(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 112131350 A
(43)申请公布日 2020.12.25

(21)申请号 202011065821.5

(22)申请日 2020.09.30

(71)申请人 腾讯科技(深圳)有限公司
地址 518057 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 刘刚

(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 徐立

(51)Int.Cl.
G06F 16/33(2019.01)
G06F 16/9535(2019.01)
G06F 16/783(2019.01)
G06F 40/30(2020.01)
G06F 40/289(2020.01)

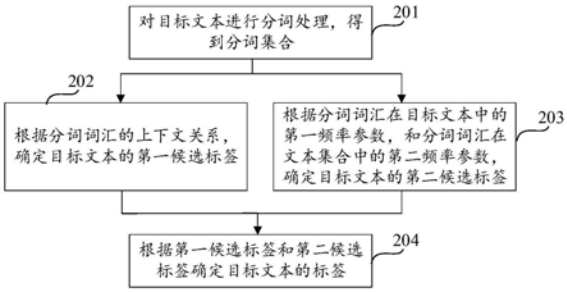
权利要求书2页 说明书13页 附图6页

(54)发明名称

文本标签确定方法、装置、终端及可读存储介质

(57)摘要

本申请是关于一种文本标签确定方法、装置、终端及可读存储介质，属于标签挖掘领域。所述方法包括：对目标文本进行分词处理，得到分词集合，所述分词集合中包括所述目标文本分词得到的分词词汇，所述目标文本为待确定标签的文本；根据所述分词词汇的上下文关系，确定所述目标文本的第一候选标签；根据所述分词词汇在所述目标文本中的第一频率参数，和所述分词词汇在文本集合中的第二频率参数，确定所述目标文本的第二候选标签；根据所述第一候选标签和所述第二候选标签确定所述目标文本的标签。本申请解决了标签确定过程中未考虑词汇在上下文语义环境，导致的标签确定的准确度不高的问题，提高了获取标签的准确度。



CN 112131350 A

1.一种文本标签确定方法,其特征在于,所述方法包括:

对目标文本进行分词处理,得到分词集合,所述分词集合中包括所述目标文本分词得到的分词词汇,所述目标文本为待确定标签的文本;

根据所述分词词汇的上下文关系,确定所述目标文本的第一候选标签;

根据所述分词词汇在所述目标文本中的第一频率参数,和所述分词词汇在文本集合中的第二频率参数,确定所述目标文本的第二候选标签;

根据所述第一候选标签和所述第二候选标签确定所述目标文本的标签。

2.根据权利要求1所述的方法,其特征在于,所述根据所述第一候选标签和所述第二候选标签确定所述目标文本的标签,包括:

响应于所述第一候选标签和所述第二候选标签之间存在交集的情况,对所述第一候选标签和所述第二候选标签取交集,得到所述目标文本的标签。

3.根据权利要求1所述的方法,其特征在于,所述方法还包括:

响应于所述第一候选标签和所述第二候选标签之间不存在交集的情况,根据预设的选择规则,从所述第一候选标签和所述第二候选标签中确定所述目标文本的标签。

4.根据权利要求1至3任一所述的方法,其特征在于,所述根据所述第一频率参数和所述第二频率参数,确定所述目标文本的第二候选标签,包括:

根据所述第一频率参数和所述第二频率参数,确定所述分词词汇对应的词汇频率;

将所述词汇频率符合频率要求的分词词汇,确定为所述目标文本的第二候选标签。

5.根据权利要求4所述的方法,其特征在于,所述根据所述第一频率参数和所述第二频率参数,确定所述分词词汇对应的词汇频率,包括:

确定所述第一频率参数和所述第二频率参数之积,作为所述分词词汇对应的所述词汇频率。

6.根据权利要求1至3任一所述的方法,其特征在于,所述根据所述分词词汇的上下文关系,确定所述目标文本的第一候选标签,包括:

对所述分词词汇进行特征提取,得到所述分词词汇的词汇向量;

对所述词汇向量结合上下文词汇向量进行特征分析,得到所述分词词汇对应的实体概率;

根据所述实体概率从所述分词词汇中确定所述目标文本的第一候选标签。

7.根据权利要求6所述的方法,其特征在于,所述实体概率中包括第一概率、第二概率和第三概率,所述第一概率表示所述分词词汇属于标签实体的概率,所述第二概率表示所述分词词汇不属于所述标签实体的概率,所述第三概率表示所述分词词汇属于标签实体内对应实体的概率;

所述对所述词汇向量结合上下文词汇向量进行特征分析,得到所述分词词汇对应的实体概率,包括:

对所述词汇向量结合上下文词汇向量进行特征分析,得到所述词汇向量对应的所述第一概率、所述第二概率以及所述第三概率;

所述根据所述实体概率从所述分词词汇中确定所述目标文本的第一候选标签,包括:

从所述第一概率、所述第二概率以及所述第三概率中,将数值最高的概率确定为所述分词词汇的所述实体概率;

将所述实体概率对应的所述分词词汇作为所述第一候选标签。

8.根据权利要求7所述的方法，其特征在于，所述方法，还包括：

滤除所述实体概率对应为所述第二概率的所述分词词汇；

根据所述第一概率和所述第三概率确定所述实体概率对应的所述分词词汇，得到所述第一候选标签。

9.一种文本标签确定装置，其特征在于，所述装置包括：

处理模块，用于对目标文本进行分词处理，得到分词集合，所述分词集合中包括所述目标文本分词得到的分词词汇，所述目标文本为待确定标签的文本；

确定模块，用于根据所述分词词汇的上下文关系，确定所述目标文本的第一候选标签；

所述确定模块，还用于根据所述分词词汇在所述目标文本中的第一频率参数，和所述分词词汇在文本集合中的第二频率参数，确定所述目标文本的第二候选标签；

所述确定模块，还用于根据所述第一候选标签和所述第二候选标签确定所述目标文本的标签。

10.根据权利要求9所述的装置，其特征在于，所述确定模块，还用于响应于所述第一候选标签和所述第二候选标签之间存在交集的情况，对所述第一候选标签和所述第二候选标签取交集，得到所述目标文本的标签。

11.根据权利要求9所述的装置，其特征在于，所述确定模块，还用于响应于所述第一候选标签和所述第二候选标签之间不存在交集的情况，根据预设的选择规则，从所述第一候选标签和所述第二候选标签中确定所述目标文本的标签。

12.根据权利要求9至11任一所述的装置，其特征在于，所述确定模块还包括：

第一确定单元，用于根据所述第一频率参数和所述第二频率参数，确定所述分词词汇对应的词汇频率；

所述第一确定单元，还用于将所述词汇频率符合频率要求的分词词汇，确定为所述目标文本的第二候选标签。

13.根据权利要求9至11任一所述的装置，其特征在于，所述确定模块还包括：

提取单元，用于对所述分词词汇进行特征提取，得到所述分词词汇的词汇向量；

分析单元，用于对所述词汇向量结合上下文词汇向量进行特征分析，得到所述分词词汇对应的实体概率；

第二确定单元，用于根据所述实体概率从所述分词词汇中确定所述目标文本的第一候选标签。

14.一种计算机设备，其特征在于，所述计算机设备包括处理器和存储器，所述存储器中存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或指令集由所述处理器加载并执行以实现如权利要求1至8任一所述的文本标签确定方法。

15.一种计算机可读存储介质，其特征在于，所述存储介质中存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或指令集由处理器加载并执行以实现如权利要求1至8任一所述的文本标签确定方法。

文本标签确定方法、装置、终端及可读存储介质

技术领域

[0001] 本申请涉及标签挖掘领域，特别涉及一种文本标签确定方法、装置、终端及可读存储介质。

背景技术

[0002] 标签被定义为能够代表内容的最重要的关键词，在信息流内容分发过程当中，无论是图文还是视频，标签信息都非常重要，当内容有了标签后，就可以按照不同标签组织和展示内容，也可以根据标签和用户画像进行匹配，实现更精准的内容推荐。

[0003] 在相关技术中，文章内容的标签提取方法包括基于TF-IDF统计特征来确定当前内容的标签，该方法倾向于过滤文章中的常见词语，保留重要词语。

[0004] 但该基于统计的方法未考虑文章中词语与词语之间、词语与文档之间的关系，所获取到的标签与内容表达的实际语义存在偏差，获取的标签准确度不高。

发明内容

[0005] 本申请提供了一种文本标签确定方法、装置、终端及可读存储介质，能够提高标签确定的准确度。所述技术方案如下：

[0006] 一方面，提供了一种文本标签确定方法，所述方法包括：

[0007] 对目标文本进行分词处理，得到分词集合，所述分词集合中包括所述目标文本分词得到的分词词汇，所述目标文本为待确定标签的文本；

[0008] 根据所述分词词汇的上下文关系，确定所述目标文本的第一候选标签；

[0009] 根据所述分词词汇在所述目标文本中的第一频率参数，和所述分词词汇在文本集合中的第二频率参数，确定所述目标文本的第二候选标签；

[0010] 根据所述第一候选标签和所述第二候选标签确定所述目标文本的标签。

[0011] 另一方面，提供了一种文本标签确定装置，所述装置包括：

[0012] 处理模块，用于对目标文本进行分词处理，得到分词集合，所述分词集合中包括所述目标文本分词得到的分词词汇，所述目标文本为待确定标签的文本；

[0013] 确定模块，用于根据所述分词词汇的上下文关系，确定所述目标文本的第一候选标签；

[0014] 所述确定模块，还用于根据所述分词词汇在所述目标文本中的第一频率参数，和所述分词词汇在文本集合中的第二频率参数，确定所述目标文本的第二候选标签；

[0015] 所述确定模块，还用于根据所述第一候选标签和所述第二候选标签确定所述目标文本的标签。

[0016] 另一方面，提供了一种计算机设备，所述计算机设备包括处理器和存储器，所述存储器中存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或指令集由所述处理器加载并执行以实现如上述本申请实施例中任一所述的文本标签确定方法。

