(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 111680132 A
(43)申请公布日 2020.09.18

(21)申请号 202010654254.0

(22)申请日 2020.07.08

(71)申请人 中国人民解放军国防科技大学
地址 410073 湖南省长沙市开福区德雅路109号

(72)发明人 张翀 何春辉 谭真 葛斌

(74)专利代理机构 长沙国科天河知识产权代理有限公司 43225
代理人 邱轶

(51)Int.Cl.
G06F 16/33(2019.01)
G06F 16/335(2019.01)
G06F 16/35(2019.01)
G06F 40/30(2020.01)
G06F 16/9536(2019.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

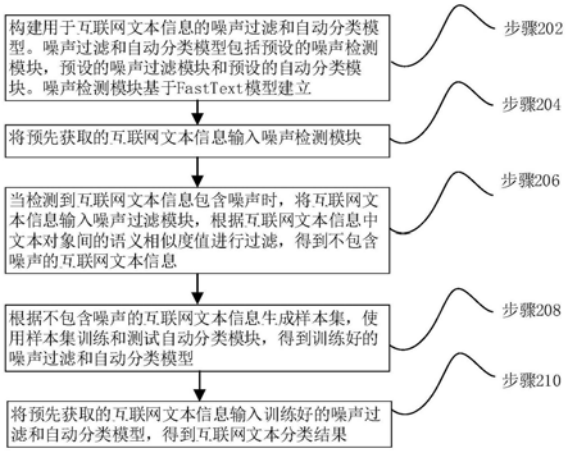
权利要求书2页 说明书11页 附图3页

(54)发明名称

一种用于互联网文本信息的噪声过滤和自动分类方法

(57)摘要

本申请涉及一种用于互联网文本信息的噪声过滤和自动分类方法。所述方法包括:构建互联网文本信息的噪声过滤和自动分类模型,包括噪声检测模块、噪声过滤模块和自动分类模块。将获取到的互联网文本信息输入噪声检测模块,检测到包含噪声时将其输入噪声过滤模块,根据文本对象间的语义相似度值进行过滤,输出不包含噪声的互联网文本信息。使用不包含噪声的互联网文本信息训练和测试自动分类模块,使用训练好的模型获得互联网文本信息分类结果。采用本方法能同时实现文本噪声过滤和分类,避免噪声检测和过滤过分依赖特征选取的问题,及其对训练数据集质量要求较高的问题,还能消除噪声信息对文本分类结果的影响,提高文本分类结果的准确性。



CN 111680132 A

1.一种用于互联网文本信息的噪声过滤和自动分类方法,所述方法包括:

构建用于互联网文本信息的噪声过滤和自动分类模型;所述噪声过滤和自动分类模型包括预设的噪声检测模块,预设的噪声过滤模块和预设的自动分类模块;所述噪声检测模块基于FastText模型建立;

将预先获取的互联网文本信息输入所述噪声检测模块;

当检测到所述互联网文本信息包含噪声时,将所述互联网文本信息输入所述噪声过滤模块,根据所述互联网文本信息中文本对象间的语义相似度值进行过滤,得到不包含噪声的互联网文本信息;

根据所述不包含噪声的互联网文本信息生成样本集,使用所述样本集训练和测试所述自动分类模块,得到训练好的噪声过滤和自动分类模型;

将预先获取的互联网文本信息输入所述训练好的噪声过滤和自动分类模型,得到互联网文本分类结果。

2.根据权利要求1所述的方法,其特征在于,构建所述噪声检测模块的方式包括:

根据预设的规则标注预先获取的互联网文本数据中的噪声数据和非噪声数据,得到用于模型训练的噪声二分类数据集;

将所述噪声二分类数据集输入预设的FastText文本识别模型,得到训练好的噪声检测模块。

3.根据权利要求1所述的方法,其特征在于,所述噪声过滤模块基于BERT模型建立;

所述当检测到所述互联网文本信息包含噪声时,将所述互联网文本信息输入所述噪声过滤模块,根据所述互联网文本信息中文本对象间的语义相似度值进行过滤,输出不包含噪声的互联网文本信息的步骤包括:

当检测到所述互联网文本信息包含噪声时,获取所述互联网文本信息中的标题文本和正文文本,将所述正文文本按照预设的规则拆分为正文段落文本;

将所述标题文本和所述正文段落文本依次输入所述噪声过滤模块,计算所述标题文本和所述正文段落文本间的语义相似度值,当所述正文段落文本和所述标题文本间的语义相似度值低于预设值时,将所述正文段落文本标记为噪声;

按照所述正文文本中所述正文段落文本的先后顺序,拼接未标记为噪声的所述正文段落文本,输出不包含噪声的互联网文本信息。

4.根据权利要求3所述的方法,其特征在于,将所述标题文本和所述正文段落文本依次输入所述噪声过滤模块,计算所述标题文本和所述正文段落文本间的语义相似度值,当所述正文段落文本和所述标题文本间的语义相似度值低于预设值时,将所述正文段落文本标记为噪声的步骤包括:

将所述标题文本和所述正文段落文本输入所述噪声过滤模块,根据预设的余弦相似度算法计算所述标题文本和所述正文段落文本间的语义相似度值;

当所述正文段落文本和所述标题文本间的语义相似度值低于预设值时,将所述正文段落文本标记为噪声。

5.根据权利要求1所述的方法,其特征在于,所述自动分类模块基于卷积神经网络,包括输入层、词嵌入层、卷积层、最大池化层、全连接层和输出层;

构建所述自动分类模块的方式包括:

使用反向传播方法确定所述自动分类模块的卷积层参数。

6.根据权利要求5所述的方法,所述根据所述不包含噪声的互联网文本信息生成样本集,使用所述样本集训练和测试所述自动分类模块,得到训练好的噪声过滤和自动分类模型的步骤包括:

根据所述不包含噪声的互联网文本信息生成样本集,将所述样本集通过所述输入层输入所述自动分类模块;

由所述词嵌入层、所述卷积层和所述最大池化层提取文本特征向量,由所述全连接层通过所述输出层输出互联网文本信息分类结果;

根据所述自动分类模块输出的文本自动分类结果和对应的文本分类概率值,得到训练好的噪声过滤和自动分类模型。

7.一种用于互联网文本信息的噪声过滤和自动分类装置,其特征在于,所述装置包括:

模型构建单元,用于构建用于互联网文本信息的噪声过滤和自动分类模型;所述噪声过滤和自动分类模型包括预设的噪声检测模块,预设的噪声过滤模块和预设的自动分类模块;所述噪声检测模块基于FastText模型建立;

