(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110377695 B
(45)授权公告日 2022.11.22

(21)申请号 201910522043.9
(22)申请日 2019.06.17
(65)同一申请的已公布的文献号
申请公布号 CN 110377695 A
(43)申请公布日 2019.10.25
(73)专利权人 广州艾媒数聚信息咨询股份有限公司
地址 510006 广东省广州市番禺区小谷围街青蓝街26号701房
(72)发明人 张毅
(74)专利代理机构 广州嘉权专利商标事务所有限公司 44205
专利代理师 黎扬鹏
(51)Int.Cl.
G06F 16/33(2019.01)
G06F 16/35(2019.01)
(56)对比文件
CN 105955965 A,2016.09.21
CN 107832467 A,2018.03.23
CN 109710728 A,2019.05.03
CN 109189934 A,2019.01.11
曾庆山等.基于距离阈值的自适应K-均值聚类算法.《郑州大学学报》.2016,
审查员 何华

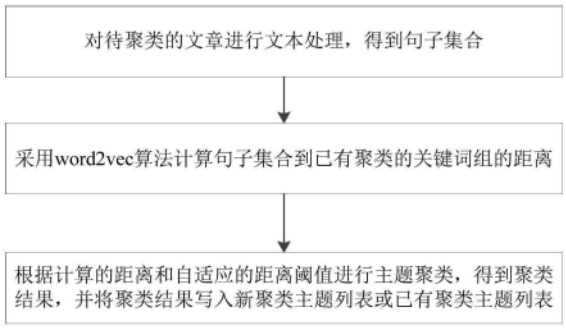
权利要求书2页 说明书9页 附图3页

(54)发明名称

一种舆情主题数据聚类方法、装置及存储介质

(57)摘要

本发明公开了一种舆情主题数据聚类方法、装置及存储介质，方法包括：对待聚类的文章进行文本处理，得到句子集合，文本处理包括分割；采用word2vec算法计算句子集合到已有聚类的关键词组的距离；根据计算的距离和自适应的距离阈值进行主题聚类，得到聚类结果，并将聚类结果写入新聚类主题列表或已有聚类主题列表，所述已有聚类主题列表由已有聚类组成。本发明通过自适应的距离阈值提供了通用的、可自动学习调整的分类阈值标准，适用性强；结合了已有聚类这一历史聚类成果来进行主题聚类，优化了聚类的结果；采用了word2vec算法这一神经网络学习方法配合关键词组的距离特征，提升了聚类的速度和准确度，可广泛应用于舆情监控领域。



CN 110377695 B

1.一种舆情主题数据聚类方法,其特征在于:包括以下步骤:

对待聚类的文章进行预处理,所述预处理包括切词、词性标记、去停用词、计算词频和去重,所述待聚类的文章为预设周期内获取的文章;

对预处理后的文章进行杂质过滤;从杂质过滤后的文章中抽取文章标题和摘要,并将所述杂质过滤后的文章分割为句子集合;

提取各个句子的关键词组;

计算各个句子集合之间的相互距离;

提取已有聚类的关键词组;

通过Skip-gram算法提取各个句子的关键词组的特征向量作为第一向量;

通过所述Skip-gram算法提取已有聚类的关键词组的特征向量作为第二向量;

对第一向量与第二向量进行多维距离计算;

将第一向量与第二向量之间的多维距离降维处理为一维距离,从而得到各个句子的关键词组到各已有聚类的关键词组的距离;

合并计算所述各个句子集合到所述已有聚类的关键词组的距离;

根据计算的距离和自适应的距离阈值进行主题聚类,得到聚类结果,并将聚类结果写入新聚类主题列表或已有聚类主题列表,所述已有聚类主题列表由已有聚类组成。

2.根据权利要求1所述的一种舆情主题数据聚类方法,其特征在于:所述根据计算的距离和自适应的距离阈值进行主题聚类,得到聚类结果,并将聚类结果写入新聚类主题列表或已有聚类主题列表这一步骤,具体包括:

将句子集合的相互距离小于等于第一距离阈值的文章主题合并为新分类;

将新分类中与已有聚类的关键词组距离最大的新分类确定为可能分类;

将可能分类与各个已有聚类进行距离对比,从而将与已有聚类的距离小于第二距离阈值的可能分类归入该已有聚类所在的已有聚类主题列表中,并更新该已有聚类的关键词组。

3.根据权利要求2所述的一种舆情主题数据聚类方法,其特征在于:所述根据计算的距离和自适应的距离阈值进行主题聚类,得到聚类结果,并将聚类结果写入新聚类主题列表或已有聚类主题列表这一步骤,还具体包括:

确定待聚类的文章的各个句子的关键词组与已有聚类的关键词组相同或相似度大于预设相似度阈值时,直接将待聚类的文章主题归入已有聚类所在的已有聚类主题列表中。

4.一种舆情主题数据聚类装置,其特征在于:包括:

文本处理模块,用于对待聚类的文章进行预处理,所述预处理包括切词、词性标记、去停用词、计算词频和去重;对预处理后的文章进行杂质过滤;从杂质过滤后的文章中抽取文章标题和摘要,并将所述杂质过滤后的文章分割为句子集合,所述待聚类的文章为预设周期内获取的文章;

距离计算模块,用于提取各个句子的关键词组;计算各个句子集合之间的相互距离;提取已有聚类的关键词组;通过Skip-gram算法提取各个句子的关键词组的特征向量作为第一向量;通过所述Skip-gram算法提取已有聚类的关键词组的特征向量作为第二向量;对第一向量与第二向量进行多维距离计算;将第一向量与第二向量之间的多维距离降维处理为一维距离,从而得到各个句子的关键词组到各已有聚类的关键词组的距离;