[0017] 另一方面，提供了一种计算机可读存储介质，所述存储介质中存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或指令集由处理器加载并执行以实现如上述本申请实施例中任一所述的文本标签确定方法。

[0018] 另一方面，提供了一种计算机程序产品或计算机程序，该计算机程序产品或计算机程序包括计算机指令，该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令，处理器执行该计算机指令，使得该计算机设备执行上述实施例中任一所述的文本标签确定方法。

[0019] 本申请实施例提供的技术方案带来的有益效果至少包括：

[0020] 在通过对目标文本的每个分词词汇的上下文关系，及每个分词词汇在目标文本中的第一频率参数、在文本集合中的第二频率参数共同确定目标文本的标签，从深层语义和浅层频率两方面共同确定目标文本的标签，提高了目标文本标签确定的准确度。

附图说明

[0021] 为了更清楚地说明本申请实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0022] 图1是本申请一个示例性实施例提供的实施环境示意图；

[0023] 图2是本申请一个示例性实施例提供的文本标签确定方法的流程图；

[0024] 图3是本申请另一个示例性实施例提供的文本标签确定方法的流程图；

[0025] 图4是本申请一个示例性实施例提供的基于双向长短时记忆神经网络的文本标签确定模型；

[0026] 图5是本申请一个示例性实施例提供的注意力计算机制示意图；

[0027] 图6是本申请一个示例性实施例提供的注意力计算流程图；

[0028] 图7是本申请一个示例性实施例提供的条件随机场确定标签的方法流程图；

[0029] 图8是本申请一个示例性实施例提供的文本标签确定方法的系统流程示意图；

[0030] 图9是本申请一个示例性实施例提供的标签确定的系统流程示意图；

[0031] 图10是本申请一个示例性实施例提供的文本标签确定装置的结构框图；

[0032] 图11是本申请另一个示例性实施例提供的文本标签确定装置的结构框图；

[0033] 图12是本申请一个示例性实施例提供的服务器的结构框图。

具体实施方式

[0034] 为使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请实施方式作进一步地详细描述。

[0035] 首先，针对本申请实施例中涉及的名词进行简单介绍：

[0036] 人工智能(Artificial Intelligence，AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说，人工智能是计算机科学的一个综合技术，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能

也就是研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。

[0037] 人工智能技术是一门综合学科，涉及领域广泛，既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

[0038] 计算机视觉技术(Computer Vision,CV)计算机视觉是一门研究如何使机器“看”的科学，更进一步的说，就是指用摄影机和电脑代替人眼对目标进行识别、跟踪和测量等机器视觉，并进一步做图形处理，使电脑处理成为更适合人眼观察或传送给仪器检测的图像。作为一个科学学科，计算机视觉研究相关的理论和技术，试图建立能够从图像或者多维数据中获取信息的人工智能系统。计算机视觉技术通常包括图像处理、图像识别、图像语义理解、图像检索、光学字符识别(Optical Character Recognition,OCR)、视频处理、视频语义理解、视频内容/行为识别、三维物体重建、三维技术、虚拟现实、增强现实、同步定位与地图构建等技术，还包括常见的人脸识别、指纹识别等生物特征识别技术。

[0039] 语音技术(Speech Technology)的关键技术有自动语音识别技术(Automatic Speech Recognition,ASR)和语音合成技术(Text To Speech,TTS)以及声纹识别技术。让计算机能听、能看、能说、能感觉，是未来人机交互的发展方向，其中语音成为未来最被看好的人机交互方式之一。

[0040] 自然语言处理(Nature Language Processing,NLP)是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。因此，这一领域的研究将涉及自然语言，即人们日常使用的语言，所以它与语言学的研究有着密切的联系。自然语言处理技术通常包括文本处理、语义理解、机器翻译、机器人问答、知识图谱等技术。

[0041] 机器学习(Machine Learning,ML)是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、示教学习等技术。

[0042] 词向量(word embedding)，是嵌入式自然语言处理中的一组语言建模和特征学习技术的统称，其中来自词汇表的单词或短语被映射到实数的向量。从概念上讲，它涉及从每个单词一维的空间到具有更低维度的连续向量空间的数学嵌入。生成这种映射的方法包括神经网络，单词共生矩阵的降维，概率模型，可解释的知识库方法，和术语的显式表示单词出现的背景。当用作底层输入表示时，单词和短语嵌入已经被证明可以提高自然语言处理任务的性能，例如语法分析和情感分析。

[0043] 标签，在推荐系统中，标签被定义为能够代表文章语义的最重要的关键词，并且适合用于用户画像和内容的匹配项，相比于分类和主题，是更细粒度的语义。存在于推荐系统各个环节，内容画像维度、用户画像维度、召回模型特征、排序模型特征、多样性打散等。

[0044] 本申请实施例提供的方案涉及人工智能的自然语言处理和机器学习等技术，其

中，本申请提供的文本标签确定方法可以应用于如下场景中的至少一种：

[0045] 第一，该文本标签确定方法应用于文章阅读平台的推荐系统中，该文章包括新闻文章、公众号文章、个人原创文章、书籍内容等。服务器接收用户或合作方通过终端上传的文章内容，推荐系统对内容进行标签挖掘，根据标签将内容分发至相关用户。例如，一篇报道最近房价的变化趋势和政府实施新的相关政策的新闻，服务器的推荐系统将标签确定为“房价、政策”，将其推送至最近关注过房价变化趋势的用户对应应用程序的首页或其他页面，增加该新闻被点击的可能性。

[0046] 第二，该文本标签确定方法应用于社交平台的推荐系统中，该社交平台包括博客、微型博客、社交软件的公共互动模块。服务器接收用户通过终端上传的内容，推荐系统对内容进行标签挖掘，根据标签将内容进行分类并定向推送。在一个示例中，在一个社交应用程序的“广场”上，用户可以以文字、图片、视频等形式发布动态内容，服务器的推荐系统对其标签进行挖掘，例如发布内容为“今天到北京旅游，逛了故宫和鸟巢，北京烤鸭真好吃”，推荐系统将标签确定为“北京、旅游”，并将其推荐给定位为北京的用户或推荐给也曾发布过“旅游”相关动态的用户，得以推荐更精准垂直的目标用户。

[0047] 第三，该文本标签确定方法应用于视频平台的推荐系统中，该视频平台可以是提供普通视频、短视频等视频内容。服务器接收用户通过终端上传内容或供应商提供的视频内容，对视频内容进行语音转文字处理或光学字符识别字幕获取视频对应的文本，针对文本进行标签挖掘，并根据用户画像进行推荐。例如，从视频中提取到的文字内容为“苹果公司即将上线iphone12，手机支持……”，推荐系统对标签进行确定为“苹果、手机”，并针对苹果和手机的关系，确定该视频属于电子产品领域，对该视频进行推荐，将其推荐给近期关注过手机买卖的用户，并不会将其推荐给常关注农业领域的用户。

[0048] 值得注意的是，上述应用场景仅为示意性的举例，本申请提供的文本标签确定可以应用于其他标签确定的应用场景中。

[0049] 结合上述名词简介和应用场景的介绍，对本申请实施例的实施环境进行说明。

[0050] 示意性的，请参考图1，以该文本标签确定方法应用于服务器中为例进行说明，该文本标签确定方法的实施环境包括：终端110、服务器120以及通信网络130；

[0051] 终端110中安装有相关的应用程序，该相关的应用程序111可以是文章阅读平台对应的应用程序、社交平台对应的应用程序或视频平台对应的应用程序，用户可以通过终端110上传想要发布的内容，也可以通过终端110接收从服务器120获取的内容，也即，终端110是内容生产端也是内容消费端。终端110与服务器120之间通过通信网络130进行数据传输。

[0052] 服务器120中包括内容推荐模块140和存储模块150，其中，内容推荐模块140包括机器处理模块141、内容调度模块142、人工审核模块143、排重模块144、接口模块145。

