(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117493704 A
(43)申请公布日 2024.02.02

(21)申请号 202311675070.2

(22)申请日 2023.12.07

(71)申请人 中国农业银行股份有限公司
地址 100005 北京市东城区建国门内大街69号

(72)发明人 栾吉海 魏依鹤 李娟 宋志刚

(74)专利代理机构 北京品源专利代理有限公司 11332
专利代理师 陈金忠

(51)Int.Cl.
G06F 16/9536(2019.01)
G06F 16/951(2019.01)
G06F 16/955(2019.01)
G06F 16/958(2019.01)
G06F 40/169(2020.01)
G06F 40/289(2020.01)
G06F 18/23213(2023.01)

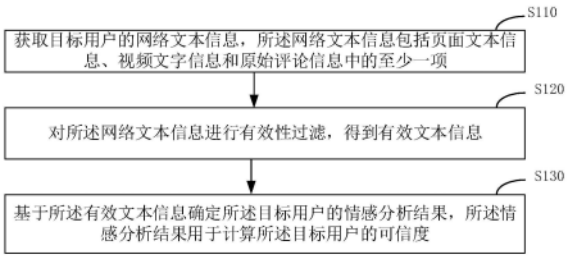
权利要求书2页 说明书12页 附图6页

(54)发明名称

一种用户可信度的计算方法、装置、电子设备及介质

(57)摘要

本发明公开了一种用户可信度的计算方法、装置、电子设备及介质,所述方法包括:获取目标用户的网络文本信息,所述网络文本信息包括页面文本信息、视频文字信息和原始评论信息中的至少一项;对所述网络文本信息进行有效性过滤,得到有效文本信息;基于所述有效文本信息确定所述目标用户的情感分析结果,所述情感分析结果用于计算所述目标用户的可信度。该方法通过对网络文本信息进行有效性过滤,为后续计算用户可信度提供了信息基础,进而提高了用户可信度的准确率。



CN 117493704 A

1.一种用户可信度的计算方法,其特征在于,所述方法包括:

获取目标用户的网络文本信息,所述网络文本信息包括页面文本信息、视频文字信息和原始评论信息中的至少一项;

对所述网络文本信息进行有效性过滤,得到有效文本信息;

基于所述有效文本信息确定所述目标用户的情感分析结果,所述情感分析结果用于计算所述目标用户的可信度。

2.根据权利要求1所述的方法,其特征在于,所述获取目标用户的网络文本信息,包括:

采用基于搜索引擎的主题爬虫器下载目标网页;

基于文本密度指标与符号密度指标提取所述目标网页的正文内容;

采用改进的k-means聚类算法对所述正文内容中的文本进行聚类,得到目标簇数据;

对所述目标簇数据进行主题判别,得到相似度满足预设相似阈值的页面文本信息。

3.根据权利要求2所述的方法,其特征在于,所述获取目标用户的网络文本信息,还包括:

采用目标网站爬虫器对所述目标用户的目标网站进行聚焦网络爬虫和增量爬虫,得到所述目标网站的第一评论信息,其中,所述主题爬虫器和所述目标网站爬虫器采用改进的Best-First搜索策略。

4.根据权利要求1所述的方法,其特征在于,所述获取目标用户的网络文本信息,包括:

采用视频爬虫器获取目标用户的视频信息;

提取所述视频信息的视频文字信息和第二评论信息。

5.根据权利要求1所述的方法,其特征在于,所述网络文本信息包括页面文本信息、视频文字信息和原始评论信息,所述对所述网络文本信息进行有效性过滤,得到有效文本信息,包括:

提取第一网络信息的目标关键词,所述第一网络信息包括页面文本信息、视频文字信息和第三评论信息,所述第三评论信息为原始评论信息中支持率低于预设阈值的评论信息;

将所述目标关键词和第四评论信息确定为所述网络文本信息的有效文本信息,所述第四评论信息为原始评论信息中支持率高于预设阈值的评论信息。

6.根据权利要求5所述的方法,其特征在于,所述提取第一网络信息的目标关键词,包括:

对第一网络信息进行分句处理,得到至少一个目标语句;

基于预设网络模型生成所述第一网络信息的目标词义结构,所述预设网络模型由至少一个目标语句训练得到;

对所述目标词义结构进行关键词抽取排序,得到第一网络信息的目标关键词。

7.根据权利要求1所述的方法,其特征在于,所述基于所述有效文本信息确定所述目标用户的情感分析结果,包括:

基于第一有效信息对长短期记忆模型进行模型训练,得到所述目标用户对应的情感分析模型,所述第一有效信息为对所述有效文本信息中训练文本信息进行预处理后的信息;

将第二有效信息输入至所述情感分析模型中,得到所述目标用户的情感分析结果,所述第二有效信息为对所述有效文本信息中测试文本信息进行预处理后的信息。

8.一种用户可信度的计算装置,其特征在于,包括:

获取模块,用于获取目标用户的网络文本信息,所述网络文本信息包括页面文本信息、视频文字信息和原始评论信息中的至少一项;

过滤模块,用于对所述网络文本信息进行有效性过滤,得到有效文本信息;

确定模块,用于基于所述有效文本信息确定所述目标用户的情感分析结果,所述情感分析结果用于计算所述目标用户的可信度。

9.一种电子设备,其特征在于,所述电子设备包括:

至少一个处理器;以及

与所述至少一个处理器通信连接的存储器;其中,

所述存储器存储有可被所述至少一个处理器执行的计算机程序,所述计算机程序被所述至少一个处理器执行,以使所述至少一个处理器能够执行权利要求1-7中任一项所述的用户可信度的计算方法。

10.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质存储有计算机指令,所述计算机指令用于使处理器执行时实现权利要求1-7中任一项所述的用户可信度的计算方法。

一种用户可信度的计算方法、装置、电子设备及介质

技术领域

[0001] 本发明涉及网络技术领域,尤其涉及一种用户可信度的计算方法、装置、电子设备及介质。

背景技术

[0002] 互联网产业迅速发展导致互联网信息爆炸增长,为金融行业评估用户可信度以进行风险管理提供了便利。

[0003] 现有用户可信度的计算方法中的信息来源较为复杂,对进行情感分析存在一定的干扰,从而降低了用户可信度的准确率。

发明内容

[0004] 本发明提供了一种用户可信度的计算方法、装置、电子设备及介质,以提高用户可信度的准确率。

[0005] 根据本发明的一方面,提供了一种用户可信度的计算方法,所述方法包括:

[0006] 获取目标用户的网络文本信息,所述网络文本信息包括页面文本信息、视频文字信息和原始评论信息中的至少一项;

[0007] 对所述网络文本信息进行有效性过滤,得到有效文本信息;

[0008] 基于所述有效文本信息确定所述目标用户的情感分析结果,所述情感分析结果用于计算所述目标用户的可信度。

[0009] 根据本发明的另一方面,提供了一种用户可信度的计算装置,包括:

[0010] 获取模块,用于获取目标用户的网络文本信息,所述网络文本信息包括页面文本信息、视频文字信息和原始评论信息中的至少一项;

[0011] 过滤模块,用于对所述网络文本信息进行有效性过滤,得到有效文本信息;

[0012] 确定模块,用于基于所述有效文本信息确定所述目标用户的情感分析结果,所述情感分析结果用于计算所述目标用户的可信度。

[0013] 根据本发明的另一方面,提供了一种电子设备,所述电子设备包括:

[0014] 至少一个处理器;以及

[0015] 与所述至少一个处理器通信连接的存储器;其中,

[0016] 所述存储器存储有可被所述至少一个处理器执行的计算机程序,所述计算机程序被所述至少一个处理器执行,以使所述至少一个处理器能够执行本发明任一实施例所述的用户可信度的计算方法。

