(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 108710654 A
(43)申请公布日 2018.10.26

(21)申请号 201810441355.2

(22)申请日 2018.05.10

(71)申请人 新华智云科技有限公司
地址 310012 浙江省杭州市西湖区文一西路460号文娱中心430室

(72)发明人 丁治宇

(74)专利代理机构 上海百一领御专利代理事务所(普通合伙) 31243
代理人 陈贞健 邵栋

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

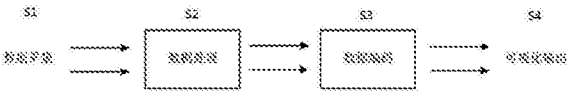
权利要求书2页 说明书9页 附图4页

(54)发明名称

一种舆情数据可视化方法及设备

(57)摘要

本申请提供了一种舆情数据可视化方法及设备，该方案在采集舆情数据之后，先确定每条舆情数据的情感类别和时间信息，然后根据所述舆情数据的情感类别和时间信息，确定每个预设时间段内属于目标情感类别的舆情数据的数量，并获取每个预设时间段内属于目标情感类别的舆情数据的关键词，进而根据每个预设时间段内属于目标情感类别的舆情数据的数量，生成关于每个预设时间段的可视化图形，并在所述可视化图形中添加对应的关键词。由于可视化图形中直观地表现了预设时间段内舆情数据的情感倾向、数量以及涉及的关键词，因此用户通过查看可视化图形即可全面、高效地了解预设时间段内的舆论情况。



CN 108710654 A

1.一种舆情数据可视化方法，其中，该方法包括：

采集舆情数据；

确定每条舆情数据的情感类别和时间信息；

根据所述舆情数据的情感类别和时间信息，确定每个预设时间段内属于目标情感类别的舆情数据的数量，并获取每个预设时间段内属于目标情感类别的舆情数据的关键词；

根据每个预设时间段内属于目标情感类别的舆情数据的数量，生成关于每个预设时间段的可视化图形，并在所述可视化图形中添加对应的关键词。

2.根据权利要求1所述的方法，其中，确定每条舆情数据的情感类别，包括：

根据每条舆情数据的内容，识别并标记每条舆情数据的情感类别，其中，所述情感类别包括正面情感和负面情感。

3.根据权利要求1所述的方法，其中，确定每条舆情数据的时间信息，包括：

若舆情数据中包含发布时间，则将所述发布时间确定为所述舆情数据的时间信息；

若舆情数据中不包含发布时间，则将根据采集所述舆情数据的时间，确定所述舆情数据的时间信息。

4.根据权利要求1所述的方法，其中，根据每个预设时间段内属于目标情感类别的舆情数据的数量，生成关于每个预设时间段的可视化图形，包括：

根据每个预设时间段内属于目标情感类别的舆情数据的数量，确定关于每个预设时间段的可视化图形的图形属性；

根据所述图形属性，在目标情感类别的展示区域内生成所述可视化图形。

5.根据权利要求4所述的方法，其中，根据每个预设时间段内属于目标情感类别的舆情数据的数量，生成关于每个预设时间段的可视化图形之前，还包括：

确定可视化布局，其中，所述可视化布局至少包括目标情感类别的展示区域和可视化图形的形式。

6.根据权利要求1所述的方法，其中，在所述可视化图形中添加对应的关键词，包括：

统计每个预设时间段内属于目标情感类别的舆情数据的关键词的词频；

根据所述关键词的词频，以词云的形式在所述可视化图形中添加对应的关键词，其中，所述关键词的显示尺寸与所述关键词的词频相关。

7.根据权利要求1所述的方法，其中，在所述可视化图形中添加对应的关键词，包括：

根据所述关键词的时间信息，确定所述关键词在所述可视化图形中的位置信息，其中，所述关键词的时间信息为该关键词所属的舆情数据的时间信息；

根据所述位置信息，在所述可视化图形中的相应位置处添加对应的关键词。

8.根据权利要求1所述的方法，其中，该方法还包括：

为所述可视化图形和所述关键词添加色彩，其中，所述色彩与所述可视化图形对应的情感类别相关。

9.一种舆情数据可视化设备，其中，该设备包括：

数据采集装置，用于采集舆情数据；

数据处理装置，用于确定每条舆情数据的情感类别和时间信息，以及根据所述舆情数据的情感类别和时间信息，确定每个预设时间段内属于目标情感类别的舆情数据的数量，并获取每个预设时间段内属于目标情感类别的舆情数据的关键词；

数据编码装置，用于根据每个预设时间段内属于目标情感类别的舆情数据的数量，生成关于每个预设时间段的可视化图形，并在所述可视化图形中添加对应的关键词。

10.根据权利要求9所述的设备，其中，所述数据处理装置，用于根据每条舆情数据的内容，识别并标记每条舆情数据的情感类别，其中，所述情感类别包括正面情感和负面情感。

11.根据权利要求9所述的设备，其中，所述数据处理装置，用于在舆情数据中包含发布时间时，将所述发布时间确定为所述舆情数据的时间信息；以及在舆情数据中不包含发布时间时，将根据采集所述舆情数据的时间，确定所述舆情数据的时间信息。

12.根据权利要求9所述的设备，其中，所述数据编码装置，用于根据每个预设时间段内属于目标情感类别的舆情数据的数量，确定关于每个预设时间段的可视化图形的图形属性，以及根据所述图形属性，在目标情感类别的展示区域内生成所述可视化图形。

13.根据权利要求12所述的设备，其中，所述数据编码装置，还用于在根据每个预设时间段内属于目标情感类别的舆情数据的数量，生成关于每个预设时间段的可视化图形之前，确定可视化布局，其中，所述可视化布局至少包括目标情感类别的展示区域和可视化图形的形式。

14.根据权利要求9所述的设备，其中，所述数据处理装置，还用于统计每个预设时间段内属于目标情感类别的舆情数据的关键词的词频；

所述数据编码装置，用于根据所述关键词的词频，以词云的形式在所述可视化图形中添加对应的关键词，其中，所述关键词的显示尺寸与所述关键词的词频相关。

15.根据权利要求9所述的设备，其中，所述数据编码装置，用于根据所述关键词的时间信息，确定所述关键词在所述可视化图形中的位置信息，以及根据所述位置信息，在所述可视化图形中的相应位置处添加对应的关键词，其中，所述关键词的时间信息为该关键词所属的舆情数据的时间信息。

