(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111160019 B
(45)授权公告日 2023.08.15

(21)申请号 201911404334.4

(22)申请日 2019.12.30

(65)同一申请的已公布的文献号
申请公布号 CN 111160019 A

(43)申请公布日 2020.05.15

(73)专利权人 中国联合网络通信集团有限公司
地址 100033 北京市西城区金融大街21号
专利权人 联通系统集成有限公司
联通(黑龙江)产业互联网有限公司

(72)发明人 董浩俊 胡坤 房啟麾 赵文奇

(74)专利代理机构 北京同立钧成知识产权代理有限公司 11205
专利代理师 张宁 刘芳

(51)Int.Cl.
G06F 40/279(2020.01)
G06F 40/169(2020.01)
G06F 16/35(2019.01)

(56)对比文件
CN 110188337 A,2019.08.30
CN 107544988 A,2018.01.05
CN 103544255 A,2014.01.29
CN 106844786 A,2017.06.13
CN 107491548 A,2017.12.19
CN 108959383 A,2018.12.07
CN 110334300 A,2019.10.15
CN 109992661 A,2019.07.09
CN 109271512 A,2019.01.25
CN 109145215 A,2019.01.04
CN 106294619 A,2017.01.04
CN 107315778 A,2017.11.03
CN 109684646 A,2019.04.26
CN 109325165 A,2019.02.12
CN 104965847 A,2015.10.07
CN 105824959 A,2016.08.03
CN 110069623 A,2019.07.30
CN 104184750 A,2014.12.03
胡坤.基于社交关系强度的社区发现及商品推荐模型.《CNKI中国知网》.2018,全文.

审查员 彭帜

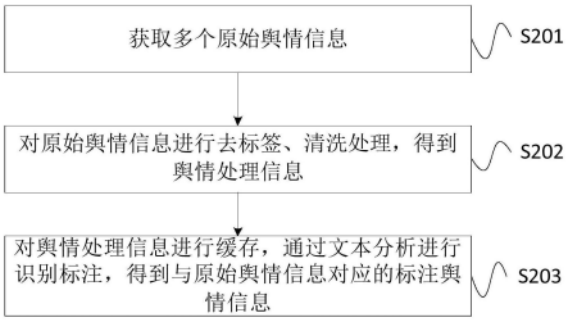
权利要求书2页 说明书12页 附图4页

(54)发明名称

一种舆情监测的方法、装置及系统

(57)摘要

本发明提供一种舆情监测的方法、装置及系统,该方法,包括:获取多个原始舆情信息;对所述原始舆情信息进行去标签、清洗处理,得到舆情处理信息;将所述舆情处理信息进行缓存,通过文本分析进行识别标注,得到与所述原始舆情信息对应的标注舆情信息。减少了人工成本,提高了舆情监测的准确率、有效性,极大的提高了舆情监测的效率。



CN 111160019 B

1.一种舆情监测的方法,其特征在于,包括:

获取多个原始舆情信息;

对所述原始舆情信息进行去标签、清洗处理,得到舆情处理信息;

将所述舆情处理信息进行缓存,通过文本分析进行识别标注,得到与所述原始舆情信息对应的标注舆情信息;

所述通过文本分析进行识别标注,得到与所述原始舆情信息对应的标注舆情信息包括:对舆情处理信息进行分词,将分词后的舆情处理信息在地域词典中进行匹配,若匹配成功则对所述舆情处理信息进行地域标注,得到地域舆情处理信息;根据所述地域舆情处理信息出现的位置以及频次,获得所述地域舆情处理信息对应的评分;根据所述评分的大小依次进行排序,并将最高评分对应的所述地域舆情处理信息进行地域标注,得到与原始舆情信息对应的标注舆情信息;所述地域词典通过获取地域词汇,并将所述地域词汇整理构建获得;

获取舆情处理信息中的摘要文本信息,提取、标注所述摘要文本信息中的转折句,得到与所述原始舆情信息对应的标注舆情信息;

对所述摘要文本信息中每个摘要语句求取相似性;

获取最高相似性对应的摘要语句并删除,得到保留摘要语句并进行标注,得到与所述原始舆情信息对应的标注舆情信息;

其中,所述对所述摘要文本信息中每个摘要语句求取相似性包括:采用关键字提取TextRank公式和相似程度计算公式对每个摘要语句求取相似性;

所述TextRank公式为:

$$WS(V_i)=(1-d)+d*\sum_{V_j\in In(V_i)}\frac{w_{ji}}{\sum_{V_k\in Out(V_j)}w_{jk}}WS(V_j)$$

式中,TextRank公式左边表示一个摘要句子的权重(WS是weight_sum的缩写),右侧的求和表示每个相邻摘要句子对本摘要句子的贡献程度,wji表示两个句子的相似程度,WS(Vj)代表上次迭代j的权重,Vi表示某个网页,Vj表示链接到Vi的网页,In(Vi)表示网页Vi的所有入链的集合,Out(Vj)表示网页的所有出链的集合,d表示阻尼系数;

所述相似程度计算公式为:

$$Score(Q,d)=\sum_{i}^{n}IDF(q_i)\cdot\frac{f_i\cdot(k_1+1)}{f_i+k_1\cdot\left(1-b+b\cdot\frac{dl}{avgdl}\right)}$$

式中,fi是词在文章中的出现次数,dl是文章长度,avgdl是文章平均长度,IDF为逆向文本频率表示词普遍重要性的度量,Q表示查询Query,qi表示Q解析之后的一个语素,d表示一个搜索结果文本,k1和b均为人为设置的调节因子。

2.根据权利要求1所述的方法,其特征在于,通过文本分析进行识别标注,得到与所述原始舆情信息对应的标注舆情信息,包括:

将所述舆情处理信息划分为第一类文本信息、第二类文本信息;

对所述第二类文本信息进行干扰项排除处理,通过机器学习模型对所述第二类文本信

息进行监督分类,获取正向舆情信息并标注,得到与所述原始舆情信息对应的标注舆情信息;其中机器学习模型采用情感词库作为训练数据集进行监督训练,以输出分类为正向舆情信息的机器学习模型。

3.根据权利要求1所述的方法,其特征在于,通过文本分析进行识别标注,得到与所述原始舆情信息对应的标注舆情信息,包括:

获取舆情处理信息中多个文本信息,且所述文本信息包括多个主题信息;

所述文本信息、所述主题信息的分布参数分别服从Dirichlet分布;

根据所述文本信息的分布参数服从Dirichlet分布,生成对应的主题信息;

根据所述主题信息的分布参数服从Dirichlet分布,生成对应的词信息;

遍历所述的文本信息、主题信息生成所述主题信息对应的词信息;得到与原始舆情信息对应的标注舆情信息。

4.根据权利要求1所述的方法,其特征在于,通过文本分析进行识别标注,得到与所述原始舆情信息对应的标注舆情信息,包括:

将舆情处理信息中对应的文本信息以及标题信息进行分词处理,得到与所述文本信息、所述标题信息分别对应的词袋向量;

将所述词袋向量作为特征计算所有文本信息的相似性,通过聚类删除相似性低于预设阈值的文本信息以及标题信息,并将保留的文本信息以及标题信息进行标注,得到与所述原始舆情信息对应的标注舆情信息。

5.根据权利要求1-4中任一项所述的方法,其特征在于,获取多个原始舆情信息,包括:

按照预设规则,从多个网络资源中获取原始舆情信息;

或者通过API接口获取原始舆情信息。

6.根据权利要求1所述的方法,其特征在于,在得到与所述原始舆情信息对应的标注舆情信息之后,还包括:

将所述标注舆情信息与所述原始舆情信息存储于全文搜索引擎中,以便在交互界面上搜索显示所述标注舆情信息对应的原始舆情信息。

一种舆情监测的方法、装置及系统

技术领域

[0001] 本发明涉及计算机技术领域，尤其涉及一种舆情监测的方法、装置及系统。

背景技术

[0002] 目前，通过面向各大运营商进行舆情分析，可基于舆情分析的结果为各大运营商的运营起到辅助决策的作用。舆情分析就是根据特定问题的需要，对针对这个问题的舆情进行深层次的思维加工和分析研究，得到相关结论的过程。

[0003] 现有技术中，在进行舆情分析时，一般采用人工研判为主系统判断为辅的方式进行分析，而且，对于部分舆情所属业务的识别是采用基于关键词简单匹配的方式来确定业务方式。

[0004] 然而现有技术中，基于人工研判为主的方式进行地分析，将会导致现有分析过程在舆情处理的实效性较差，而且采用关键词简单匹配的方法来确定业务方式时，由于采用的匹配方式较为简单，从而将会造成舆情分析的准确率较低。

发明内容

[0005] 本发明提供一种舆情监测的方法、装置及系统，以减少了人工成本，提高舆情监测的准确率、有效性，极大的提高了舆情监测的效率。

[0006] 第一方面，本发明实施例提供的一种舆情监测的方法，包括：

[0007] 获取多个原始舆情信息；

[0008] 对所述原始舆情信息进行去标签、清洗处理，得到舆情处理信息；

[0009] 将所述舆情处理信息进行缓存，通过文本分析进行识别标注，得到与所述原始舆情信息对应的标注舆情信息。

[0010] 在一种可能的设计中，通过文本分析进行识别标注，得到与所述原始舆情信息对应的标注舆情信息，包括：

[0011] 对舆情处理信息进行分词，将分析后的舆情处理信息在地域词典中进行匹配，若匹配成功则对舆情信息进行地域标识，得到地域舆情处理信息；

[0012] 根据所述地域舆情处理信息出现的位置以及频次，获得所述地域舆情处理信息对应的评分；

[0013] 根据所述评分的大小依次进行排序，并将最高评分对应的所述地域舆情处理信息进行地域标注，得到与原始舆情信息对应的标注舆情信息。

[0014] 在一种可能的设计中，所述地域词典通过获取地域词汇，并将所述地域词汇整理构建获得。

[0015] 在一种可能的设计中，通过文本分析进行识别标注，得到与所述原始舆情信息对应的标注舆情信息，还包括：

[0016] 获取舆情处理信息中的摘要文本信息，提取、标注所述摘要文本信息中的转折句，得到与所述原始舆情信息对应的标注舆情信息。

[0017] 在一种可能的设计中，所述方法，还包括：

[0018] 对所述摘要文本信息中每个摘要语句求取相似性；

[0019] 获取最高相似性对应的摘要语句并删除，得到保留摘要语句并进行标注，得到与所述原始舆情信息对应的标注舆情信息。

[0020] 在一种可能的设计中，通过文本分析进行识别标注，得到与所述原始舆情信息对应的标注舆情信息，包括：

[0021] 将所述舆情处理信息划分为第一类文本信息、第二类文本信息；

[0022] 对所述第二类文本信息进行干扰项排除处理，通过机器学习模型对所述第二类文本信息进行监督分类，获取正向舆情信息并标注，得到与所述原始舆情信息对应的标注舆情信息；其中机器学习模型采用情感词库作为训练数据集进行监督训练，以输出分类为正向舆情信息的机器学习模型。

[0023] 在一种可能的设计中，通过文本分析进行识别标注，得到与所述原始舆情信息对应的标注舆情信息，包括：

[0024] 获取舆情处理信息中多个文本信息，且所述文本信息包括多个主题信息；

[0025] 所述文本信息、所述主题信息的分布参数分别服从Dirichlet分布；

[0026] 根据所述文本信息的分布参数服从Dirichlet分布，生成对应的主题信息；

[0027] 根据所述主题信息的分布参数服从Dirichlet分布，生成对应的词信息；

[0028] 遍历所述的文本信息、主题信息生成所述主题信息对应的词信息；得到与原始舆情信息对应的标注舆情信息。

[0029] 在一种可能的设计中，通过文本分析进行识别标注，得到与所述原始舆情信息对应的标注舆情信息，包括：

[0030] 将舆情处理信息中对应的文本信息以及标题信息进行分词处理，得到与所述文本信息、所述标题信息分别对应的词袋向量；

[0031] 将所述词袋向量作为特征计算所有文本信息的相似性，通过聚类删除相似性低于预设阈值的文本信息以及标题信息，并将保留的文本信息以及标题信息进行标注，得到与所述原始舆情信息对应的标注舆情信息。

[0032] 在一种可能的设计中，获取多个原始舆情信息，包括：

[0033] 按照预设规则，从多个网络资源中获取原始舆情信息；

[0034] 或者通过API接口获取原始舆情信息。

[0035] 在一种可能的设计中，在得到与所述原始舆情信息对应的标注舆情信息之后，还包括：

[0036] 将所述标注舆情信息与所述原始舆情信息存储于全文搜索引擎中，以便在交互界面上搜索显示所述标注舆情信息对应的原始舆情信息。

[0037] 第二方面，本发明实施例提供的一种舆情监测的装置，包括：

[0038] 获取模块，用于获取多个原始舆情信息；

[0039] 得到模块，用于对所述原始舆情信息进行去标签、清洗处理，得到舆情处理信息；

[0040] 标注模块，用于将所述舆情处理信息进行缓存，通过文本分析进行识别标注，得到与所述原始舆情信息对应的标注舆情信息。

[0041] 在一种可能的设计中，通过文本分析进行识别标注，得到与所述原始舆情信息对

应的标注舆情信息，包括：

[0042] 对舆情处理信息进行分词，将分析后的舆情处理信息在地域词典中进行匹配，若匹配成功则对舆情信息进行地域标识，得到地域舆情处理信息；

[0043] 根据所述地域舆情处理信息出现的位置以及频次，获得所述地域舆情处理信息对应的评分；

[0044] 根据所述评分的大小依次进行排序，并将最高评分对应的所述地域舆情处理信息进行地域标注，得到与原始舆情信息对应的标注舆情信息。

[0045] 在一种可能的设计中，所述地域词典通过获取地域词汇，并将所述地域词汇整理构建获得。

[0046] 在一种可能的设计中，通过文本分析进行识别标注，得到与所述原始舆情信息对应的标注舆情信息，还包括：

[0047] 获取舆情处理信息中的摘要文本信息，提取、标注所述摘要文本信息中的转折句，得到与所述原始舆情信息对应的标注舆情信息。

[0048] 在一种可能的设计中，还包括：

[0049] 对所述摘要文本信息中每个摘要语句求取相似性；

[0050] 获取最高相似性对应的摘要语句并删除，得到保留摘要语句并进行标注，得到与所述原始舆情信息对应的标注舆情信息。

[0051] 在一种可能的设计中，通过文本分析进行识别标注，得到与所述原始舆情信息对应的标注舆情信息，包括：

[0052] 将所述舆情处理信息划分为第一类文本信息、第二类文本信息；

[0053] 对所述第二类文本信息进行干扰项排除处理，通过机器学习模型对所述第二类文本信息进行监督分类，获取正向舆情信息并标注，得到与所述原始舆情信息对应的标注舆情信息；其中机器学习模型采用情感词库作为训练数据集进行监督训练，以输出分类为正向舆情信息的机器学习模型。