[0053] 内容推荐模块140用于抽取内容有效的实体信息作为标签，将内容标签与用户画像进行匹配，实现更有针对性的内容分发和推荐。

[0054] 其中，机器处理模块141包括标签确定模块和标签确定模型，用于对输入的内容进行标签确定，并将处理结果发送至内容调度模块142。

[0055] 内容调度模块142用于负责内容流转的整个调度过程，接收接口模块145的内容，从内容数据库模块153中获取内容的元信息；内容调度模块142还用于与内容存储模块151交互，从内容存储模块151存储内容或读取内容；内容调度模块142还用于调度人工审核模

块143和机器处理模块141，控制调度的顺序和优先级；内容调度模块142还用于获取通过人工审核模块143审核的内容，通过接口模块145中的出口单元分发至终端110。

[0056] 人工审核模块143是人工服务能力的载体，用于审核过滤政治敏感、色情、法律不允许等机器无法确定判断的内容。

[0057] 排重模块144用于和内容调度模块142通讯，用于标题去重，封面图的图片去重，内容正文去重及视频指纹和音频指纹去重。在一个示例中，将图文内容标题和正文向量化，对图片向量去重，对于视频内容抽取视频指纹和音频指纹构建向量，然后计算向量之间的距离，如欧式距离来确定是否重复。

[0058] 接口模块145包括入口单元和出口单元，入口单元用于接收终端110的内容输入，出口单元用于向终端110输出内容。

[0059] 存储模块150中包括内容数据库模块153、语料库152、内容存储模块151、爬虫模块154。

[0060] 其中，内容数据库模块153用于存储内容的元信息，上述元信息包括文件大小数据、封面图链接信息、码率数据、文件格式、标题信息、发布时间数据、作者信息、视频文件大小数据、视频格式信息中的至少一个数据信息；内容数据库模块153还用于存储人工审核模块143对内容的审核结果和审核状态；内容数据库模块153还用于存储通过内容调度模块142传输的机器处理模块141、排重模块144的处理结果。

[0061] 内容存储模块151用于存储内容的元信息之外的内容实体信息，例如，视频源文件和图文内容的图片源文件；内容存储模块151还用于在视频内容标签确定的时候，为机器处理模块141提供视频源文件，包括源文件中间的抽帧内容、视频文件的语音转文字和光学字符识别提供原始素材的输入。

[0062] 爬虫模块154用于通过互联网上获取外部领域的语料，作为领域语料信息，存储在语料库152中，语料库152中存储有分词序列与词向量的对应关系，可选的，词向量通过用来产生词向量的相关模型(word to vector，word2vec)从大规模语料中无监督学习得到。

[0063] 图2是本申请一个示例性实施例提供的文本标签确定方法的流程图，本申请实施例中，以该方法应用在服务器中为例进行说明，请参考图2，该方法包括：

[0064] 步骤201，对目标文本进行分词处理，得到分词集合。

[0065] 在本申请实施例中，服务器获取终端上传的文字、图片或视频资源。对于文字资源，直接生成目标文本。对于图片资源，服务器通过光学字符识别获取图片资源中的文字信息，生成目标文本。对于视频资源，服务器通过提取视频中的音频，对音频进行转文字处理，生成目标文本；或，抽取视频帧，对视频帧进行光学字符识别，生成目标文本。该目标文本为待确定标签的文本。

[0066] 上述分词集合中包括对目标文本分词得到的分词词汇，也即，分词集合是通过对一段文本或一句话分割成几个单独的词汇得到的集合，例如“从吴晓波到罗振宇，知识付费IP有哪些脆弱点?”进行分词得到“从”、“吴晓波”、“到”、“罗振宇”、“，”、“知识付费”、“IP”、“有”、“哪些”、“脆弱点”、“?”11个分词词汇，这11个分词词汇组成分词集合。

[0067] 在分词过程中，存在多种分词方式，在相同字数的情况下，总词数越少，说明语义单元越少，相对于单个语义单元的权重越大，准确度越高，示意性的“知识付费”可以分词为“知识”和“付费”两个分词或“知识付费”一个分词，而在“从吴晓波到罗振宇，知识付费IP有

哪些脆弱点?”的语义理解中,“知识付费”作为一个分词来作为标签提取,标签的指向性会更准确。

[0068] 可选的,在分词过程中,使用基于词典的分词方法。给定一个分词的最大长度,以该长度进行切分,比对得到的分词是否在词典中出现,如果出现的话,该分词就是切词的结果,否则缩短词的长度。其中,切分方式分为正向切割和反向切割,分别得到正向切割结果和反向切割结果,选择长度最小的结果作为分词词汇集合。

[0069] 可选的,在分词过程中,使用基于隐含马尔可夫模型(Hiden Markov Model,HMM)的分词方法。获取每个状态之间转移的概率和每个状态产生该词概率,以及最初的状态,也即,每个状态在句首的概率;从最初状态开始,求状态序列集合{S,B,M,E}中四个状态分别可以产生第一个字符串的概率,并记录每个状态下的概率;判断第二个字符,首先由上一个时间状态下的状态转移到当前时间状态,并且取当前时间下每个状态能得到当前位置字符概率最大的那个,然后在记录一下上一个状态;依次处理每个字符;最后根据状态序列进行分词,得到分词词汇集合。

[0070] 可选的,在分词过程中,使用基于二元语法的分词方法。遍历所有以当前位置为结尾的词,其中,词的长度有限制,找到该词之后再往前找到前一个词,也即,这两个词之间得到一个切分结果,最后求所有的切分结果中概率最大值,根据其对应的切分结果得到分词词汇集合。

[0071] 步骤202,根据分词词汇的上下文关系,确定目标文本的第一候选标签。

[0072] 在本申请实施例中,对分词词汇进行预设处理,得到分词词汇基于上下文关系的表示,也即,同时考虑该分词词汇前的上下文信息,和该分词词汇后的上下文信息,对人类用户一般读完整个句子才更好的确定词的重要度进行很好的拟合,从而能够更有效进行第一候选标签的确定。

[0073] 可选的,将分词词汇输入机器学习模型中,输出得到目标文本的第一候选标签。该机器学习模型包括双向长短时记忆(Bi Long Short Term Memory,Bi-LSTM)神经网络、条件随机场(Conditional Random Field,CRF)、注意力机制(Attention Mechanism)。

[0074] 可选的,对分词词汇基于词图模型(Text Rank,TR)进行标签提取,其中,该方法是由通过互联网中的超链接关系来确定一个网页排名方法(Page Rank,PR)改进而来,PR是一种基于投票思想设计的方法,当计算网页A的PR值时,需要知道有哪些网页链接到网页A,也就是要首先得到网页A的入链,然后通过入链给网页A的投票来计算网页A的PR值。请参考公式一:

[0075] 公式一:$$S(V_i)=(1-d)+d*\sum_{j\in In(V_i)}\frac{1}{|Out(V_j)|}S(V_j)$$

[0076] 其中,V_i表示某个网页,V_j表示链接到V_i的网页(即V_i的入链),$S(V_i)$表示网页V_i的PR值,$In(V_i)$表示网页V_i的所有入链的集合,$Out(V_j)$表示网页,d表示阻尼系数,是用来克服这个公式中“d*”后面的部分的固有缺陷用的:如果仅仅有求和的部分,那么该公式将无法处理没有入链的网页的PR值,因为这时,根据该公式这些网页的PR值为0,但实际情况却不是这样,所以加入了一个阻尼系数来确保每个网页都有一个大于0的PR值,在一个实例中,在0.85的阻尼系数下,大约100多次迭代PR值就能收敛到一个稳定的值,而当阻尼系数接近1时,需要的迭代次数会陡然增加很多,且排序不稳定。公式中$S(V_j)$前面的分数指的是V_j所

有出链指向的网页应该平分V_j的PR值，这样才算是把当前票分给了当前链接到的网页。

[0077] TR是由PR改进而来，在计算过程中，仅多了一个权重项W_{ji}，用来表示两个节点之间的边连接有不同的重要程度。该方法包括：(1)将目标文本按照完整的句子进行分割；(2)对于每个句子，进行分词和词性标注处理，并过滤停用词，只保留指定词性的单词，如名词、动词、形容词，即，得到句子集合，其中集合中的元素为保留的候选标签；(3)构建候选关键词图G＝(V,E)，其中V为节点集，由(2)生成的候选关键词组成，然后采用共现关系构造任两点之间的边，两个节点之间存在边仅当它们对应的词汇在长度为m的窗口中共现，m表示窗口大小，即最多共现m个单词；(4)计算迭代传播各节点的权重，直至收敛；(5)对节点权重进行倒序排序，从而得到最重要的T个词汇，作为候选标签；(6)在原始的目标文本中标记(5)中得到的T个词汇，若形成相邻词组，则组合成多词标签。

[0078] 步骤203，根据分词词汇在目标文本中的第一频率参数，和分词词汇在文本集合中的第二频率参数，确定目标文本的第二候选标签。

[0079] 在本申请实施例中，根据第一频率参数$tf_{i,j}$和第二频率参数idf_i，确定分词词汇对应的词汇频率$tfidf_{i,j}$，其中，第一频率参数$tf_{i,j}$的计算方法请参考公式二，第二频率参数idf_i的计算方法请参考公式三：