互联网文本信息输入单元,用于将预先获取的互联网文本信息输入所述噪声检测模块;

互联网文本信息噪声检测与过滤单元,用于当检测到所述互联网文本信息包含噪声时,将所述互联网文本信息输入所述噪声过滤模块,根据所述互联网文本信息中文本对象间的语义相似度值进行过滤,得到不包含噪声的互联网文本信息;

模型训练单元,用于根据所述不包含噪声的互联网文本信息生成样本集,使用所述样本集训练和测试所述自动分类模块,得到训练好的噪声过滤和自动分类模型;

互联网文本信息分类单元,用于将预先获取的互联网文本信息输入所述训练好的噪声过滤和自动分类模型,得到互联网文本分类结果。

8.根据权利要求7所述的装置,其特征在于,所述噪声过滤模块基于BERT模型建立;

所述互联网文本信息噪声检测与过滤单元用于:

当检测到所述互联网文本信息包含噪声时,获取所述互联网文本信息中的标题文本和正文文本,将所述正文文本按照预设的规则拆分为正文段落文本;

将所述标题文本和所述正文段落文本依次输入所述噪声过滤模块,计算所述标题文本和所述正文段落文本间的语义相似度值,当所述正文段落文本和所述标题文本间的语义相似度值低于预设值时,将所述正文段落文本标记为噪声;

按照所述正文文本中所述正文段落文本的先后顺序,拼接未标记为噪声的所述正文段落文本,输出不包含噪声的互联网文本信息。

9.一种计算机设备,包括存储器和处理器,所述存储器存储有计算机程序,其特征在于,所述处理器执行所述计算机程序时实现权利要求1至6中任一项所述方法的步骤。

10.一种计算机可读存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现权利要求1至6中任一项所述的方法的步骤。

一种用于互联网文本信息的噪声过滤和自动分类方法

技术领域

[0001] 本申请涉及互联网文本信息处理技术领域，特别是涉及一种用于互联网文本信息的噪声过滤和自动分类方法。

背景技术

[0002] 互联网文本信息作为互联网信息传递的一种方式，在信息共享中起着举足轻重的作用。然而互联网文本中通常包含大量与主题无关的内容，如许多网页新闻中都会夹杂广告、插图简介、网站推荐内容等，这些与主题无关的内容被称为噪声信息。噪声信息会对互联网文本内容的分类产生干扰，因此如何过滤噪声并提纯互联网文本内容以提高这些文本的分类准确率具有重要意义。

[0003] 现有方法大多将文本噪声识别(或噪声过滤)和文本分类作为两个独立的任务进行分别的建模和处理。目前的文本噪声识别与过滤方法主要分为两大类：第一类是结合词袋模型和传统机器学习进行噪声识别与过滤的方法，这类方法过分依赖特征的选取，且对噪声识别的准确率不高；第二类是基于深度学习方法来实现噪声识别与过滤，这类方法的识别准确率比较高，但是对人工标注数据集的质量要求较高，且这类方法大多将待分类的文本内容作为语料直接用于分类模型的训练和测试，然而文本语料中包含噪声信息会干扰文本分类的结果。

发明内容

[0004] 基于此，有必要针对上述技术问题，提供能够识别并过滤噪声文本信息并能够提高文本分类准确度的一种用于互联网文本信息的噪声过滤和自动分类方法。

[0005] 一种用于互联网文本信息的噪声过滤和自动分类方法，所述方法包括：

[0006] 构建用于互联网文本信息的噪声过滤和自动分类模型。噪声过滤和自动分类模型包括预设的噪声检测模块，预设的噪声过滤模块和预设的自动分类模块。噪声检测模块基于FastText模型建立。

[0007] 将预先获取的互联网文本信息输入噪声检测模块。

[0008] 当检测到互联网文本信息包含噪声时，将互联网文本信息输入噪声过滤模块，根据互联网文本信息中文本对象间的语义相似度值进行过滤，得到不包含噪声的互联网文本信息。

[0009] 根据不包含噪声的互联网文本信息生成样本集，使用样本集训练和测试自动分类模块，得到训练好的噪声过滤和自动分类模型。

[0010] 将预先获取的互联网文本信息输入训练好的噪声过滤和自动分类模型，得到互联网文本分类结果。

[0011] 其中一个实施例中，构建所述噪声检测模块的方式包括：

[0012] 根据预设的规则标注预先获取的互联网文本数据中的噪声数据和非噪声数据，得到用于模型训练的噪声二分类数据集。

[0013] 将噪声二分类数据集输入预设的FastText文本识别模型，得到训练好的噪声检测模块。

[0014] 其中一个实施例中，噪声过滤模块基于BERT模型建立，当检测到互联网文本信息包含噪声时，将互联网文本信息输入噪声过滤模块，根据互联网文本信息中文本对象间的语义相似度值进行过滤，输出不包含噪声的互联网文本信息的步骤包括：

[0015] 当检测到互联网文本信息包含噪声时，获取互联网文本信息中的标题文本和正文文本，将正文文本按照预设的规则拆分为正文段落文本。

[0016] 将标题文本和正文段落文本依次输入噪声过滤模块，计算标题文本和正文段落文本间的语义相似度值，当正文段落文本和标题文本间的语义相似度值低于预设值时，将该正文段落文本标记为噪声。

[0017] 按照正文文本中正文段落文本的先后顺序，拼接未标记为噪声的正文段落文本，输出不包含噪声的互联网文本信息。

[0018] 其中一个实施例中，将标题文本和所述正文段落文本依次输入噪声过滤模块，计算标题文本和正文段落文本间的语义相似度值，当正文段落文本和标题文本间的语义相似度值低于预设值时，将该正文段落文本标记为噪声的步骤包括：