16.根据权利要求9所述的设备，其中，所述数据编码装置，还用于为所述可视化图形和所述关键词添加色彩，其中，所述色彩与所述可视化图形对应的情感类别相关。

17.一种舆情数据可视化设备，其中，该设备包括：

处理器；以及

存储有机器可读指令的一个或多个机器可读介质，当所述处理器执行所述机器可读指令时，使得所述设备执行如权利要求1至8中任一项所述的方法。

一种舆情数据可视化方法及设备

技术领域

[0001] 本申请涉及信息技术领域，尤其涉及一种舆情数据可视化方法及设备。

背景技术

[0002] 互联网大数据时代下，自媒体等各种新媒体形式层出不穷，针对社会事件和新闻报道的各种评论、文章内容良莠不齐，鱼龙混杂，更有别有用心的势力恶意中伤造谣，刻意发布虚假信息，或对不实报道推波助澜，企图借助快速而多样的互联网传播途误导舆论，由此会导致大规模的舆情事件的发生，给国家和社会造成巨大损失。由于目前互联网上的新闻报道、事件消息、评论等舆情数据发布渠道多，信息量巨大，还没有一种能够让用户能够全面、高效地了解当前舆情倾向的方式。

[0003] 申请内容

[0004] 本申请的一个目的是提供一种舆情数据可视化方案，能够让用户全面、高效地了解当前舆情倾向，进而达成对舆情的监控。

[0005] 为实现上述目的，本申请提供了一种舆情数据可视化方法，该方法包括：

[0006] 采集舆情数据；

[0007] 确定每条舆情数据的情感类别和时间信息；

[0008] 根据所述舆情数据的情感类别和时间信息，确定每个预设时间段内属于目标情感类别的舆情数据的数量，并获取每个预设时间段内属于目标情感类别的舆情数据的关键词；

[0009] 根据每个预设时间段内属于目标情感类别的舆情数据的数量，生成关于每个预设时间段的可视化图形，并在所述可视化图形中添加对应的关键词。

[0010] 基于本申请的另一方面，还提供一种舆情数据可视化设备，其中，该设备包括：

[0011] 数据采集装置，用于采集舆情数据；

[0012] 数据处理装置，用于确定每条舆情数据的情感类别和时间信息，以及根据所述舆情数据的情感类别和时间信息，确定每个预设时间段内属于目标情感类别的舆情数据的数量，并获取每个预设时间段内属于目标情感类别的舆情数据的关键词；

[0013] 数据编码装置，用于根据每个预设时间段内属于目标情感类别的舆情数据的数量，生成关于每个预设时间段的可视化图形，并在所述可视化图形中添加对应的关键词。

[0014] 此外，本申请还提供了一种舆情数据可视化设备，该设备包括：

[0015] 处理器；以及

[0016] 存储有机器可读指令的一个或多个机器可读介质，当所述处理器执行所述机器可读指令时，使得所述设备执行前述的舆情数据可视化方法。

[0017] 本申请提供的方案中，在采集舆情数据之后，先确定每条舆情数据的情感类别和时间信息，然后根据所述舆情数据的情感类别和时间信息，确定预设时间段内属于目标情感类别的舆情数据的数量，并获取所述预设时间段内属于目标情感类别的舆情数据的关键词，进而根据预设时间段内属于目标情感类别的舆情数据的数量，生成关于所述预设时间

段的可视化图形，并在所述可视化图形中添加对应的关键词。由于可视化图形中直观地表现了预设时间段内舆情数据的情感倾向、数量以及涉及的关键词，因此用户通过查看可视化图形即可全面、高效地了解预设时间段内的舆论情况。

附图说明

[0018] 通过阅读参照以下附图所作的对非限制性实施例所作的详细描述，本申请的其它特征、目的和优点将会变得更明显：

[0019] 图1为本申请实施例提供的一种舆情数据可视化方法的处理流程图；

[0020] 图2为本申请实施例中涉及的一种可视化布局的示意图；

[0021] 图3为本申请实施例采用本申请实施例提供的方案生成舆情数据可视化图形时的整体流程图；

[0022] 图4示出了本申请一种实施例中数据处理步骤的详细处理流程；

[0023] 图5示出了本申请一种实施例中数据编码步骤的详细处理流程；

[0024] 图6为本申请实施例提供的一种舆情数据可视化设备的结构示意图；

[0025] 图7为本申请实施例提供的另一种舆情数据可视化设备的结构示意图；

[0026] 附图中相同或相似的附图标记代表相同或相似的部件。

具体实施方式

[0027] 下面结合附图对本申请作进一步详细描述。

[0028] 在本申请一个典型的配置中，终端、服务网络的设备均包括一个或多个处理器(CPU)、输入/输出接口、网络接口和内存。

[0029] 内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器(RAM)和/或非易失性内存等形式，如只读存储器(ROM)或闪存(flash RAM)。内存是计算机可读介质的示例。

[0030] 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体，可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的装置或其他数据。计算机的存储介质的例子包括，但不限于相变内存(PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、其他类型的随机存取存储器(RAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、快闪记忆体或其他内存技术、只读光盘(CD-ROM)、数字多功能光盘(DVD)或其他光学存储、磁盒式磁带，磁带磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。

[0031] 本申请实施例提供了一种舆情数据可视化方法，该方法可以根据一段时间内收集的舆情数据生成可视化图形，并在该可视化图形中直观地表现预设时间段内舆情数据的情感倾向、数量以及涉及的关键词，使得用户通过查看可视化图形即可全面、高效地了解某一时间段内的舆论情况。在实际场景中，该方法的执行主体可以包括但不限于网络主机、单个网络服务器、多个网络服务器集或基于云计算的计算机集合等。在此，云由基于云计算(Cloud Computing)的大量主机或网络服务器构成，其中，云计算是分布式计算的一种，由一群松散耦合的计算机集组成的一个虚拟计算机。