[0017] 根据本发明的另一方面,提供了一种计算机可读存储介质,所述计算机可读存储介质存储有计算机指令,所述计算机指令用于使处理器执行时实现本发明任一实施例所述的用户可信度的计算方法。

[0018] 本发明实施例提供了一种用户可信度的计算方法、装置、电子设备及介质,所述方法包括:获取目标用户的网络文本信息,所述网络文本信息包括页面文本信息、视频文字信

息和原始评论信息中的至少一项;对所述网络文本信息进行有效性过滤,得到有效文本信息;基于所述有效文本信息确定所述目标用户的情感分析结果,所述情感分析结果用于计算所述目标用户的可信度。利用上述技术方案,通过对网络文本信息进行有效性过滤,为后续计算用户可信度提供了信息基础,进而提高了用户可信度的准确率。

[0019] 应当理解,本部分所描述的内容并非旨在标识本发明的实施例的关键或重要特征,也不用于限制本发明的范围。本发明的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0020] 为了更清楚地说明本发明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的前提下,还可以根据这些附图获得其他的附图。

[0021] 图1是根据本发明实施例一提供的一种用户可信度的计算方法的流程图;

[0022] 图2是根据本发明实施例一提供的一种获取网络文本信息的流程图;

[0023] 图3是根据本发明实施例一提供的一种聚焦网络爬虫过程的流程图;

[0024] 图4是根据本发明实施例一提供的一种正文提取的流程图;

[0025] 图5是根据本发明实施例一提供的一种文本聚类的流程图;

[0026] 图6是根据本发明实施例一提供的一种得到有效文本信息的流程图;

[0027] 图7是根据本发明实施例二提供的一种用户可信度的计算方法的流程图;

[0028] 图8是根据本发明实施例二提供的一种得到有效文本信息的流程图;

[0029] 图9是根据本发明实施例二提供的一种用户可信度的计算方法的整体架构图;

[0030] 图10是根据本发明实施例二提供的一种网页信息提取的流程图;

[0031] 图11是根据本发明实施例三提供的一种用户可信度的计算装置的结构示意图;

[0032] 图12是根据本发明实施例四提供的一种电子设备的结构示意图。

具体实施方式

[0033] 为了使本技术领域的人员更好地理解本发明方案,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,所描述的实施例仅仅是本发明一部分的实施例,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,都应当属于本发明保护的范围。

[0034] 需要说明的是,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”等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。

[0035] 实施例一

[0036] 图1是根据本发明实施例一提供的一种用户可信度的计算方法的流程图,本实施例可适用于对用户可信度进行计算的情况,该方法可以由用户可信度的计算装置来执行,该用户可信度的计算装置可以采用硬件和/或软件的形式实现,该用户可信度的计算装置可配置于电子设备中。如图1所示,该方法包括:

[0037] S110、获取目标用户的网络文本信息,所述网络文本信息包括页面文本信息、视频文字信息和原始评论信息中的至少一项。

[0038] 网络文本信息可以认为是在公开网络上与目标用户相关的文本信息,网络文本信息的来源不限,如网络文本信息可以包括页面文本信息、视频文字信息和原始评论信息中的至少一项,进一步地,本实施例可以通过不同的获取方式来获取目标用户不同的网络文本信息。

[0039] 在一个实施例中,所述获取目标用户的网络文本信息,包括:

[0040] 采用基于搜索引擎的主题爬虫器下载目标网页;

[0041] 基于文本密度指标与符号密度指标提取所述目标网页的正文内容;

[0042] 采用改进的k-means聚类算法对所述正文内容中的文本进行聚类,得到目标簇数据;

[0043] 对所述目标簇数据进行主题判别,得到相似度满足预设相似阈值的页面文本信息。

[0044] 在一个实施方式中,可以通过采用基于搜索引擎的主题爬虫器来获取得到目标用户的网络文本信息。

[0045] 图2是根据本发明实施例一提供的一种获取网络文本信息的流程图,如图2所示,可以通过选定的搜索引擎作为数据获取源,输入目标用户的关键词进行检索,然后可以获取检索结果的前几页,提取其链接目标到网页进行页面下载,随后可以采用特定的算法提取下载网页中的正文,再通过聚类算法对所提取的正文进行聚类,将用户所需要的文本进行存储。进一步地,针对特定主题的关键词,可以进行面向全网的主题爬虫,例如可以首先获取种子URL目标网页进行页面下载,然后同样采用特定的算法提取下载网页中的正文,然后对正文进行主题判别,将相似度高的文本存储,随着网页的不断更新,已有的主题词库已无法准确获取全部相关文本,因此需要进行主题词库扩充,除此之外,还需要特定的搜索策略使爬虫有次序、有目的地搜索,直至最后将获取的文本返回给电子设备。

[0046] 图3是根据本发明实施例一提供的一种聚焦网络爬虫过程的流程图,如图3所示,本实施例爬取规则可以选择聚焦网络爬虫,首先可以将目标用户关键字输入搜索引擎来获取URL链接,爬取该页面并获取新的URL,可以从新的链接中过滤与目标用户公司无关的网页,并将已经爬取的URL存到列表中进行去重与过滤,之后可以将过滤后的链接放到URL队列中,确定爬取顺序流程进行下一个URL的读取,直到满足停止条件后则可完成爬取。

[0047] 在爬取得到目标网页后需要进行正文提取,本实施例可以通过正文提取技术只保留网页中正文部分的内容,剔除掉网页中的噪音部分内容,否则噪音内容会影响下面一步的文本聚类以及对后续其他项目使用数据时造成影响。图4是根据本发明实施例一提供的一种正文提取的流程图,如图4所示,首先可以解析对象结点,如获取目标网页中的body主体,遍历body中的每一个结点,将结点进行hash处理。

[0048] 其中,对于结点i,首先可以获取结点i下的所有文本组成一个字符串texti,字符

串的长度则为结点i的字符串字数Ti,如果结点包含<class>标签,且<class>标签中包含['content','article','news_txt','post_text']这些属性关键词,则可以给结点i的字符串的长度增加权重(如乘以2);可以通过获取结点下的所有<a>标签下的所有文本得到带链接的字符串字数LTi;获取结点i所有标签数则为其标签数TGi;结点i下的<a>标签数可以为其带链接的标签数LTGi,从而可以计算出结点i的文本密度。

[0049] 然后可以计算结点的符号密度,常见的中英文符号可以有(”'!,。?、;:“”“《》%(),.?:;'″!%()”'),可以遍历字符串texti,统计出字符串里共有多少个标点符号Sbi,从而计算出结点i的符号密度SbDi。其次可以计算结点的综合分数TSbDi,并将结点根据分数进行排序,返回结果集合,最后可以将分数最高的结点的文本内容作为正文提取出来。

[0050] 获取页面文本信息的中间一步是通过聚类排除由于关键词歧义而不需要的类别数据。具体的,本实施例可以采用改进的k-means聚类算法进行聚类。图5是根据本发明实施例一提供的一种文本聚类的流程图,如图5所示,首先可以对原始文本(即正文内容)进行预处理,如可以通过jieba分词进行切词以及去除停用词、标点和数字,再提取预处理后文本中的所有词语形成文本向量。

[0051] 然后进行文本表示,本实施例的文本表示模型可以采用VSM(Vector Space Model)向量空间模型,并通过TF-IDF模型计算文档的特征,TF指的词频,即TF(i,j)指i词在第j篇文档中出现的频率,IDF指的逆文档频率,即包含词i的文档数量Ni越少,IDF(i)越大。