[0080] 公式二：$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}$$

[0081] 公式三：$$idf_i=\lg\frac{|D|}{|\{j:t_i\in d_j\}|}$$

[0082] 其中，$tf_{i,j}$表示第一频率参数，idf_i表示第二频率参数，$n_{i,j}$表示当前分词词汇在文本集合j中出现的个数，$|D|$表示语料库中文本个数，$|\{j:t_i\in d_j\}|$表示包含分词词汇的文本集合个数。

[0083] 再确定的第一频率参数和第二频率参数之积，作为分词词汇对应的词汇频率$tfidf_{i,j}$，该词汇频率的计算公式请参考公式四：

[0084] 公式四：$tfidf_{i,j}=tf_{i,j}\times idf_i$

[0085] 将词汇频率符合频率要求的分词词汇，确定为目标文本的第二候选标签。

[0086] 步骤204，根据第一候选标签和第二候选标签确定目标文本的标签。

[0087] 在本申请实施例中，可选的，响应于所述第一候选标签和所述第二候选标签之间存在交集的情况，对第一候选标签和第二候选标签取交集，得到目标文本的标签。响应于第一候选标签和第二候选标签之间不存在交集的情况，根据预设的选择规则，从第一候选标签和第二候选标签中确定目标文本的标签。在一个示例中，该预设的选择规则为给第一候选标签和第二候选标签根据目标文本属于的领域分配不同权重，当第一候选标签的权重大于第二候选标签的权重时，选取第一候选标签作为目标文本的标签，当第一候选标签的权重小于第二候选标签的权重时，选取第二候选标签作为目标文本的标签。

[0088] 综上所述，本实施例提供的文本标签确定方法，在通过对目标文本的每个分词词汇的上下文关系，及每个分词词汇在目标文本中的第一频率参数、在文本集合中的第二频率参数共同确定目标文本的标签，从深层语义和浅层频率两方面共同确定目标文本的标签，提高了目标文本标签确定的准确度，且通过机器学习模型有效地代替人工特征识别工

程,提高了目标文本标签确定的效率。

[0089] 结合上述实施例,以通过机器学习模型得到基于分词词汇的上下文关系的第一候选标签为例,对本申请实施例提供的文本标签确定方法进行说明,请参考图3,其示出了一个示例性实施例提供的文本标签确定方法的流程图,该方法包括:

[0090] 步骤301,对目标文本进行分词处理,得到分词集合。

[0091] 分词集合的获取过程的相关内容请参考上述步骤201,此处不再赘述。

[0092] 步骤302,对分词词汇进行特征提取,得到分词词汇的词汇向量。

[0093] 在本申请实施例中,示意性的,通过爬虫实时获取互联网的大规模语料,并通过无监督方式学习得到分词对应的词汇向量,根据目标文本分解得到的分词集合,得到分词序列,将分词序列输入到语料库中查询得到对应的词汇向量。

[0094] 步骤303,对词汇向量结合上下文词汇向量进行特征分析,得到分词词汇对应的实体概率。

[0095] 在本申请实施例中,对所述词汇向量结合上下文词汇向量进行特征分析,得到分词词汇对应的实体概率,在一个示例中,将词汇向量输入双向长短时记忆神经网络,其中,实体概率中包括第一概率、第二概率和第三概率,第一概率表示分词词汇属于标签实体的概率,第二概率表示分词词汇不属于标签实体的概率,第三概率表示分词词汇属于标签实体内对应实体的概率。例如,若输入的分词词汇为实体,例如人名、地名或机构名等,则其第一概率高;若输入的分词词汇为非实体,例如介词、动词或形容词等,则其第二概率高;若输入的分词词汇为实体且其前一分词词汇为实体,则其第三概率高。

[0096] 之后得到词汇向量对应的第一概率、第二概率以及第三概率。从第一概率、第二概率以及第三概率中,将数值最高的概率确定为分词词汇的实体概率。

[0097] 步骤304,根据实体概率从分词词汇中确定目标文本的第一候选标签。

[0098] 在本申请实施例中,滤除实体概率对应为第二概率的分词词汇,可选的,第二概率对应的分词词汇不属于标签实体,也即,是不可作为标签的分词词汇,如词性为介词的分词词汇。再根据第一概率和第三概率确定实体概率对应的分词词汇,得到第一候选标签。

[0099] 在一个示例中,请参考图4,将词汇向量410输入至双向长短时记忆神经网络(即图4中的BiLSTM)420,分词词汇411与词汇向量410对应,每个输入的词汇向量410都会输出三个预测分值,该三个预测分值分别代表第一概率B、第二概率O和第三概率I430,可选的,将预测分值输入至条件随机场(即图4中的CRF)440,得到预测标签(即图4中各个分词词汇对应的B标签、I标签或O标签)450,根据预测标签得到第一候选标签。

[0100] 可选的,在对词汇向量结合上下文词汇向量进行特征分析之后,增加对分词词汇的自注意力计算,计算机制请参考图5,自注意力的计算方法包括:图中的Value501是双向长短时记忆神经网络的输出词汇向量,Key502是不同于Value501的参数矩阵,注意力函数的本质可以被描述为一个查询元素(即图中的Query,对应当中自注意力计算层)503得到到一系列(键Key502-值Value501)对的映射,即通过计算Query503和各个Key502的相似性或者相关性,得到每个Key502对应Value501的权重系数,然后对Value501进行加权求和,即得到了最终的注意力数值504Attention,以公式表达为如下公式五:

[0101] 公式五:$$Attention(Query,source)=\sum_{i=1}^{L_x}Similarity(Qurey,Key_i)*Value_i$$

[0102] 其中，Attention(Qurey,Source)代表当前句子(Source)和当前元素对应的注意力值，Source代表输入句子，Query代表元素，Key代表参数矩阵，L_x代表句子的长度，Similarity(Qurey,Key_i)参见公式七。

[0103] 计算机制请参考图6，将其归纳为两个过程：第一个过程是根据Query603和Key602计算权重系数，第二个过程根据权重系数对Value601进行加权求和。而第一个过程又可以细分为两个阶段：第一个阶段610根据Query603和Key602计算两者的相似性或者相关性得到中间向量；第二个阶段620对第一阶段610的原始分值进行归一化处理。其中，在第一个阶段610，可以引入不同的函数和计算机制，根据Query603和某个Key602，计算两者的相似性或者相关性，最常见的方法包括：求两者的向量点积，得到第一中间向量(即图6中的S1、S2、S3、S4)612，计算方法请参考公式六611(即图中的S(Q,K))：

[0104] 公式六：$Similarity(Qurey,Key_i) = Qurey \cdot Key_i$

[0105] 其中，Similarty代表向量点积，Query代表元素，Key_i代表第i个参数矩阵。

[0106] 第一阶段610产生的第一中间向量根据具体产生的方法不同其数值取值范围也不一样，第二阶段620引入类似SoftMax归一化621的计算方式对第一阶段的得分进行数值转换得到第二中间向量(即图中的a1、a2、a3、a4)622，一方面可以进行归一化，将原始计算分值整理成所有元素权重之和为1的概率分布；另一方面也可以通过SoftMax的内在机制更加突出重要元素的权重，计算方法请参考公式七：

[0107] 公式七：$$a_i = Soft\max(Sim_i) = \frac{e^{Sim_i}}{\sum_{j=1}^{L_x} e^{Sim_j}}$$

[0108] 其中，a_i为第二中间向量，Softmax()为softmax函数，Sim_i为第一阶段求出的第一中间向量，L_x表示当前句子(Source)的长度。

[0109] 第三阶段630的计算结果即为对应的权重系数，然后进行加权求和即可得到注意力数值604，请参考公式八：

[0110] 公式八：$$Attention(Qurey,Source) = \sum_{i=1}^{L_x} a_i \cdot Value_i$$

[0111] 其中，Attention(Qurey,Source)代表当前句子(Source)和当前元素对应的注意力值，L_x表示当前句子(Source)的长度，a_i为第二中间向量，$Value_i$为双向长短时记忆神经网络的输出向量。

[0112] 步骤305，根据分词词汇在目标文本中的第一频率参数，和分词词汇在文本集合中的第二频率参数，确定目标文本的第二候选标签。

[0113] 可选的，第二候选标签还可以通过词库特征或领域特征得到，示意性的，领域特征包括特定垂类领域的词汇编码，例如苹果在电子设备领域代表一个品牌，在农业领域代表一种水果。

[0114] 第二候选标签的确定方法的相关内容请参考上述步骤203，此处不再赘述。

[0115] 步骤306，根据第一候选标签和第二候选标签确定目标文本的标签。

[0116] 可选的，还可以将从双向长短时记忆神经网络输出的数据进行自注意力计算得到的结果，与第二候选标签共同输入条件随机场模型。使用条件随机场确定标签的方法，请参考图7，其中，特征包括：字、词性、分词边界、特征词、词库，词库包括人名、地名、机构、影视、