[0032] 图1示出了本申请实施例提供的一种舆情数据可视化方法的处理流程，包括以下

处理步骤：

[0033] S101，采集舆情数据。本申请实施中，所述舆情数据是指有关于舆论情况的数据，可以来源于新闻报道、论坛、博客、微博、社区评论等各类媒体，能够反映出社会公众对于社会事件的发生、发展和变化所持有的态度和情感倾向。在互联网场景下，可以采用爬虫程序(web crawler)根据预设的规则从互联网上获取各类舆情数据，例如定期从各类门户网站获取各类新闻及对于该新闻的评论。

[0034] S102，确定每条舆情数据的情感类别和时间信息，以用于对这些舆情数据进行分类，将特定时间段内的同一情感类别的舆情数据作为一个集合。

[0035] 在确定舆情数据的情感类别时，可以根据每条舆情数据的内容，识别并标记每条舆情数据的情感类别。其中，情感类别可以根据实际分类的需求进行设定，例如本申请实施例中将情感类别设定为两个类别，即正面情感和负面情感，在实际场景中也可以根据实际需求对正面情感和负面情感进行进一步的细分，以获得更多的情感类别。在本申请实施例中，若本次采集到的舆情数据有1至N条，依据情感正负进行分类之后，其中1至M条为归类为正面情感，M+1到N打标归类为负面情感。

[0036] 舆情数据的情感类别可以基于舆情数据的内容来确定，例如某一新闻的内容为报道某地民众见义勇为的事件、并对该行为予以肯定，则该条新闻的情感类别会被分类为正面情感。在实际场景中，可以采用机器学习算法进行自动识别，首先采用标记过情感类别的训练集对机器学习的分类模型进行训练，在完成训练之后，该分类模型即可以对新输入的舆情数据的情感类别进行识别。其中，机器学习的具体算法可以根据实际场景的需求进行选择，例如逻辑回归、决策树、朴素贝叶斯等算法。

[0037] 在确定每条舆情数据的时间信息时，可以将所述发布时间确定为所述舆情数据的时间信息，例如某一舆情数据的发布时间为2018-4-22,20:22:22，则可以确定其时间信息为该发布时间。但是在一些特殊情况下，例如网站不开放相应的接口或者未对发布时间进行记录，则无法获取到发布时间，此时可以将根据采集所述舆情数据的时间，确定所述舆情数据的时间信息。例如对于更新较快的网站，爬虫程序设置为每隔10s对其获取一次该网站的舆情数据，每次获取到舆情数据的时间即设定为该舆情数据的时间信息。由此，在本申请的一些实施例中，若舆情数据中包含发布时间，则将所述发布时间确定为所述舆情数据的时间信息；若舆情数据中不包含发布时间，则将根据采集所述舆情数据的时间，确定所述舆情数据的时间信息。

[0038] 步骤S103，根据所述舆情数据的情感类别和时间信息，确定每个预设时间段内属于目标情感类别的舆情数据的数量，并获取每个预设时间段内属于目标情感类别的舆情数据的关键词。

[0039] 预设时间段作为可视化图形生成的时间单位，例如预设时间段为1个小时，则本申请实施例的方案在进行处理时以1个小时作为统计的时间间隔，来统计1个小时之内的特定情感类别的舆情数据的数量。例如，采集的舆情数据可能包含了最近5个小时内的数据，此时按照1个小时时间间隔划分为5个预设时间段，统计每个预设时间段内舆情数据的数量和关键词。目标情感类别是指需要在可视化图形中体现的情感类别，可以包含所有情感类别，也可以在所有情感类别中选取部分用户需要关注的情感类别，例如本申请实施例中，标情感类别可以是正面情感和负面情感。以预设时间段00:00:01-01:00:00为例，若采集到的舆

情数据的时间信息在该预设时间段之内的有200条，其中，情感类别为正面情感的有130条，情感类别为负面情感的有70条，由此可以确定预设时间段00:00:01-01:00:00内属于目标情感类别的舆情数据的数量为：正面情感130条，负面情感70条。

[0040] 对于采集到的舆情数据，可以采用关键词提取算法提取关键词。本申请实施例进行关键词提取时所使用的算法可以采用任意适用于舆情数据处理场景的算法，例如TF-IDF、KEA等算法。对于提取到的关键词，可以为该关键词添加标志信息，以区分内容相同、但来自于不同预设时间段的舆情数据的关键词。例如，对于舆情数据1，可以提取得到关键词1和关键词2，在记录时可以关键词添加标志信息，例如附带时间戳或者其它能够用于区分关键词所属预设时间段的信息。在本申请的一种实施例中，可以采用如下方式形式记录：关键词_时间段_情感类别，如关键词1_时间段1_正面、关键词1_时间段2_正面等。其中，时间段即为关键词1对应的舆情数据所属的预设时间段，而情感类别则是关键词1对应的舆情数据的情感类别。

[0041] 步骤S104，根据每个预设时间段内属于目标情感类别的舆情数据的数量，生成关于每个预设时间段的可视化图形，并在所述可视化图形中添加对应的关键词。

[0042] 在生成关于每个预设时间段的可视化图形时，可以先根据每个预设时间段内属于目标情感类别的舆情数据的数量，确定关于每个预设时间段的可视化图形的图形属性，该图形属性可以是与可视化图形的视觉形象相关的参数，根据实际采用的图形样式具体确定，例如可以是可视化图形的面积、高度、宽度、直径、曲率中任意一项或者多项的组合。在确定图形属性之后，可以根据所述图形属性，在目标情感类别的展示区域内生成所述可视化图形。例如，在可视化图形为柱状图时，图形属性可以是柱状图的高度，可视化图形为折线图时，图形属性可以是折线图所包围的封闭图形的面积等。

[0043] 本申请的一些实施例中，在根据每个预设时间段内属于目标情感类别的舆情数据的数量，生成关于每个预设时间段的可视化图形之前，可以先确定可视化布局，其中，所述可视化布局至少包括目标情感类别的展示区域和可视化图形的形式。例如，图2示出了本申请实施例中的一种可视化布局，采用镜像布局的样式，即以时间为横轴，作为划分两种情感类别展示区域的基线，将上下划分为两种情感类别的展示区域，基线上方为正面情感的展示区域，在基线下方为负面情感的展示区域，预设时间段内属于目标情感类别的舆情数据的数量作为纵轴，可视化图形的形式为折线图。

