(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 107533698 A
(43)申请公布日 2018.01.02

(21)申请号 201680026874.1

(22)申请日 2016.05.02

(30)优先权数据
62/158,609 2015.05.08 US
62/186,419 2015.06.30 US

(85)PCT国际申请进入国家阶段日
2017.11.08

(86)PCT国际申请的申请数据
PCT/US2016/030357 2016.05.02

(87)PCT国际申请的公布数据
WO2016/182774 EN 2016.11.17

(71)申请人 汤森路透全球资源无限公司
地址 瑞士巴尔

(72)发明人 S·沙哈 X·刘 Q·李 R·蔡
A·诺巴卡西 Q·李 R·方

(74)专利代理机构 北京润平知识产权代理有限公司 11283
代理人 金旭鹏 肖冰滨

(51)Int.Cl.
G06Q 10/10(2012.01)
G06Q 50/00(2012.01)

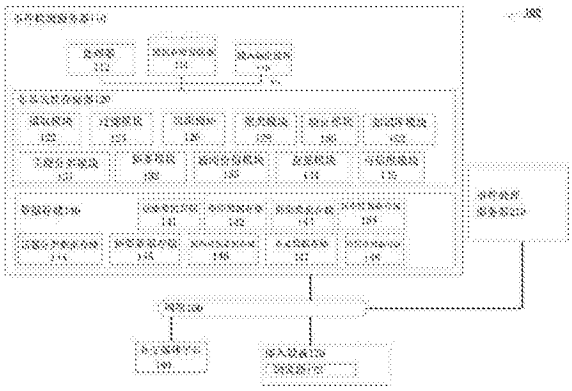
权利要求书2页 说明书13页 附图22页

(54)发明名称

社交媒体事件的检测与验证

(57)摘要

本发明公开了用于检测和验证社交媒体事件的系统和技术。该系统和技术允许处理社交媒体数据以及时提取可能有价值的信息并确定检测的信息的真实性。本公开的一个实施方式涉及事件检测。事件检测涉及社交媒体数据的摄取和处理。本公开的另一实施方式涉及对检测事件的验证并生成验证分数。



CN 107533698 A

1.一种系统,包括:

事件检测服务器,所述事件检测服务器包括处理器和存储器,所述存储器存储指令,所述指令响应于从至少一个数据源接收社交媒体数据而使所述处理器:

将成组过滤器模块应用于所述社交媒体数据以生成数据存储,所述数据存储包括所述社交媒体数据的一组识别概念和对应属性;

使用与所述社交媒体数据的所述属性相关联的对应阈值从所述数据存储中选择所述一组识别概念中的一个识别概念;并且

利用选择的识别概念生成事件检测群集。

2.如权利要求1所述的系统,其中所述存储器存储指令,所述指令响应于接收所述社交媒体数据而使所述处理器将所述选择的识别概念从所述数据存储中删除。

3.如权利要求1所述的系统,其中所述一组过滤器模块中的一个过滤器模块检测所述社交媒体数据的语言并删除不是英语的所述社交媒体数据。

4.如权利要求1所述的系统,其中所述一组过滤器模块中的一个过滤器模块检测在所述社交媒体数据中使用的脏话,并移除包含所述检测的脏话的所述社交媒体数据。

5.如权利要求1所述的系统,其中所述一组过滤器模块中的一个过滤器模块检测所述社交媒体数据中的垃圾信息、聊天和广告中的至少一种,并移除包含所述至少一个检测的垃圾信息、聊天和广告的所述社交媒体数据。

6.如权利要求1所述的系统,其中所述一组过滤器模块中的一个过滤器模块应用所述社交媒体数据的词性标注。

7.如权利要求1所述的系统,其中所述一组过滤器模块中的一个过滤器模块分析所述社交媒体数据中的语义和句法结构以确定所述社交媒体数据中的识别概念。

8.如权利要求1所述的系统,其中所述对应阈值与和所述识别概念有关的所述社交媒体数据的三个不同的属性相关联。

9.如权利要求1所述的系统,其中所述社交媒体数据的所述属性之一是作者身份值,并且所述对应阈值表示为与不同作者身份值相关联的三个相似的识别概念。

10.如权利要求1所述的系统,还包括被配置为生成所述事件检测群集的主题类别的主题分类模块。

11.如权利要求1所述的系统,还包括摘要模块,其被配置成生成用于所述事件检测群集的摘要。

12.如权利要求1所述的系统,还包括新闻价值模块,其被配置成生成用于所述事件检测群集的新闻价值分数。

13.如权利要求1所述的系统,还包括意见模块,其被配置为识别与所述事件检测群集相关联的社交媒体数据的每一项是意见还是事实。

14.如权利要求1所述的系统,还包括可信度模块,其被配置成生成用于与所述事件检测群集相关联的社交媒体数据的每一项的可信度分数。

15.如权利要求1所述的系统,还包括验证模块,其被配置成生成用于与所述事件检测群集相关联的社交媒体数据的每一项的验证分数,所述验证分数指示所述事件检测群集的真实性。

16.如权利要求15所述的系统,还包括事件处理服务器,其被配置成在图形用户界面上

向所述用户提供事件群集和所述验证分数。

17.如权利要求15所述的系统，其中所述验证分数通过分析用户类别、社交媒体级别和事件特征来确定。

18.如权利要求17所述的系统，其中所述用户类别包括作者姓名、作者描述、作者URL、作者位置、作者位置、与所述事件的位置匹配的所述作者的位置、作为所述事件见证人的作者、所述作者的账户的保护级别或所述作者的验证。

19.如权利要求17所述的系统，其中所述社交媒体级别包括与所述社交媒体数据相关联的多媒体、url、拉长的单词或来自新闻来源的url、或者词语感情色彩中的至少一个。

20.如权利要求17所述的系统，其中所述事件特征包括所述事件的至少一个主题以及所述社交媒体否认、相信或质疑与所述社交媒体数据的每一项相关联的所述事件的部分。

21.如权利要求17所述的系统，其中所述社交媒体级别是推特数据，并且所述事件特征还包括以下各项中的至少一项：与所述社交媒体数据的每一项相关联的转推最多的推文的次数、转推的推文的频率或井字标签的频率。