主题聚类模块，用于根据计算的距离和自适应的距离阈值进行主题聚类，得到聚类结果，并将聚类结果写入新聚类主题列表或已有聚类主题列表，所述已有聚类主题列表由已有聚类组成。

5.根据权利要求4所述的一种舆情主题数据聚类装置，其特征在于：所述主题聚类模块具体包括：

新分类合并单元，用于将句子集合的相互距离小于等于第一距离阈值的文章主题合并为新分类；

可能分类确定单元，用于将新分类中与已有聚类的关键词组距离最大的新分类确定为可能分类；

聚类单元，用于将可能分类与各个已有聚类进行距离对比，从而将与已有聚类的距离小于第二距离阈值的可能分类归入该已有聚类所在的已有聚类主题列表中，并更新该已有聚类的关键词组。

6.一种舆情主题数据聚类装置，其特征在于：包括：

至少一个处理器；

至少一个存储器，用于存储至少一个程序；

当所述至少一个程序被所述至少一个处理器执行，使得所述至少一个处理器实现如权利要求1-3所述的一种舆情主题数据聚类方法。

7.一种存储介质，其中存储有处理器可执行的指令，其特征在于：所述处理器可执行的指令在由处理器执行时用于实现如权利要求1-3任一项所述的一种舆情主题数据聚类方法。

一种舆情主题数据聚类方法、装置及存储介质

技术领域

[0001] 本发明涉及舆情监控领域，尤其是一种舆情主题数据聚类方法、装置及存储介质。

背景技术

[0002] 舆情监控，整合了互联网信息采集技术及信息智能处理技术，通过对互联网海量信息自动抓取、自动分类聚类、主题检测、专题聚焦，实现用户的网络舆情监测和新闻专题追踪等信息需求，形成简报、报告、图表等分析结果，为客户全面掌握群众思想动态，做出正确舆论引导，提供分析依据。

[0003] 在舆情监控中，舆情数据聚类是话题发现的重要手段之一，目前的舆情主题数据聚类方法包括以下步骤：对社交网络中的当前数据进行冗余过滤，以获取非冗余数据；对所述非冗余数据进行分析，以在所述非冗余数据中确定相关舆情数据；对所述相关舆情数据进行聚类，以在所述相关舆情数据中确定目标舆情数据。这种方法的问题在于聚类结果没有一个通用的、可自动学习调整的分类阈值标准，同时不可继承历史聚类成果，长期监控过程中舆情文章的巨大增量也对聚类计算带来持续增长的压力。

发明内容

[0004] 为解决上述技术问题，本发明实施例的目的在于：提供一种舆情主题数据聚类方法、装置及存储介质。

[0005] 本发明实施例所采取的第一技术方案是：

[0006] 一种舆情主题数据聚类方法，包括以下步骤：

[0007] 对待聚类的文章进行文本处理，得到句子集合，所述文本处理包括分割；

[0008] 采用word2vec算法计算句子集合到已有聚类的关键词组的距离；

[0009] 根据计算的距离和自适应的距离阈值进行主题聚类，得到聚类结果，并将聚类结果写入新聚类主题列表或已有聚类主题列表，所述已有聚类主题列表由已有聚类组成。

[0010] 进一步，所述对待聚类的文章进行文本处理，得到句子集合这一步骤，具体包括：

[0011] 对待聚类的文章进行预处理，所述预处理包括切词、词性标记、去停用词、计算词频和去重；

[0012] 对预处理后的文章进行杂质过滤；

[0013] 从杂质过滤后的文章中抽取文章标题和摘要，并将杂质过滤后的文章分割为句子集合。

[0014] 进一步，所述采用word2vec算法计算句子集合到已有聚类的关键词组的距离这一步骤，具体包括：

[0015] 提取各个句子的关键词组；

[0016] 计算各个句子集合之间的相互距离；

[0017] 提取已有聚类的关键词组；

[0018] 采用word2vec算法分别计算各个句子的关键词组到各已有聚类的关键词组的距

离；

[0019] 合并计算各个句子集合到已有聚类的关键词组的距离。

[0020] 进一步，所述采用word2vec算法分别计算各个句子的关键词组到各已有聚类的关键词组的距离这一步骤，具体包括：

[0021] 提取各个句子的关键词组的特征向量作为第一向量；

[0022] 提取已有聚类的关键词组的特征向量作为第二向量；

[0023] 对第一向量与第二向量进行多维距离计算；

[0024] 将第一向量与第二向量之间的多维距离降维处理为一维距离，从而得到各个句子的关键词组到各已有聚类的关键词组的距离。

[0025] 进一步，所述根据计算的距离和自适应的距离阈值进行主题聚类，得到聚类结果，并将聚类结果写入新聚类主题列表或已有聚类主题列表这一步骤，具体包括：

[0026] 将句子集合的相互距离小于等于第一距离阈值的文章主题合并为新分类；

[0027] 将新分类中与已有聚类的关键词组距离最大的新分类确定为可能分类；

[0028] 将可能分类与各个已有聚类进行距离对比，从而将与已有聚类的距离小于第二距离阈值的可能分类归入该已有聚类所在的已有聚类主题列表中，并更新该已有聚类的关键词组。

[0029] 进一步，所述根据计算的距离和自适应的距离阈值进行主题聚类，得到聚类结果，并将聚类结果写入新聚类主题列表或已有聚类主题列表这一步骤，还具体包括：

[0030] 确定待聚类的文章的各个句子的关键词组与已有聚类的关键词组相同或相似度大于预设相似度阈值时，直接将待聚类的文章主题归入已有聚类所在的已有聚类主题列表中。