[0044] 本申请的一些实施例中，在所述可视化图形中添加对应的关键词时，可以先统计每个预设时间段内属于目标情感类别的舆情数据的关键词的词频，然后根据所述关键词的词频，以词云的形式在所述可视化图形中添加对应的关键词。通过采用词云的形式，使得关键词在可视化图形中的显示尺寸与其出现的词频相关，例如，出现词频高的关键词显示尺寸较大，而出现词频低的关键词显示尺寸则相对较小。

[0045] 本申请实施例中，提取关键词后可以采用“关键词_时间段_情感类别”的形式记录该关键词的统计信息。根据该统计信息，可以进一步统计出每个时间段内关键词的词频，例如，对于关键词1，统计信息一共有关于关键词1的记录60条，其中“关键词1_时间段1_正面”的记录有10条，“关键词1_时间段2_正面”的记录有20条，而“关键词1_时间段4_正面”的记录有30条。在构建词云时，显示尺寸的大小与关键词出现的词频相关，由此可以按照比例确定关键词在每个预设时间段内的显示尺寸，例如，对于前述的关键词1，其在预设时间段1中

出现了10次，在预设时间段2中出现了20次，在预设时间段4中出现了30次，由此在生成词云时，在预设时间段1中显示尺寸最小，在预设时间段2中次之，而在预设时间段4中显示尺寸最大。

[0046] 进一步地，也可以采用同样的方式确定其它关键词的词频，进而基于所有关键词的词频统一确定词云中不同关键词在不同预设时间段中显示尺寸的大小。

[0047] 此外，由于可视化图形对应于预设时间段，与其显示位置会与时间这一属性相关，因此在添加关键词时，关键词的显示位置也可以与时间进行关联，用户可以从添加有关键词的可视化图形中获取更多的信息，例如通过某一时间所对应的位置上的关键词，确定该时刻舆情主要关注的话题。由此，在所述可视化图形中添加对应的关键词时，还可以根据所述关键词的时间信息，确定所述关键词在所述可视化图形中的位置信息，然后根据所述位置信息，在所述可视化图形中的相应位置处添加对应的关键词，其中，所述关键词的时间信息为该关键词所属的舆情数据的时间信息。

[0048] 进一步地，本申请的一些实施例中，还可以为所述可视化图形和所述关键词添加色彩，且所述色彩与所述可视化图形对应的情感类别相关，由此可以使得用户可以从色彩上更加直观的感受每一类情感类别的舆情数据的整体情况。在实际场景中，色彩的选择可以根据用户的使用习惯来设定，例如对于本申请实施例中的两个情感类别，可以使用红、黄等暖色系的色彩标记正面情感，使用蓝、青等冷色系的色彩来标记负面情感。

[0049] 图3、图4和图5示出了采用本申请实施例提供的方案生成舆情数据可视化图形时的处理流程，包括以下处理步骤：

[0050] 步骤S1，数据采集。

[0051] 步骤S2，数据处理。

[0052] 其中，步骤S2包含3个部分的内容，如图4所示。首先，将采集来的舆情数据依据情感正负进行打标分类处理，假设数据内容有1到N条，依据情感正负将内容数据1到M打标归类为正面情感数据，将M+1到N打标归类为负面情感数据。其次，设定一个时间间隔，作为预设时间段，比如1小时，再逐个时间段内进行舆情数据的数量统计。最后，在每一个预设时间段内，依据情感类别的正负分类，统计该预设时间段内所出现的关键词的词频。

[0053] 步骤S3，数据编码。

[0054] 其中，步骤S3如下3个部分的内容，如图5所示。首先，确定用于数据展示的可视化图形的可视化布局。以镜像布局作为布局样式，即以时间为横轴、舆情数据的数量为纵轴，舆情数据的数量为0作为基线，在基线上方为正面情感的展示区域，在基线下方为负面情感的展示区域。

[0055] 其次，按情感类别和确定好的时间段，对舆情数据的数量进行编码，在情感正负区域，分别逐时间段按照统计的数量以折线图形式进行编码。编码形式可以是折线图，也可以是曲线图等其他能和基线构成封闭区域的形式，折线图或曲线图与基线所构成的面积大小表示数量，情感类别的正负可以使用不同颜色进行编码以示区分，例如可以使用红色编码负面情感，使用蓝色编码正面情感。

[0056] 最后，对关键词进行编码，按照情感类别，在各个时间段内，对处理好的关键词及词频数据进行词云编码，用词云形式展示关键词数据，其中词的大小编码关键词的词频数据，情感类别和其时间信息决定其所属的空间位置，依据舆情正负可以使用和各自所属舆

情情感相同颜色的色调编码关键词的颜色以示情感类别。

[0057] 步骤S4，可视化输出。将按照上述过程处理并编码完成的数据，使用计算机或其他手段以可视化图形的形式呈现出来，即可得到一套针对耦合了舆情数据情感变化趋势数据、数量、舆情关键词以及词频数据的多维时序数据的可视化展示方案，利用该方案能够帮助掌握热点舆情、事件的整体发展态势，公众及网民对其的主要评论、看法、观点以及情感倾向，进而达成对互联网舆情的监控及管理，亦可作为舆情事件分析研判的基础。

[0058] 基于同一发明构思，本申请实施例中还提供了一种舆情数据可视化设备，所述设备对应的方法是前述实施例中的舆情数据可视化方法，并且其解决问题的原理与该方法相似。

[0059] 本申请实施例提供的一种舆情数据可视化设备可以根据一段时间内收集的舆情数据生成可视化图形，并在该可视化图形中直观地表现预设时间段内舆情数据的情感倾向、数量以及涉及的关键词，使得用户通过查看可视化图形即可全面、高效地了解某一时间段内的舆论情况。在实际场景中，该设备的具体实现可以包括但不限于网络主机、单个网络服务器、多个网络服务器集或基于云计算的计算机集合等。在此，云由基于云计算(Cloud Computing)的大量主机或网络服务器构成，其中，云计算是分布式计算的一种，由一群松散耦合的计算机集组成的一个虚拟计算机。