社交媒体事件的检测与验证

[0001] 版权声明

[0002] 本专利文件公开的部分包含受版权保护的材料。当其出现在专利和商标局专利文件或记录中时，该版权的所有者对任何人拷贝复制专利文件或专利公开都无异议，但是无论如何要以其它方式保留所有的版权权利。以下通知适用于本文件：©2015汤森路透版权所有。

[0003] 相关申请的交叉引用

[0004] 本申请要求申请日为2015年5月8日、名称为“在新闻机构公开之前揭穿推特上的谣言”的第62/158,609号美国临时专利申请，以及申请日为2015年6月30日、名称为“用于自动检测和验证社交媒体事件的系统和方法”的第62/186,419号美国临时专利申请的优先权。本段中所提到的每项申请都通过全文引用的方式被结合于此。

技术领域

[0005] 本申请涉及事件检测与验证，更具体地，涉及用于检测和验证社交媒体数据中事件的方法和系统。

背景技术

[0006] 社交媒体平台——如Twitter®或Facebook®——已经影响了新闻采集。在每一分钟里，世界各地的人都在发布图片、视频、推文并且除此以外还就各种事件进行交流。例如，人们可以对他们在事故现场看到的事情发表实时评论。由于在地理位置上接近事件的人是突发新闻的宝贵来源，因此他们发表的信息可能是非常有价值的。但是，利用这些信息是非常困难的。

[0007] 根据Twitter®网站的统计，全球约有3.2亿推特用户，包括6500万美国用户和2.54亿国际用户(Twitter的2015年第四季度盈利报告，第4页)。每分钟也有大约35万条推文。与一次可用的全部社交媒体数据相比，有价值的信息的百分比非常小。人们已经注意到，社交媒体数据主要包括谣言、噪音、垃圾信息以及大多数对专业消费者无用的信息。因此，可能有用的信息是很难发现的。此外，发现有用的信息并不能保证所声称事件的真实性。

[0008] 目前，市场上的工具采取自下而上的方式来处理从社交媒体中的信息提取。对利基信息感兴趣的用户可以通过关键字进行搜索，或者维护庞大的人员数据库，以期从社交媒体数据中获取有用的信息。这种自下而上的信息抽取方法需要进行推测工作并对列表和关键字进行持续维护。

[0009] 因此，需要一种改进的系统和技术来检测社交媒体数据级别的新兴趋势，并验证此新兴趋势的真实性。

发明内容

[0010] 本申请公开了用于检测和验证社交媒体事件的系统和技术。该系统和技术允许处理社交媒体数据以及时提取可能有价值的信息并判断检测的信息的真实性。

[0011] 本公开的一个方面涉及事件检测。事件检测涉及社交媒体数据的摄取和处理。例如，根据一个方面，一种方法包括：由事件检测服务器从至少一个数据源接收社交媒体数据，并且由事件检测服务器向该社交媒体数据应用一组过滤器以生成数据存储(即数据库或散列映射)，数据存储包括一组识别概念和该社交媒体数据的对应属性。该方法还包括由事件检测服务器使用与该社交媒体数据的属性相关联的对应阈值从数据库中选择一组识别概念中的一个，并由事件检测服务器利用选择的识别概念生成事件群集。该方法还可以包括由事件检测服务器将选择的识别概念从数据库中删除。

[0012] 在一个实施方式中，该方法还包括检测该社交媒体数据的语言并移除非英语的社交媒体数据。在另一个实施方式中，该方法还包括检测该社交媒体数据中使用的脏话并移除含有所检测出的脏话的社交媒体数据。在又一个实施方式中，该方法可包括检测该社交媒体数据中的垃圾信息、聊天和广告中的至少一种，并移除包含至少一个检测出的垃圾信息、聊天和广告的社交媒体数据。

[0013] 在进一步的实施方式中，该方法包括应用该社交媒体数据的词性标注。在替代实施方式中，该方法可以包括分析该社交媒体数据中的语义和句法结构以确定该社交媒体数据中的识别概念。

[0014] 阈值可以用于从数据库中选择一组识别概念中的一个，并且可以与和该识别概念相关的社交媒体数据的可选数量的不同属性(即，三个不同的属性)相关联。在一个实施方式中，该社交媒体数据的属性之一是作者身份值(即，用户)，并且对应阈值表示与不同作者身份值(即，不同用户)相关联的预定数量(即，三个)的类似识别概念。

[0015] 在又一个实施方式中，该方法包括但不限于为每个群集及其相应数据生成新闻价值分数、主题分类、摘要和可信度分数。

[0016] 在一个实施方式中，例如，该方法还包括为每个群集及其相应数据生成验证分数，该验证分数表示群集中每一说法的真实性或准确性。真实性分数和事件群集可以在图形用户界面上提供给用户。

[0017] 在一个实施方式中，通过分析用户类别、社交媒体级别和事件特征来确定真实性分数。

[0018] 用户类别包括但不限于确定与社交媒体数据的每一项相关联的作者姓名、作者描述、作者URL、作者位置、与事件的位置相匹配的作者的位置、作为事件见证者的作者、作者的账号的保护级别以及作者的验证中的至少一个。

[0019] 社交媒体级别包括但不限于确定与该社交媒体数据相关联的多媒体、url、拉长的单词和来自新闻来源的url、或者词语感情色彩中的至少一个。

[0020] 事件特征包括但不限于确定事件的主题和社交媒体否认、相信或质疑与社交媒体数据的每一项相关联的事件的部分中的至少一个。

[0021] 在社交媒体数据为推特数据的进一步的实施方式中，事件特征还包括确定与社交媒体数据的每一项相关联的转推最多的推文的次数、转推的推文的频率和井字标签的频率中的至少一个。