小说、音乐、医疗、网络舆情热词等词库。特征收集好之后需要配置特征模版，特征模版需要配置同一个特征不同位置组合，同一个位置不同特征组合，不同位置不同特征组合。在一个示例中，特征确定举例为：(1)如果当前位置701分词702的(词性703为名词)且(是否词库704＝1)且(标签705＝1)，则让特征模板中的t1＝1否则t1＝0，如果权重λ1越大，模型越倾向把词库中的词当成标签。(2)如果上一个位置的分词的词性703是标点，下一个词性703是动词且当前分词是名词，则让特征模板中的t2＝1，如果权重λ2越大，模型越倾向把夹在标点和动词之间的名词当作标题实体。

[0117] 综上所述，本实施例提供的文本标签确定方法，在通过对目标文本的每个分词词汇的上下文关系，及每个分词词汇在目标文本中的第一频率参数、在文本集合中的第二频率参数共同确定目标文本的标签，从深层语义和浅层频率两方面共同确定目标文本的标签，提高了目标文本标签确定的准确度，且通过机器学习模型有效地代替人工特征识别工程，提高了目标文本标签确定的效率。

[0118] 在一个可选的实施例中，以本申请实施例提供的文本标签确定方法应用于短视频应用程序中为例进行说明，对该方法的系统过程进行示意性说明，请参考图8，该系统流程800包括四个主要过程，该四个主要过程包括视频获取810、内容整理820、标签确定830以及视频分发840。

[0119] 视频获取810执行过程中，接收用户通过客户端上传的视频文件，可选的，该客户端对应的应用程序可以是与视频分发840执行过程中接收视频的应用程序相同，两者也可以不同。

[0120] 内容整理820中包括：语料收集821、视频文本转换822、分词提取823；

[0121] 其中，语料收集821在执行过程中，通过爬虫实时获取互联网的大规模语料，并通过无监督方式学习得到分词对应的词汇向量；视频文本转换822在执行过程中，通过抽取视频文件中音频，将音频转换为文字，得到目标文本，或抽取视频帧，通过光学字符识别视频中的文字得到目标文本；分词提取823在执行过程中，将视频文本转换822得到的目标文本分解为多个分词词汇集合。

[0122] 标签确定830在执行过程中，获取内容整理820得到的分词词汇，对分词词汇进行处理确定目标文本的标签，将标签与目标文本的对应关系输出到视频分发840。

[0123] 视频分发840在执行过程中，根据标签和目标文本的对应关系确定标签和视频的对应关系，根据标签对视频进行分类，将视频分发至应用程序对应视频分类的接口中，或，根据用户画像，将视频分发至可能对该视频感兴趣的用户客户端中。

[0124] 其中，请参考图9，标签确定930包括分词映射向量层931、中间向量层932、自注意力计算层933、分词映射外部特征层934、全连接中间层935、全连接层936、多分类层937、约束层938、目标文本分词词汇900。

[0125] 分词词汇分别输入分词映射向量层931和分词映射外部特征层934。在分词映射向量层931中，分词词汇映射成词汇向量，之后输入双向长短时记忆神经网络，也即走前向循环神经网络(Recurrent Neural Network，RNN)层9311和后向循环神经网络层9312，结果输入到中间向量层932，并重复上述循环神经网络的过程，从中间向量层932输出多个线性映射关系9321，并将其输入至自注意力计算层933，经过计算后结果输入全连接层936；在分词映射外部特征层934中，分词词汇输入分词映射外部特征层934，再将结果输出至全连接中

间层935进行连接，将结果输出至全连接层936，与自注意力计算层933输出的结果进行全连接，再将结果输出至多分类层937，对分类结果输入约束层938，最后输出目标文本的标签。

[0126] 图10是本申请一个示例性实施例提供的文本标签确定装置的结构框图，该装置包括：

[0127] 处理模块1010，用于对目标文本进行分词处理，得到分词集合，所述分词集合中包括所述目标文本分词得到的分词词汇，所述目标文本为待确定标签的文本；

[0128] 确定模块1020，用于根据所述分词词汇的上下文关系，确定所述目标文本的第一候选标签；

[0129] 所述确定模块1020，还用于根据所述分词词汇在所述目标文本中的第一频率参数，和所述分词词汇在文本集合中的第二频率参数，确定所述目标文本的第二候选标签；

[0130] 所述确定模块1020，还用于根据所述第一候选标签和所述第二候选标签确定所述目标文本的标签。

[0131] 在一个可选的实施例中，所述确定模块1020，还用于响应于所述第一候选标签和所述第二候选标签之间存在交集的情况，对所述第一候选标签和所述第二候选标签取交集，得到所述目标文本的标签。

[0132] 在一个可选的实施例中，所述确定模块，还用于响应于所述第一候选标签和所述第二候选标签之间不存在交集的情况，根据预设的选择规则，从所述第一候选标签和所述第二候选标签中确定所述目标文本的标签。

[0133] 在一个可选的实施例中，请参考图11，所述确定模块1020还包括：

[0134] 第一确定单元1021，用于根据所述第一频率参数和所述第二频率参数，确定所述分词词汇对应的词汇频率；

[0135] 所述第一确定单元1021，还用于将所述词汇频率符合频率要求的分词词汇，确定为所述目标文本的第二候选标签。

[0136] 在一个可选的实施例中，所述确定模块1020还包括：

[0137] 提取单元1022，用于对所述分词词汇进行特征提取，得到所述分词词汇的词汇向量；

[0138] 分析单元1023，用于对所述词汇向量结合上下文词汇向量进行特征分析，得到所述分词词汇对应的实体概率；

[0139] 第二确定单元1024，用于根据所述实体概率从所述分词词汇中确定所述目标文本的第一候选标签。

[0140] 需要说明的是：上述实施例提供的文本标签确定装置，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，即将设备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标签确定装置与文本标签确定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。

[0141] 本申请还提供了一种服务器，该服务器包括处理器和存储器，存储器中存储有至少一条指令，至少一条指令由处理器加载并执行以实现上述各个方法实施例提供的文本极性识别方法。需要说明的是，该服务器可以是如下图12所提供的服务器。

[0142] 请参考图12，其示出了本申请一个示例性实施例提供的服务器的结构示意图。具

体来讲:所述服务器1200包括中央处理单元(Central Processing Unit,CPU)1201、包括随机存取存储器(Random Access Memory,RAM)1202和只读存储器(Read Only Memory,ROM)1203的系统存储器1204,以及连接系统存储器1204和中央处理单元1201的系统总线1205。所述服务器1200还包括帮助计算机内的各个器件之间传输信息的基本输入/输出系统1206,和用于存储操作系统1213、应用程序1214和其他程序模块1215的大容量存储设备1207。

[0143] 所述基本输入/输出系统1206包括有用于显示信息的显示器1208和用于用户输入信息的诸如鼠标、键盘之类的输入设备1209。其中所述显示器1208和输入设备1209都通过连接到系统总线1205的输入输出控制器1210连接到中央处理单元1201。所述基本输入/输出系统1206还可以包括输入输出控制器1210以用于接收和处理来自键盘、鼠标、或电子触控笔等多个其他设备的输入。类似地,输入输出控制器1210还提供输出到显示屏、打印机或其他类型的输出设备。

[0144] 所述大容量存储设备1207通过连接到系统总线1205的大容量存储控制器(未示出)连接到中央处理单元1201。所述大容量存储设备1207及其相关联的计算机可读介质为服务器1200提供非易失性存储。也就是说,所述大容量存储设备1207可以包括诸如硬盘或者CD-ROM驱动器之类的计算机可读介质(未示出)。

[0145] 不失一般性,所述计算机可读介质可以包括计算机存储介质和通信介质。计算机存储介质包括以用于存储诸如计算机可读指令、数据结构、程序模块或其他数据等信息的任何方法或技术实现的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括RAM、ROM、可擦除可编程只读存储器(Erasable Programmable Read Only Memory,EPROM)、带电可擦可编程只读存储器(Electrically Erasable Programmable Read Only Memory,EEPROM)、闪存或其他固态存储其技术,CD-ROM、数字通用光盘(Digital Versatile Disc,DVD)或其他光学存储、磁带盒、磁带、磁盘存储或其他磁性存储设备。当然,本领域技术人员可知所述计算机存储介质不局限于上述几种。上述的系统存储器1204和大容量存储设备1207可以统称为存储器。

[0146] 存储器存储有一个或多个程序,一个或多个程序被配置成由一个或多个中央处理单元1201执行,一个或多个程序包含用于实现上述文本极性识别方法的指令,中央处理单元1201执行该一个或多个程序实现上述各个方法实施例提供的文本极性识别方法。

[0147] 根据本申请的各种实施例,所述服务器1200还可以通过诸如因特网等网络连接到网络上的远程计算机运行。也即服务器1200可以通过连接在所述系统总线1205上的网络接口单元1211连接到网络1212,或者说,也可以使用网络接口单元1211来连接到其他类型的网络或远程计算机系统(未示出)。