[0052] 随后进行文本聚类,即可以选取作为初始中心的k个点{c1,c2,...,ck};计算每个文本di与各个中心cj的相似度,将文本di分入到最大相似度中心所在的簇中去,然后可以得到聚类{C1,C2,...,Ck}并将计算的相似度记录下来;计算各个簇Ci内部的平均相似度,计作meansSim;针对每个簇,可以选择簇内部相似度大于(1+μ)meansSim的文本集合{d1,d2,...,dm};可以计算{d1,d2,...,dm}的均值点作为所在簇的新的中心;重复b-e步直到簇中心不再发生变化。从而可以选取合适的那一簇数据存入数据库。

[0053] 获取页面文本信息的最后一步是将输出的正文进行主题判别,判断页面是否与主题相关,如可以使用基于机器学习与向量空间模型相结合的方式进行主题判别,通过向量空间训练数据,机器学习准确的筛选数据。另外,使用TF-IDF(词语词频-逆文档频率)公式计算特征权值,两文档之间相似度可以用其对应的向量之间的夹角余弦表示,其中N为所有文档的数目、N_i为含有词条t_i的文档数目,将相似度大于某一阈值的文档作为正例集,小于某一阈值的为反例集,使用LSTM模型进行训练,训练之后将后续爬取的文本通过此模型进行筛选。

[0054] 在一个实施例中,所述获取目标用户的网络文本信息,还包括:

[0055] 采用目标网站爬虫器对所述目标用户的目标网站进行聚焦网络爬虫和增量爬虫,得到所述目标网站的第一评论信息,其中,所述主题爬虫器和所述目标网站爬虫器采用改进的Best-First搜索策略。

[0056] 在一个实施方式中,可以采用目标网站爬虫器对目标用户的目标网站进行爬虫,以此得到目标网站的第一评论信息,示例性的,针对特殊网页,本实施例可以采用增量爬虫的方式,对该类型网站进行监控,检测其数据更新情况,增量爬虫的核心主要是进行去重。在访问起始URL发送请求之前,可以判断该URL是否被爬取过;在解析内容时判断这部分内容是否被解析过;写入数据库的时候判断是否已存入数据库中,如可以遍历列表是否存在

Redis中,若不存在,则可以向符合条件的URL发送请求,解析页面对内容生成唯一标识,如可通过MD5、sha等摘要算法生成数据指纹作为唯一标识;遍历生成的唯一标识是否存在Redis中,若不存在,则可以存入数据库中,即得到目标网站的第一评论信息。

[0057] 进一步地,本实施例所使用的主题爬虫器和目标网站爬虫器可以采用改进的Best-First搜索策略,即采用的链接爬取策略可以为一种基于链接内容评价的主题搜索策略,链接的价值由三部分组成,第一部分是链接本身的地址内容,从当前链接地址与父页面链接地址进行结构上的比对,获得链接的站内地址价值,该价值能一定程度上评估当前连接的加制。第二部分是链接继承至父页面的价值,如果链接所在父页面主题相关度高,那么可以认为该链接与主题相关的可能性也较大。最后一部分则是与链接相关的锚文本内容,因为一般锚文本内容是对页面内容的高度概括。相对现有技术只考虑第二部分链接父页面的价值,本实施例采用改进的Best-First搜索策略,能够综合考虑三部分的价值,提高网络文本信息的准确性。

[0058] 在一个实施例中,所述获取目标用户的网络文本信息,包括:

[0059] 采用视频爬虫器获取目标用户的视频信息;

[0060] 提取所述视频信息的视频文字信息和第二评论信息。

[0061] 其中,视频文字信息可以是指从视频信息中所提取的文本,第二评论信息则为视频信息中的评论区信息。

[0062] 在一个实施方式中,可以采用视频爬虫器抓取移动端短视频数据,并对抓取的视频提取其中的文本与评论区信息,如可以下载模拟器对短视频抓包,利用程序实现翻页、滑动取下一批评论的功能,并进行文字的提取,如可以通过将视频截成一张一张的图片,再将每张图片中的文字提取出来得到视频文字信息。

[0063] S120、对所述网络文本信息进行有效性过滤,得到有效文本信息。

[0064] 由于获取到的网络文本信息存在包含不实信息的情况,本实施例可以对网络文本信息进行有效性过滤,以过滤掉不实、含引导性等信息。具体过滤的手段不限,只要能得到有效文本信息即可。

[0065] 图6是根据本发明实施例一提供的一种得到有效文本信息的流程图,如图6所示,可以通过将原始数据(即网络文本信息)输入到信息处理单元来获取得到有效评论数据,具体的,可以对网络文本信息中的页面信息进行分析,并进行分句处理;将所有渠道获取的原始评论信息、分句信息进行汇总并进行数据清洗;对数据清洗后的信息进行可信度分析,如可以包括时效性分析、有用性投票处理、评论内容挖掘等操作,最终获取有效的评论数据。

[0066] 其中,评论时效性采集可以是采用当前评论与最早评论发布时间的差值作为衡量可信度的时效性指标,间隔的天数越多,说明评论距当前阅读时间越近,时效性越强。评论内容挖掘可以采用汉语分词系统,通过中文分词、词性标注、关键词统计等功能对评论进行信息挖掘,包括特征词、情感词的统计。并通过特征词、情感词、评论总长度进行计算判断其评论有效长度。此处不对可信度分析的具体过程作进一步赘述。

[0067] S130、基于所述有效文本信息确定所述目标用户的情感分析结果,所述情感分析结果用于计算所述目标用户的可信度。

[0068] 通过上述步骤得到有效文本信息之后,则可以对目标用户的情感分析结果进行确定,如可以将有效文本信息输入至某预设分析模型中来直接输出得到目标用户的情感分析

结果,也可以通过训练得到目标用户对应的情感分析模型,再基于有效文本信息并结合情感分析模型来得到目标用户的情感分析结果,本实施例对比不作限定。

[0069] 在一个实施例中,所述基于所述有效文本信息确定所述目标用户的情感分析结果,包括:

[0070] 基于第一有效信息对长短期记忆模型进行模型训练,得到所述目标用户对应的情感分析模型,所述第一有效信息为对所述有效文本信息中训练文本信息进行预处理后的信息;

[0071] 将第二有效信息输入至所述情感分析模型中,得到所述目标用户的情感分析结果,所述第二有效信息为对所述有效文本信息中测试文本信息进行预处理后的信息。

[0072] 在本实施例中,可以将有效文本信息切分为训练文本信息和测试文本信息两部分,并分别进行预处理操作得到第一有效信息和第二有效信息,预处理操作例如可以包括文本预处理和向量化表示等,示例性的,首先可以进行标点符号的修正、异常数据的去除、重复数据的删除;再进行中文分词、停用词处理,停用词主要是去除无意义的助词、介词,本实施例可以采用哈工大停用词词表;并进行通用情感词典的构建。向量化表示例如可以使用Word2vec进行文本向量化处理。

[0073] 随后,可以采用基于LSTM(即长短期记忆模型)的方法对模型进行训练,如基于第一有效信息对长短期记忆模型进行模型训练,得到目标用户对应的情感分析模型,然后将第二有效信息经过训练好的模型中得到目标用户的情感分析结果,从而可将情感分析结果用于计算目标用户的可信度,实现大众角度下展示目标用户的消极、积极的想法,依据评分进行情感分类,并展示词云图。

[0074] 本发明实施例一提供的一种用户可信度的计算方法,获取目标用户的网络文本信息,所述网络文本信息包括页面文本信息、视频文字信息和原始评论信息中的至少一项;对所述网络文本信息进行有效性过滤,得到有效文本信息;基于所述有效文本信息确定所述目标用户的情感分析结果,所述情感分析结果用于计算所述目标用户的可信度。利用该方法,通过对网络文本信息进行有效性过滤,为后续计算用户可信度提供了信息基础,进而提高了用户可信度的准确率。