[0022] 本申请公开了这样一种系统、设备以及物品，所述系统、设备和物品包括存储用于实现各种技术的机读指令的机读介质。下面将更详细地讨论各种实施方式的细节。

[0023] 一个优点在于准确性和速度。例如，在一个实施方式中，通过使用上述系统和技

术，集体用户对事件的真实性预测可以达到约85％的准确度，并能够以比主流媒体更快的速度确认同一信息。

[0024] 结合下列详细描述、所附附图以及权利要求，本发明的其他特征和优点将变得显而易见。

附图说明

[0025] 图1是系统的示例性的架构图；

[0026] 图2是示例性的事件处理服务器；

[0027] 图3a是本公开中一个实施方式的示例性的流程图；

[0028] 图3b是本公开中另一个实施方式的示例性的流程图；

[0029] 图4a示出了真实性计算中的示例性的要素；

[0030] 图4b示出了在替代验证计算中的示例性的要素；

[0031] 图5a示出了社交媒体数据项的示例性的处理；

[0032] 图5b示出了将关键概念映射到各自的社交媒体数据的示例性表格表示；

[0033] 图5c示出了与图5a的示例性社交媒体数据相关的示例数据库表示；

[0034] 图5d示出了示例性单位群集；

[0035] 图5e示出了示例性的摄取数据；

[0036] 图5f-5k是图5e中摄取数据的示例性元数据；

[0037] 图5l-5n是将图5e的摄取数据作为相关单元数据之一的事件检测群集的示例性元数据；

[0038] 图6a示出了通过示例性图形用户界面(GUI)可查看的默认事件检测群集；

[0039] 图6b示出了通过示例性图形用户界面(GUI)可查看的示例性事件检测群集；

[0040] 图6c示出了通过示例性图形用户界面(GUI)可查看的选定事件检测群集；以及

[0041] 图7a-7e示出了通过示例性图形用户界面(GUI)可用的事件检测群集的附加过滤器。

具体实施方式

[0042] 以下将结合所述附图进行说明，所述附图构成说明书的一部分，并且其中实施本公开的具体实施方式作为示例出现。应该理解，在不脱离本公开的范围的情况下，可以利用其他实施方式并且可以进行结构上的改变。

[0043] 图1示出了一种用于检测和验证来自社交媒体数据的事件的示例性系统100。如图1所示，在一个实施方式中，该系统100被配置为包括事件检测服务器110，该事件检测服务器110通过网络160与社交媒体平台180通信。系统100还包括接入设备170，该接入设备170通过网络160与事件处理服务器210通信。关于示例性的事件处理服务器210的更多细节在图2中示出。事件检测服务器110通过网络160与事件处理服务器210通信。接入设备170可以包括个人计算机、膝上型计算机或其他类型的电子设备，诸如移动电话、智能电话、平板电脑、PDA或PDA电话。在一个实施方式中，例如，接入设备170耦合到输入输出设备(未示出)，该输入输出设备包括与诸如鼠标的指向设备相结合的键盘，用于向事件处理服务器210发送事件请求。优选地，接入设备170被配置为包括用于从事件处理服务器210请求和接收信

息的浏览器172。接入设备170的浏览器172与事件处理服务器210之间的通信可以采用一个或多个网络协议，该网络协议可以包括HTTP、HTTPS、RTSP或RTMP。虽然在图1中示出了一个接入设备170，但是系统100能够支持一个或多个接入设备。

[0044] 网络160可以包括连接在内联网、外联网或因特网配置中的各种设备，例如路由器、服务器和交换元件。在一些实施方式中，网络160使用有线通信以在接入设备170和事件处理服务器210、社交媒体平台180和事件检测服务器110之间传送信息。在另一个实施方式中，网络160采用无线通信协议。在又一实施方式中，网络160将有线和无线技术结合使用。

[0045] 如图1所示，在一个实施方式中，事件检测服务器110可以是专用服务器，并且优选地包括处理器112，诸如中央处理单元('CPU')、随机存取存储器('RAM')114、诸如显示设备(未示出)之类的输入输出设备116以及非易失性存储器120，其全部通过公共总线111互连并由处理器112控制。

[0046] 在一个实施方式中，如图1示例所示，非易失性存储器120被配置为包括用于从社交媒体平台180接收社交媒体数据的摄取模块122。示例性的社交媒体平台是但不限于是Twitter®、Reddit®、Facebook®、Instagram®或LinkedIn®。此处所用的术语"摄取数据"是指从社交媒体平台180接收到的社交媒体数据，其可以是但不限于推文和/或在线消息。

[0047] 非易失性存储器120还包括用于处理摄取数据的过滤模块124。在一个实施方式中，摄取数据的处理可以包括但不限于检测摄取数据的语言并且滤除包含脏话、垃圾信息、聊天和广告的摄取数据。

[0048] 非易失性存储器120还被配置为包括用于分析摄取数据中的语义和句法结构的组织模块126。在一个实施方式中，组织模块126可以应用摄取数据的词性标注。在另一个实施方式中，组织模块126检测包含在摄取数据中的关键概念。

[0049] 如图1示例所示，非易失性存储器120还可以被配置为包括聚类模块128，聚类模块用于将由组织模块126识别的关键概念存储到数据库(其实例可以是但不限于散列映射)中，以及在达到包含共同关键概念的不同摄取数据的阈值时生成事件检测群集。

[0050] 非易失性存储器120还被配置为包括：主题分类模块131，用于按照主题对事件检测群集进行分类；摘要模块132，用于为事件检测群集选取代表性描述；以及新闻价值模块133，用于确定新闻价值分数以指示事件检测群集的重要性。

[0051] 非易失性存储器120还被配置为包括意见模块134，用于检测事件检测群集中的每一摄取数据是特定人的意见还是事实(例如，非意见性的语调)，以及可信度模块135，用于确定摄取数据的可信度分数。在一个实施方式中，可信度分数与三个因素相关：用户/来源的可信度，即谁提供了该信息，群集的可信度，即该信息是什么，以及推文的可信度，即该信息与其他信息如何相关。

[0052] 非易失性存储器120被进一步配置为包括用于确定事件检测群集的准确性的验证模块150。在一个实施方式中，验证可以通过生成真实性分数的真实性算法来完成。在另一个实施方式中，验证模块150可以为基于从摄取数据收集的证据认定为真的断言生成一个概率分数。