[0148] 所述存储器还包括一个或者一个以上的程序,所述一个或者一个以上程序存储于存储器中,所述一个或者一个以上程序包含用于进行本申请实施例提供的文本极性识别方法中由服务器所执行的步骤。

[0149] 本申请实施例还提供一种计算机设备,该计算机设备包括存储器和处理器,存储器中存储有至少一条指令、至少一段程序、代码集或指令集,至少一条指令、至少一段程序、代码集或指令集由处理器加载并实现上述实施例中任一所述的文本标签确定方法。

[0150] 本申请实施例还提供一种计算机可读存储介质,该可读存储介质中存储有至少一

条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或指令集由所述处理器加载并执行以实现上述实施例中任一所述的文本标签确定方法。

[0151] 本申请还提供了一种计算机程序产品或计算机程序，该计算机程序产品或计算机程序包括计算机指令，该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令，处理器执行该计算机指令，使得该计算机设备执行上述实施例中任一所述的文本标签确定方法。

[0152] 本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通过程序来指令相关的硬件来完成，该程序可以存储于一计算机可读存储介质中，该计算机可读存储介质可以是上述实施例中的存储器中所包含的计算机可读存储介质；也可以是单独存在，未装配入终端中的计算机可读存储介质。该计算机可读存储介质中存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或指令集由所述处理器加载并执行以实现本申请实施例中任一所述的文本标签确定方法。

[0153] 可选地，该计算机可读存储介质可以包括：只读存储器(ROM，Read Only Memory)、随机存取记忆体(RAM，Random Access Memory)、固态硬盘(SSD，Solid State Drives)或光盘等。其中，随机存取记忆体可以包括电阻式随机存取记忆体(ReRAM，Resistance Random Access Memory)和动态随机存取存储器(DRAM，Dynamic Random Access Memory)。上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。

[0154] 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。

[0155] 以上所述仅为本申请的可选实施例，并不用以限制本申请，凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

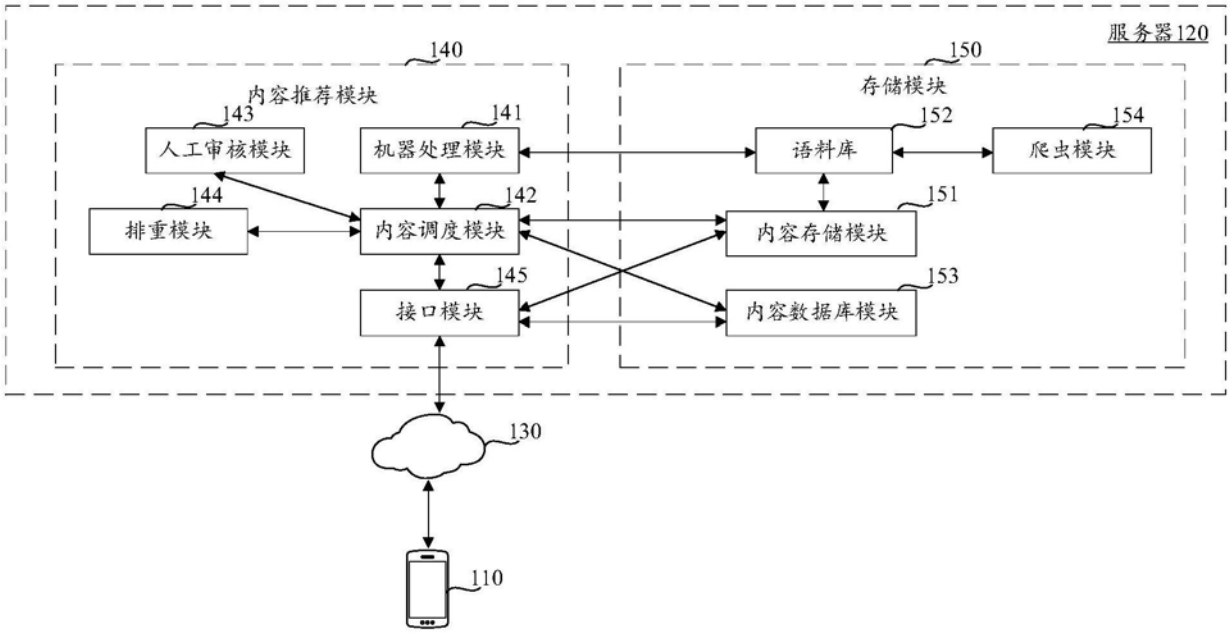

服务器120
140
内容推荐模块
143
人工审核模块
141
机器处理模块
144
排重模块
142
内容调度模块
145
接口模块
150
存储模块
152
语料库
154
爬虫模块
151
内容存储模块
153
内容数据库模块
130
110

图1

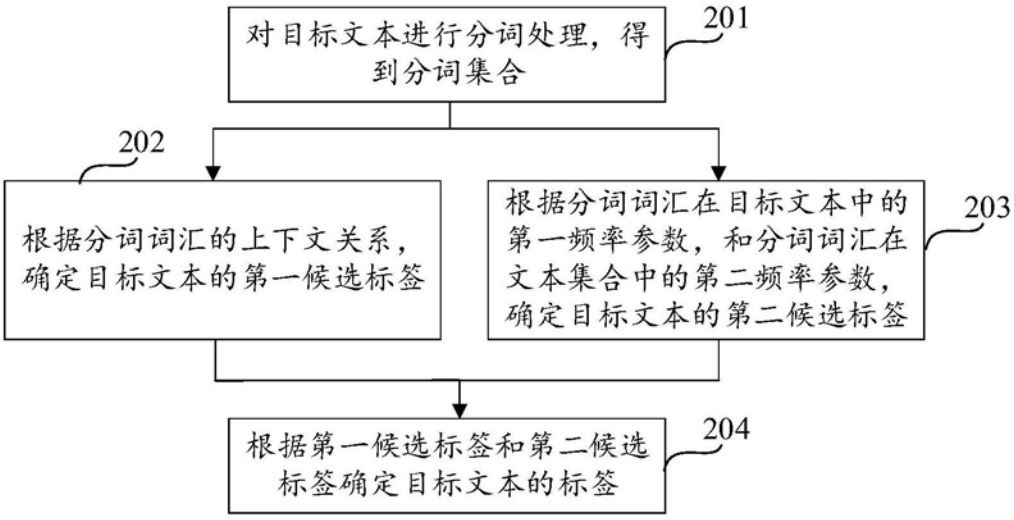

对目标文本进行分词处理，得到分词集合
201
202
根据分词词汇的上下文关系，确定目标文本的第一候选标签
根据分词词汇在目标文本中的第一频率参数，和分词词汇在文本集合中的第二频率参数，确定目标文本的第二候选标签
203
根据第一候选标签和第二候选标签确定目标文本的标签
204

图2

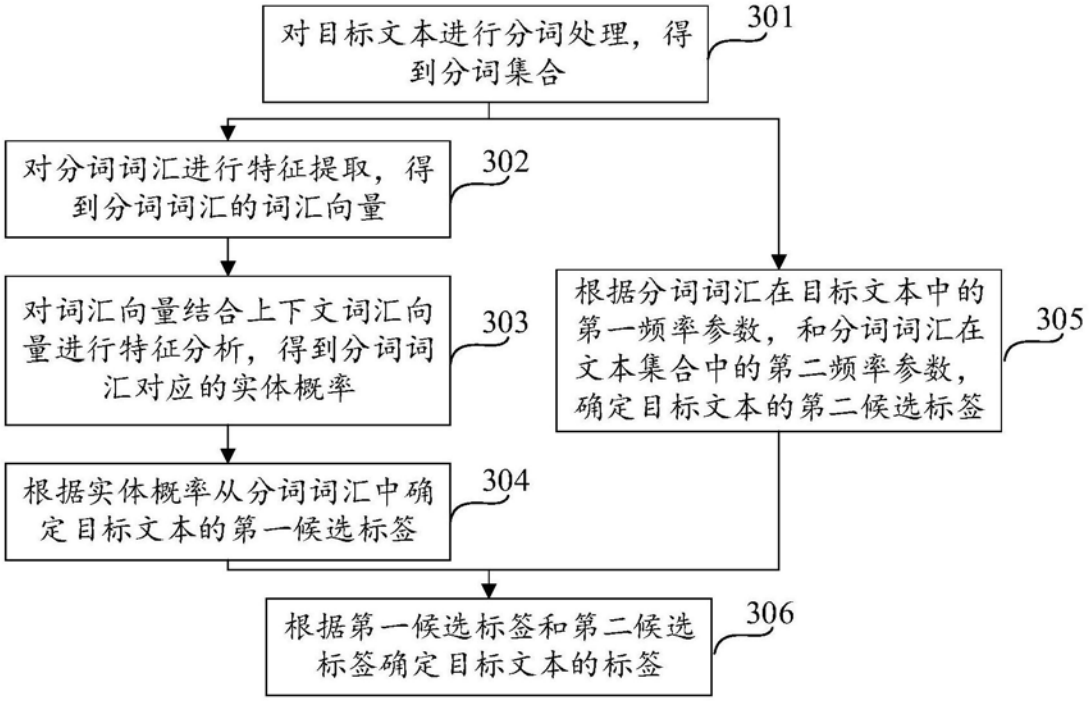

对目标文本进行分词处理，得到分词集合
301
对分词词汇进行特征提取，得到分词词汇的词汇向量
302
对词汇向量结合上下文词汇向量进行特征分析，得到分词词汇对应的实体概率
303
根据实体概率从分词词汇中确定目标文本的第一候选标签
304
根据分词词汇在目标文本中的第一频率参数，和分词词汇在文本集合中的第二频率参数，确定目标文本的第二候选标签
305
根据第一候选标签和第二候选标签确定目标文本的标签
306