[0031] 本发明实施例所采取的第二技术方案是：

[0032] 一种舆情主题数据聚类装置，包括：

[0033] 文本处理模块，用于对待聚类的文章进行文本处理，得到句子集合，所述文本处理包括分割；

[0034] 距离计算模块，用于采用word2vec算法计算句子集合到已有聚类的关键词组的距离；

[0035] 主题聚类模块，用于根据计算的距离和自适应的距离阈值进行主题聚类，得到聚类结果，并将聚类结果写入新聚类主题列表或已有聚类主题列表，所述已有聚类主题列表由已有聚类组成。

[0036] 进一步，所述主题聚类模块具体包括：

[0037] 新分类合并单元，用于将句子集合的相互距离小于等于第一距离阈值的文章主题合并为新分类；

[0038] 可能分类确定单元，用于将新分类中与已有聚类的关键词组距离最大的新分类确定为可能分类；

[0039] 聚类单元，用于将可能分类与各个已有聚类进行距离对比，从而将与已有聚类的距离小于第二距离阈值的可能分类归入该已有聚类所在的已有聚类主题列表中，并更新该已有聚类的关键词组。

[0040] 本发明实施例所采取的第三技术方案是：

[0041] 一种舆情主题数据聚类装置,包括:

[0042] 至少一个处理器;

[0043] 至少一个存储器,用于存储至少一个程序;

[0044] 当所述至少一个程序被所述至少一个处理器执行,使得所述至少一个处理器实现如本发明所述的一种舆情主题数据聚类方法。

[0045] 本发明实施例所采取的第四技术方案是:

[0046] 一种存储介质,其中存储有处理器可执行的指令,所述处理器可执行的指令在由处理器执行时用于实现如本发明所述的一种舆情主题数据聚类方法。

[0047] 上述本发明实施例中的一个或多个技术方案具有如下优点:本发明实施例先通过文本处理将待聚类的文章分割为句子集合,然后采用word2vec算法计算句子集合到已有聚类的关键词组的距离,最后根据计算的距离和自适应的距离阈值进行主题聚类,通过自适应的距离阈值提供了通用的、可自动学习调整的分类阈值标准,适用性强;通过句子集合到已有聚类的关键词组的距离来进行主题聚类,结合了已有聚类这一历史聚类成果来进行主题聚类,优化了聚类的结果;采用word2vec算法计算句子集合到已有聚类的关键词组的距离,采用了word2vec算法这一神经网络学习方法配合关键词组的距离特征,提升了聚类的速度和准确度,减轻了聚类计算的压力。

附图说明

[0048] 图1为本发明实施例提供的一种舆情主题数据聚类方法流程图;

[0049] 图2为现有技术的聚类算法流程图;

[0050] 图3为本发明具体实施例的距离计算方法流程图;

[0051] 图4为本发明具体实施例句子集合、句子及已有主题聚类关键词的关系示意图;

[0052] 图5为本发明具体实施例获取聚类结果过程的流程图。

具体实施方式

[0053] 首先对本发明涉及的名词术语进行解释和说明:

[0054] Word2vec:是一群用来产生词向量的相关模型。这些模型为浅而双层的神经网络,用来训练以重新建构语言学之词文本。网络以词表现,并且需猜测相邻位置的输入词,在word2vec中词袋模型假设下,词的顺序是不重要的。训练完成之后,word2vec模型可用来映射每个词到一个向量,可用来表示词对词之间的关系,该向量为神经网络之隐藏层。

[0055] 词袋模型(Bag-of-words model):是个在自然语言处理和信息检索(IR)下被简化的表达模型。此模型下,像是句子或是文件这样的文字可以用一个袋子装着这些词的方式表现,这种表现方式不考虑文法以及词的顺序。最近词袋模型也被应用在计算机视觉领域。词袋模型被广泛应用在文件分类领域,词出现的频率可以用来当作训练分类器的特征。关于"词袋"这个用字的由来可追溯到泽里格·哈里斯于1954年在Distributional Structure的文章。

[0056] 统计语言模型(Statistical Language Model),是今天所有自然语言处理的基础,并且广泛应用于机器翻译、语音识别、印刷体或手写体识别、拼写纠错、汉字输入和文献查询。统计语言模型直观地解决了一个问题:一个句子是否合理,就看它的可能性大小如

何。至于可能性就用概率来衡量。

[0057] 假定S表示某一个有意义的句子，由一连串特定顺序排列的词W1，W2，…，Wn组成，这里n是句子的长度(句子中词汇的个数)。于是S出现的可能性也就是数学上所说的S的概率P(S)＝P(W1,W2,...,Wn)。

[0058] 利用条件概率公式，以上算式可以展开为：

[0059] P(W1,W2,...,Wn)＝P(W1)*P(W2|W1)*P(W3|W1,W2)...P(Wn|W1,W2,...,Wn-1)

[0060] 其中P(W1)表示第一个词W1出现的概率；P(W2|W1)是在已知第一个词的前提下，第二个词出现的概率；以此类推，词Wn出现的概率取决于它前面所有的词。俄国数学家马尔可夫(Andrey Markov)提出假设任意一个词Wi出现的概率只同它前面的词Wi-1有关，S出现的概率就变得简单了：

[0061] P(S)＝P(W1)*P(W2|W1)*P(W3|W2)...P(Wn|Wn-1)

[0062] 上述公式就是统计语言模型的二元模型(Bigram Model)。接下来的问题就是如何计算P(Wn|Wn-1)，根据概率论，该公式可以变化为：

[0063] P(Wn|Wn-1)＝P(Wn-1,Wn)/P(Wn-1)

[0064] 因为在互联网时代有大量的语料库(Corpus)可以作为训练样本，所以只要数一数Wn-1、Wn这对词在语料库中前后相邻出现了多少次，以及Wn-1本身在相同的语料库中出现了多少次，就可得到P(Wn|Wn-1)。