[0054] 在一种可能的设计中，通过文本分析进行识别标注，得到与所述原始舆情信息对应的标注舆情信息，包括：

[0055] 获取舆情处理信息中多个文本信息，且所述文本信息包括多个主题信息；

[0056] 所述文本信息、所述主题信息的分布参数分别服从Dirichlet分布；

[0057] 根据所述文本信息的分布参数服从Dirichlet分布，生成对应的主题信息；

[0058] 根据所述主题信息的分布参数服从Dirichlet分布，生成对应的词信息；

[0059] 遍历所述的文本信息、主题信息生成所述主题信息对应的词信息；得到与原始舆情信息对应的标注舆情信息。

[0060] 在一种可能的设计中，通过文本分析进行识别标注，得到与所述原始舆情信息对应的标注舆情信息，包括：

[0061] 将舆情处理信息中对应的文本信息以及标题信息进行分词处理，得到与所述文本信息、所述标题信息分别对应的词袋向量；

[0062] 将所述词袋向量作为特征计算所有文本信息的相似性，通过聚类删除相似性低于预设阈值的文本信息以及标题信息，并将保留的文本信息以及标题信息进行标注，得到与所述原始舆情信息对应的标注舆情信息。

[0063] 在一种可能的设计中，获取多个原始舆情信息，包括：

[0064] 按照预设规则，从多个网络资源中获取原始舆情信息；

[0065] 或者通过API接口获取原始舆情信息。

[0066] 在一种可能的设计中，在得到与所述原始舆情信息对应的标注舆情信息之后，还包括：

[0067] 将所述标注舆情信息与所述原始舆情信息存储于全文搜索引擎中，以便在交互界面上搜索显示所述标注舆情信息对应的原始舆情信息。

[0068] 第三方面，本发明实施例提供的一种舆情监测的系统，包括：存储器和处理器，存储器中存储有所述处理器的可执行指令；其中，所述处理器配置为经由执行所述可执行指令来执行第一方面中任一项所述的舆情监测的方法。

[0069] 第四方面，本发明实施例提供的一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现第一方面中任一项所述的舆情监测的方法。

[0070] 本发明提供一种舆情监测的方法、装置及系统，该方法，包括：获取多个原始舆情信息；对所述原始舆情信息进行去标签、清洗处理，得到舆情处理信息；将所述舆情处理信息进行缓存，通过文本分析进行识别标注，得到与所述原始舆情信息对应的标注舆情信息。减少了人工成本，提高了舆情监测的准确率、有效性，极大的提高了舆情监测的效率。

附图说明

[0071] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0072] 图1为本发明一典型应用示意图；

[0073] 图2为本发明实施例一提供的舆情监测的方法的流程图；

[0074] 图3为本发明实施例一提供的舆情监测的方法中数据采集的示意图；

[0075] 图4为本发明实施例一提供的舆情监测的方法的示意图；

[0076] 图5为本发明实施例一提供的舆情监测的方法中部分方法的示意图；

[0077] 图6为本发明实施例二提供的舆情监测的方法的流程图；

[0078] 图7为本发明实施例三提供的舆情监测的装置的结构示意图；

[0079] 图8为本发明实施例四提供的舆情监测的系统的结构示意图。

具体实施方式

[0080] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0081] 本发明的说明书和权利要求书及上述附图中的术语“第一”、“第二”、“第三”“第四”等(如果存在)是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的本发明的实施例例如能够以除

了在这里图示或描述的那些以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0082] 下面以具体地实施例对本发明的技术方案进行详细说明。下面这几个具体的实施例可以相互结合，对于相同或相似的概念或过程可能在某些实施例不再赘述。

[0083] 图1为本发明一典型应用示意图，如图1所示，终端设备11可以与互联网平台12进行通信，互联网平台可以通过互联网发布舆情信息的平台，且并不限定于一个，可以包括官方网站和非官方网站。舆情监测系统通过获取多个原始舆情信息；对原始舆情信息进行去标签、清洗处理，得到舆情处理信息；将舆情处理信息进行缓存，通过文本分析进行识别标注，得到与原始舆情信息对应的标注舆情信息。其中舆情监测系统可以设置连接于互联网平台的数据库，本发明不作限定。舆情监测系统还可以在终端设备交互界面上搜索显示标注舆情信息对应的原始舆情信息等等。终端设备11可以为智能手机、平板电脑、笔记本电脑、超级移动个人计算机(ultra-mobile personal computer，UMPC)、上网本、个人数字助理(personal digital assistant，PDA)等。减少了人工成本，提高舆情监测的准确率、有效性，极大的提高了舆情监测的效率。

[0084] 图2为本发明实施例一提供的舆情监测的方法的流程图，如图2所示，本实施例中的方法可以包括：

[0085] S201、获取多个原始舆情信息。

[0086] 本实施例中可以按照预设规则，例如每隔10分钟从多个网络资源中获取原始舆情信息；还可以通过API接口获取原始舆情信息。参考图3，图3为本发明实施例一提供的舆情监测的方法中数据采集的示意图。

[0087] 如图3所示，例如通过网络爬虫程序搜索网页文件读取舆情信息，舆情信息包括文章标题、内容、作者、发布时间、网站名称、所属版块、网站链接URL、阅读数、点赞数、评论数。网络资源可以包括新闻网站、地方门户、专业网站、纸媒(电子报)、论坛、博客、微信公众号、手机APP、新浪微博等。又例如针对定义的20余万重点站点中的新闻、论坛、博客等定向全面抓取来获得原始舆情信息。又例如通过自定义采集栏目、URL、更新时间、扫描间隔等，以便及时获取原始舆情信息。再例如，通过与新浪微博官方接口对接，例如通过官方API接口准实时获取微博类的原始舆情信息，主要包括以下信息：微博ID、微博内容、发布时间、采集时间、转发数、评论数、点赞数、微博作者名称、微博作者ID、作者性别、作者头像、是否认证、认证类型、注册省分、注册城市等。

[0088] S202、对原始舆情信息进行去标签、清洗处理，得到舆情处理信息。

[0089] 本实施例中通过SparkStreaming程序进行数据的预处理，在这一步骤需要对原始舆情信息去html标签处理，去html标签采用正则表达式来匹配规则，对于满足规则的标签，如<html></html>
<p><style/>等html标签进行去除，只保留文本信息内容，得到舆情处理信息。还可以调用噪音过滤模型对原始舆情数据进行清洗，去除提及运营商关键词但是与运营商业务不相关的舆情信息。其中噪音过滤模型基于关键词匹配的方式进行数据的清洗，因此首先整理过滤词库，对于命中过滤词的舆情信息进行标注并直接入库。