[0060] 图6示出了本申请实施例提供的一种舆情数据可视化设备的结构，包括数据采集装置610、数据处理装置620、数据编码装置630。其中，所述数据采集装置610用于采集舆情数据。本申请实施中，所述舆情数据是指有关于舆论情况的数据，可以来源于新闻报道、论坛、博客、微博、社区评论等各类媒体，能够反映出社会公众对于社会事件的发生、发展和变化所持有的态度和情感倾向。在互联网场景下，可以采用爬虫程序(web crawler)根据预设的规则从互联网上获取各类舆情数据，例如定期从各类门户网站获取各类新闻及对于该新闻的评论。

[0061] 数据处理装置620用于确定每条舆情数据的情感类别和时间信息，以用于对这些舆情数据进行分类，将特定时间段内的同一情感类别的舆情数据作为一个集合。

[0062] 在确定舆情数据的情感类别时，所述数据处理装置可以根据每条舆情数据的内容，识别并标记每条舆情数据的情感类别。其中，情感类别可以根据实际分类的需求进行设定，例如本申请实施例中将情感类别设定为两个类别，即正面情感和负面情感，在实际场景中也可以根据实际需求对正面情感和负面情感进行进一步的细分，以获得更多的情感类别。在本申请实施例中，若本次采集到的舆情数据有1至N条，依据情感正负进行分类之后，其中1至M条为归类为正面情感，M+1到N打标归类为负面情感。

[0063] 舆情数据的情感类别可以基于舆情数据的内容来确定，例如某一新闻的内容为报道某地民众见义勇为的事件、并对该行为予以肯定，则该条新闻的情感类别会被分类为正面情感。在实际场景中，可以采用机器学习算法进行自动识别，首先采用标记过情感类别的训练集对机器学习的分类模型进行训练，在完成训练之后，该分类模型即可以对新输入的舆情数据的情感类别进行识别。其中，机器学习的具体算法可以根据实际场景的需求进行选择，例如逻辑回归、决策树、朴素贝叶斯等算法。

[0064] 在确定每条舆情数据的时间信息时，所述数据处理装置可以将所述发布时间确定为所述舆情数据的时间信息，例如某一舆情数据的发布时间为2018-4-22,20:22:22，则可

以确定其时间信息为该发布时间。但是在一些特殊情况下，例如网站不开放相应的接口或者未对发布时间进行记录，则无法获取到发布时间，此时可以将根据采集所述舆情数据的时间，确定所述舆情数据的时间信息。例如对于更新较快的网站，爬虫程序设置为每隔10s对其获取一次该网站的舆情数据，每次获取到舆情数据的时间即设定为该舆情数据的时间信息。由此，在本申请的一些实施例中，若舆情数据中包含发布时间，则将所述发布时间确定为所述舆情数据的时间信息；若舆情数据中不包含发布时间，则将根据采集所述舆情数据的时间，确定所述舆情数据的时间信息。

[0065] 数据处理装置620还用于根据所述舆情数据的情感类别和时间信息，确定每个预设时间段内属于目标情感类别的舆情数据的数量，并获取每个预设时间段内属于目标情感类别的舆情数据的关键词。

[0066] 预设时间段作为可视化图形生成的时间单位，例如预设时间段为1个小时，则本申请实施例的方案在进行处理时以1个小时作为统计的时间间隔，来统计1个小时之内的特定情感类别的舆情数据的数量。例如，采集的舆情数据可能包含了最近5个小时内的数据，此时按照1个小时时间间隔划分为5个预设时间段，统计每个预设时间段内舆情数据的数量和关键词。目标情感类别是指需要在可视化图形中体现的情感类别，可以包含所有情感类别，也可以在所有情感类别中选取部分用户需要关注的情感类别，例如本申请实施例中，标情感类别可以是正面情感和负面情感。以预设时间段00:00:01-01:00:00为例，若采集到的舆情数据的时间信息在该预设时间段之内的有200条，其中，情感类别为正面情感的有130条，情感类别为负面情感的有70条，由此可以确定预设时间段00:00:01-01:00:00内属于目标情感类别的舆情数据的数量为：正面情感130条，负面情感70条。

[0067] 对于采集到的舆情数据，数据处理装置可以采用关键词提取算法提取关键词。本申请实施例进行关键词提取时所使用的算法可以采用任意适用于舆情数据处理场景的算法，例如TF-IDF、KEA等算法。对于提取到的关键词，可以为该关键词添加标志信息，以区分内容相同、但来自于不同舆情数据的关键词。例如，对于舆情数据1，可以提取得到关键词1和关键词2，在记录时可以关键词添加标志信息，例如附带时间戳或者其它能够用于区分关键词所属预设时间段的信息。在本申请的一种实施例中，可以采用如下方式形式记录：关键词_时间段_情感类别，如关键词1_时间段1_正面、关键词1_时间段2_正面等。其中，时间段即为关键词1对应的舆情数据所属的预设时间段，而情感类别则是关键词1对应的舆情数据的情感类别。

[0068] 数据编码装置630用于根据每个预设时间段内属于目标情感类别的舆情数据的数量，生成关于每个预设时间段的可视化图形，并在所述可视化图形中添加对应的关键词。

[0069] 在生成关于每个预设时间段的可视化图形时，数据编码装置可以先根据每个预设时间段内属于目标情感类别的舆情数据的数量，确定关于每个预设时间段的可视化图形的图形属性，该图形属性可以是与可视化图形的视觉形象相关的参数，根据根据实际采用的图形样式具体确定，例如可以是可视化图形的面积、高度、宽度、直径、曲率中任意一项或者多项的组合。在确定图形属性之后，可以根据所述图形属性，在目标情感类别的展示区域内生成所述可视化图形。例如，在可视化图形为柱状图时，图形属性可以是柱状图的高度，可视化图形为折线图时，图形属性可以是折线图所包围的封闭图形的面积等。

[0070] 本申请的一些实施例中，在根据预设时间段内属于目标情感类别的舆情数据的数

量，生成关于所述预设时间段的可视化图形之前，数据编码装置可以先确定可视化布局，其中，所述可视化布局至少包括目标情感类别的展示区域和可视化图形的形式。例如，图2示出了本申请实施例中的一种可视化布局，采用镜像布局的样式，即以时间为横轴，作为划分两种情感类别展示区域的基线，将上下划分为两种情感类别的展示区域，基线上方为正面情感的展示区域，在基线下方为负面情感的展示区域，预设时间段内属于目标情感类别的舆情数据的数量作为纵轴，可视化图形的形式为折线图。