[0075] 实施例二

[0076] 图7是根据本发明实施例二提供的一种用户可信度的计算方法的流程图,本实施例二在上述各实施例的基础上进行优化。在本实施例中,所述网络文本信息包括页面文本信息、视频文字信息和原始评论信息,将对所述网络文本信息进行有效性过滤,得到有效文本信息进一步具体化为:提取第一网络信息的目标关键词,所述第一网络信息包括页面文本信息、视频文字信息和第三评论信息,所述第三评论信息为原始评论信息中支持率低于预设阈值的评论信息;将所述目标关键词和第四评论信息确定为所述网络文本信息的有效文本信息,所述第四评论信息为原始评论信息中支持率高于预设阈值的评论信息。

[0077] 本实施例尚未详尽的内容请参考实施例一。

[0078] 如图7所示,该方法包括:

[0079] S210、获取目标用户的网络文本信息,所述网络文本信息包括页面文本信息、视频文字信息和原始评论信息。

[0080] S220、提取第一网络信息的目标关键词,所述第一网络信息包括页面文本信息、视

频文字信息和第三评论信息,所述第三评论信息为原始评论信息中支持率低于预设阈值的评论信息。

[0081] 在本实施例中,第一网络信息可以包括页面文本信息、视频文字信息和第三评论信息的信息,其中,第三评论信息可以为原始评论信息中支持率低于预设阈值的评论信息,预设阈值可以由经验值或者通过计算得到。目标关键词可以理解为第一网络信息中所提取的高频关键词。

[0082] 在一个实施例中,所述提取第一网络信息的目标关键词,包括:

[0083] 对第一网络信息进行分句处理,得到至少一个目标语句;

[0084] 基于预设网络模型生成所述第一网络信息的目标词义结构,所述预设网络模型由至少一个目标语句训练得到;

[0085] 对所述目标词义结构进行关键词抽取排序,得到第一网络信息的目标关键词。

[0086] 在一个实施方式中,由于原始评论信息会包括点赞信息,但页面以及视频无法获取点赞信息,因此针对无法获得支持信息的评论,本实施例可以采用基于低频关键词得评论有用性分析方法,首先可以获取原始评论信息中的低频关键词,以获得出现频率低且有用性高的评论作为样本集;此外,还可以针对可以获得点赞信息的数据获取支持度大于平均值的评论作为样本集。

[0087] 图8是根据本发明实施例二提供的一种得到有效文本信息的流程图,如图8所示,首先可以选取评论支持高于平均值的句子进入候选集,将低于平均值的评论与其他信息等一同进行处理。

[0088] 其中,可以将页面文本信息、视频文字信息和第三评论信息分割成句,进行神经网络模型训练,并获取候选特征、聚类生成关键词的词义结构;进行词义结构排序、关键词抽取,再根据评论与目标用户相关性对低频关键词进行排序,本实施例采取的排序依据可以是低频关键词中的各单词在语句中的上下文信息,关键词向量计算规则可以为式

$$V_i = \frac{\sum_{w_i \in P_i} \frac{V_{W_i}}{\| V_{W_i} \|}}{\| \sum_{w_{i \in Pi}} \frac{V_{w_i}}{\| V_{w_i} \|} \|}$$

,其中,Vi表示该关键词的向量,Pi表示当前排序的关键字,wi表示构成该关键词中的单词,Vwi代表wi在评论集上下文信息,其中,对Vi的评分可以为

$$Scoring(V_i, V_t \mid V_b) = \frac{|V_i - V_b| * |V_t - V_b|}{\| V_i - V_b \| V_t - V_b |}$$

,其中,Vt为文档聚类后人工选择的文档簇所产生的词频向量,Vb表示用全部文档集中的词频生成背景向量,分别计算每个关键词对向量Vi的评分,即可获得低频关键词的排序,最终将含排名靠前的低频关键词存入数据库作为数据集。

[0089] S230、将所述目标关键词和第四评论信息确定为所述网络文本信息的有效文本信息,所述第四评论信息为原始评论信息中支持率高于预设阈值的评论信息。

[0090] 第四评论信息可以为原始评论信息中支持率高于预设阈值的评论信息。

[0091] 综上,本步骤可以将所提取的目标关键词和第四评论信息确定为网络文本信息的有效文本信息,以进行后续情感分析结果的确定。

[0092] S240、基于所述有效文本信息确定所述目标用户的情感分析结果,所述情感分析结果用于计算所述目标用户的可信度。

[0093] 本发明实施例二提供的一种用户可信度的计算方法,获取目标用户的网络文本信息,所述网络文本信息包括页面文本信息、视频文字信息和原始评论信息;提取第一网络信息的目标关键词,所述第一网络信息包括页面文本信息、视频文字信息和第三评论信息,所述第三评论信息为原始评论信息中支持率低于预设阈值的评论信息;将所述目标关键词和第四评论信息确定为所述网络文本信息的有效文本信息,所述第四评论信息为原始评论信息中支持率高于预设阈值的评论信息;基于所述有效文本信息确定所述目标用户的情感分析结果,所述情感分析结果用于计算所述目标用户的可信度。利用该方法,通过提取页面文本信息、视频文字信息和原始评论信息中支持率低于预设阈值的评论信息的目标关键词,并将所提取的目标关键词和支持率较高的评论信息确定为网络文本信息的有效文本信息,提高了有效文本信息的有效性,进一步提高了用户可信度的准确率。

[0094] 图9是根据本发明实施例二提供的一种用户可信度的计算方法的整体架构图,如图9所示,首先可以由目标公司数据库,经过网页信息获取单元中的基于搜索引擎的主题爬虫器与特定目标网站爬虫器,获取所需要的文本信息;由短视频信息提取单元,经过短视频爬取器获取视频,再经过评论信息获取器、视频文字获取器得到文本信息;然后可以对前两个过程获取的文本信息,经过信息处理单元,分别进行页面信息处理器、评论信息处理器,如可以对文本信息进行去噪、关键内容提取,获得清洗后的数据;并将清洗后的数据输入情感分析单元,最终获取目标用户的可信程度。

[0095] 其中,图10是根据本发明实施例二提供的一种网页信息提取的流程图,如图10所示,一方面,可以通过聚焦网络爬虫手段进行基于搜索引擎的主题爬虫下载目标页面,并进行正文提取,对获取的正文进行文本聚类等得到页面文本信息;另一方面,可以对特定目标的网站进行爬虫,首先模拟用户进行登录,之后可以使用聚焦网络爬虫,并对网站进行监控当有新数据时进行增量爬虫,获取固定网站的用户评论作为原始数据的一部分。

[0096] 通过上述描述可以发现,本发明实施例提供的用户可信度的计算方法基于搜索引擎的主题爬虫与基于短视频平台爬虫获取大量评论信息,并针对重点网站进行增量爬虫。其中,原始信息经过评论信息有用性过滤,过滤掉不实、无用评论,并针对页面信息、支持度低的评论进行低频关键词分析,获取虽出现次数少却有意义的评论信息,并且进行初步的挖掘、评论时效性分析,最终获取最可能有意义的评论作为样本;之后对评论样本进行情感分析,包括文本预处理、向量化表示、模型构建等操作,最终实现从大众角度对目标公司进行情感分析,从而从大众视角上计算目标用户对应的可信程度,实现了利用大数据对企业履行承诺的能力和信誉程度进行全面评价,有助于企业防范商业危险。

[0097] 其中,通过使用多重爬虫手段,包括对短视频平台数据的爬取、基于搜索引擎的聚焦网络爬虫、基于特定网址的增量网络爬虫,尽可能全面地获取目标客户的网络文本信息,为后续计算可信度提供了数据基础。另一方面,在进行情感分析前通过进行有用性评论过滤,既能获取支持度高的评论,又使用低频关键词方式获取支持度低的评论,尽可能全面且准确的获取评论样本。