[0053] 非易失性存储器120被进一步配置为包括知识库模块152，用于开发属于可信来源的信息的数据库并将该信息存储在知识库数据存储248(图2)中。

[0054] 如图1示例所示，具有数据存储140，其被软件模块124、126、128、131、132、133、134、135、150、152中的一个或多个用于获得和存储与摄取数据有关的信息。在一个实施方式中，数据存储140为关联数据库。在另一个实施方式中，数据存储140是文件服务器。在又一实施方式中，数据存储140是事件检测服务器110的非易失性存储器120中的配置区域。虽然图1中示出的数据存储140是事件检测服务器110的一部分，但是本领域技术人员可以理解，该数据存储140可以分布在各种服务器上，并且可以通过网络160访问服务器110。如图1所示，在一个实施方式中，数据存储140被配置为包括过滤数据存储141、组织数据存储142、群集数据存储143、话题分类数据存储144、摘要数据存储145、新闻价值数据存储146、意见事实数据存储147、可信度数据存储148和真实性数据存储154。

[0055] 过滤数据存储141包括已经由过滤模块124处理的摄取数据。例如，在一个实施方式中，由过滤模块124处理的摄取数据可以是不包含脏话、广告、垃圾信息、聊天或宣传的英文推文。

[0056] 组织数据存储142包括已经由组织模块126处理的摄取数据。在一个实施方式中，组织数据存储142中的摄取数据可以包括作为摄取数据元数据的一部分而存储的词性标注符号或识别的关键概念。

[0057] 群集数据存储143包括已被过滤模块124和组织模块126处理并排队等候以形成群集的摄取数据。在进一步的实施方式中，群集数据存储143还可以包含由组织模块126识别的与对应的摄取数据相匹配的关键概念(例如，散列映射)的数据存储或数据库。如此处关于关键概念的数据库所使用的那样，摄取数据(例如，推文和/或在线消息)也可以被称为单位数据。

[0058] 主题分类数据存储144包括由主题分类模块131确定的事件检测群集的类别。示例性的主题可以包括但不限于商业/金融、技术/科学、政治、体育、娱乐、健康/医疗、危机(战争/灾难)、天气、法律/犯罪、生活/社会以及其他。

[0059] 摘要数据存储145包括代表了由摘要模块132确定的事件检测群集的所选单位数据。

[0060] 新闻价值数据存储146包括由新闻价值模块133计算的新闻价值分数。例如，较高的分数意味着从新闻标准来看事件检测群集可能是重要的。

[0061] 意见数据存储147包括这样一种信息，该信息关于由意见模块134来确定给定单位数据是包括特定个人的意见还是对事实的断言。

[0062] 可信度数据存储148包括由可信度模块135确定的可信度或置信度分数。

[0063] 真实性数据存储154包括由验证模块150生成的关于事件检测群集的准确度等级的衡量标准。在一个实施方式中，它可以是通过真实性算法确定的真实性分数。在另一个实施方式中，它可以是基于从社交媒体收集的所有证据来指示准确性的概率的验证分数。

[0064] 在进一步的实施方式中，如图1所示，事件处理服务器210包括处理器(未示出)、随机存取存储器(未示出)和非易失性存储器(未示出)，它们经由公共总线互连并由处理器控制。在一个实施方式中，事件处理服务器210负责存储由事件检测服务器110产生或待使用的处理后的信息。在另一个实施方式中，事件处理服务器210还与用户直接通信。事件处理服务器210在图2中进一步示出。

[0065] 需要说明的是，图1所示的系统100是本公开的一种实施方式。本公开的其他系统

实施方式可以包括未示出的附加结构，例如辅助存储器和附加计算设备。另外，本公开的其他各种实施方式包括的结构比图1中所示的更少。

[0066] 现在转到图2，在一个实施方式中的事件处理服务器210包含具有非易失性存储器230和UI(用户界面)模块232的网络服务器220。

[0067] UI模块232经由网络160通过浏览器172与接入设备170通信。UI模块232可以通过浏览器172向用户呈现检测的事件检测群集及其相关联的元数据。示例性的相关联的元数据可以是但不限于与一个或多个事件检测群集相关的主题、新闻价值标识和验证分数。

[0068] 事件处理服务器210可以进一步包括数据存储240以主持摄取数据存储242、生成群集数据存储244、发出数据存储246和知识库数据存储248。

[0069] 摄取数据存储242包括从社交平台180接收并由摄取模块122处理的摄取数据。

[0070] 生成群集数据存储244包括已经由模块122、124、126、128、131、132、133、134、135和150处理的事件检测群集。

[0071] 如图3a中关于步骤330-332的解释，发出数据存储246包括由聚类模块128舍弃的关键概念和对应的摄取数据。在替代实施方式中，发出数据存储可以位于事件检测服务器110中。

[0072] 知识库数据存储248包括由知识库模块152确定的可信来源的列表。

[0073] 在一个实施方式中，事件处理服务器210通过网络160与事件检测服务器110通信。在另一个实施方式中，事件处理服务器210被包括在事件检测服务器110的非易失性存储器120中。在又一个实施方式中，事件处理服务器210被配置为与事件检测服务器110直接通信。示例性的事件处理服务器210可以是但不限于MongoDB®或ElasticSearch®。

[0074] 现在参考图3，公开了一种检测和验证社交媒体事件的示例性的方法300。如图3所示，在步骤302中，来自社交媒体平台180的信息由事件检测服务器110的摄取模块122检索。在一个实施方式中，摄取模块122可以包括与社交媒体平台180应用程序API对接的脚本或代码。脚本或代码也能够从API中请求和提取信息。在另一个实施方式中，摄取模块122可以确定用户和摄取数据的位置，并将位置信息作为元数据附加到摄取数据。

[0075] 接下来在步骤304中，一旦接收到摄取数据，摄取模块122将摄取数据存储到事件处理服务器210的摄取数据存储242中。在进一步的实施方式中，元数据也可以由摄取模块122生成并在存储到摄取数据存储242中之前附加到摄取数据。