图3

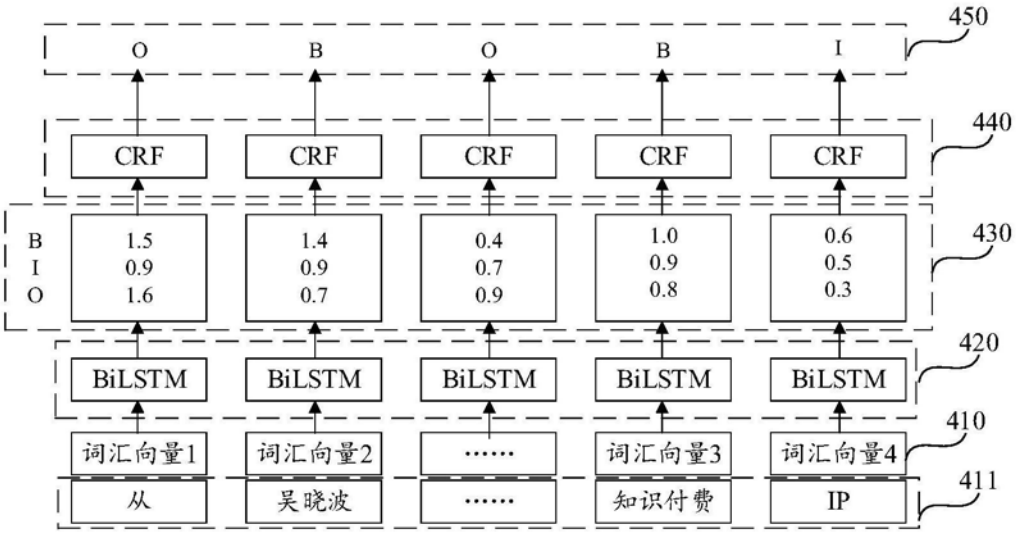

450
O B O B I
440
CRF CRF CRF CRF CRF
430
B 1.5 0.9 1.6
I 1.4 0.9 0.7
O 0.4 0.7 0.9
1.0 0.9 0.8
0.6 0.5 0.3
420
BiLSTM BiLSTM BiLSTM BiLSTM BiLSTM
410
词汇向量1 词汇向量2 …… 词汇向量3 词汇向量4
411
从 吴晓波 …… 知识付费 IP

图4

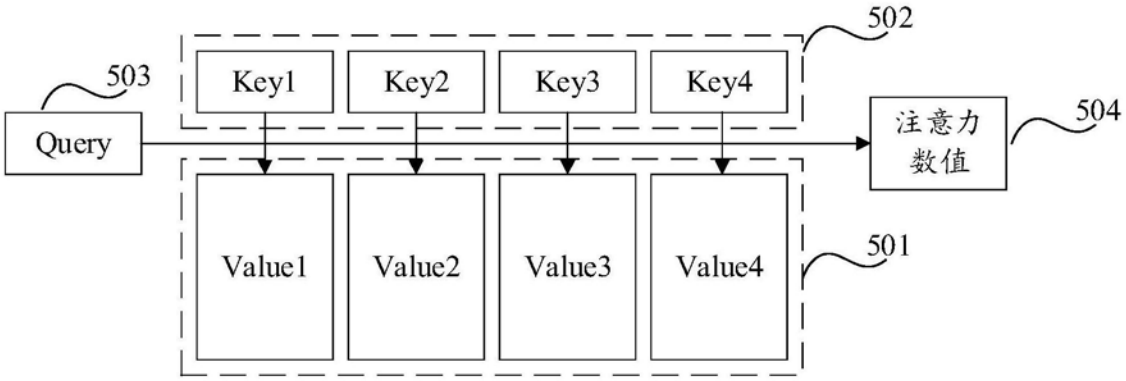

502
503
Key1
Key2
Key3
Key4
Query
注意力数值
504
Value1
Value2
Value3
Value4
501

图5

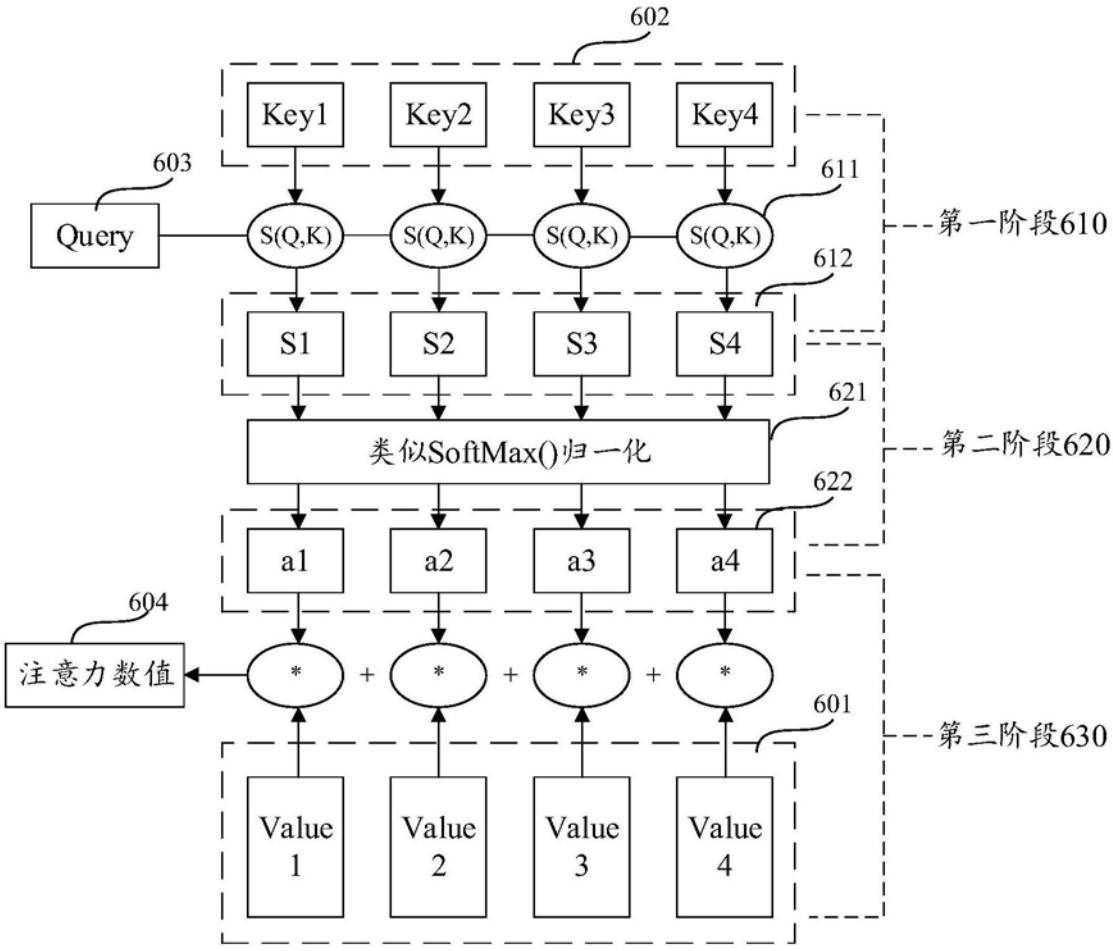

602
Key1
Key2
Key3
Key4
603
611
Query
S(Q,K)
S(Q,K)
S(Q,K)
S(Q,K)
第一阶段610
612
S1
S2
S3
S4
621
类似SoftMax()归一化
第二阶段620
622
a1
a2
a3
a4
604
注意力数值
*
+
*
+
*
+
*
601
第三阶段630
Value 1
Value 2
Value 3
Value 4

图6

位置701	分词702	词性703	是否词库704	标签705
0	从	介词	0	0
1	吴晓波	名词	1	1
2	到	介词	0	0
3	罗振宇	名词	1	1
4	，	标点	0	0
5	知识付费	名词	0	1
6	IP	名词	1	1
7	有	动词	0	0
8	哪些	代词	0	0
9	脆弱点	名词	0	0
10	？	标点	0	0