[0019] 将标题文本和正文段落文本依次输入噪声过滤模块，根据预设的余弦相似度算法计算标题文本和正文段落文本间的语义相似度值。

[0020] 当正文段落文本和标题文本间的语义相似度值低于预设值时，将正文段落文本标记为噪声。

[0021] 其中一个实施例中，自动分类模块基于卷积神经网络，包括输入层、词嵌入层、卷积层、最大池化层、全连接层和输出层。

[0022] 构建自动分类模块的方式包括：

[0023] 使用反向传播方法确定自动分类模块的卷积层参数。

[0024] 其中一个实施例中，根据不包含噪声的互联网文本信息生成样本集，使用样本集训练和测试自动分类模块，得到训练好的噪声过滤和自动分类模型的步骤包括：

[0025] 根据不包含噪声的互联网文本信息生成样本集，将样本集通过输入层输入自动分类模块。

[0026] 由词嵌入层、卷积层和最大池化层提取文本特征向量，由全连接层通过输出层输出互联网文本信息分类结果。

[0027] 根据自动分类模块输出的文本自动分类结果和对应的文本分类概率值，得到训练好的噪声过滤和自动分类模型。

[0028] 一种用于互联网文本信息的噪声过滤和自动分类装置，其特征在于，所述装置包括：

[0029] 模型构建单元，用于构建用于互联网文本信息的噪声过滤和自动分类模型。噪声过滤和自动分类模型包括预设的噪声检测模块，预设的噪声过滤模块和预设的自动分类模块。噪声检测模块基于FastText模型建立。

[0030] 互联网文本信息输入单元，用于将预先获取的互联网文本信息输入噪声检测模块。

[0031] 互联网文本信息噪声检测与过滤单元，用于当检测到互联网文本信息包含噪声

时，将互联网文本信息输入噪声过滤模块，根据互联网文本信息中文本对象间的语义相似度值进行过滤，得到不包含噪声的互联网文本信息。

[0032] 模型训练单元，用于根据不包含噪声的互联网文本信息生成样本集，使用样本集训练和测试自动分类模块，得到训练好的噪声过滤和自动分类模型。

[0033] 互联网文本信息分类单元，用于将预先获取的互联网文本信息输入训练好的噪声过滤和自动分类模型，得到互联网文本分类结果。

[0034] 其中一个实施例中，噪声过滤模块基于BERT模型建立，互联网文本信息检测与过滤单元用于：

[0035] 当检测到互联网文本信息包含噪声时，获取互联网文本信息中的标题文本和正文文本，将正文文本按照预设的规则拆分为正文段落文本。

[0036] 将标题文本和正文段落文本依次输入噪声过滤模块，计算标题文本和正文段落文本间的语义相似度值，当正文段落文本和标题文本间的语义相似度值低于预设值时，将该正文段落文本标记为噪声。

[0037] 按照正文文本中正文段落文本的先后顺序，拼接未标记为噪声的所述正文段落文本，输出不包含噪声的互联网文本信息。

[0038] 一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，所述处理器执行所述计算机程序时实现以下步骤：

[0039] 构建用于互联网文本信息的噪声过滤和自动分类模型。噪声过滤和自动分类模型包括预设的噪声检测模块，预设的噪声过滤模块和预设的自动分类模块。噪声检测模块基于FastText模型建立。

[0040] 将预先获取的互联网文本信息输入噪声检测模块。

[0041] 当检测到互联网文本信息包含噪声时，将互联网文本信息输入噪声过滤模块，根据互联网文本信息中文本对象间的语义相似度值进行过滤，得到不包含噪声的互联网文本信息。

[0042] 根据不包含噪声的互联网文本信息生成样本集，使用样本集训练和测试自动分类模块，得到训练好的噪声过滤和自动分类模型。

[0043] 将预先获取的互联网文本信息输入训练好的噪声过滤和自动分类模型，得到互联网文本分类结果。

[0044] 一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现以下步骤：

[0045] 构建用于互联网文本信息的噪声过滤和自动分类模型。噪声过滤和自动分类模型包括预设的噪声检测模块，预设的噪声过滤模块和预设的自动分类模块。噪声检测模块基于FastText模型建立。

[0046] 将预先获取的互联网文本信息输入噪声检测模块。

[0047] 当检测到互联网文本信息包含噪声时，将互联网文本信息输入噪声过滤模块，根据互联网文本信息中文本对象间的语义相似度值进行过滤，得到不包含噪声的互联网文本信息。

[0048] 根据不包含噪声的互联网文本信息生成样本集，使用样本集训练和测试自动分类模块，得到训练好的噪声过滤和自动分类模型。

[0049] 将预先获取的互联网文本信息输入训练好的噪声过滤和自动分类模型，得到互联网文本分类结果。

[0050] 上述一种用于互联网文本信息的噪声过滤和自动分类方法、装置、计算机设备和存储介质，将噪声过滤任务分成了噪声检测和噪声过滤两个阶段，根据互联网文本信息中文本对象间的语义相似度值过滤文本信息中的噪声，输出不包含噪声的互联网文本信息，能够避免依赖特征选取的噪声识别准确率不高的问题，也能克服基于深度学习的噪声识别对人工标注的训练数据集质量要求较高的问题；将无噪声的互联网文本信息输入采用无噪声的样本集训练的自动分类模块，能够消除噪声信息对文本分类结果的影响，能够提高文本分类结果的准确性。

附图说明

[0051] 图1为一个实施例中一种用于互联网文本信息的噪声过滤和自动分类方法的应用场景图；

[0052] 图2为一个实施例中一种用于互联网文本信息的噪声过滤和自动分类方法的流程示意图；

[0053] 图3为另一个实施例中一种用于互联网文本信息的噪声过滤和自动分类方法的流程示意图；

[0054] 图4为一个实施例中基于BERT的噪声过滤模块的语义相似度计算方法的流程图；

[0055] 图5为一个实施例中基于卷积神经网络的自动分类模块的框架示意图；

[0056] 图6为一个实施例中计算机设备的内部结构图。

具体实施方式

[0057] 为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

[0058] 经过深入分析发现，噪声识别和噪声过滤完全可以作为数据预处理的一部分融入文本分类任务，因此本申请提出了一种用于互联网文本信息的噪声过滤和自动分类方法，可以应用于图1所示的应用环境中，用于针对互联网爬虫实时采集的各种公开网页文本信息进行快速分类，通过实验论证证明其可以在互联网开源文本信息自动采集系统中取得与设计预期一致的理想效果。

[0059] 网页文本信息主要是指含有标题和正文部分的半结构化互联网文本信息(例如网页新闻，博客，公告等等)，当爬虫采集到互联网文本内容后，在数据入库之前，需要对采集到的文本内容进行分类，然后才能按照类别进行归档存储，这样既便于对互联网文本数据进行管理，又利于后续的深度挖掘任务。其中，设备102通过爬虫获取互联网文本信息，经过噪声识别/过滤、文本分类后将分类好的文本输出至设备104进行后续的归档存储和深度挖掘处理。设备102可以但不限于是服务器或者是多个服务器组成的服务器集群，也可以是各种个人计算机、笔记本电脑等能够提供相应计算能力的设备。