[0076] 在替代实施方式中，知识库模块152可以使用从摄取数据收集的信息来编译可信来源的列表。知识库模块152将可信来源的列表存储在知识库数据存储248中。在一个实施方式中，知识库模块152可以从摄取数据中分析用户配置文件以获取诸如待用于对可信来源的列表进行编译的用户隶属关系或地理位置的信息。在进一步的实施方式中，知识库模块152将建立的可信用户和由用户生成的评论列表用作可用于生成可信来源的列表的相关信息。例如，如果可信用户具有包含技术用户列表的技术列表，则与技术用户相关联的用户ID和相关信息(例如，与用户ID相关联的相关技术列表)也被挖掘以获取信息。随着更多社交媒体数据被摄取，知识库模块152不断更新知识库数据存储248，并且可以以预定频率评估知识库模块152以确保信息是最新的。

[0077] 继续到步骤306，过滤模块124从摄取数据存储242中检索摄取数据并处理该摄取数据。过滤模块124的示例性处理可以包括语言检测和脏话检测。在一个实施方式中，过滤

模块124确定摄取数据的语言并且消除不是英语的摄取数据。在替代实施方式中，可以针对其他语言进行摄取数据的消除。

[0078] 过滤模块124还可以检测摄取数据中的脏话词语并且标记包含脏话的该摄取数据。然后由过滤模块124消除包含脏话的摄取数据。在一个实施方式中，脏话的检测基于查询词典集的脏话词语。

[0079] 在进一步的实施方式中，过滤模块124可以采用分类算法，该分类算法移除被识别为垃圾信息、聊天或者广告的摄取数据。垃圾信息的示例性标识为称为“follow me@xyz”的摄取数据。摄取数据中的示例性聊天可能是关于日常生活的像“早上好”那样的一般闲谈。摄取数据中的示例性广告可以包含诸如“点击这里购买这件极好的T恤，只需10美元”的语言。在一个实施方式中，分类算法基于机器学习模式，该机器学习模式已经基于语言(即，用于构建数据的术语)、消息质量(即，大写字母、表情符号的存在)、用户特征(即，平均注册年龄)进行了针对诸多特征的训练。示例性的机器学习模式包括但不限于支持向量机、随机森林和回归模型。然后，经过滤的摄取数据被存储在过滤数据存储141中。

[0080] 一旦由过滤模块124完成了过滤，则在步骤308中，组织模块126从过滤数据存储141中检索现在过滤的摄取数据，并检测该摄取数据中的关键概念。在一个实施方式中，组织模块126检测该摄取数据中的语义和句法结构。

[0081] 在另一个实施方式中，组织模块126可以通过词性标注器对摄取数据应用词性标注。例如，组织模块126识别摄取数据中的动词、副词、专有名词和形容词。在进一步的实施方式中，可以具有用于由组织模块126进行识别的预定的术语列表，其包括但不限于诸如“火”、“龙卷风”或“爆炸”之类的危机术语。该预定的术语列表还可以基于这样一种概念进一步定制，所述概念不是专有名词但是很好的表示了摄取数据的主要内容。

[0082] 然后，可以将词性标注符号或识别的关键概念存储到组织数据存储142中。在一个实施方式中，可以将该词性标注符号或识别的关键概念附加到摄取数据元数据中并存储到组织数据存储142中。

[0083] 所有在摄取数据中找到的关键概念、专有名词、井字标签以及任何列表术语均被指定为‘可标记词’。在进一步的实施方式中，该可标记词可以进一步连接以形成更有意义的可标记词。例如，如果跟有“York”的“New”被标识为可标记词，则将这些术语连接以将修订后的可标记词指示为“New_York”并去除单个的“New”和“York”。

[0084] 一旦关键概念被组织模块126识别，则在步骤310中，聚类模块128从组织数据存储142获得经组织的摄取数据，并参照相应的摄取数据创建关键概念数据库。在一个实施方式中，所引用的相应的摄取数据可以是单位数据的形式。然后，该数据库存储在群集数据存储143中。

[0085] 在步骤312中，每个关键概念具有预定的时间范围以成长为形成单元群集所需的单元数据的最小计数，否则它就会被舍弃。示例性的阈值计数可以是但不限于关键概念的三(3)个单位数据。举例来说，如果众多用户(即，作者身份值)在他们的社交媒体数据中提及类似的关键概念，那么可能有一个刚发生的事件。

[0086] 一旦满足了包含共同可标记词的单位数据的阈值数量，则在步骤314中，聚类模块128生成单位群集。在进一步的实施方式中，与可标记词对应的单元数据在步骤314中生成为单元群集，并在步骤316中从数据库中移除。

[0087] 然而，如果该阈值未被满足，则在步骤330中对数据库中的可标记词进行审核。尚未超过预定时间窗(即，2小时)的可标记词和新摄取的数据一起从步骤302开始重复该过程。举例来说，这可能是其他众多用户尚未提及的最新的社交媒体信息。

[0088] 然而，在步骤332中，在预定时间窗(即，2小时)之后永远不会达到单位数据的最小阈值的可标记词从数据库中被移除。被舍弃的可标记词和单位数据可以和与其相关的其他元数据一起被发送到发出数据存储246。举例来说，无其他用户提及的社交媒体信息对专业消费者来说可能并不是重要的事件。

[0089] 返回到步骤314，一旦生成了单位群集，则在步骤316中将与其相应的可标记词和单位数据从数据库中移除。在步骤318中，针对一组先前生成的事件检测群集来检查新生成的单位群集。该组先前生成的事件检测群集可以位于群集数据存储143中。在替代实施方式中，生成的群集可以位于事件处理服务器210的生成群集数据存储244中。

[0090] 如果与该组先前生成的事件检测群集不匹配，则继续到步骤324，由聚类模块128将单元群集确定为新的事件检测群集并将其存储到群集数据存储143中。

[0091] 然而，如果与现有的已生成的事件检测群集相匹配，则在步骤320中，基于一组预定的规则，做出合并两个相似群集或将它们保持为两个单独的群集的决定。在一个实施方式中，可以基于相同的基本概念做出该合并决定。

[0092] 如果决定合并两个相似群集，则继续到步骤322，聚类模块128对群集进行合并，并将合并的事件检测群集存储，存储在聚类数据存储143中。例如，如果社交媒体信息与先前检测的事件相同，则将该社交媒体信息与先前检测的事件合并。