[0098] 实施例三

[0099] 图11是根据本发明实施例三提供的一种用户可信度的计算装置的结构示意图。如

图11所示,该装置包括:

[0100] 获取模块310,用于获取目标用户的网络文本信息,所述网络文本信息包括页面文本信息、视频文字信息和原始评论信息中的至少一项;

[0101] 过滤模块320,用于对所述网络文本信息进行有效性过滤,得到有效文本信息;

[0102] 确定模块330,用于基于所述有效文本信息确定所述目标用户的情感分析结果,所述情感分析结果用于计算所述目标用户的可信度。

[0103] 本发明实施例三提供的一种用户可信度的计算装置,通过获取模块获取目标用户的网络文本信息,所述网络文本信息包括页面文本信息、视频文字信息和原始评论信息中的至少一项;通过过滤模块对所述网络文本信息进行有效性过滤,得到有效文本信息;通过确定模块基于所述有效文本信息确定所述目标用户的情感分析结果,所述情感分析结果用于计算所述目标用户的可信度。利用该装置,通过对网络文本信息进行有效性过滤,为后续计算用户可信度提供了信息基础,进而提高了用户可信度的准确率。

[0104] 可选的,获取模块310具体用于:

[0105] 采用基于搜索引擎的主题爬虫器下载目标网页;

[0106] 基于文本密度指标与符号密度指标提取所述目标网页的正文内容;

[0107] 采用改进的k-means聚类算法对所述正文内容中的文本进行聚类,得到目标簇数据;

[0108] 对所述目标簇数据进行主题判别,得到相似度满足预设相似阈值的页面文本信息。

[0109] 可选的,获取模块310具体用于:

[0110] 采用目标网站爬虫器对所述目标用户的目标网站进行聚焦网络爬虫和增量爬虫,得到所述目标网站的第一评论信息,其中,所述主题爬虫器和所述目标网站爬虫器采用改进的Best-First搜索策略。

[0111] 可选的,获取模块310具体用于:

[0112] 采用视频爬虫器获取目标用户的视频信息;

[0113] 提取所述视频信息的视频文字信息和第二评论信息。

[0114] 可选的,所述网络文本信息包括页面文本信息、视频文字信息和原始评论信息,所述过滤模块320包括:

[0115] 提取单元,用于提取第一网络信息的目标关键词,所述第一网络信息包括页面文本信息、视频文字信息和第三评论信息,所述第三评论信息为原始评论信息中支持率低于预设阈值的评论信息;

[0116] 确定单元,用于将所述目标关键词和第四评论信息确定为所述网络文本信息的有效文本信息,所述第四评论信息为原始评论信息中支持率高于预设阈值的评论信息。

[0117] 可选的,所述提取单元具体用于:

[0118] 对第一网络信息进行分句处理,得到至少一个目标语句;

[0119] 基于预设网络模型生成所述第一网络信息的目标词义结构,所述预设网络模型由至少一个目标语句训练得到;

[0120] 对所述目标词义结构进行关键词抽取排序,得到第一网络信息的目标关键词。

[0121] 可选的,确定模块330具体用于:

[0122] 基于第一有效信息对长短期记忆模型进行模型训练,得到所述目标用户对应的情感分析模型,所述第一有效信息为对所述有效文本信息中训练文本信息进行预处理后的信息;

[0123] 将第二有效信息输入至所述情感分析模型中,得到所述目标用户的情感分析结果,所述第二有效信息为对所述有效文本信息中测试文本信息进行预处理后的信息。

[0124] 本发明实施例所提供的用户可信度的计算装置可执行本发明任意实施例所提供的用户可信度的计算方法,具备执行方法相应的功能模块和有益效果。

[0125] 实施例四

[0126] 图12是根据本发明实施例四提供的一种电子设备的结构示意图。电子设备旨在表示各种形式的数字计算机,诸如,膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置,诸如,个人数字处理、蜂窝电话、智能电话、可穿戴设备(如头盔、眼镜、手表等)和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例,并且不意在限制本文中描述的和/或者要求的本发明的实现。

[0127] 如图12所示,电子设备10包括至少一个处理器11,以及与至少一个处理器11通信连接的存储器,如只读存储器(ROM)12、随机访问存储器(RAM)13等,其中,存储器存储有可被至少一个处理器执行的计算机程序,处理器11可以根据存储在只读存储器(ROM)12中的计算机程序或者从存储单元18加载到随机访问存储器(RAM)13中的计算机程序,来执行各种适当的动作和处理。在RAM 13中,还可存储电子设备10操作所需的各种程序和数据。处理器11、ROM 12以及RAM 13通过总线14彼此相连。输入/输出(I/O)接口15也连接至总线14。

[0128] 电子设备10中的多个部件连接至I/O接口15,包括:输入单元16,例如键盘、鼠标等;输出单元17,例如各种类型的显示器、扬声器等;存储单元18,例如磁盘、光盘等;以及通信单元19,例如网卡、调制解调器、无线通信收发机等。通信单元19允许电子设备10通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

[0129] 处理器11可以是各种具有处理和计算能力的通用和/或专用处理组件。处理器11的一些示例包括但不限于中央处理单元(CPU)、图形处理单元(GPU)、各种专用的人工智能(AI)计算芯片、各种运行机器学习模型算法的处理器、数字信号处理器(DSP)、以及任何适当的处理器、控制器、微控制器等。处理器11执行上文所描述的各个方法和处理,例如用户可信度的计算方法。

[0130] 在一些实施例中,用户可信度的计算方法可被实现为计算机程序,其被有形地包含于计算机可读存储介质,例如存储单元18。在一些实施例中,计算机程序的部分或者全部可以经由ROM 12和/或通信单元19而被载入和/或安装到电子设备10上。当计算机程序加载到RAM 13并由处理器11执行时,可以执行上文描述的用户可信度的计算方法的一个或多个步骤。备选地,在其他实施例中,处理器11可以通过其他任何适当的方式(例如,借助于固件)而被配置为执行用户可信度的计算方法。

[0131] 本文中以上描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、芯片上系统的系统(SOC)、负载可编程逻辑设备(CPLD)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括:实施在一个或者多个计算机程序中,该一个或者多个计算

机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释,该可编程处理器可以是专用或者通用可编程处理器,可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令,并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0132] 用于实施本发明的方法的计算机程序可以采用一个或多个编程语言的任何组合来编写。这些计算机程序可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器,使得计算机程序当由处理器执行时使流程图和/或框图中所规定的功能/操作被实施。计算机程序可以完全在机器上执行、部分地在机器上执行,作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

[0133] 在本发明的上下文中,计算机可读存储介质可以是有形的介质,其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的计算机程序。计算机可读存储介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备,或者上述内容的任何合适组合。备选地,计算机可读存储介质可以是机器可读信号介质。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或快闪存储器)、光纤、便捷式紧凑盘只读存储器(CD-ROM)、光学储存设备、磁储存设备、或上述内容的任何合适组合。

[0134] 为了提供与用户的交互,可以在电子设备上实施此处描述的系统和技术,该电子设备具有:用于向用户显示信息的显示装置(例如,CRT(阴极射线管)或者LCD(液晶显示器)监视器);以及键盘和指向装置(例如,鼠标或者轨迹球),用户可以通过该键盘和该指向装置来将输入提供给电子设备。其它种类的装置还可以用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的传感反馈(例如,视觉反馈、听觉反馈、或者触觉反馈);并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0135] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如,作为数据服务器)、或者包括中间件部件的计算系统(例如,应用服务器)、或者包括前端部件的计算系统(例如,具有图形用户界面或者网络浏览器的用户计算机,用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括:局域网(LAN)、广域网(WAN)、区块链网络和互联网。