[0060] 在一个实施例中，如图2所示，提供了一种用于互联网文本信息的噪声过滤和自动分类方法，以该方法应用于图1中的设备102为例进行说明，包括以下步骤：

[0061] 步骤202:构建用于互联网文本信息的噪声过滤和自动分类模型。噪声过滤和自动分类模型包括预设的噪声检测模块,预设的噪声过滤模块和预设的自动分类模块。噪声检测模块基于FastText模型建立。

[0062] 步骤204:将预先获取的互联网文本信息输入噪声检测模块。

[0063] 具体地,噪声检测模块基于FastText模型建立。FastText文本分类算法是脸书人工智能研究院(FAIR,Facebook AI Research)提出的一种简单的模型。实验表明一般情况下,FastText算法能获得和深度学习模型相同的精度,但是其计算时间却要远远小于深度学习模型。FastText可以作为一个文本分类模型的基础。因此,与采用支持向量机、贝叶斯、决策树等传统的机器学习方法,基于FastText实现噪声检测能够提高噪声检测的速度。噪声检测模块对互联网文本信息进行噪声检测,仅将包含噪声的互联网文本信息输入噪声过滤模块,以提高噪声过滤过程的效率。

[0064] 步骤206:当检测到互联网文本信息包含噪声时,将互联网文本信息输入噪声过滤模块,根据互联网文本信息中文本对象间的语义相似度值进行过滤,输出不包含噪声的互联网文本信息。

[0065] 输入噪声过滤模块的是确定包含噪声的文本信息,因此可以不依靠预设的噪声特征库,也不需要大规模的高质量模型训练样本集就能够实现噪声过滤。具体地,可以采用word2vec或者glove等预训练词嵌入表示技术获取互联网文本信息中不同文本对象间的向量,以欧氏距离、马氏距离等度量向量间的距离,以获得文本对象(如标题和正文段落)之间的相似度,根据相似度的值过滤噪声,得到不包含噪声的互联网文本信息。

[0066] 步骤208:根据不包含噪声的互联网文本信息生成样本集,使用样本集训练和测试自动分类模块,得到训练好的噪声过滤和自动分类模型。

[0067] 具体地,自动分类模块可以采用支持向量机、贝叶斯等传统机器学习算法实现,也可以使用卷积神经网络、循环神经网络模型来实现。

[0068] 值得注意的是,本申请提供的一种用于互联网文本信息的噪声过滤和自动分类方法有严格的逻辑顺序,即先对从互联网获取的原始文本信息进行噪声识别和过滤,然后使用本身不含有噪声或去噪后的文本信息训练自动分类模块,并使用训练好的自动分类模块对不含有噪声的文本信息进行分类。通过这一逻辑顺序可以达到提纯原始语料的目的,且能有效的减少语料长度,从而降低模型的计算复杂度。

[0069] 步骤210:将预先获取的互联网文本信息输入训练好的噪声过滤和自动分类模型,得到互联网文本分类结果。

[0070] 上述一种用于互联网文本信息的噪声过滤和自动分类方法,可以同时实现噪声过滤与文本分类两个不同的任务。该方法将噪声过滤任务分成了噪声检测和噪声过滤两个阶段,根据互联网文本信息中文本对象间的语义相似度值过滤文本信息中的噪声,输出不包含噪声的互联网文本信息,能够避免依赖特征选取的噪声识别准确率不高的问题,也能克服基于深度学习的噪声识别对训练数据集质量要求较高的问题;将无噪声的互联网文本信息输入采用无噪声的样本集训练的自动分类模块,能够消除噪声信息对文本分类结果的影响,能够提高文本分类结果的准确性。

[0071] 其中一个实施例中,构建噪声检测模块的方式包括:

[0072] 根据预设的规则标注预先获取的互联网文本数据中的噪声数据和非噪声数据,得

到用于模型训练的噪声二分类数据集。

[0073] 将噪声二分类数据集输入预设的FastText文本识别模型，得到训练好的噪声检测模块。

[0074] 本实施例利用一个标注了噪声和非噪声的互联网文本二分类数据集作为样本集，去训练一个FastText噪声识别模型，然后利用该模型来实现对互联网文本内容的噪声识别任务。基于此，本实施例能够基于FastText模型的特性，快速识别噪声文本，并且能够提供准确的文本噪声识别结果。

[0075] 其中一个实施例中，噪声过滤模块基于BERT模型建立。当检测到互联网文本信息包含噪声时，将互联网文本信息输入噪声过滤模块，根据互联网文本信息中文本对象间的语义相似度值进行过滤，输出不包含噪声的互联网文本信息的步骤包括：

[0076] 当检测到互联网文本信息包含噪声时，获取互联网文本信息中的标题文本和正文文本，将正文文本按照预设的规则拆分为正文段落文本。

[0077] 将标题文本和正文段落文本依次输入噪声过滤模块，根据预设的余弦相似度算法计算标题文本和正文段落文本间的语义相似度值。

[0078] 当正文段落文本和标题文本间的语义相似度值低于预设值时，将正文段落文本标记为噪声。

[0079] 按照正文文本中正文段落文本的先后顺序，拼接未标记为噪声的正文段落文本，输出不包含噪声的互联网文本信息。

[0080] BERT的全称为Bidirectional Encoder Representation from Transformers，是一个预训练的语言表征模型。它强调了不再像以往一样采用传统的单向语言模型或者把两个单向语言模型进行浅层拼接的方法进行预训练，而是采用新的掩码语言模型(MLM)，以致能生成深度的双向语言表征。BERT模型的主要优点包括：预训练后，只需要添加一个额外的输出层进行微调，就将BERT模型应用于各种任务，而不需要针对不同的任务修改模型的结构。

[0081] 本实施例提供的用于互联网文本信息的噪声过滤和自动分类方法的流程图如图3所示。具体地，使用基于BERT模型的噪声过滤模块，根据互联网文本中文本对象间的语义相似度值进行过滤，输出不包含噪声的互联网文本信息的步骤包括：

[0082] 获取原始内容中的标题T和所有段落P，然后按顺序将它们添加到列表中。

[0083] 采用BERT向量转换工具将标题T和所有段落内容Pi转换为固定长度的向量Te和Pie，然后使用余弦相似度算法计算Te和每个段落Pie之间的语义相似度，详细的计算公式如下：

[0084] $$S_i(T_e, P_{ie}) = cosine(T_e, V_i) = \frac{\mathbf{T}_e^T V_i}{\|T_e\| \|V_i\|}$$