[0065] 统计语言模型称为N元模型(N-Gram Model)。如果N＝2，那么就是上面的二元模型公式。而在实际中应用最多的是N＝3的三元模型，更高阶的模型就很少使用了。N取值一般较小，这主要是因为复杂度，当N从1到2，再从2到3时，模型的效果上升显著。而当模型从3到4时，效果的提升就不是很显著了，而资源的耗费却增加得非常快。Google的罗塞塔翻译系统和语音搜索系统，使用的是四元模型，该模型存储于500台以上的Google服务器中。

[0066] kip-gram模型：一个简单但却非常实用的模型，用于使用当前词来预测上下文词汇。在自然语言处理中，语料的选取是一个相当重要的问题:第一，语料必须充分。一方面词典的词量要足够大，另一方面要尽可能多地包含反映词语之间关系的句子，例如，只有“鱼在水中游”这种句式在语料中尽可能地多，模型才能够学习到该句中的语义和语法关系，这和人类学习自然语言一个道理，重复的次数多了，也就会模仿了；第二，语料必须准确。也就是说所选取的语料能够正确反映该语言的语义和语法关系，这一点似乎不难做到，例如中文里，《人民日报》的语料比较准确。但是，更多的时候，并不是语料的选取引发了对准确性问题的担忧，而是处理的方法。n元模型中，因为窗口大小的限制，导致超出窗口范围的词语与当前词之间的关系不能被正确地反映到模型之中，如果单纯扩大窗口大小又会增加训练的复杂度。Skip-gram模型的提出很好地解决了这些问题。顾名思义，Skip-gram就是“跳过某些符号”，例如，句子“中国足球踢得真是太烂了”有4个3元词组，分别是“中国足球踢得”、“足球踢得真是”、“踢得真是太烂”、“真是太烂了”，可是我们发现，这个句子的本意就是“中国足球太烂”可是上述4个3元词组并不能反映出这个信息。Skip-gram模型却允许某些词被跳过，因此可以组成“中国足球太烂”这个3元词组。如果允许跳过2个词，即2-Skip-gram。

[0067] 词向量：具有良好的语义特性，是表示词语特征的常用方式。词向量每一维的值代表一个具有一定的语义和语法上解释的特征。所以，可以将词向量的每一维称为一个词语特征。词向量具有多种形式，distributed representation是其中一种。一个distributed

representation是一个稠密、低维的实值向量。distributed representation的每一维表示词语的一个潜在特征，该特征捕获了有用的句法和语义特性。可见，distributed representation中的distributed一词体现了词向量这样一个特点：将词语的不同句法和语义特征分布到它的每一个维度去表示。

[0068] 下面结合说明书附图和具体实施例对本发明作进一步解释和说明。

[0069] 参照图1，本发明实施例提供了一种舆情主题数据聚类方法，包括以下步骤：

[0070] 对待聚类的文章进行文本处理，得到句子集合，所述文本处理包括分割；

[0071] 采用word2vec算法计算句子集合到已有聚类的关键词组的距离；

[0072] 根据计算的距离和自适应的距离阈值进行主题聚类，得到聚类结果，并将聚类结果写入新聚类主题列表或已有聚类主题列表，所述已有聚类主题列表由已有聚类组成。

[0073] 具体地，待聚类的文章是一定周期内文章，其可以通过互联网从自媒体网站、新闻门户网站等获取。文本处理主要包括切词、词性标记、去停用词、计算词频、去重、过滤和分割等。分割用于将文章分割为若干个句子的集合。

[0074] Word2vec算法，属于无监督的机器学习算法的一种，不需要预先进行人工标注，能提升训练效率和降低人工成本。已有聚类是指已经过聚类计算，确定所属类型的主题类型。

[0075] 为了降低长期监控过程中舆情文章的巨大增量对聚类计算的压力，本实施例抽取了文章的关键词组的距离特征来进行聚类计算，与传统通过文章的全部特征来进行聚类计算的方式相比，效率更高。距离可以是马氏距离等。

[0076] 自适应的距离阈值，是指距离阈值可以自动学习调整。例如，自适应的距离阈值可以是判断不同新文章之间的相似性大小的距离阈值，也可以是判断新文章与历史聚类内容(即已有聚类)之间的相似性大小的距离阈值。

[0077] 新聚类主题列表，用于存储经聚类计算后识别为新聚类主题的文章主题。新聚类主题一般与所有已有聚类主题的距离大于预设的距离阈值。

[0078] 已有聚类主题列表能累加经聚类计算后识别为已有聚类主题的文章主题并更新。文章主题属于已有聚类主题时，其一般与某个已有聚类主题的距离小于等于预设的距离阈值。

[0079] 由此可见，本实施例通过句子集合到已有聚类的关键词组的距离来进行主题聚类，结合了已有聚类这一历史聚类成果来进行主题聚类，优化了聚类的结果；采用了word2vec算法这一神经网络学习方法配合关键词组的距离特征，提升了聚类的速度和准确度，减轻了聚类计算的压力；通过自适应的距离阈值提供了通用的、可自动学习调整的分类阈值标准，适用性强。

[0080] 进一步作为优选的实施方式，所述对待聚类的文章进行文本处理，得到句子集合这一步骤，具体包括：

[0081] 对待聚类的文章进行预处理，所述预处理包括切词、词性标记、去停用词、计算词频和去重；

[0082] 对预处理后的文章进行杂质过滤；

[0083] 从杂质过滤后的文章中抽取文章标题和摘要，并将杂质过滤后的文章分割为句子集合。

[0084] 具体地，本实施例通过预处理、杂质过滤、抽取和分割等文本处理操作，为后续的

距离计算和聚类做好了准备。

[0085] 进一步作为优选的实施方式,所述采用word2vec算法计算句子集合到已有聚类的关键词组的距离这一步骤,具体包括:

[0086] 提取各个句子的关键词组;

[0087] 计算各个句子集合之间的相互距离;

[0088] 提取已有聚类的关键词组;

[0089] 采用word2vec算法分别计算各个句子的关键词组到各已有聚类的关键词组的距离;

[0090] 合并计算各个句子集合到已有聚类的关键词组的距离。

[0091] 具体地,一个句子集合可以包括若干个句子。在存在多篇文章且每篇文章只有1个句子集合时,各个句子集合之间的相互距离可以反映这些文章的相似性大小。

[0092] 本实施例可通过Jieba切词中的KeywordExtract方法来提取各个句子的关键词组。已有聚类的关键词组可以分布式存储,提取时可以通过Hadoop的方式来实现。分别计算各个(即单个)句子的关键词组到各已有聚类的关键词组的距离之后,可通过累加(即合并计算)来得到某个句子集合到已有聚类的关键词组的距离。

[0093] 进一步作为优选的实施方式,所述采用word2vec算法分别计算各个句子的关键词组到各已有聚类的关键词组的距离这一步骤,具体包括:

[0094] 提取各个句子的关键词组的特征向量作为第一向量;

[0095] 提取已有聚类的关键词组的特征向量作为第二向量;

[0096] 对第一向量与第二向量进行多维距离计算;

[0097] 将第一向量与第二向量之间的多维距离降维处理为一维距离,从而得到各个句子的关键词组到各已有聚类的关键词组的距离。

[0098] 传统的聚类计算方法,会获取文本的全部特征向量后计算其的相似度,以判断是否聚类;本实施例使用了新的距离计算方法,仅抽取文本中的关键词组获取特征向量,并将特征向量距离降维计算为一维距离后再进行其它计算,极大地降低了计算复杂度,提高了计算效率。

[0099] 优选地,本实施例的第一向量和第二向量可以是通过word2vec的Skip-gram算法获取的词向量。这些词向量已经包含了上下文的信息,而且数据规模与初始相比也得到较大压缩,能进一步提升聚类计算效率。

[0100] 进一步作为优选的实施方式,所述根据计算的距离和自适应的距离阈值进行主题聚类,得到聚类结果,并将聚类结果写入新聚类主题列表或已有聚类主题列表这一步骤,具体包括:

[0101] 将句子集合的相互距离小于等于第一距离阈值的文章主题合并为新分类;

[0102] 将新分类中与已有聚类的关键词组距离最大的新分类确定为可能分类;

[0103] 将可能分类与各个已有聚类进行距离对比,从而将与已有聚类的距离小于第二距离阈值的可能分类归入该已有聚类所在的已有聚类主题列表中,并更新该已有聚类的关键词组。

[0104] 具体地,分类是针对可能是主题但未确定是何种主题的情况,聚类是针对确定为何种主题的情况。本实施例在进行主题聚类时,先判断是否将不同的文章主题合并为新分

类，再在新分类中确定可能分类，最后根据可能分类与已有聚类的距离来确定可能分类是属于已有聚类还是新聚类。

[0105] 句子集合的相互距离小于等于第一距离阈值的文章表明这些文章的相似度较高，可以归入同一个新分类中，此过程用于判断不同的新文章之间是否适合合并到同一个分类中。

[0106] 合并为新分类后，可结合已有聚类确定好新分类是否为可能分类(即候选的目标分类)，根据先验知识，可能分类一般与已有聚类的距离较大。

[0107] 确定可能分类后，再将该可能分类与所有已有聚类的距离阈值做对比，大于距离阈值则可以确定为新聚类主题；若与一个已有聚类主题的距离小于距离阈值，则将该可能分类归入该已有聚类主题列表，并更新该聚类主题的关键词组。

[0108] 本实施例结合新文章之间的相似性以及与已有聚类间的距离来进行主题聚类，提升了聚类计算的效率和准确度。

[0109] 进一步作为优选的实施方式，所述根据计算的距离和自适应的距离阈值进行主题聚类，得到聚类结果，并将聚类结果写入新聚类主题列表或已有聚类主题列表这一步骤，还具体包括：

[0110] 确定待聚类的文章的各个句子的关键词组与已有聚类的关键词组相同或相似度大于预设相似度阈值时，直接将待聚类的文章主题归入已有聚类所在的已有聚类主题列表中。

[0111] 具体地，本实施例在检测到待聚类的文章的各个句子的关键词组与已有聚类的关键词相同或相似时，直接将该文章主题归入该已有聚类主题中，从而跳过该文章后续的聚类计算过程，进一步提升了聚类的效率。

[0112] 为了将一定周期内文本内容高度相近的文章聚类作为主题，本具体实施例提出了一种用于舆情监控的神经网络主题聚类方法。该方法利用了神经网络学习方法，通过对每日新语料内容的词向量关系计算，不仅提高了内容聚类的处理速度和准确度，同时还以词向量的方式将每日新的语料内容聚类关系加入历史结果，以机器学习方法持续地自动训练聚类模型，优化了聚类结果。该方法主要包括以下步骤：

[0113] S1、文本处理。

[0114] 文本处理的过程可进一步细分为：

[0115] S11、提取文本预处理；

[0116] S12、过滤杂质信息；

[0117] S13、抽取文章标题和摘要并分割为句子集合。

[0118] S2、聚类距离计算。

[0119] 具体地，可利用word2vec算法对句子和已有聚类的对应关键词组做距离计算。如图2所示，传统的计算方法，使用全部切词结果来获取文本特征向量，然后计算其的相似度以判断是否聚类。如图3所示，本具体实施例使用了新的距离计算方法，仅抽取文本中的关键词组来获取特征向量距离结果，并将距离结果降维计算为一维距离后再进行其它计算，极大地降低了计算复杂度，提高计算效率。如图3和图4所示，该新的距离计算方法具体步骤如下：