[0136] 计算系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。服务器可以是云服务器,又称为云计算服务器或云主机,是云计算服务体系中的一项主机产品,以解决了传统物理主机与VPS服务中,存在的管理难度大,业务扩展性弱的缺陷。

[0137] 应该理解,可以使用上面所示的各种形式的流程,重新排序、增加或删除步骤。例如,本发明中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行,只要能够实现本发明的技术方案所期望的结果,本文在此不进行限制。

[0138] 上述具体实施方式,并不构成对本发明保护范围的限制。本领域技术人员应该明

白的是,根据设计要求和其他因素,可以进行各种修改、组合、子组合和替代。任何在本发明的精神和原则之内所作的修改、等同替换和改进等,均应包含在本发明保护范围之内。

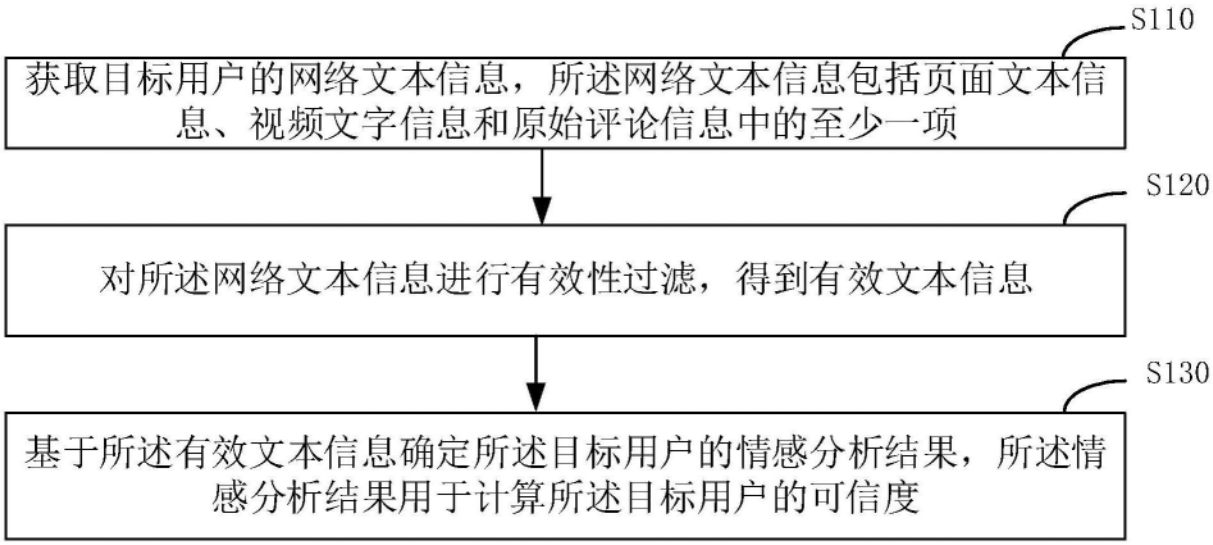

S110
获取目标用户的网络文本信息，所述网络文本信息包括页面文本信息、视频文字信息和原始评论信息中的至少一项
S120
对所述网络文本信息进行有效性过滤，得到有效文本信息
S130
基于所述有效文本信息确定所述目标用户的情感分析结果，所述情感分析结果用于计算所述目标用户的可信度

图1

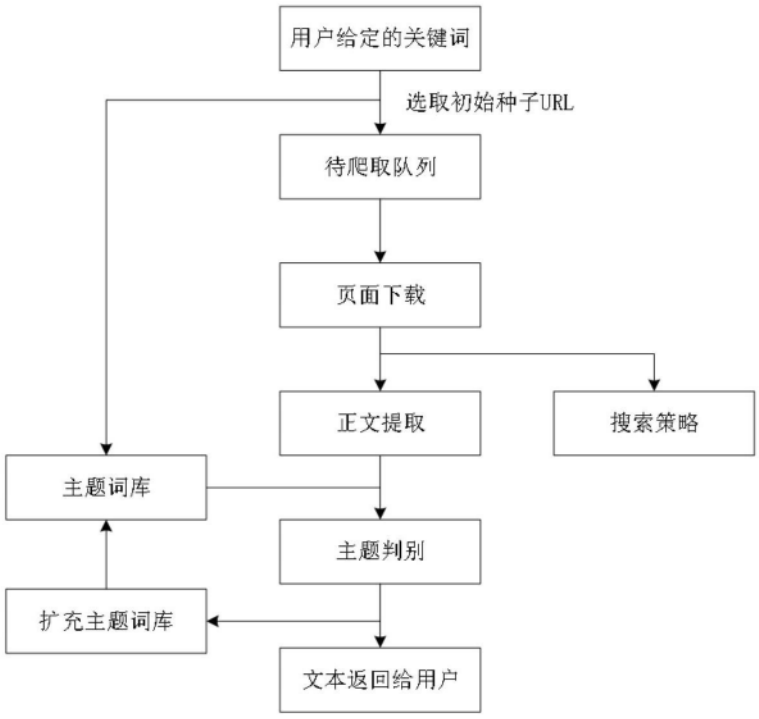

用户给定的关键词
选取初始种子URL
待爬取队列
页面下载
正文提取
搜索策略
主题词库
主题判别
扩充主题词库
文本返回给用户

图2

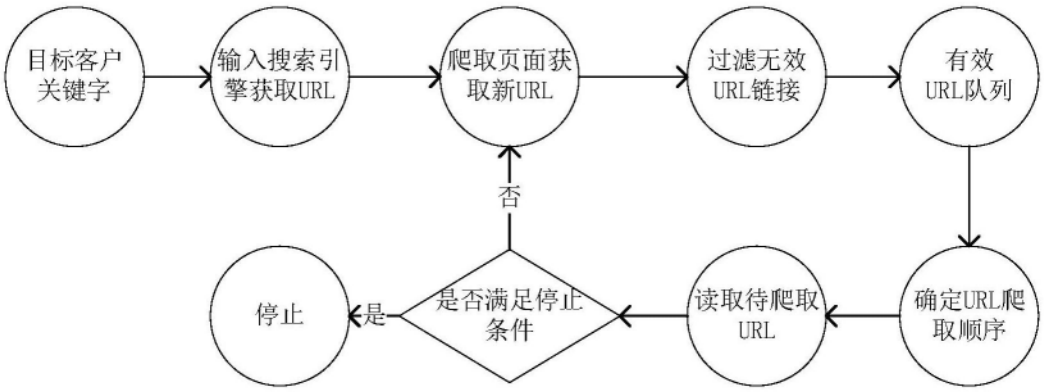

目标客户关键字
输入搜索引擎获取URL
爬取页面获取新URL
过滤无效URL链接
有效URL队列
确定URL爬取顺序
读取待爬取URL
是否满足停止条件
否
是
停止