[0090] S203、将舆情处理信息进行缓存，通过文本分析进行识别标注，得到与原始舆情信

息对应的标注舆情信息。

[0091] 具体的，将舆情处理信息写入到Kafka消息中间件中，进而通过文本分析进行识别标注，得到与原始舆情信息对应的标注舆情信息。可以对舆情处理信息进行分词并进行地域标注，得到地域舆情处理信息；可以获取舆情处理信息中的摘要文本信息等等，对其中的转折句进行标注，得到与原始舆情信息对应的标注舆情信息；还可以对摘要文本信息中每个摘要语句求取相似性，得到保留摘要语句并进行标注，得到与原始舆情信息对应的标注舆情信息；或者获取正向舆情信息并标注，生成文本信息中主题信息对应的词信息，通过聚类删除相似性低于预设阈值的文本信息以及标题信息，并将保留的文本信息以及标题信息进行标注等等，得到与原始舆情信息对应的标注舆情信息。

[0092] 下面依次详细介绍通过文本分析进行识别标注，得到与原始舆情信息对应的标注舆情信息可以参考图4，图4为本发明实施例一提供的舆情监测的方法的示意图。

[0093] 在一种可选的实施中，对舆情处理信息进行分词，将分词后的舆情处理信息在地域词典中进行匹配，若匹配成功则对舆情信息进行地域标注，得到地域舆情处理信息；根据地域舆情处理信息出现的位置以及频次，获得地域舆情处理信息对应的评分；根据评分的大小依次进行排序，并将最高评分对应的地域舆情处理信息进行地域标注，得到与原始舆情信息对应的标注舆情信息。其中，地域词典通过获取地域词汇，并将地域词汇整理构建获得。

[0094] 例如，基于命名实体识别的算法，通过获取地域词汇，将地域词汇整理构建生成地域词典。对舆情处理信息进行分词，将分词后的舆情处理信息在地域词典中进行匹配，若在地域词典中匹配成功则表示该舆情处理信息中包括地域词典中的地域词汇，则对该舆情处理信息进行地域标注，从而得到地域舆情处理信息；根据该地域舆情处理信息出现的位置(例如标题、首段等等)以及出现的频率次数，综合得出该地域舆情处理信息对应的评分，对所有的匹配成功的地域舆情处理信息按照评分大小由高到低依次进行排序，并选择最高评分对应的地域舆情处理信息进行地域标注，并作为与原始舆情信息对应的地域，得到与原始舆情信息对应的标注舆情信息。

[0095] 在一种可选的实施例中，获取舆情处理信息中的摘要文本信息，提取、标注摘要要文本信息中的转折句，得到与原始舆情信息对应的标注舆情信息。

[0096] 为能够获取舆情处理信息文章主旨的概括，例如摘要，以方便阅读，基于Google开源Pagerank算法形成的Textrank算法获取舆情处理信息中的摘要文本信息，并提取摘要文本信息中的关键转折词，例如，但是、却、然而、可是、只是、不过、不料、竟然、偏偏、可惜、岂知等转折词。根据包含该转折词的转折句，可以获得关于该舆情处理信息更多的信息量，故将摘要文本信息中的转折句进行前置处理，即将该转折句前置于该摘要文本信息适当的位置，以方便阅读。

[0097] 在一种可选的实施例中，还可以对摘要文本信息中每个摘要语句求取相似性，获取最高相似性对应的摘要语句并删除，得到保留摘要语句并进行标注，得到与原始舆情信息对应的标注舆情信息。

[0098] 如果摘要文本信息中包括有相似摘要语句，认为关联性高需删除后，引入下一个摘要语句。通过拟定一个权重的评分标准，可以计算相似性，采用如下TextRank公式、相似程度计算公式(一)。

[0099] $$WS(V_i)=(1-d)+d*\sum_{V_j\in In(V_i)}\frac{w_{ji}}{\sum_{V_k\in Out(V_j)}w_{jk}}WS(V_j)$$ （TextRank 公式）

[0100] 其中，TextRank公式左边表示一个摘要句子的权重(WS是weight_sum的缩写)，右侧的求和表示每个相邻摘要句子对本摘要句子的贡献程度，一般认为一篇文本信息中全部的摘要句子都是相邻的。求和的分母wji表示两个句子的相似程度，分母又是一个weight_sum，而WS(Vj)代表上次迭代j的权重，整个公式是一个迭代的过程。其中Vi表示某个网页，Vj表示链接到Vi的网页(即Vi的入链)，S(Vi)表示网页Vi的PR(即PageRank)值，In(Vi)表示网页Vi的所有入链的集合，Out(Vj)表示网页，d表示阻尼系数，一个网页被很多其他网页链接到的话说明这个网页比较重要，也就是PageRank值会相对较高。

[0101] 公式一具体如下：

[0102] $$Score(Q,d)=\sum_{i}^{n}IDF(q_i)\cdot\frac{f_i\cdot(k_1+1)}{f_i+k_1\cdot\left(1-b+b\cdot\frac{dl}{avgdl}\right)}$$ （公式一）

[0103] 其中，fi是词在文章中的出现次数，dl是文章长度，avgdl是文章平均长度，可以看出其他因素不变时，dl越大，相似程度越低。通过除以一个avgdl，以免dl取值过大。IDF(inverse document frequency)逆向文本频率表示词普遍重要性的度量。Q表示Query，qi表示Q解析之后的一个语素(对中文而言，我们可以把对Query的分词作为语素分析，每个词看成语素qi。)；d表示一个搜索结果文本。k1，b为调节因子，通常根据经验设置，一般k1＝2，b＝0.75，且b的作用是调整文本长度对相关性影响的大小。b越大，文本长度的对相关性影响越大，反之越小。而文本的相对长度越长，K值将越大，则相关性得分会越小。这可以理解为，当文本较长时，包含qi的机会越大，因此，同等fi的情况下，长文本与qi的相关性应该比短文本与qi的相关性弱。

[0104] 进而获取最高相似性对应的摘要语句并删除，得到保留摘要语句并进行标注，得到与原始舆情信息对应的标注舆情信息。

[0105] 在一种可选的实施例中，将舆情处理信息划分为第一类文本信息、第二类文本信息；对第二类文本信息进行干扰项排除处理，通过机器学习模型对第二类文本信息进行监督分类，获取正向舆情信息并标注，得到与原始舆情信息对应的标注舆情信息；其中机器学习模型采用情感词库作为训练数据集进行监督训练，以输出分类为正向舆情信息的机器学习模型。

[0106] 例如将舆情处理信息划分为第一类文本信息和第二类文本信息，其中第一类文本信息可以包括长文本数据，第二类文本信息可以包括短文本数据，在一种可选的实施例中，第一类文本信息和第二类文本信息可以通过预设字数阈值，当舆情处理信息的字数大于预设字数阈值时，则划分为第一类文本信息，；当舆情处理信息的字数不大于预设字数阈值时，则划分为第二类文本信息。还可以对第一类文本信息和第二类文本信息进行不同的处理，例如第一类文本信息主要偏向语义识别，第二类文本信息偏向情感分类。采用机器学习模型将情感词库作为训练数据集进行监督训练，以输出分类为正向舆情信息的机器学习模型。通过预设正向情感概率，当检测第二类文本信息通过机器学习模型的预测概率不小于预设正向情感概率时，输出第二类文本信息分类为正向舆情信息；当检测第二类文本信息