[0085] 其中，T_e和V_i是标题和第i个段落的固定长度向量表示结果，详细的语义相似度计算过程如图4所示。

[0086] 根据$S_i(T_e, P_{ie})$的计算结果，将相似度得分小于预设值(如0.1)的段落标记为噪声。

[0087] 最后，将未标记为噪声的段落按照原始的顺序拼接起来，和标题一起作为待分类语料一起输入到自动分类模块中。

[0088] 本实施例利用了BERT模型的性能优势，可以提高文本噪声过滤的效果，为自动分类模块提供更好的无噪训练数据集，以及为自动分类模块提供更好的无噪待分类语料，改进最终的分类效果。

[0089] 其中一个实施例中，自动分类模块基于卷积神经网络，包括输入层、词嵌入层、卷积层、最大池化层、全连接层和输出层。

[0090] 构建自动分类模块的方式包括：

[0091] 使用反向传播方法确定自动分类模块的卷积层参数。

[0092] 其中一个实施例中，根据不包含噪声的互联网文本信息生成样本集，使用样本集训练自动分类模块，得到训练好的噪声过滤和自动分类模型的步骤包括：

[0093] 根据不包含噪声的互联网文本信息生成样本集，将样本集通过输入层输入自动分类模块。

[0094] 由词嵌入层、卷积层和最大池化层提取文本特征向量，由全连接层通过输出层输出互联网文本信息分类结果。

[0095] 根据自动分类模块输出的文本自动分类结果和对应的文本分类概率值，得到训练好的噪声过滤和自动分类模型。

[0096] 具体地，本实施例基于google开源的Tensorflow框架构建了卷积神经网络模型。该模型共包含输入层，词嵌入层，卷积层，最大池化层，全连接层和输出层，模型框架如图5所示。

[0097] 图5所示的模型中，将词嵌入层分为四个区域，分别用unigram，bigram，trigram和4-gram序列表示不同的词嵌入方式，以表示四个不同的特征。然后使用加权函数来获得固定长度的向量，作为整个输入语料库的向量表示。如果输入的文本信息中包含unigram，bigram，trigram和4-gram序列的M个有效词语序列，则其对应的词向量表示为：

[0098] $$x = x_1 \oplus x_2 \oplus x_3 \oplus x_4$$

[0099] 其中$\oplus$是连接运算符，x_1、x_2、x_3、x_4分别为四个序列对应的有效词语序列的向量。

[0100] 卷积层由多个单元组成，每个卷积单元的参数通过反向传播过程获得。设$x_{i:j}$为词向量$x_i, x_{i+1}, \cdots, x_{i+j}$的连接，卷积核是$w \in R^{s*d}$，其中s是卷积窗口大小，d是词向量维度。本实施例中卷积窗口大小设置为s＝2、3和4(3个不同的卷积层)，每个卷积层包括128个单元。卷积层生成的特征向量Fi为：

[0101] $F_i = f(w \cdot x_{i:i+s} + b)$

[0102] 其中b是偏置向量，f是激活函数，在我们的实验中使用Relu函数。之后，将卷积核应用于每个可能的窗口$\{x_{1:s}, x_{2:s+1}, \cdots x_{M-s+1:M}\}$，最后生成特征图：

[0103] $F = [F_1, F_2, \cdots, F_{M-s+1}]$

[0104] 池化层用于减小特征的尺寸并提高模型的容错性。在本发明中，采用了最大的池化策略作为池化方法。通过最大池化操作给出映射F，以获得特征$\hat{F} = Max(F)$。

[0105] 通过卷积层和池化层，将获得的特征图按行顺序展开并连接成向量，然后将其传递到全连接层，通过输出层给出对应的文本自动分类结果和对应的文本分类概率值。

[0106] 通过实验测试，本实施例提供的噪声过滤和自动分类模型，其噪声识别任务的平均F_1值达到了93.07％，文本分类任务的平均F_1值达到了95.61％。

[0107] 应该理解的是，虽然图2-3的流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，这些步骤可以以其它的顺序执行。而且，图2-3中的至少一部分步骤可以包括多个子步骤或者多个阶段，这些子步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，这些子步骤或者阶段的执行顺序也不必然是依次进行，而是可以与其它步骤或者其它步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

[0108] 一种用于互联网文本信息的噪声过滤和自动分类装置，其特征在于，所述装置包括：

[0109] 模型构建单元，用于构建用于互联网文本信息的噪声过滤和自动分类模型。噪声过滤和自动分类模型包括预设的噪声检测模块，预设的噪声过滤模块和预设的自动分类模块。噪声检测模块基于FastText模型建立。

[0110] 互联网文本信息输入单元，用于将预先获取的互联网文本信息输入噪声检测模块。

[0111] 互联网文本信息噪声检测与过滤单元，用于当检测到互联网文本信息包含噪声时，将互联网文本信息输入噪声过滤模块，根据互联网文本信息中文本对象间的语义相似度值进行过滤，得到不包含噪声的互联网文本信息。

[0112] 模型训练单元，用于根据不包含噪声的互联网文本信息生成样本集，使用样本集训练自动分类模块，得到训练好的噪声过滤和自动分类模型。

[0113] 互联网文本信息分类单元，用于将预先获取的互联网文本信息输入训练好的噪声过滤和自动分类模型，得到互联网文本分类结果。

[0114] 其中一个实施例中，还包括噪声检测模块构建单元，用于：根据预设的规则标注预先获取的互联网文本数据中的噪声数据和非噪声数据，得到用于模型训练的噪声二分类数据集。将噪声二分类数据集输入预设的FastText文本识别模型，得到训练好的噪声检测模块。

[0115] 其中一个实施例中，噪声过滤模块基于BERT模型建立，噪声检测与过滤单元用于：当检测到互联网文本信息包含噪声时，获取互联网文本信息中的标题文本和正文文本，将正文文本按照预设的规则拆分为正文段落文本。将标题文本和正文段落文本依次输入噪声过滤模块，计算标题文本和正文段落文本间的语义相似度值，当正文段落文本和标题文本间的语义相似度值低于预设值时，将该正文段落文本标记为噪声。按照正文文本中正文段落文本的先后顺序，拼接未标记为噪声的所述正文段落文本，输出不包含噪声的互联网文本信息。

[0116] 其中一个实施例中，噪声检测与过滤单元用于：将标题文本和正文段落文本依次输入噪声过滤模块，根据预设的余弦相似度算法计算标题文本和正文段落文本间的语义相似度值。当正文段落文本和标题文本间的语义相似度值低于预设值时，将正文段落文本标记为噪声。