图3

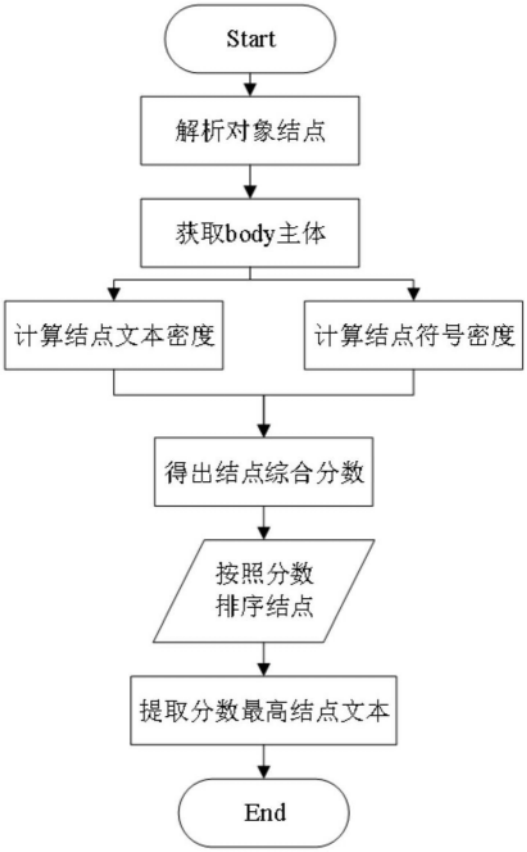

Start
解析对象结点
获取body主体
计算结点文本密度
计算结点符号密度
得出结点综合分数
按照分数排序结点
提取分数最高结点文本
End

图4

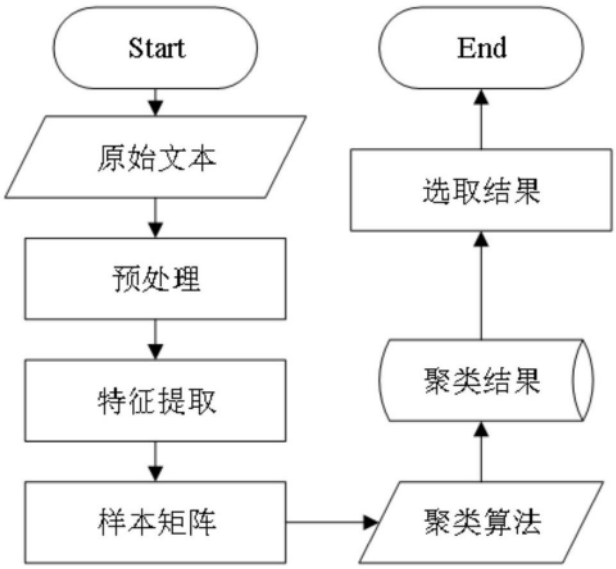

Start
原始文本
预处理
特征提取
样本矩阵
聚类算法
聚类结果
选取结果
End

图5

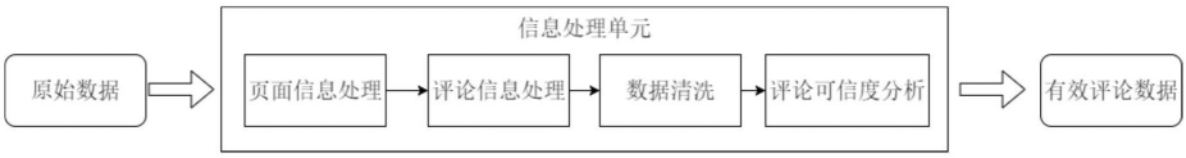

信息处理单元
原始数据
页面信息处理
评论信息处理
数据清洗
评论可信度分析
有效评论数据

图6

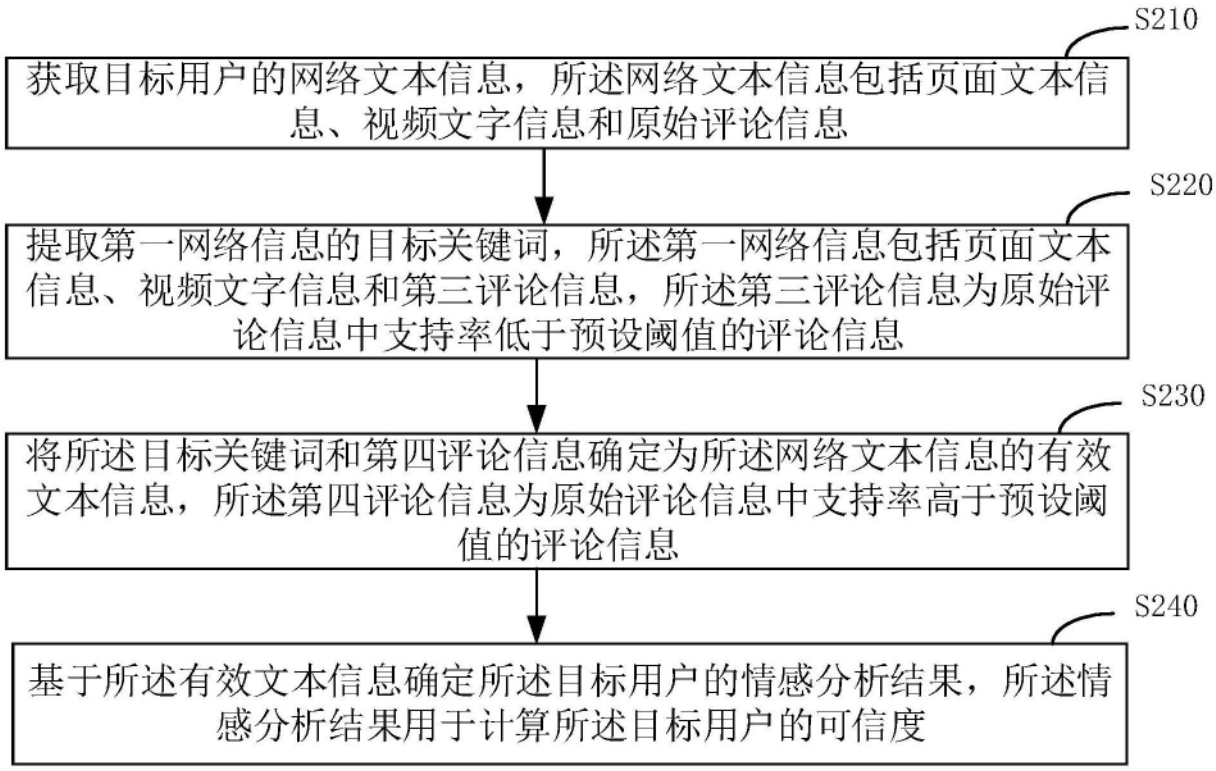

S210
获取目标用户的网络文本信息，所述网络文本信息包括页面文本信息、视频文字信息和原始评论信息
S220
提取第一网络信息的目标关键词，所述第一网络信息包括页面文本信息、视频文字信息和第三评论信息，所述第三评论信息为原始评论信息中支持率低于预设阈值的评论信息
S230
将所述目标关键词和第四评论信息确定为所述网络文本信息的有效文本信息，所述第四评论信息为原始评论信息中支持率高于预设阈值的评论信息
S240
基于所述有效文本信息确定所述目标用户的情感分析结果，所述情感分析结果用于计算所述目标用户的可信度