[0120] S21、使用Jieba切词中的KeywordExtract方法来对单个句子进行处理，提取关键

词组；

[0121] S22、计算各句子集合之间的相互距离；

[0122] S23、使用Hadoop提取分布式存储的已有聚类的关键词组；

[0123] S24、单独计算各句子的关键词组到各已有聚类的关键词组的距离；

[0124] S25、合并计算各单个句子集合到已有聚类关键词组的距离。

[0125] S3、得出聚类结果，归入聚类主题列表。

[0126] 如图5所示，该过程可进一步细分为如下步骤：

[0127] S31、根据句子集合的相互距离判断是否合并为新分类；

[0128] S32、将与其它已有聚类关键词组距离最大的新分类确定为可能分类；

[0129] S33、与各个已有聚类的距离阈值做对比，大于阈值则可以确定为新聚类主题；若与一个已有聚类主题的距离小于阈值，则将可能分类归入该已有聚类主题，并更新该已有聚类主题的关键词组。

[0130] 上述计算过程S31～S33中，每个句子的关键词组可以同时对全量的聚类主题关键词进行同步计算，每次计算过的文本(或文章)写入已计算列表，下次计算将跳过这些文本(或文章)，提高了同步计算处理量与计算效率。分类结果也分别归入相应的聚类主题列表。

[0131] 本具体实施例采用了改进的距离计算算法改良优化了聚类结果，提高了计算处理量和极大地缩短了处理时间，使得聚类结果更加快速优质；同时继承积累了历史舆情文本聚类计算的结果，运用Word2vec算法不断自动训练优化其聚类模型，在舆情监控领域具有广阔的应用前景。

[0132] 与图1的方法相对应，本发明实施例还提供了一种舆情主题数据聚类装置，包括：

[0133] 文本处理模块，用于对待聚类的文章进行文本处理，得到句子集合，所述文本处理包括分割；

[0134] 距离计算模块，用于采用word2vec算法计算句子集合到已有聚类的关键词组的距离；

[0135] 主题聚类模块，用于根据计算的距离和自适应的距离阈值进行主题聚类，得到聚类结果，并将聚类结果写入新聚类主题列表或已有聚类主题列表，所述已有聚类主题列表由已有聚类组成。

[0136] 进一步作为优选的实施方式，所述主题聚类模块具体包括：

[0137] 新分类合并单元，用于将句子集合的相互距离小于等于第一距离阈值的文章主题合并为新分类；

[0138] 可能分类确定单元，用于将新分类中与已有聚类的关键词组距离最大的新分类确定为可能分类；

[0139] 聚类单元，用于将可能分类与各个已有聚类进行距离对比，从而将与已有聚类的距离小于第二距离阈值的可能分类归入该已有聚类所在的已有聚类主题列表中，并更新该已有聚类的关键词组。

[0140] 上述方法实施例中的内容均适用于本装置实施例中，本装置实施例所具体实现的功能与上述方法实施例相同，并且达到的有益效果与上述方法实施例所达到的有益效果也相同。

[0141] 与图1的方法相对应，本发明实施例还提供了一种舆情主题数据聚类装置，包括：

[0142] 至少一个处理器；

[0143] 至少一个存储器，用于存储至少一个程序；

[0144] 当所述至少一个程序被所述至少一个处理器执行，使得所述至少一个处理器实现如本发明所述的一种舆情主题数据聚类方法。

[0145] 上述方法实施例中的内容均适用于本装置实施例中，本装置实施例所具体实现的功能与上述方法实施例相同，并且达到的有益效果与上述方法实施例所达到的有益效果也相同。

[0146] 与图1的方法相对应，本发明实施例还提供了一种存储介质，其中存储有处理器可执行的指令，所述处理器可执行的指令在由处理器执行时用于实现如本发明所述的一种舆情主题数据聚类方法。

[0147] 上述方法实施例中的内容均适用于存储介质实施例中，本存储介质实施例所具体实现的功能与上述方法实施例相同，并且达到的有益效果与上述方法实施例所达到的有益效果也相同。

[0148] 以上是对本发明的较佳实施进行了具体说明，但本发明并不限于所述实施例，熟悉本领域的技术人员在不违背本发明精神的前提下还可做作出种种的等同变形或替换，这些等同的变形或替换均包含在本申请权利要求所限定的范围内。

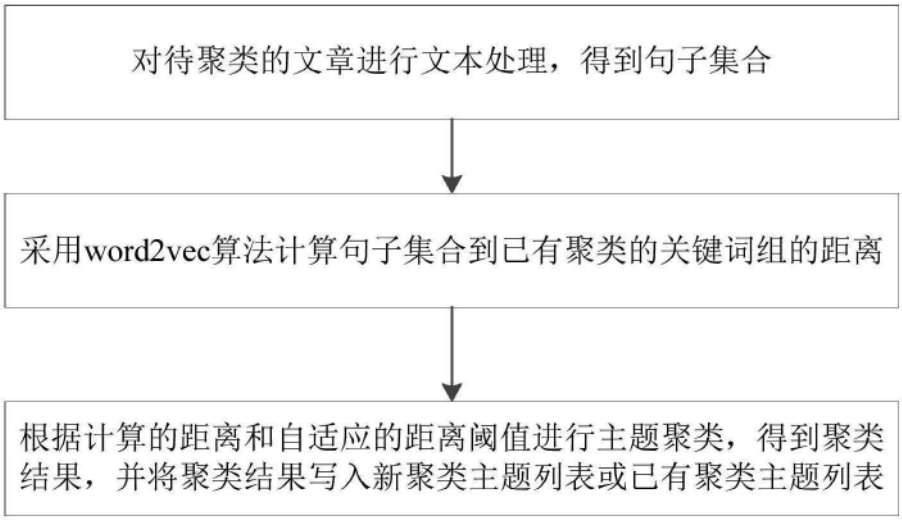

对待聚类的文章进行文本处理，得到句子集合
采用word2vec算法计算句子集合到已有聚类的关键词组的距离
根据计算的距离和自适应的距离阈值进行主题聚类，得到聚类结果，并将聚类结果写入新聚类主题列表或已有聚类主题列表