[0117] 其中一个实施例中，自动分类模块基于卷积神经网络，包括输入层、词嵌入层、卷积层、最大池化层、全连接层和输出层。所述装置还包括自动分类模块构建单元，用于使用反向传播方法确定自动分类模块的卷积层参数。

[0118] 其中一个实施例中，模型训练单元用于：

[0119] 根据不包含噪声的互联网文本信息生成样本集，将样本集通过输入层输入自动分类模块。

[0120] 由词嵌入层、卷积层和最大池化层提取文本特征向量，由全连接层通过输出层输出互联网文本信息分类结果。

[0121] 根据自动分类模块输出的文本自动分类结果和对应的文本分类概率值，得到训练好的噪声过滤和自动分类模型。

[0122] 关于一种用于互联网文本信息的噪声过滤和自动分类装置的具体限定可以参见上文中对于一种用于互联网文本信息的噪声过滤和自动分类方法的限定，在此不再赘述。上述一种用于互联网文本信息的噪声过滤和自动分类装置中的各个单元可全部或部分通过软件、硬件及其组合来实现。上述各单元可以硬件形式内嵌于或独立于计算机设备中的处理器中，也可以以软件形式存储于计算机设备中的存储器中，以便于处理器调用执行以上各个单元对应的操作。

[0123] 在一个实施例中，提供了一种计算机设备，该计算机设备可以是服务器，其内部结构图可以如图6所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口和数据库。其中，该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统、计算机程序和数据库。该内存储器为非易失性存储介质中的操作系统和计算机程序的运行提供环境。该计算机设备的数据库用于存储互联网文本信息、噪声检测模块、噪声过滤模块、自动分类模块以及一种用于互联网文本信息的噪声过滤和自动分类方法的文本处理过程数据。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机程序被处理器执行时以实现一种用于互联网文本信息的噪声过滤和自动分类方法。

[0124] 本领域技术人员可以理解，图6中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定，具体的计算机设备可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

[0125] 在一个实施例中，提供了一种计算机设备，包括存储器和处理器，该存储器存储有计算机程序，该处理器执行计算机程序时实现以下步骤：

[0126] 构建用于互联网文本信息的噪声过滤和自动分类模型。噪声过滤和自动分类模型包括预设的噪声检测模块，预设的噪声过滤模块和预设的自动分类模块。噪声检测模块基于FastText模型建立。

[0127] 将预先获取的互联网文本信息输入噪声检测模块。

[0128] 当检测到互联网文本信息包含噪声时，将互联网文本信息输入噪声过滤模块，根据互联网文本信息中文本对象间的语义相似度值进行过滤，得到不包含噪声的互联网文本信息。

[0129] 根据不包含噪声的互联网文本信息生成样本集，使用样本集训练自动分类模块，得到训练好的噪声过滤和自动分类模型。

[0130] 将预先获取的互联网文本信息输入训练好的噪声过滤和自动分类模型，得到互联网文本分类结果。

[0131] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：根据预设的规则标

注预先获取的互联网文本数据中的噪声数据和非噪声数据，得到用于模型训练的噪声二分类数据集。将噪声二分类数据集输入预设的FastText文本识别模型，得到训练好的噪声检测模块。

[0132] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：当检测到互联网文本信息包含噪声时，获取互联网文本信息中的标题文本和正文文本，将正文文本按照预设的规则拆分为正文段落文本。将标题文本和正文段落文本依次输入噪声过滤模块，计算标题文本和正文段落文本间的语义相似度值，当正文段落文本和标题文本间的语义相似度值低于预设值时，将该正文段落文本标记为噪声。按照正文文本中正文段落文本的先后顺序，拼接未标记为噪声的正文段落文本，输出不包含噪声的互联网文本信息。

[0133] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：将标题文本和正文段落文本输入预设的噪声过滤模块，根据预设的余弦相似度算法计算标题文本和正文段落文本间的语义相似度值。当正文段落文本和标题文本间的语义相似度值低于预设值时，将正文段落文本标记为噪声。

[0134] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：使用反向传播方法确定自动分类模块的卷积层参数。

[0135] 其中一个实施例中，处理器执行计算机程序时还实现以下步骤：根据不包含噪声的互联网文本信息生成样本集，将样本集通过输入层输入自动分类模块。由词嵌入层、卷积层和最大池化层提取文本特征向量，由全连接层通过输出层输出互联网文本信息分类结果。根据自动分类模块输出的文本自动分类结果和对应的文本分类概率值，得到训练好的噪声过滤和自动分类模型。

[0136] 在一个实施例中，提供了一种计算机可读存储介质，其上存储有计算机程序，计算机程序被处理器执行时实现以下步骤：

[0137] 构建用于互联网文本信息的噪声过滤和自动分类模型。噪声过滤和自动分类模型包括预设的噪声检测模块，预设的噪声过滤模块和预设的自动分类模块。噪声检测模块基于FastText模型建立。

[0138] 将预先获取的互联网文本信息输入噪声检测模块。

[0139] 当检测到互联网文本信息包含噪声时，将互联网文本信息输入噪声过滤模块，根据互联网文本信息中文本对象间的语义相似度值进行过滤，得到不包含噪声的互联网文本信息。

[0140] 根据不包含噪声的互联网文本信息生成样本集，使用样本集训练和测试自动分类模块，得到训练好的噪声过滤和自动分类模型。

[0141] 将预先获取的互联网文本信息输入训练好的噪声过滤和自动分类模型，得到互联网文本分类结果。

[0142] 在一个实施例中，计算机程序被处理器执行时还实现以下步骤：根据预设的规则标注预先获取的互联网文本数据中的噪声数据和非噪声数据，得到用于模型训练的噪声二分类数据集。将噪声二分类数据集输入预设的FastText文本识别模型，得到训练好的噪声检测模块。

[0143] 在一个实施例中，计算机程序被处理器执行时还实现以下步骤：当检测到互联网文本信息包含噪声时，获取互联网文本信息中的标题文本和正文文本，将正文文本按照预

设的规则拆分为正文段落文本。将标题文本和正文段落文本依次输入噪声过滤模块，计算标题文本和正文段落文本间的语义相似度值，当正文段落文本和标题文本间的语义相似度值低于预设值时，将该正文段落文本标记为噪声。按照正文文本中正文段落文本的先后顺序，拼接未标记为噪声的正文段落文本，输出不包含噪声的互联网文本信息。