通过机器学习模型的预测概率小于预设正向情感概率时，输出第二类文本信息分类为负向舆情信息，其中机器学习模型可以包括朴素贝叶斯分类学习模型等。例如对干扰项进行排除处理，例如第二类文本信息微博中的博主名字“世界不美好”为负向情感，而博文内容打分结果为正向情感，则排除博主名字的干扰，最终得出结论为正向舆情信息，进而将获取正向舆情信息并标注，得到与原始舆情信息对应的标注舆情信息。

[0107] 在一种可选的实施例中，获取舆情处理信息中多个文本信息，且文本信息包括多个主题信息；文本信息、主题信息的分布参数分别服从Dirichlet分布；根据文本信息的分布参数服从Dirichlet分布，生成对应的主题信息；根据主题信息的分布参数服从Dirichlet分布，生成对应的词信息；遍历的文本信息、主题信息生成主题信息对应的词信息；得到与原始舆情信息对应的标注舆情信息。

[0108] 为从舆情处理信息中通过运营商识别获得运营商信息，例如中国移动、中国联通以及中国电信等等信息，或者通过大督查问题识别、专业线分析获得反映网络、服务、业务等信息，例如网络信息可以包括上网速度、网络稳定性、信息覆盖，业务信息可以包括计费争议、套餐设计与价格、订购办理、促销宣传与产品，服务信息可以包括服务人员态度与技能、业务办理方便快捷、信息查询与告知、问题解决与服务等等。主要通过LDA主题识别算法，对文本信息进行降维处理，生成若干个具有特征向量的主题分布，再根据对应的特征向量中相关主题的概率得到对应的词信息，以表示该文本信息的主题。

[0109] 具体参考图5，图5为本发明实施例一提供的舆情监测的方法中部分方法的示意图，例如获取舆情处理信息中的M个文本信息，且这些文本信息一共涉及有K个主体信息；且每个文本信息(例如长度为Nm)都有各自的主题分布，主题分布是多项分布，该多项分布的参数服从Dirichlet分布，该Dirichlet分布的参数为α；每个主题都有各自的词分布，词分布为多项分布，该多项分布的参数服从Dirichlet分布，该Dirichlet的参数为β；对于某个文本信息中的第n个词，首先从该文本信息的主题分布中采样一个主题，然后在这个主题对应的词分布中采样一个词。

[0110] 不断重复上述这个随机生成过程，直到m篇文章全部完成上述过程，最终得到各文本信息的主题。其中α和β为先验分布的参数，可以预先设置，α表示不同文本信息之间主题是否关联性较高，β度量有多少个近义词能够属于同一类别。例如取0.1的对称Dirichlet分布，采用θ表示文本信息对应的主题分布，即为K维的主题向量。从α控制的Dirichlet分布的概率密度函数中采取一个对应的K维主题分布即θm(例如第m个文本信息的主题)。从β控制的Dirichlet分布的概率密度函数中生成K个对应的V维的词分布，例如ψk。Zm,n代表第m个文本cinch的第n个主题，例如当n＝2时代表采到第m个文本信息的第二个主题，对应的β生成的第二个主题的词分布ψk(对应的第几个主题的词分布)。从ψk中随机挑选一个词作为Wm,n作为主题，即对应第m个文本信息第n个主题对应的词，循环执行上述步骤直到得到每个主题对应的词，并进行标注，即得到与原始舆情信息对应的标注舆情信息。在一种可选的实施例中，主题数目为K，词数目为W，则α＝50/K,β＝200/w。

[0111] 在一种可选的实施例中，将舆情处理信息中对应的文本信息以及标题信息进行分词处理，得到与文本信息、标题信息分别对应的词袋向量；

[0112] 将词袋向量作为特征计算所有文本信息的相似性，通过聚类删除相似性低于预设阈值的文本信息以及标题信息，并将保留的文本信息以及标题信息进行标注，得到与原始

舆情信息对应的标注舆情信息。

[0113] 例如将舆情处理信息中对应的文本信息以及其标题信息进行分词处理切分成单个的词，得到与文本信息、标题信息分别对应的词袋向量，即通过构造词条列表，为文本信息、或者标题信息在词条列表中赋值，即词袋向量的值可以通过统计文本信息或者标题信息中的词在词条列表中出现的次数。通过将词袋向量作为特征计算向量的余弦距离，以表达文本信息的相似性，例如余弦距离最小时即表示相似性最高。进而通过聚类删除相似性低于预设阈值的文本信息以及标题信息，并将保留的文本信息以及标题信息进行标注，得到与原始舆情信息对应的标注舆情信息。其中聚类算法可以包括二分K-均值聚类算法。

[0114] 上述示例的多个可选的实施例，可以设置先后次序执行，也可以不设置先后执行次序，以到达更好的实施效果，本发明中不作具体限定。

[0115] 图6为本发明实施例二提供的舆情监测的方法的流程图，本实施例可以在图2基础上增加步骤S204，如图6所示，本实施例中舆情监测的方法可以包括：

[0116] S201、获取多个原始舆情信息；

[0117] S202、对原始舆情信息进行去标签、清洗处理，得到舆情处理信息；

[0118] S203、将舆情处理信息进行缓存，通过文本分析进行识别标注，得到与原始舆情信息对应的标注舆情信息

[0119] 本实施例中，步骤S201～S203具体实现过程和技术原理请参见图2所示的方法中步骤S201～步骤S203中的相关描述，此处不再赘述。

[0120] S204、将标注舆情信息与原始舆情信息存储于全文搜索引擎中，以便在交互界面上搜索显示标注舆情信息对应的原始舆情信息。

[0121] 本实施例中将上述实施例得到的标注舆情信息与原始舆情信息存储于全文搜索引擎中，可以基于全文搜索引擎与面向业务的各应用模块以提供应用服务，例如包括24小时最新舆情、热点资讯、自助数据分析、全量信息、专业线分析等应用服务。可以在交互界面上搜索显示标注舆情信息对应的原始舆情信息。

[0122] 其中24小时最新舆情应用服务可以向用户展示全网、移动、电信、联通相关的最新舆情信息，展示的维度包括24小时最新非敏感舆情趋势图、24小时最新敏感舆情趋势图、24小时最新非敏感舆情信息top10、24小时最新敏感舆情信息top10、前一日行业热点信息。

[0123] 热点资讯应用服务可以向用户展示近一日、近三日、近七日、近三十日的行业热点舆情信息，包括运营商行业热点、中国移动热点、中国电信热点、中国联通热点信息，便于业务人员能够快速掌握行业最新动态与热点情况

[0124] 自助数据分析可以为用户提供制定自定义监测方案的功能，用户自定义关键词，系统根据关键词提取匹配的舆情信息，提取结果包括舆情信息的展示、全网舆情分析报告、微博舆情分析报告。

[0125] 全量信息应用服务可以向用户展示库中全量的所有舆情信息。

[0126] 专业线分布分析应用服务可以向用户展示专业线整体的声量、敏感声量、负评率，并展示中国移动、中国电信、中国联通的一级专业线(网络、业务、服务)的近七天的声量发展趋势。