图1

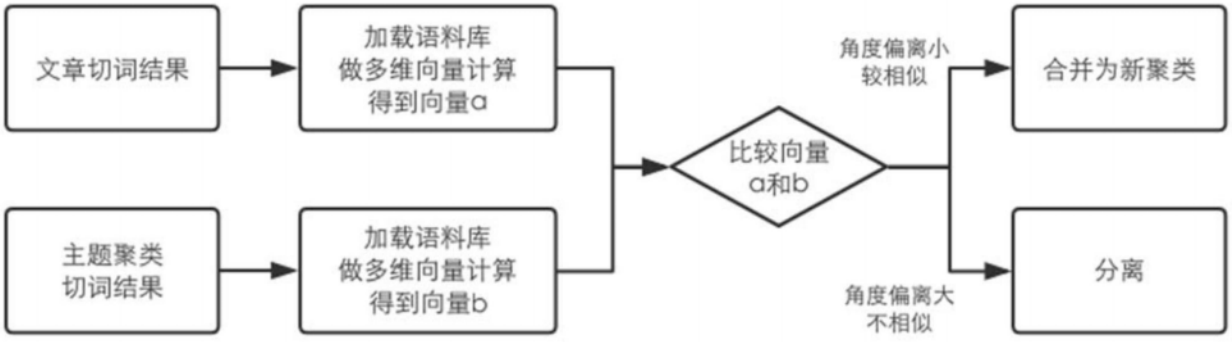

文章切词结果
加载语料库
做多维向量计算
得到向量a
主题聚类
切词结果
加载语料库
做多维向量计算
得到向量b
比较向量
a和b
角度偏离小
较相似
合并为新聚类
角度偏离大
不相似
分离