图7

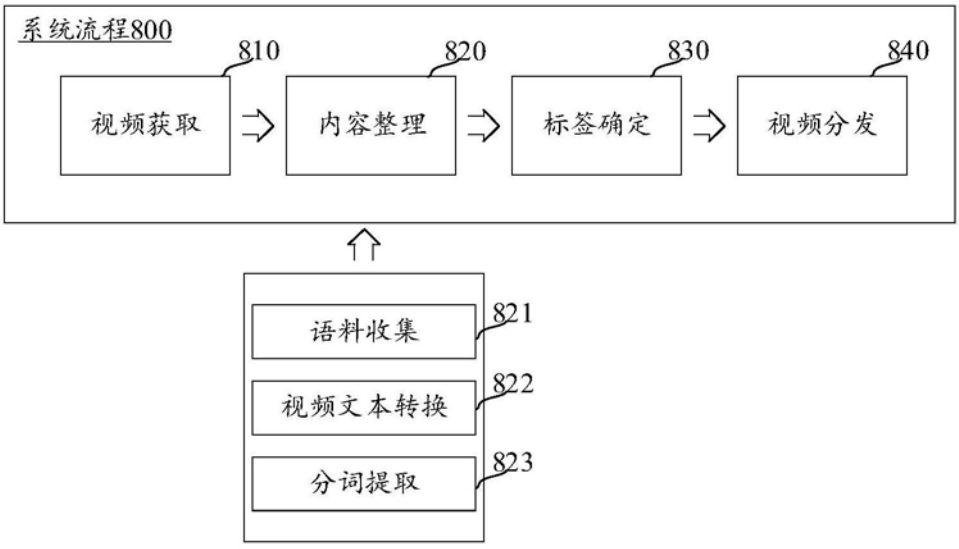


图8

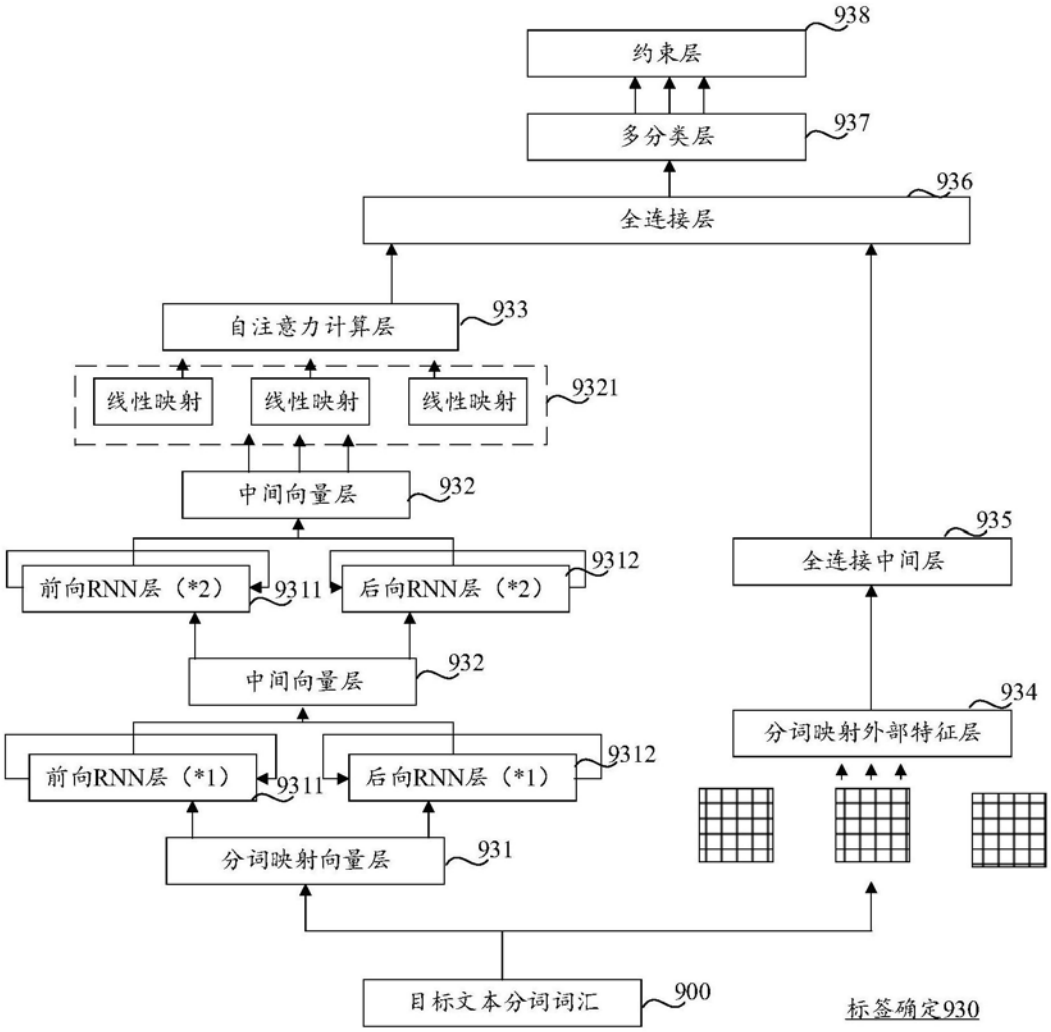

938
约束层
937
多分类层
936
全连接层
933
自注意力计算层
9321
线性映射
线性映射
线性映射
932
中间向量层
9312
前向RNN层（*2）
9311
后向RNN层（*2）
935
全连接中间层
932
中间向量层
934
分词映射外部特征层
9312
前向RNN层（*1）
9311
后向RNN层（*1）
931
分词映射向量层
900
目标文本分词词汇
标签确定930

图9

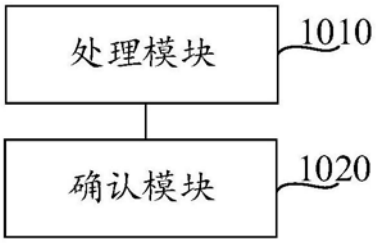

处理模块
1010
确认模块
1020

图10

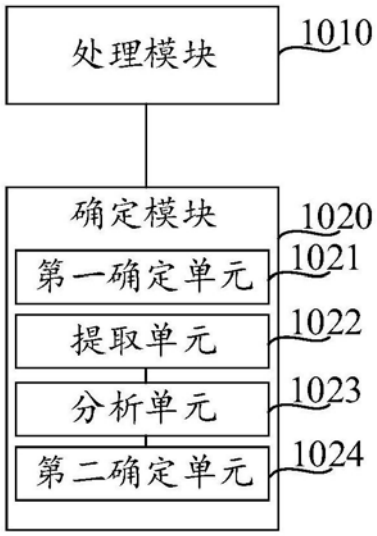

处理模块
1010
确定模块
1020
第一确定单元
1021
提取单元
1022
分析单元
1023
第二确定单元
1024

图11

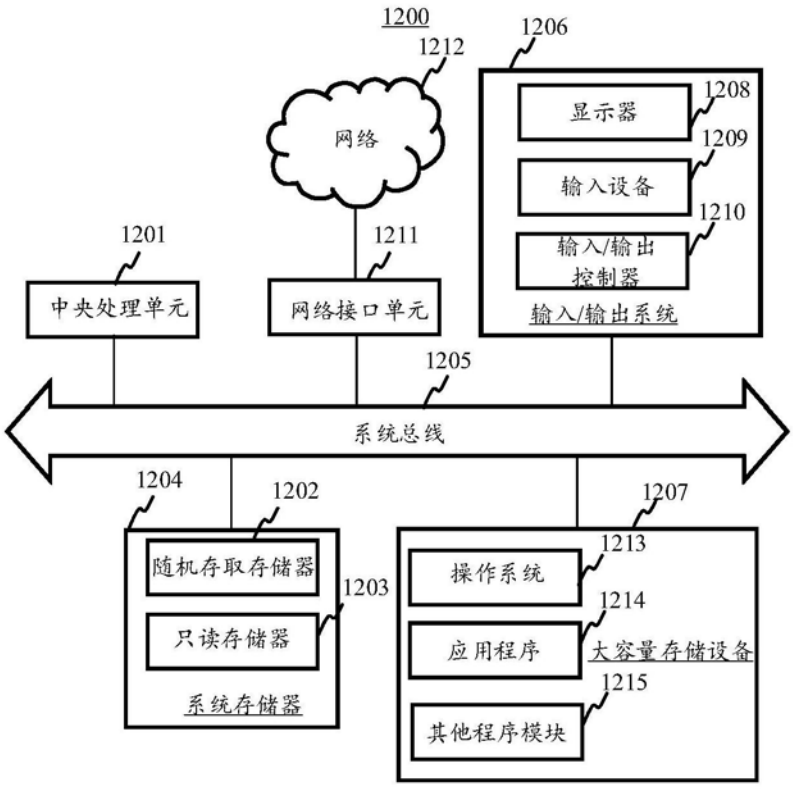

1200
1212
网络
1206
显示器
1208
输入设备
1209
输入/输出
控制器
1210
输入/输出系统
1201
中央处理单元
1211
网络接口单元
1205
系统总线
1204
1202
随机存取存储器
只读存储器
1203
系统存储器
1207
操作系统
1213
应用程序
1214
大容量存储设备
其他程序模块
1215

图12