[0093] 然而，如果群集仍有不同，则继续到步骤324，将单位群集确定为新的事件检测群集并将其存储到群集数据存储143中。例如，与先前检测的事件不同的社交媒体信息对于专业消费者来说可能是重要的事件，并且其本身也应该值得注意，因此该单位群集被聚类模块128认定为事件检测群集。

[0094] 现在转到图3b，在进一步的实施方式中，在步骤342中，在存储了事件检测群集之后，可以对事件检测群集进行补充。示例性的补充是但不限于主题分类、摘要、新闻、观点和可信度。

[0095] 如前所述，主题分类模块131可以确定事件检测群集的一个或多个类别。该类别可以是预定类别的分类(即，政治、娱乐)。该类别被添加到事件检测群集的元数据。

[0096] 摘要模块132可以在事件检测群集中选择对群集进行了最佳描述的单位数据。选定的单位数据被用作该事件检测群集的摘要。在进一步的实施方式中，摘要模块132也可以采用下列度量标准，诸如在生成事件检测群集的摘要过程中的最早的单位数据或人气高的单位数据。该摘要被添加到事件检测群集的元数据中。

[0097] 新闻价值模块133使用新闻价值算法来计算新闻价值分数。该新闻价值分数表示新闻标准意义上的事件检测群集的重要性。例如，与作为突发新闻事件的飞机坠毁有关的事件检测群集被认为比与病毒式传播的名人照片有关的群集更重要。在一个实施方式中，新闻价值算法是一种监督式机器学习算法，其已就具有新闻价值的摄取数据组进行训练，并为经它计算的任何摄取数据预测新闻价值分数。该新闻价值分数被添加到事件检测群集的元数据中。

[0098] 意见模块134判断事件检测群集中的每个单元数据是包含特定人的意见还是对事

实的断言。在一个实施方式中，对于作为对事实的断言的单位数据，还给单位数据分配一个将断言指示为事实的分数，并且对于意见也同样如此。在另一个实施方式中，意见模块134在两阶段过程中执行。在第一阶段，应用规则导向分类器，该规则导向分类器使用简单的规则来识别意见，该简单的规则系基于存在/不存在某些类型的意见/情感性词语和/或人称代词的使用。在第二阶段，所有指示为非意见的单位数据都通过一个经过专门训练的词包分类器来识别事实断言。然后，对事实或意见的判断被存储作为事件检测群集元数据的一部分。

[0099] 可信度模块135确定事件检测群集中的每个单位数据的置信度分数。在一个实施方式中，置信度分数与三个因素相关：来源可信度、群集可信度和推文可信度。然后，由该因素生成的分数和信息被存储作为事件检测群集元数据的一部分。

[0100] 来源可信度与单位数据的来源有关。如果来源是可信来源，例如，像白宫这样的对事件进行说明的权威机构比随意一个未知用户更可信。在一个实施方式中，来源可信度通过算法来测量，所述算法采用诸如但不限于用户的年龄、描述以及社交媒体账户的个人资料图像之类的特征。

[0101] 群集可信度与信息是什么有关。通常情况下，包含真实事件的检测事件检测群集可具有与虚假的事件检测群集不同的增长模式，例如，虚假事件可能由负面动机驱动，如故意传播谣言。基于历史数据使用监督式学习模式，历史数据基于增长模式来识别事件检测群集为真或为假的可能性。

[0102] 推文可信度与单元数据中的单个推文的内容以及其中提及的语言有关。在一个实施方式中，针对在可信和不可信的单元数据上训练的一组文本词语来对单元数据进行评估。

[0103] 接下来，在步骤344中，验证模块150分析应用于事件检测群集及其相关单元数据的补充以确定事件检测群集的准确度水平。在一个实施方式中，验证模块150可以基于来自单元数据的三个类别——用户、推文级别或社交媒体数据级别和事件——来生成真实性计算。在另一个实施方式中，验证模块150可以使用提取的语言、用户和来自事件检测群集及其相关单元数据的其他元数据特征来计算正在传播的谣言为真的概率。结合图4a和4b对验证进行更更详细的解释。

[0104] 最后，在步骤346中，经补充的事件检测群集被存储到事件处理服务器210的生成群集数据存储244中。

[0105] 图4a示出了在真实性计算中使用的类别的示例性描述。供考虑的第一类别与用户类别有关。在一个实施方式中，用户特征402a采用布尔值，并且可以包括但不限于图4a中示出的名称、描述、url、位置、匹配群集位置、见证人、受保护(即，是否为隐私)、已验证。该用户类别采集了从用户的社交媒体配置文件收集的用户具体信息。像位置或网址这样的示例性特征可以衡量用户的可信度。例如，如果用户隐藏了他们的位置，则很难确定他们的发言的准确性。但是，如果他们的位置与检测事件检测群集的位置相匹配，则从摄取数据中收集的事件可被更顺利地视为是准确的。

[0106] 供考虑的第二类别与社交媒体级别有关。在一个实施方式中，布尔型的社交媒体特征402b可以包括但不限于多媒体、拉长的单词、url和新闻url，如图4a所示。社交媒体类别可以进一步包括数字类型：数字情感正面词语、数字情感负面词语以及数字类型的情感

分数。例如，如果用户在所报导的事件中附加了图片或多媒体，则那可以清楚地指示社交媒体数据上该报导的准确性。在另一个实施方式中，用户所使用的词语的类型——特别是拉长的单词，即“OMMMMMMGGG!!”——可以传达该用户对该事件所表达的震惊，并使其更加可信。但是，如果用户在社交媒体数据中使用url，用户可能会通过复制进行分享。在进一步的实施方式中，还对摄取数据的感情色彩进行检查。可以针对一组正面和负面的表达情感的词语来检查摄取数据。举例来说，如果事件检测群集是关于一场灾难，则摄取数据的总体语调应为负面的。