[0071] 本申请的一些实施例中，在所述可视化图形中添加对应的关键词时，数据处理装置可以先统计每个预设时间段内属于目标情感类别的舆情数据的关键词的词频，然后数据编码装置根据所述关键词的词频，以词云的形式在所述可视化图形中添加对应的关键词。通过采用词云的形式，使得关键词在可视化图形中的显示尺寸与其出现的词频相关，例如，出现词频高的关键词显示尺寸较大，而出现词频低的关键词显示尺寸则相对较小。

[0072] 本申请实施例中，提取关键词后可以采用“关键词_时间段_情感类别”的形式记录该关键词的统计信息。根据该统计信息，可以进一步统计出每个时间段内关键词的词频，例如，对于关键词1，统计信息一共有关于关键词1的记录60条，其中“关键词1_时间段1_正面”的记录有10条，“关键词1_时间段2_正面”的记录有20条，而“关键词1_时间段4_正面”的记录有30条。在构建词云时，显示尺寸的大小与关键词出现的词频相关，由此可以按照比例确定关键词在每个预设时间段内的显示尺寸，例如，对于前述的关键词1，其在预设时间段1中出现了10次，在预设时间段2中出现了20次，在预设时间段4中出现了30次，由此在生成词云时，在预设时间段1中显示尺寸最小，在预设时间段2中次之，而在预设时间段4中显示尺寸最大。

[0073] 进一步地，也可以采用同样的方式确定其它关键词的词频，进而基于所有关键词的词频统一确定词云中不同关键词在不同预设时间段中显示尺寸的大小。

[0074] 此外，由于可视化图形对应于预设时间段，与其显示位置会与时间这一属性相关，因此在添加关键词时，关键词的显示位置也可以与时间进行关联，用户可以从添加有关键词的可视化图形中获取更多的信息，例如通过某一时间所对应的位置上的关键词，确定该时刻舆情主要关注的话题。由此，在所述可视化图形中添加对应的关键词时，数据编码装置还可以根据所述关键词的时间信息，确定所述关键词在所述可视化图形中的位置信息，然后根据所述位置信息，在所述可视化图形中的相应位置处添加对应的关键词，其中，所述关键词的时间信息为该关键词所属的舆情数据的时间信息。

[0075] 进一步地，本申请的一些实施例中，数据编码装置还可以为所述可视化图形和所述关键词添加色彩，且所述色彩与所述可视化图形对应的情感类别相关，由此可以使得用户可以从色彩上更加直观的感受每一类情感类别的舆情数据的整体情况。在实际场景中，色彩的选择可以根据用户的使用习惯来设定，例如对于本申请实施例中的两个情感类别，可以使用红、黄等暖色系的色彩标记正面情感，使用蓝、青等冷色系的色彩来标记负面情感。

[0076] 另外，本申请的一部分可被应用为计算机程序产品，例如计算机程序指令，当其被计算机执行时，通过该计算机的操作，可以调用或提供根据本申请的方法和/或技术方案。而调用本申请的方法的程序指令，可能被存储在固定的或可移动的记录介质中，和/或通过广播或其他信号承载媒体中的数据流而被传输，和/或被存储在根据程序指令运行的计算

机设备的工作存储器中。在此，根据本申请的一个实施例包括一个如图7所示的设备，该设备包括存储有机器可读指令的一个或多个机器可读介质710和用于执行机器可读指令的处理器720，其中，当该机器可读指令被该处理器执行时，使得所述设备执行基于前述根据本申请的多个实施例的方法和/或技术方案。

[0077] 此外，本申请的一些实施例还提供了一种计算机可读介质，其上存储有计算机程序指令，所述计算机可读指令可被处理器执行以实现前述本申请的多个实施例的方法和/或技术方案。

[0078] 需要注意的是，本申请可在软件和/或软件与硬件的组合体中被实施，例如，可采用专用集成电路(ASIC)、通用目的计算机或任何其他类似硬件设备来实现。在一个实施例中，本申请的软件程序可以通过处理器执行以实现上文步骤或功能。同样地，本申请的软件程序(包括相关的数据结构)可以被存储到计算机可读记录介质中，例如，RAM存储器，磁或光驱动器或软磁盘及类似设备。另外，本申请的一些步骤或功能可采用硬件来实现，例如，作为与处理器配合从而执行各个步骤或功能的电路。

[0079] 对于本领域技术人员而言，显然本申请不限于上述示范性实施例的细节，而且在不背离本申请的精神或基本特征的情况下，能够以其他的具体形式实现本申请。因此，无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本申请的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本申请内。不应将权利要求中的任何附图标记视为限制所涉及的权利要求。此外，显然“包括”一词不排除其他单元或步骤，单数不排除复数。装置权利要求中陈述的多个单元或装置也可以由一个单元或装置通过软件或者硬件来实现。第一，第二等词语用来表示名称，而并不表示任何特定的顺序。