[0144] 在一个实施例中，计算机程序被处理器执行时还实现以下步骤：将标题文本和正文段落文本依次输入噪声过滤模块，根据预设的余弦相似度算法计算标题文本和正文段落文本间的语义相似度值。当正文段落文本和标题文本间的语义相似度值低于预设值时，将正文段落文本标记为噪声。

[0145] 在一个实施例中，计算机程序被处理器执行时还实现以下步骤：使用反向传播方法确定自动分类模块的卷积层参数。

[0146] 其中一个实施例中，计算机程序被处理器执行时还实现以下步骤：根据不包含噪声的互联网文本信息生成样本集，将样本集通过输入层输入自动分类模块。由词嵌入层、卷积层和最大池化层提取文本特征向量，由全连接层通过输出层输出互联网文本信息分类结果。根据自动分类模块输出的文本自动分类结果和对应的文本分类概率值，得到训练好的噪声过滤和自动分类模型。

[0147] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，所述的计算机程序可存储于一非易失性计算机可读取存储介质中，该计算机程序在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用，均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限，RAM以多种形式可得，诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双数据率SDRAM(DDRSDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink)DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)等。

[0148] 以上实施例的各技术特征可以进行任意的组合，为使描述简洁，未对上述实施例中的各个技术特征所有可能的组合都进行描述，然而，只要这些技术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

[0149] 以上所述实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对发明专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干变形和改进，这些都属于本申请的保护范围。因此，本申请专利的保护范围应以所附权利要求为准。

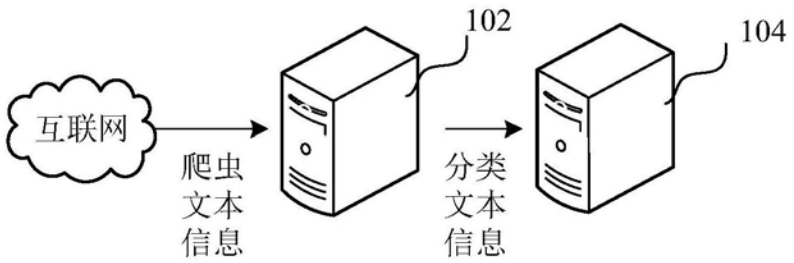

互联网
爬虫
文本
信息
102
分类
文本
信息
104

图1

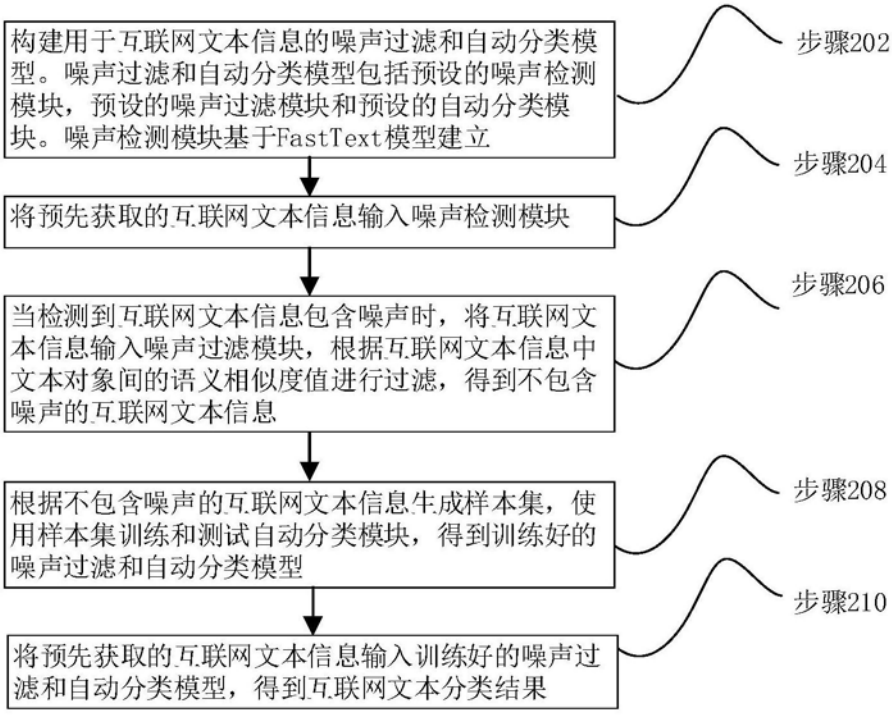

构建用于互联网文本信息的噪声过滤和自动分类模型。噪声过滤和自动分类模型包括预设的噪声检测模块，预设的噪声过滤模块和预设的自动分类模块。噪声检测模块基于FastText模型建立
步骤202
将预先获取的互联网文本信息输入噪声检测模块
步骤204
当检测到互联网文本信息包含噪声时，将互联网文本信息输入噪声过滤模块，根据互联网文本信息中文本对象间的语义相似度值进行过滤，得到不包含噪声的互联网文本信息
步骤206
根据不包含噪声的互联网文本信息生成样本集，使用样本集训练和测试自动分类模块，得到训练好的噪声过滤和自动分类模型
步骤208
将预先获取的互联网文本信息输入训练好的噪声过滤和自动分类模型，得到互联网文本分类结果
步骤210

图2

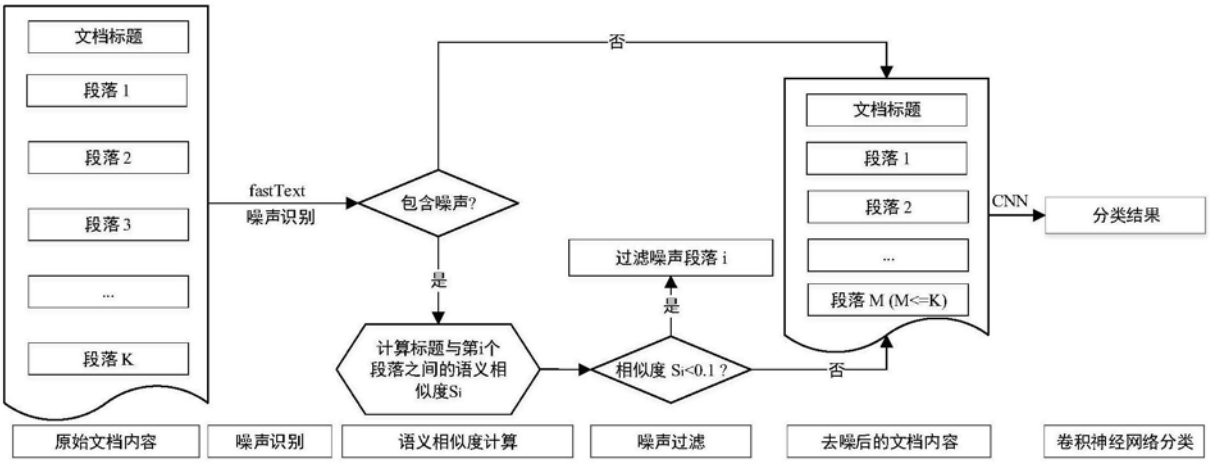

文档标题
段落 1
段落 2
段落 3
...
段落 K
fastText
噪声识别
包含噪声?
否
是
计算标题与第i个段落之间的语义相似度Si
相似度 Si<0.1 ?
是
过滤噪声段落 i
否
文档标题
段落 1
段落 2
...
段落 M (M<=K)
CNN
分类结果
原始文档内容
噪声识别
语义相似度计算
噪声过滤
去噪后的文档内容
卷积神经网络分类