[0107] 供考虑的第三类别是事件特征。在一个实施方式中，如图4a，事件特征402c可以包括：事件主题，其可以是分类类型的，以及最高转推次数、转推总和、井字标签总和、反对部分、支持部分、质疑部分，其可以是数字类型的。在一个实施方式中，如果摄取数据是twitter推文，则在假定转推次数与对准确性具有更高权重的事件的人气相关的情况下，估算转推次数和总和。在另一个实施方式中，井字标签也可以作为事件的指示符。例如，运动相关的摄取数据可包含许多井字标签，而与灾难有关的摄取数据可能没有很多井字标签，因为当灾难在用户位置处发生时，可能没有时间列出如此多的井字标签。在又一实施方式中，该算法还考虑摄取数据中否认、相信或质疑该事件的部分。

[0108] 验证模块150生成一个基于三个类别被聚合的矩阵，以对应于从虚假的谣言到真实的叙事的范围生成位于-1和1之间的真实性分数。在一个实施方式中，如图5n所示，可以将真实性评分550添加到事件检测群集的元数据中。在进一步的实施方式中，如图6b所示，真实性分数614可以以圆形表示的形式呈现给用户。

[0109] 图4b示出了基于从社交媒体收集的信息由验证模块150确定事件检测群集为真的概率分数。在图4b的实施例中，Twitter被用作一个示例性的社交媒体平台。在一个实施方式中，验证模块150首先判断事件检测群集的单元数据是专家型断言还是见证型断言。

[0110] 专家型断言是可仅由被认为是对断言具有权威性的人或组织做出的断言。示例性的专家型断言可以是 Apple® 公司声称他们将发布新的iPhone®。验证模块150可以调用知识库模块152来确定单元数据(即，Apple®)的识别用户是否是可信来源，并且如果该单元数据来自于可信来源，则授予较高分数。

[0111] 在进一步的实施方式中，如果单元数据的用户来自于由知识库模块152确定为对该主题具有权威性的可信来源的列表，则给出较高的分数。如果摄取数据的用户不具权威性，则知识库模块152转而考虑其他专家及其最近发表的推文以收集或否定用户断言。

[0112] 见证型断言是随机用户可能做出的断言。这些包括危机型事件(例如，用户123声称某一特定区域发生了爆炸)。在一个实施方式中，验证模块150将单元数据中的主题或地理位置与来自相同地理区域的其他单元数据进行比较。如果其他用户在同一时间段内没有提及相同的断言，则可以分配一个较低的分数。

[0113] 在进一步的实施方式中，也可以考虑由知识库模块152确定的有关组织的知识库。来自有关组织的集体知识库的社交媒体数据也可以由事件检测服务器110处理，以确定他们是否正在讨论类似的断言，并且该社交媒体数据被用于与当前的单元数据进行比较以确定真实性的等级。

[0114] 然后，验证模块150分配一个可以指示其真假可能性的概率。在一个实施方式中，验证模块可以在算法上计算介于-1和1之间的分数，其中0是中性的，表示我们缺乏关于此

事的信息，1描述了真实断言的置信度的最高值，-1是虚假断言的置信度的最高值。例如，如果来自非常可信的来源的信息证实断言是真实的，那么它的分数可能是1。但是，对于我们无法找到具体证据来证明其确实性或真实性的准确度的情况，则根据所收集的证据类型将分数降到-1和1之间。当新的证据被纳入评估时，可以对该置信度进行重新评估。

[0115] 现在参考图5a，图5a示出了示例性的摄取数据。在一个实施方式中，该摄取数据可以是但不限于推文。组织模块126分析该摄取数据中的语义和句法结构以识别关键概念。在一个实施例中，术语502a-502d，例如“confederate flag”、“rally”、“Linn Park”、“Birmingham”，被组织模块126识别为关键概念。尽管在这一实施例中识别了四个关键概念，但是由组织模块126识别的术语的数量可以为n个。在一个实施方式中，如图5b所示，关键概念被存储在数据库500中，同时该关键概念被指定为“可标记词”并且相应的原始摄取数据被指定为“单元数据”。如图5b所示，为n个可标记词提供一个列504，每个标记具有对应于n个单位数据的对应列506。在一个实施方式中，数据库可以是散列表或散列映射。

[0116] 转到图5c，公开了使用来自图5a的信息的数据库的示例。在这个实施例中，图5a中的摄取数据被表示为单元数据1。识别的关键概念502a-502d在可标记词列504中被标记为可标记词508a-508d，并且作为单元数据1的原始摄取数据也被标记在相应的列506中。当根据图3a的步骤302-310处理另外的摄取数据时，每个第x个摄取数据被表示为“单位数据x”。例如，第二摄取数据可以表示为“单元数据2”。如果“单元数据2”也包含可标记词“Linn Park”，则可将其添加到数据库500中的Linn Park那一行，并且“单元数据2”与“单元数据1”一起记录在相应的列506中。一旦含有可标记词的单位数据增长并达到预定的阈值，则将其作为事件检测群集发出。换句话说，这表明多个用户正在报导类似的事件，因此可能是一个刚出现的事件。

[0117] 转到图5d，示出了示例性的单位群集。在一个实施方式中，如果聚类模块128确定尚未存在一个既有的群集，则单位群集变为事件检测群集，或者如果只是基于预定规则存在一个既有的群集，则聚类模块128确定不与既有的群集进行合并。单位群集包括n个单位数据(例如，3个单位群集)的阈值数量n。

[0118] 图5e是另一示例性的推文形式的摄取数据。该摄取数据是来自与“穆加贝：是外国公司偷了钻石：津巴布韦总统罗伯特·穆加贝指责外国采矿公司……”有关的示例性事件检测群集的许多单位数据之一。这一摄取数据也被摘要模块132选为事件检测群集的代表性摘要。

[0119] 图5f-5k是图5e中摄取数据的示例性的元数据。该摄取数据包括由如图5f-5h和5k所示的社交媒体平台生成的默认元数据(即，twitter元数据)。事件检测服务器产生附加的元数据并被附加到上述的摄取数据的元数据上，并且在图5i和5j中示出。

[0120] 现在参考图5i，附加的元数据包括但不限于由可信度模块135确定的可信度分数535；由意见模块134确定的意见分数534；由过滤模块124确定的脏话指示符524和由组织模块126确定的可标记词526。