[0127] 本发明舆情监测的方法减少了人工成本，提高舆情监测的准确率、有效性，极大的提高了舆情监测的效率。

[0128] 图7为本发明实施例三提供的舆情监测的装置的结构示意图，如图7所示，本实施例的舆情监测的装置可以包括：

[0129] 获取模块31，用于获取多个原始舆情信息；

[0130] 得到模块32，用于对原始舆情信息进行去标签、清洗处理，得到舆情处理信息；

[0131] 标注模块33，用于将舆情处理信息进行缓存，通过文本分析进行识别标注，得到与原始舆情信息对应的标注舆情信息。

[0132] 在一种可能的设计中，通过文本分析进行识别标注，得到与原始舆情信息对应的标注舆情信息，包括：

[0133] 对舆情处理信息进行分词，将分析后的舆情处理信息在地域词典中进行匹配，若匹配成功则对舆情信息进行地域标识，得到地域舆情处理信息；

[0134] 根据地域舆情处理信息出现的位置以及频次，获得地域舆情处理信息对应的评分；

[0135] 根据评分的大小依次进行排序，并将最高评分对应的地域舆情处理信息进行地域标注，得到与原始舆情信息对应的标注舆情信息。

[0136] 在一种可能的设计中，地域词典通过获取地域词汇，并将地域词汇整理构建获得。

[0137] 在一种可能的设计中，通过文本分析进行识别标注，得到与原始舆情信息对应的标注舆情信息，还包括：

[0138] 获取舆情处理信息中的摘要文本信息，提取、标注摘要文本信息中的转折句，得到与原始舆情信息对应的标注舆情信息。

[0139] 在一种可能的设计中，装置，还包括：

[0140] 对摘要文本信息中每个摘要语句求取相似性；

[0141] 获取最高相似性对应的摘要语句并删除，得到保留摘要语句并进行标注，得到与原始舆情信息对应的标注舆情信息。

[0142] 在一种可能的设计中，通过文本分析进行识别标注，得到与原始舆情信息对应的标注舆情信息，包括：

[0143] 将舆情处理信息划分为第一类文本信息、第二类文本信息；

[0144] 对第二类文本信息进行干扰项排除处理，通过机器学习模型对第二类文本信息进行监督分类，获取正向舆情信息并标注，得到与原始舆情信息对应的标注舆情信息；其中机器学习模型采用情感词库作为训练数据集进行监督训练，以输出分类为正向舆情信息的机器学习模型。

[0145] 在一种可能的设计中，通过文本分析进行识别标注，得到与原始舆情信息对应的标注舆情信息，包括：

[0146] 获取舆情处理信息中多个文本信息，且文本信息包括多个主题信息；

[0147] 文本信息、主题信息的分布参数分别服从Dirichlet分布；

[0148] 根据文本信息的分布参数服从Dirichlet分布，生成对应的主题信息；

[0149] 根据主题信息的分布参数服从Dirichlet分布，生成对应的词信息；

[0150] 遍历的文本信息、主题信息生成主题信息对应的词信息；得到与原始舆情信息对应的标注舆情信息。

[0151] 在一种可能的设计中，通过文本分析进行识别标注，得到与原始舆情信息对应的

标注舆情信息，包括：

[0152] 将舆情处理信息中对应的文本信息以及标题信息进行分词处理，得到与文本信息、标题信息分别对应的词袋向量；

[0153] 将词袋向量作为特征计算所有文本信息的相似性，通过聚类删除相似性低于预设阈值的文本信息以及标题信息，并将保留的文本信息以及标题信息进行标注，得到与原始舆情信息对应的标注舆情信息。

[0154] 在一种可能的设计中，获取多个原始舆情信息，包括：

[0155] 按照预设规则，从多个网络资源中获取原始舆情信息；

[0156] 或者通过API接口获取原始舆情信息。

[0157] 在一种可能的设计中，在得到与原始舆情信息对应的标注舆情信息之后，还包括：

[0158] 将标注舆情信息与原始舆情信息存储于全文搜索引擎中，以便在交互界面上搜索显示标注舆情信息对应的原始舆情信息。

[0159] 本实施例的舆情监测的装置，可以执行图2、图6所示方法中的技术方案，其具体实现过程和技术原理参见图2、图6所示方法中的相关描述，此处不再赘述。

[0160] 图8为本发明实施例四提供的舆情监测的系统的结构示意图，如图8所示，本实施例的舆情监测的系统40可以包括：处理器41和存储器42。

[0161] 存储器42，用于存储计算机程序(如实现上述舆情监测的方法的应用程序、功能模块等)、计算机指令等；

[0162] 上述的计算机程序、计算机指令等可以分区存储在一个或多个存储器42中。并且上述的计算机程序、计算机指令、数据等可以被处理器41调用。

[0163] 处理器41，用于执行存储器42存储的计算机程序，以实现上述实施例涉及的方法中的各个步骤。

[0164] 具体可以参见前面方法实施例中的相关描述。

[0165] 处理器41和存储器42可以是独立结构，也可以是集成在一起的集成结构。当处理器41和存储器42是独立结构时，存储器42、处理器41可以通过总线43耦合连接。

[0166] 本实施例的服务器可以执行图2、图6所示方法中的技术方案，其具体实现过程和技术原理参见图2、图6所示方法中的相关描述，此处不再赘述。

[0167] 此外，本申请实施例还提供一种计算机可读存储介质，计算机可读存储介质中存储有计算机执行指令，当用户设备的至少一个处理器执行该计算机执行指令时，用户设备执行上述各种可能的方法。

[0168] 其中，计算机可读介质包括计算机存储介质和通信介质，其中通信介质包括便于从一个地方向另一个地方传送计算机程序的任何介质。存储介质可以是通用或专用计算机能够存取的任何可用介质。一种示例性的存储介质耦合至处理器，从而使处理器能够从该存储介质读取信息，且可向该存储介质写入信息。当然，存储介质也可以是处理器的组成部分。处理器和存储介质可以位于ASIC中。另外，该ASIC可以位于用户设备中。当然，处理器和存储介质也可以作为分立组件存在于通信设备中。

[0169] 本领域普通技术人员可以理解：实现上述各方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成。前述的程序可以存储于一计算机可读取存储介质中。该程序在执行时，执行包括上述各方法实施例的步骤；而前述的存储介质包括：ROM、RAM、磁碟或

者光盘等各种可以存储程序代码的介质。

[0170] 最后应说明的是:以上各实施例仅用以说明本发明的技术方案,而非对其限制;尽管参照前述各实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分或者全部技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明各实施例技术方案的范围。