图3

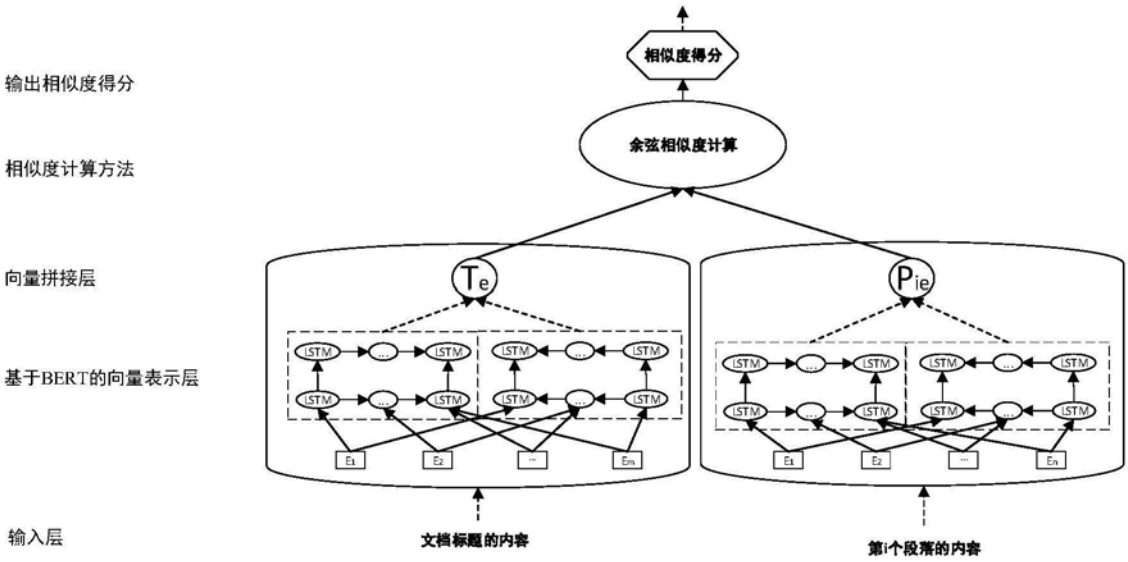

输出相似度得分
相似度计算方法
向量拼接层
基于BERT的向量表示层
输入层
相似度得分
余弦相似度计算
Te
Pie
LSTM
E1
E2
Em
En
文档标题的内容
第i个段落的内容

图4

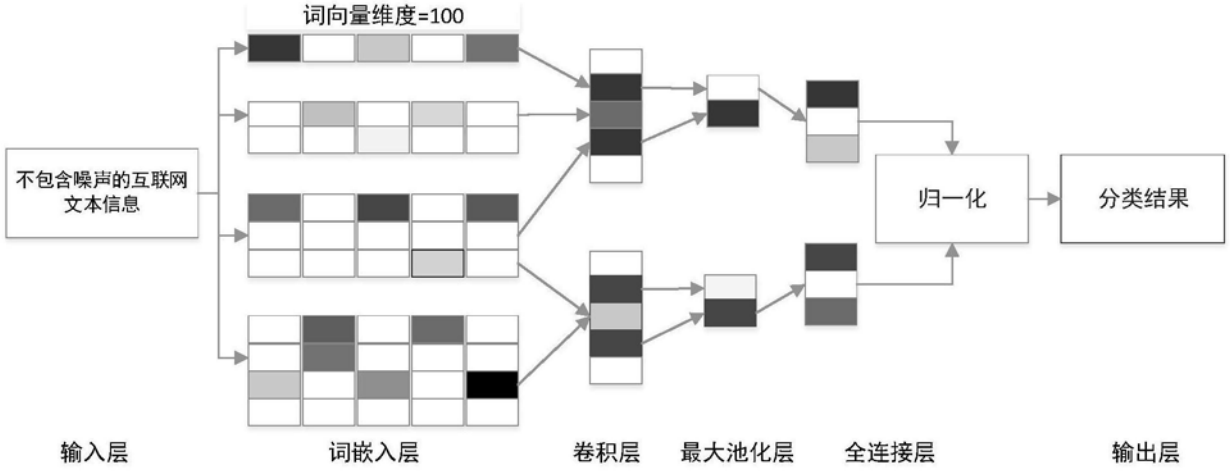

词向量维度=100
不包含噪声的互联网文本信息
归一化
分类结果
输入层
词嵌入层
卷积层
最大池化层
全连接层
输出层

图5

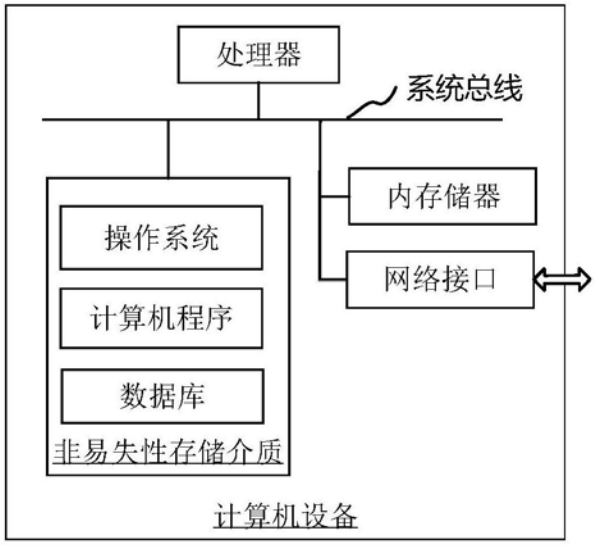

处理器
系统总线
内存储器
网络接口
操作系统
计算机程序
数据库
非易失性存储介质
计算机设备

图6