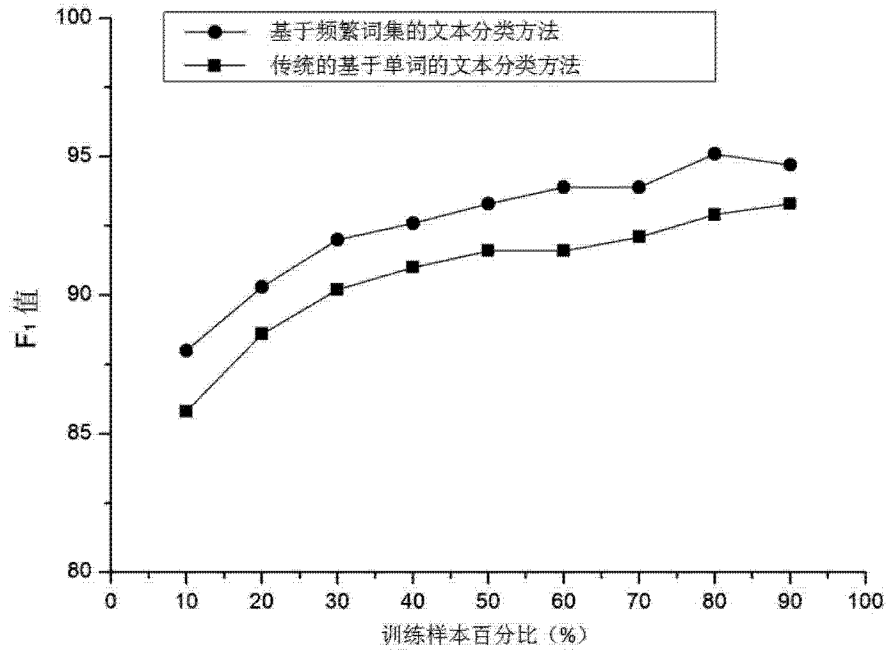

基于频繁词集的文本分类方法
传统的基于单词的文本分类方法
F1 值
训练样本百分比（%）
100
95
90
85
80
0
10
20
30
40
50
60
70
80
90
100

图 3

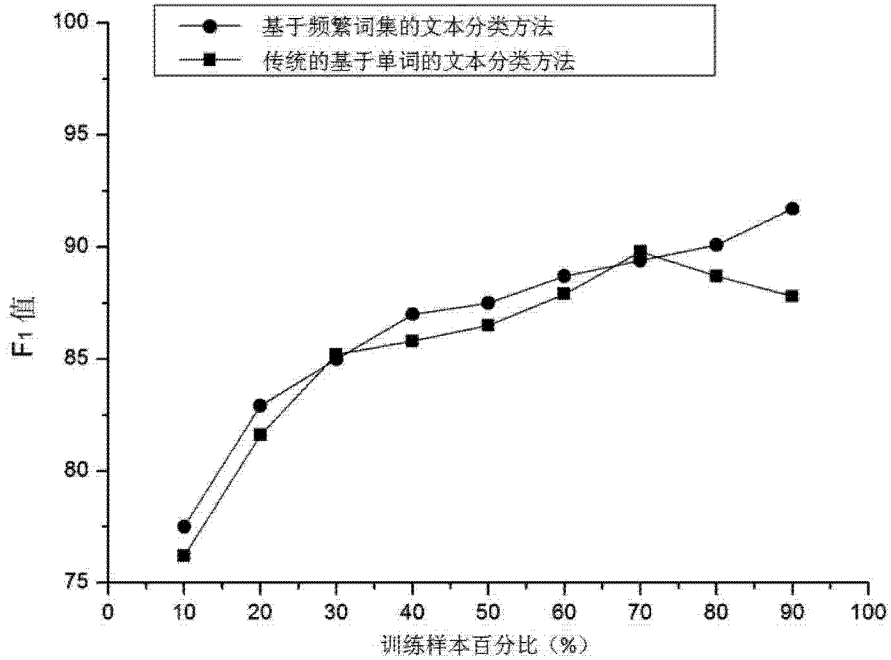

基于频繁词集的文本分类方法
传统的基于单词的文本分类方法
F1 值
训练样本百分比（%）
100
95
90
85
80
75
0
10
20
30
40
50
60
70
80
90
100

图 4