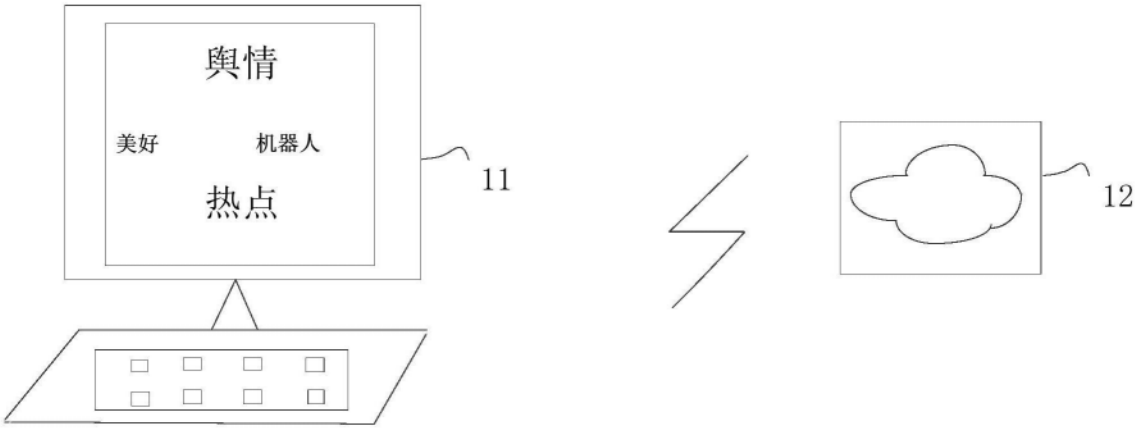

舆情
美好
机器人
热点
11
12

图1

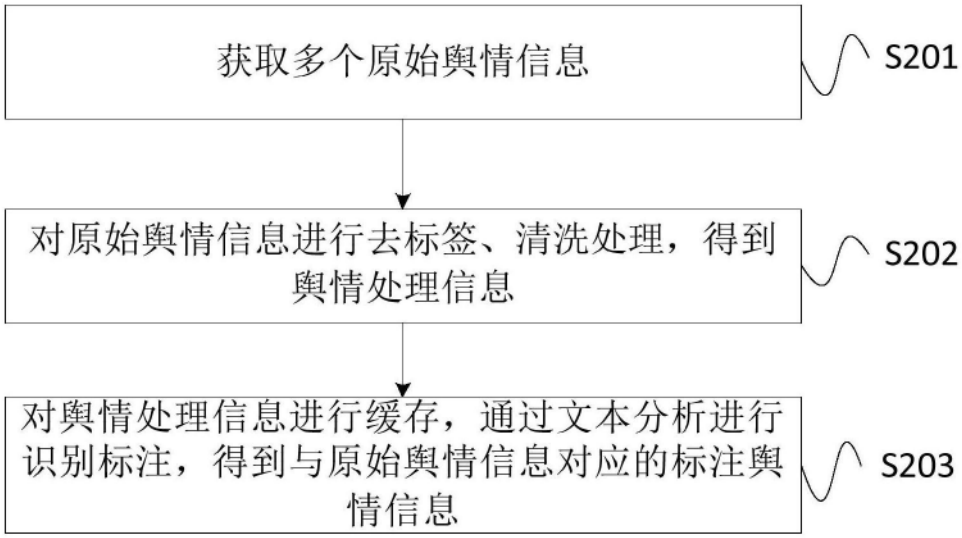

获取多个原始舆情信息
S201
对原始舆情信息进行去标签、清洗处理，得到舆情处理信息
S202
对舆情处理信息进行缓存，通过文本分析进行识别标注，得到与原始舆情信息对应的标注舆情信息
S203

图2

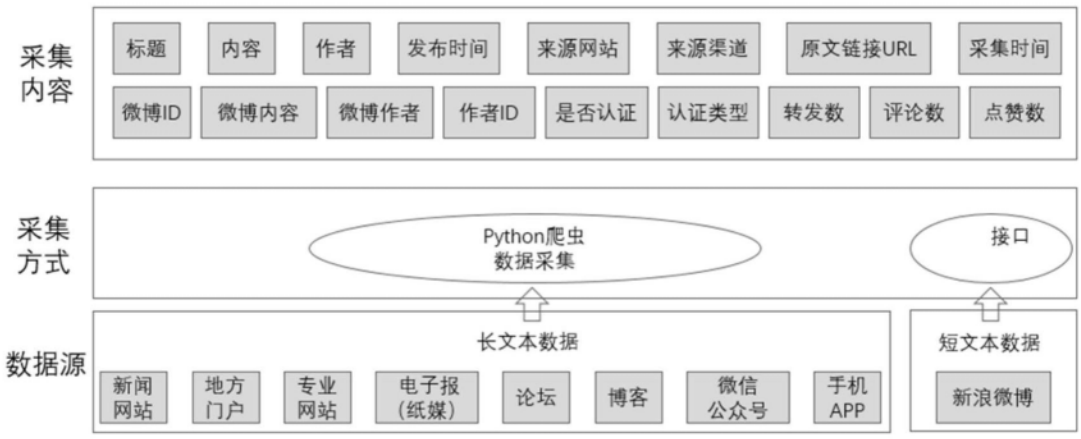

采集内容
标题
内容
作者
发布时间
来源网站
来源渠道
原文链接URL
采集时间
微博ID
微博内容
微博作者
作者ID
是否认证
认证类型
转发数
评论数
点赞数
采集方式
Python爬虫数据采集
接口
数据源
长文本数据
新闻网站
地方门户
专业网站
电子报(纸媒)
论坛
博客
微信公众号
手机APP
短文本数据
新浪微博

图3

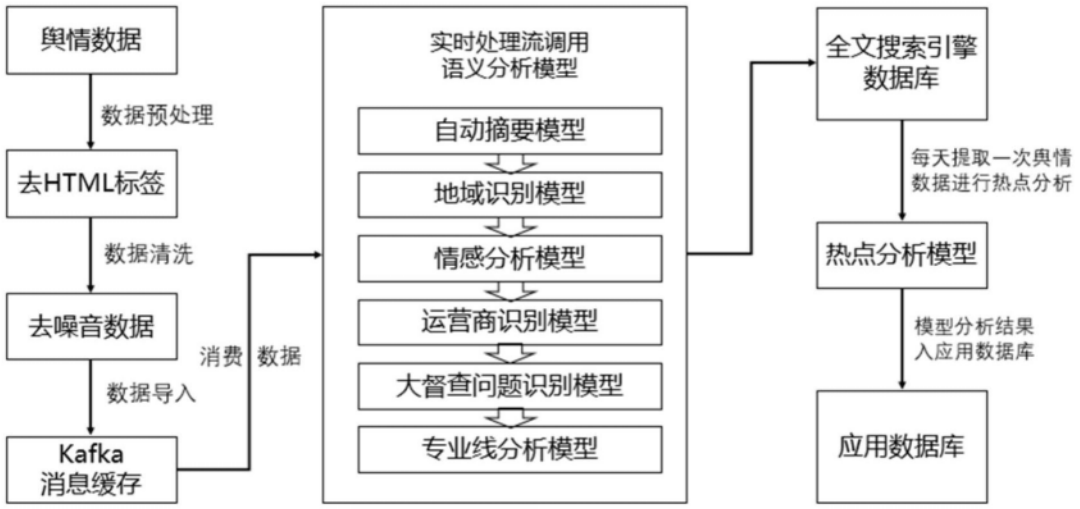

舆情数据
数据预处理
去HTML标签
数据清洗
去噪音数据
数据导入
Kafka
消息缓存
消费
数据
实时处理流调用
语义分析模型
自动摘要模型
地域识别模型
情感分析模型
运营商识别模型
大督查问题识别模型
专业线分析模型
全文搜索引擎
数据库
每天提取一次舆情
数据进行热点分析
热点分析模型
模型分析结果
入应用数据库
应用数据库