图7

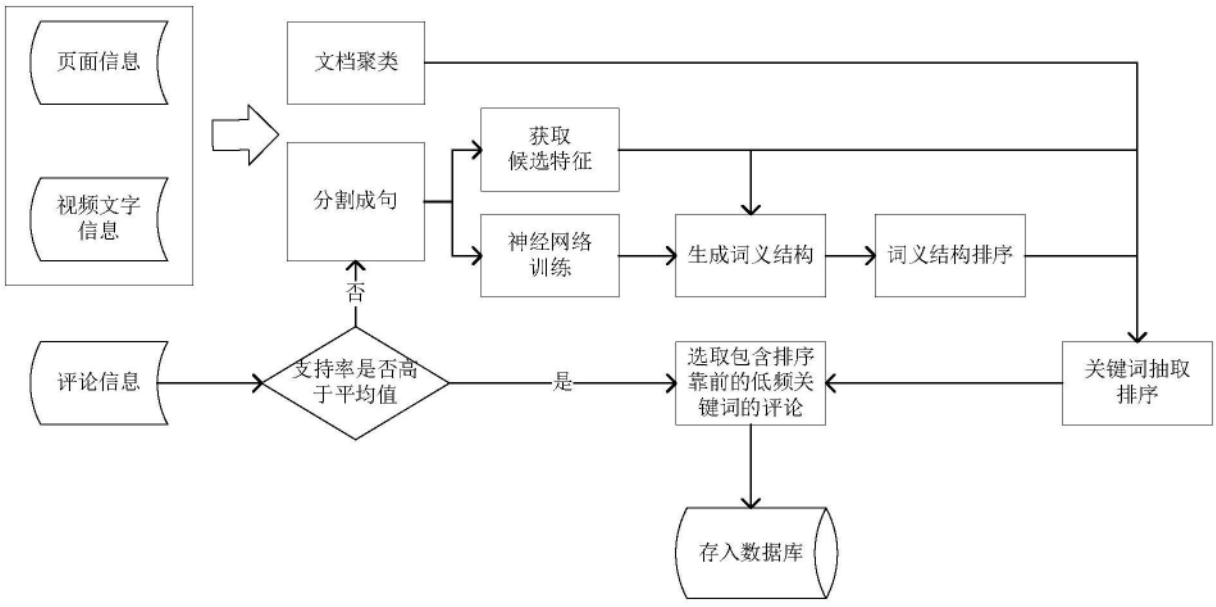

页面信息
视频文字信息
文档聚类
分割成句
获取候选特征
神经网络训练
生成词义结构
词义结构排序
否
评论信息
支持率是否高于平均值
是
选取包含排序靠前的低频关键词的评论
关键词抽取排序
存入数据库

图8

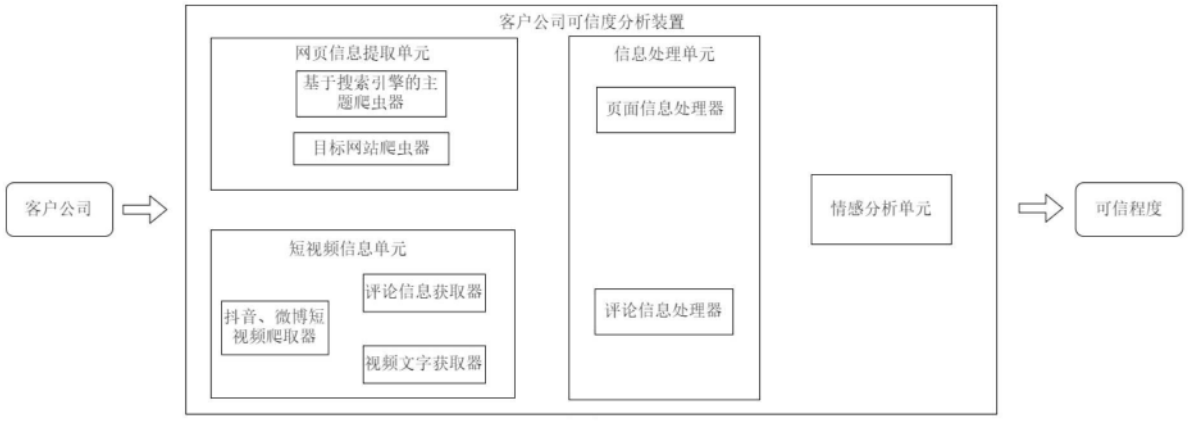

客户公司可信度分析装置
网页信息提取单元
基于搜索引擎的主题爬虫器
目标网站爬虫器
短视频信息单元
抖音、微博短视频爬取器
评论信息获取器
视频文字获取器
信息处理单元
页面信息处理器
评论信息处理器
情感分析单元
客户公司
可信程度

图9

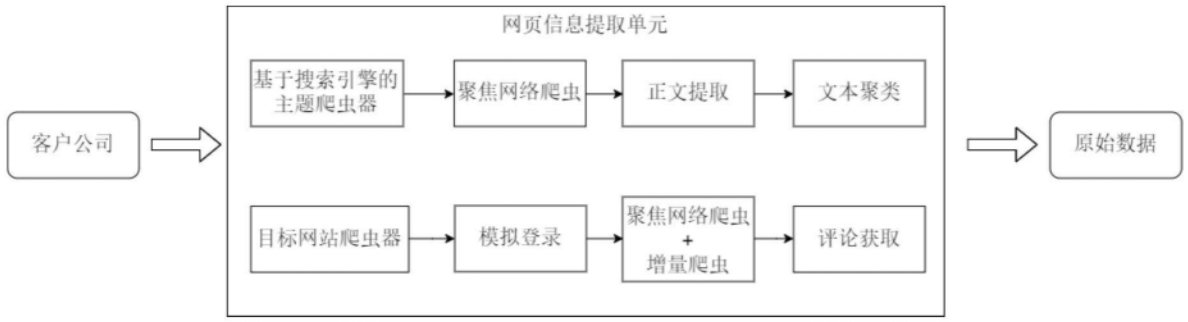

网页信息提取单元
基于搜索引擎的主题爬虫器
聚焦网络爬虫
正文提取
文本聚类
目标网站爬虫器
模拟登录
聚焦网络爬虫+增量爬虫
评论获取
客户公司
原始数据

图10

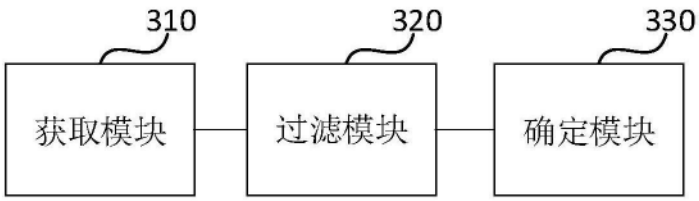

310
320
330
获取模块
过滤模块
确定模块

图11

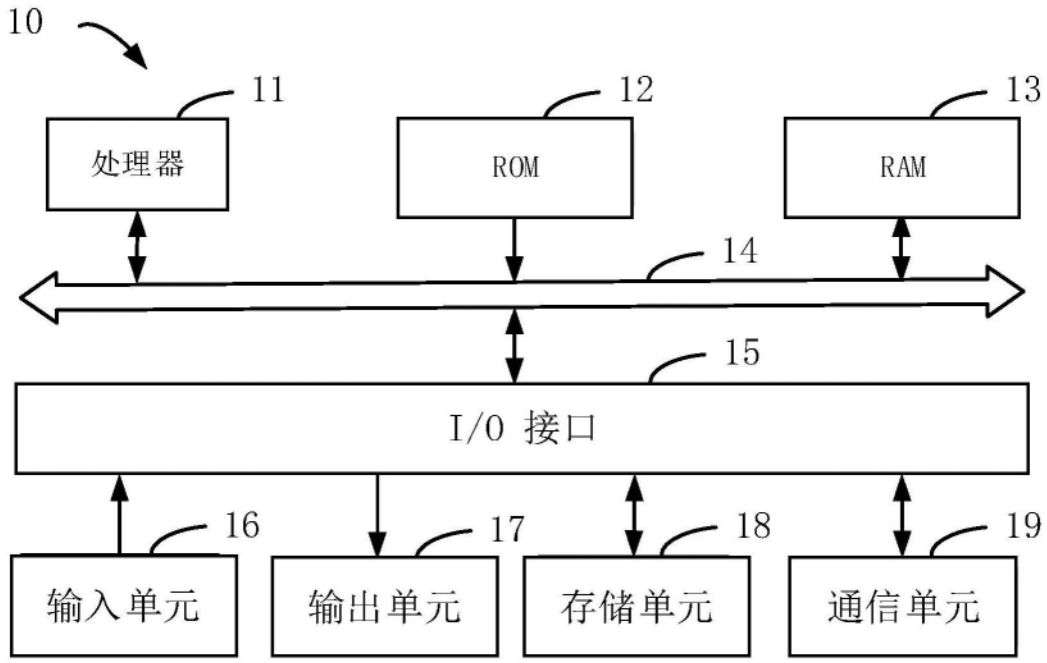


图12