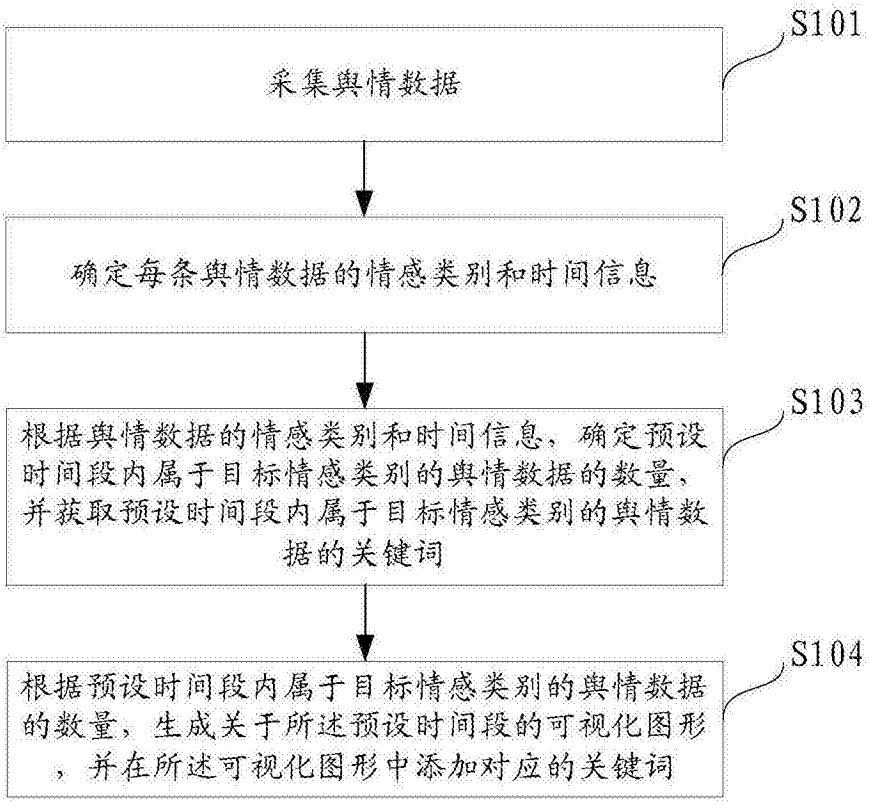

S101
采集舆情数据
S102
确定每条舆情数据的情感类别和时间信息
S103
根据舆情数据的情感类别和时间信息，确定预设时间段内属于目标情感类别的舆情数据的数量，并获取预设时间段内属于目标情感类别的舆情数据的关键词
S104
根据预设时间段内属于目标情感类别的舆情数据的数量，生成关于所述预设时间段的可视化图形，并在所述可视化图形中添加对应的关键词