图4

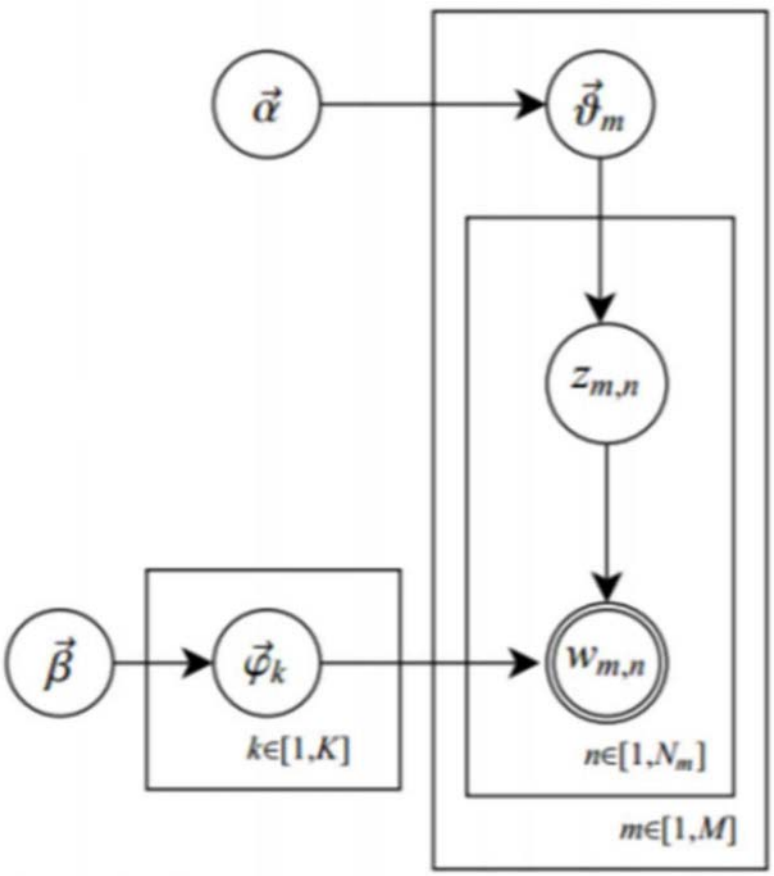

$\vec{\alpha}$
$\vec{\vartheta}_m$
$z_{m,n}$
$\vec{\beta}$
$\vec{\varphi}_k$
$w_{m,n}$
$k\in[1,K]$
$n\in[1,N_m]$
$m\in[1,M]$

图5

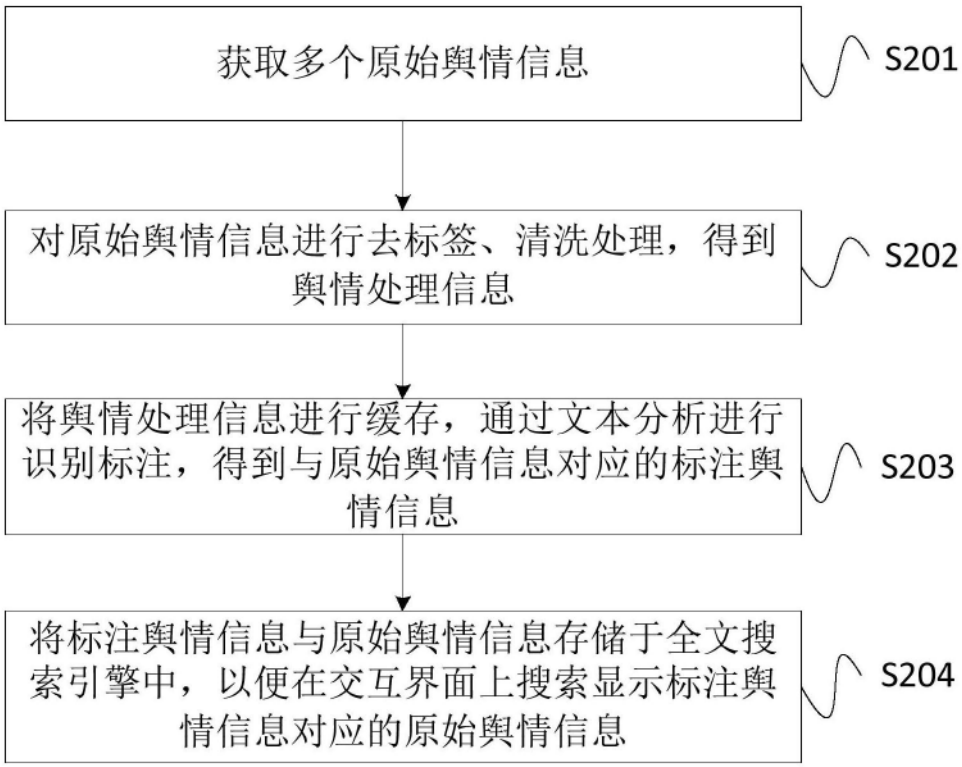

获取多个原始舆情信息
S201
对原始舆情信息进行去标签、清洗处理，得到舆情处理信息
S202
将舆情处理信息进行缓存，通过文本分析进行识别标注，得到与原始舆情信息对应的标注舆情信息
S203
将标注舆情信息与原始舆情信息存储于全文搜索引擎中，以便在交互界面上搜索显示标注舆情信息对应的原始舆情信息
S204

图6

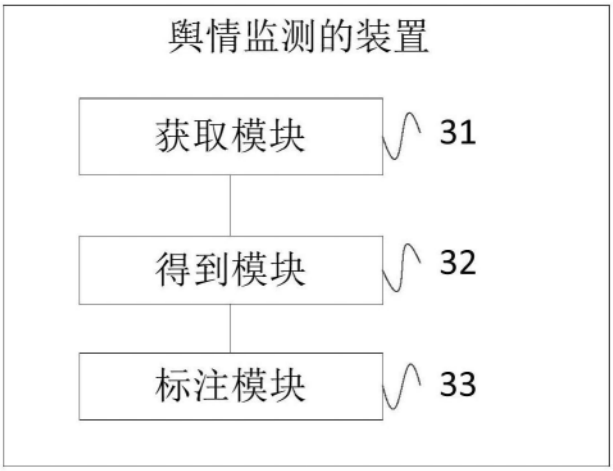

舆情监测的装置
获取模块
31
得到模块
32
标注模块
33

图7

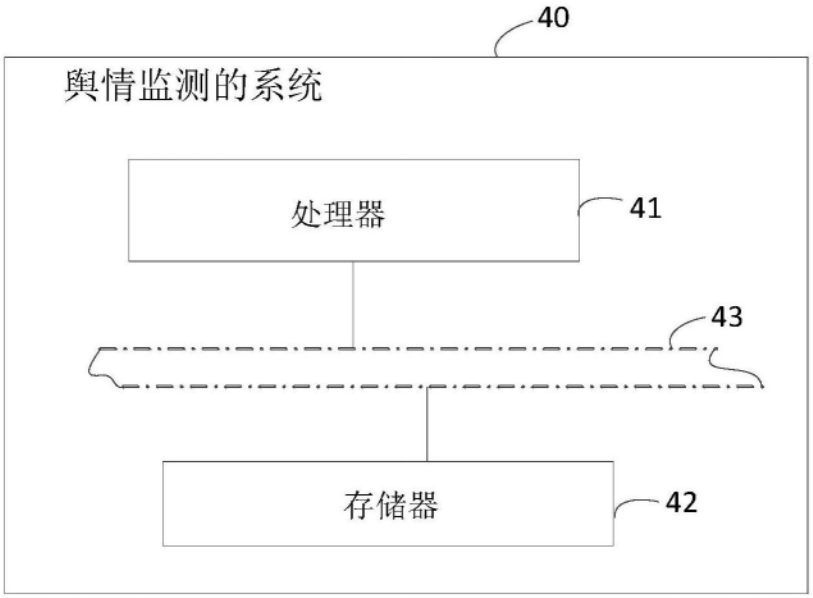


图8