图2

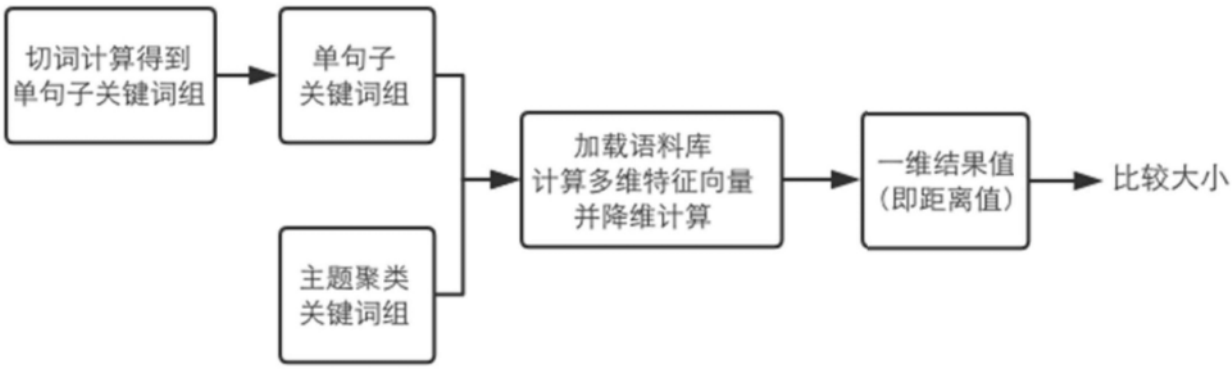

切词计算得到
单句子关键词组
单句子
关键词组
主题聚类
关键词组
加载语料库
计算多维特征向量
并降维计算
一维结果值
(即距离值)
比较大小

图3

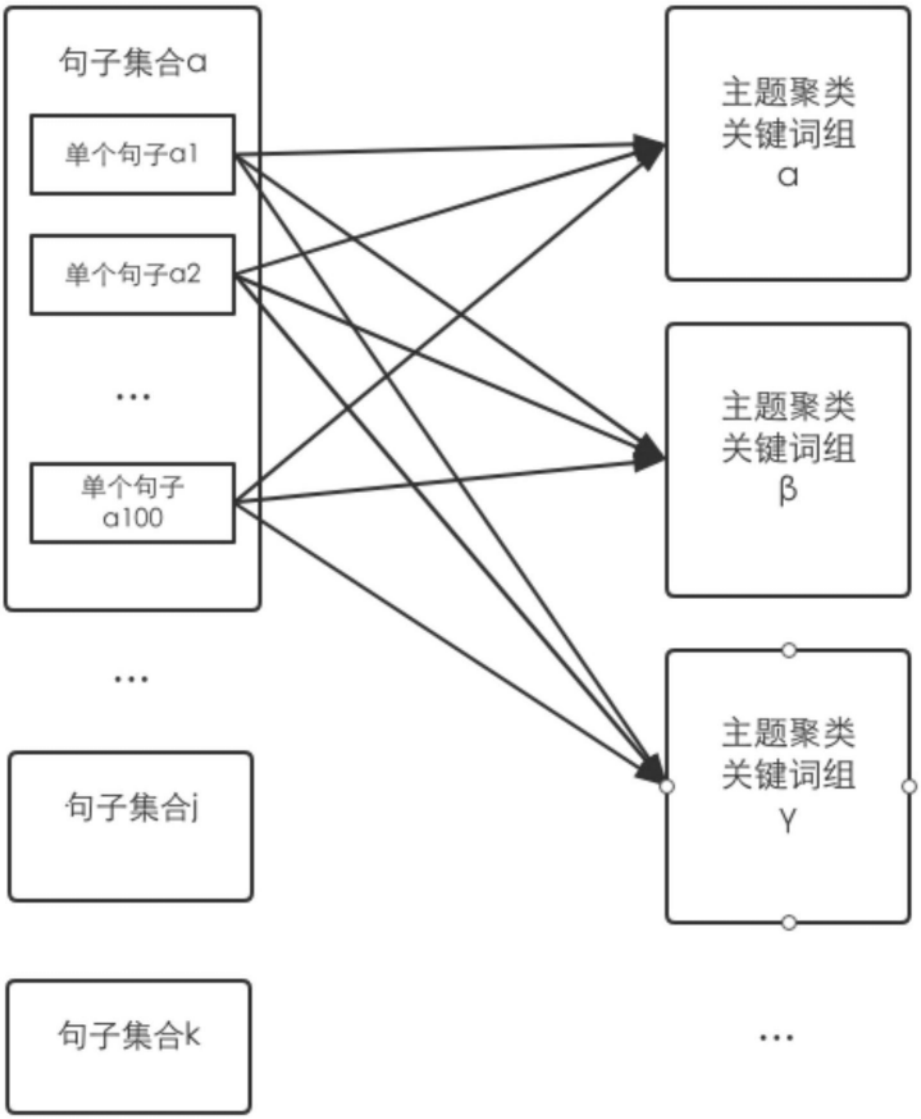

句子集合a
单个句子a1
单个句子a2
...
单个句子
a100
...
句子集合j
句子集合k
主题聚类
关键词组
α
主题聚类
关键词组
β
主题聚类
关键词组
γ
...

图4

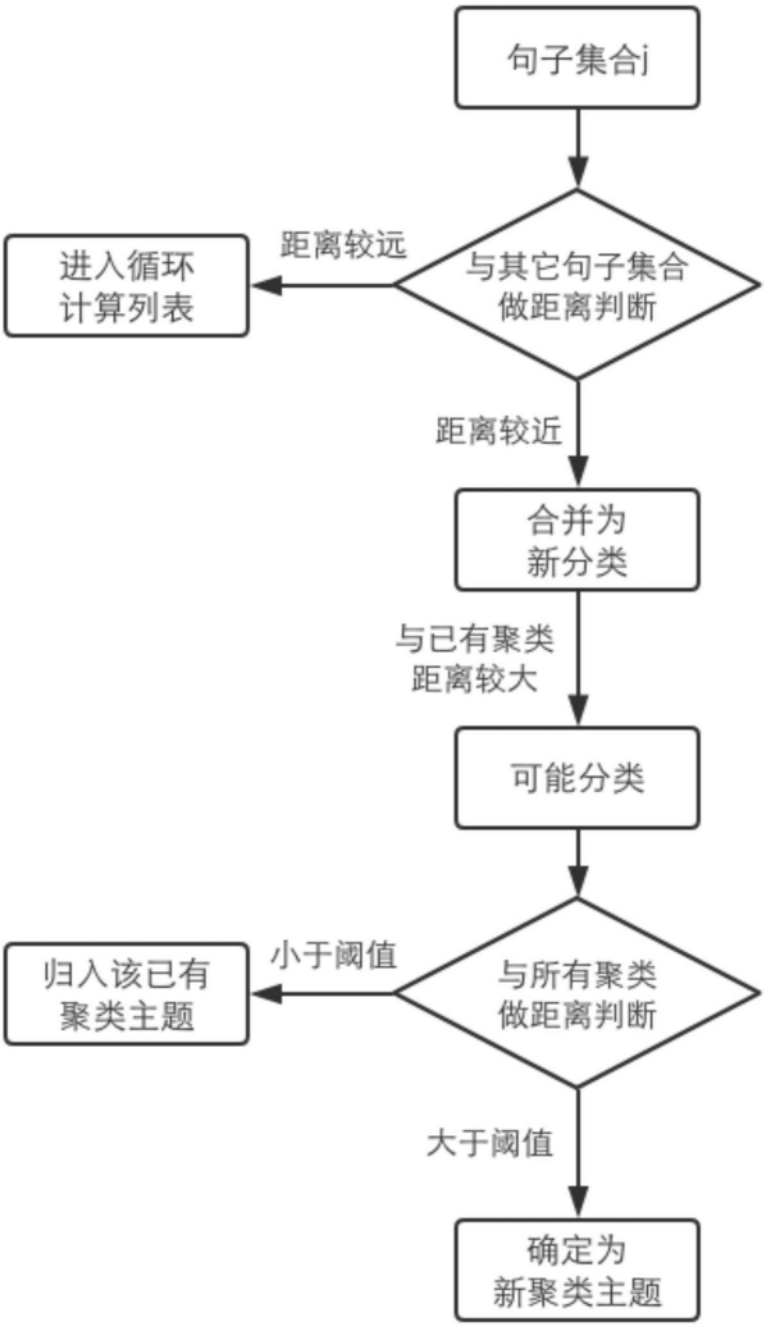

句子集合j
与其它句子集合
做距离判断
距离较远
进入循环
计算列表
距离较近
合并为
新分类
与已有聚类
距离较大
可能分类
与所有聚类
做距离判断
小于阈值
归入该已有
聚类主题
大于阈值
确定为
新聚类主题

图5