图1

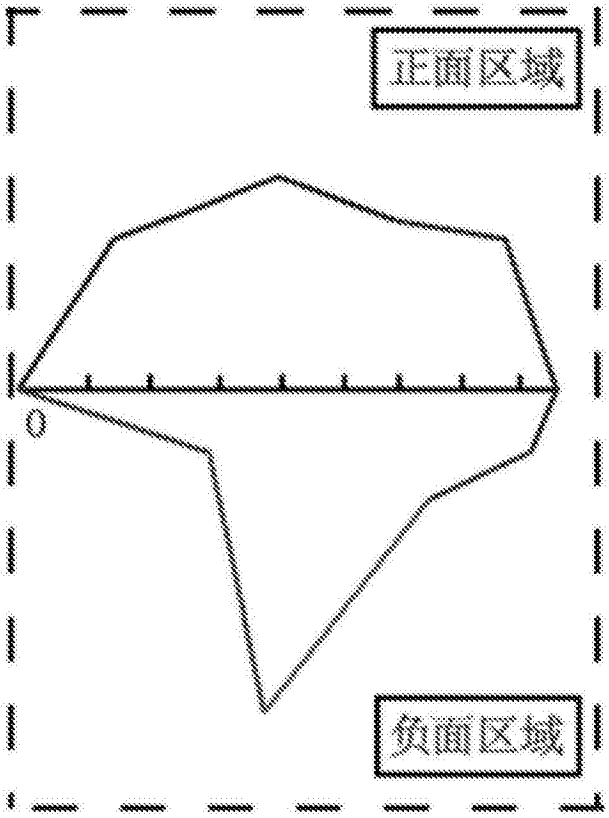


图2

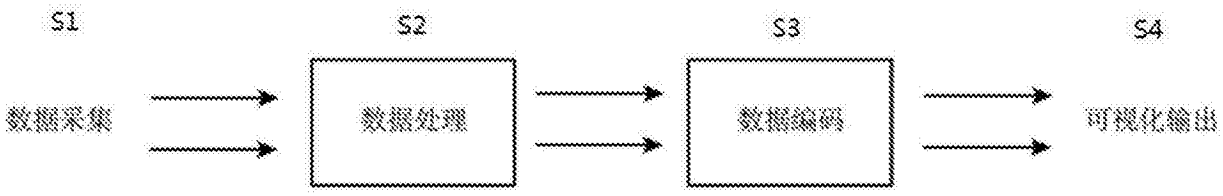


图3

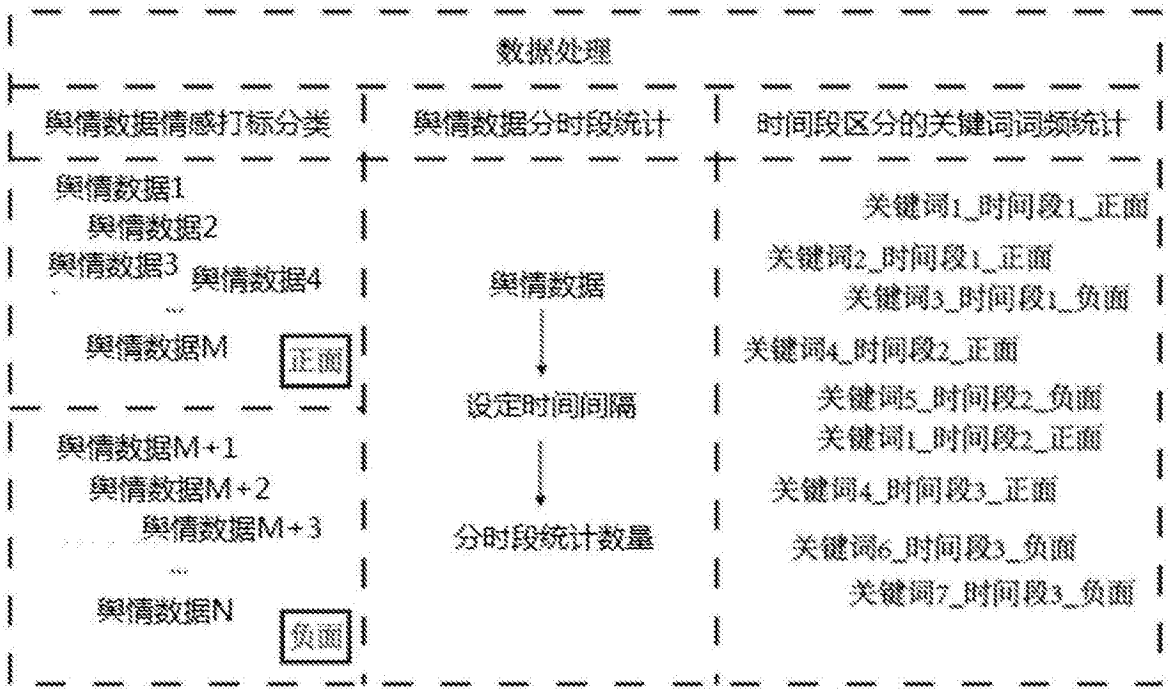

数据处理
舆情数据情感打标分类
舆情数据分时段统计
时间段区分的关键词词频统计
舆情数据1
舆情数据2
舆情数据3
舆情数据4
...
舆情数据M
正面
舆情数据M+1
舆情数据M+2
舆情数据M+3
...
舆情数据N
负面
舆情数据
设定时间间隔
分时段统计数量
关键词1_时间段1_正面
关键词2_时间段1_正面
关键词3_时间段1_负面
关键词4_时间段2_正面
关键词5_时间段2_负面
关键词1_时间段2_正面
关键词4_时间段3_正面
关键词6_时间段3_负面
关键词7_时间段3_负面

图4

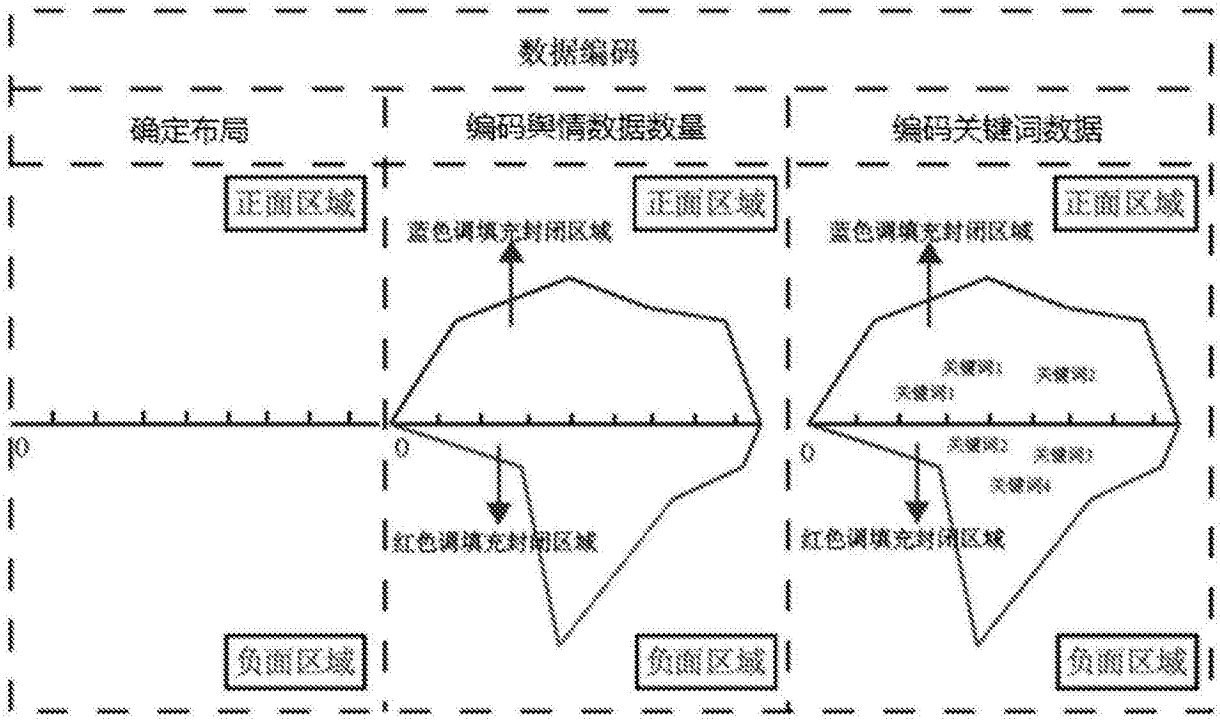

数据编码
确定布局
编码舆情数据数量
编码关键词数据
正面区域
蓝色调填充封闭区域
0
红色调填充封闭区域
负面区域

图5

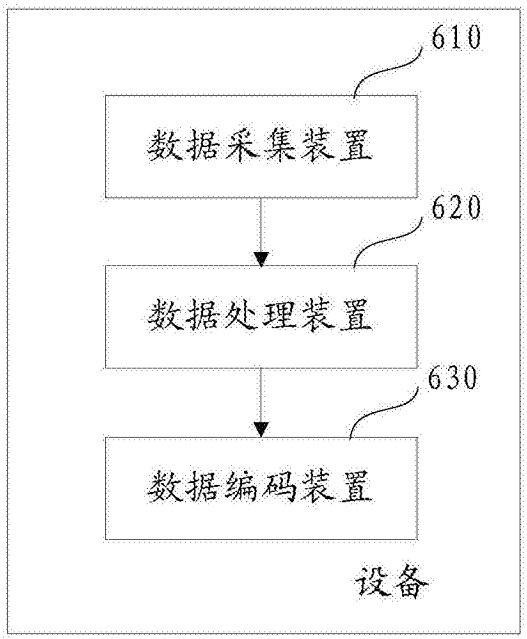


图6

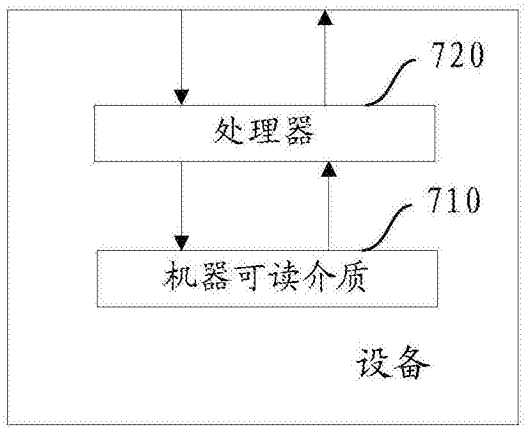


图7