[0121] 图5l-5n是具有作为相关单元数据之一的图5e的摄取数据的事件检测群集的示例性元数据。

[0122] 在图5l-5m中，群集元数据包括但不限于由新闻价值模块133确定的新闻价值分数533；由主题分类模块131确定的主题531；由摘要模块132确定的摘要532以及由组织模块

126标识在单元数据中并被选择以形成事件检测群集的可标记词504a。每个可标记词504a还可以包括相应的单元数据506a信息。

[0123] 继续到图5n，群集元数据包括但不限于形成事件检测群集的单位数据506b以及由验证模块150计算的真实性分数550。

[0124] 现在转到图6a，公开了通过接入设备170的浏览器172可用的示例性的图形用户界面(GUI)。在一个实施方式中，浏览器172包括应用界面600，该应用界面包括用于查看属于频道602的事件检测群集的列表的多个列。每个频道内具有与该频道的话题有关的事件检测群集。

[0125] 在一个实施方式中，在图6a的实施例中，可能有用于“最新”的频道602a和用于“热门”的另一频道602b。然而，尽管在本实施例中在应用界面600中仅向用户呈现了两个频道，但是在应用界面600上可以显示n个频道。由应用界面600提供的默认频道允许用户不必按关键词进行搜索即可获得有关可能为新事件或热门事件的通知。

[0126] 在另一个实施方式中，继续到图6b，用户通过接入设备170的浏览器172可以在搜索框601中输入搜索词，以定制应用界面600来满足他们的需求。然后，事件处理服务器210的UI模块232将从生成群集数据存储244中检索与用户的搜索词相匹配的任何事件检测群集。该结果由UI模块232给出并通过浏览器172在程序界面600的频道602a下呈现给用户，该搜索词呈现于所述频道中。如图6b的示例所示，展现搜索词“GOP”的频道602和展现“Democrats”的频道602d可以显示出来以供浏览。

[0127] 在一个实施方式中，在事件检测群集的文本之前提供的指示604描绘事件检测群集中的单元数据的数量。在进一步的实施方式中，可存在额外的标识605，其基于面向专业消费者的主题(例如与危机、冲突(政治或地缘政治)或犯罪活动有关的主题)指示事件检测群集的重要性。

[0128] 在进一步的实施方式中，事件检测群集还可以呈现为具有由主题分类模块131确定的主题606；类别608，其可以是定制的术语；摘要模块132确定的摘要616。事件检测群集还可以包含概念610，该概念610是由组织模块126确定的来自于形成事件检测群集的单元数据中的可标记词。

[0129] 事件检测群集还可以呈现为具有在由组织模块126检测的摄取数据中使用的井字标签612，以及由新闻价值模块133确定的新闻价值标识618。在一个实施方式中，新闻价值指示618可以被刻画成一个实心星星。

[0130] 事件检测群集还可以呈现为具有由验证模块150确定的真实性分数614。在一个实施方式中，该真实性分数可以呈实心圆形的形式以指示真实性判断的强度，其中5个实心圆为接近准确。

[0131] 在又一个实施方式中，用户可以基于事件检测群集中的概念来选择创建新的频道620。新创建的频道系基于识别概念610。

[0132] 以关键的事件检测群集为例，该群集的选择如图6c所示。对应于所选的事件检测群集631呈现了一组单元数据632a-632n。在进一步的实施方式中，用户可以利用链接634来查看具体的单位数据。

[0133] 返回到图6b，在另一个实施方式中，频道选项622允许对由UI模块232呈现在接入设备170的浏览器172上的事件检测群集的结果进行过滤。UI模块232接收由用户在应用界

面600中选择的过滤指示，并根据图7a-7e所示的过滤器来处理该请求。

[0134] 在一个实施方式中，如图7a所示，过滤可以基于主题710、排序方法720、类别730和高级740过滤来进行。

[0135] 图7b示出了示例性的主题过滤器710。该主题过滤器710包含主题过滤器712a-712n的列表。它们可以是但不限于下列有关的主题：商业/金融、危机、娱乐、重大新闻、健康/医疗、法律/犯罪、生活/社会、政治、体育、技术、天气或其他主题分类模块131识别的内容。

[0136] 图7c示出了示例性的分类过滤器720。该分类过滤器720包含选项722a-722n，并且它们可以是但不限于按照下列内容进行分类：最新的、更新的、最流行的、热门的、有新闻价值的和准确的。

[0137] 图7d示出了示例性的类别过滤器730。该类别过滤器730包含类别过滤器732a-732n的列表。类别选项可以是但不限于：突发新闻、冲突、灾难、道琼斯、金融风险、地缘政治风险、法律、法律风险、市场、石油、政治、枪击、美国选举。

[0138] 图7e是在应用界面600上选择高级740时的高级选项。在一个实施方式中，所选频道的高级选项可以是重置默认值744、具有时间范围选择的时间线746、最少帖数748的计数和对应于事实750、新闻价值752和真实性754的三个层次，即，严格760、中等762或宽松764。

[0139] 图1至图7e是允许对本公开进行解释的概念性图例。系统的各种特征可以用硬件、软件或硬件和软件的组合来实现。例如，系统的一些特征可以在可编程计算机上执行的一个或多个计算机程序中实现。每个程序可以以高阶程序语言或面向对象的编程语言来实现，从而与计算机系统或其他机器进行通信。此外，每个这样的计算机程序可以存储在诸如通用或专用可编程计算机或处理器可读的只读存储器(ROM)之类的存储介质上，以用于配置和操作计算机以执行上面所描述的功能。

[0140] 值得注意的是，上面的附图和示例并不意味着将本公开的范围限制为单个的实施方式，因为其他实施方式可以通过部分或全部所描述或示出的要素的互换来实现。此外，在本公开的某些要素可以使用已知组件部分地或完全地实现的情况下，仅描述对于理解本公开内容是必需的这些已知组件的那些部分，并且省略了这些已知组件的其他部分的详细描述以免对本公开产生混淆。在本说明书中，除非在本文中明确地另行指出，否则示出单个组件的实施方式不应必然限于包括多个相同组件的其他实施方式，反之亦然。此外，申请人不打算将说明书或权利要求中的任何术语归于不常见的或特殊的含义，除非明确阐述如此。

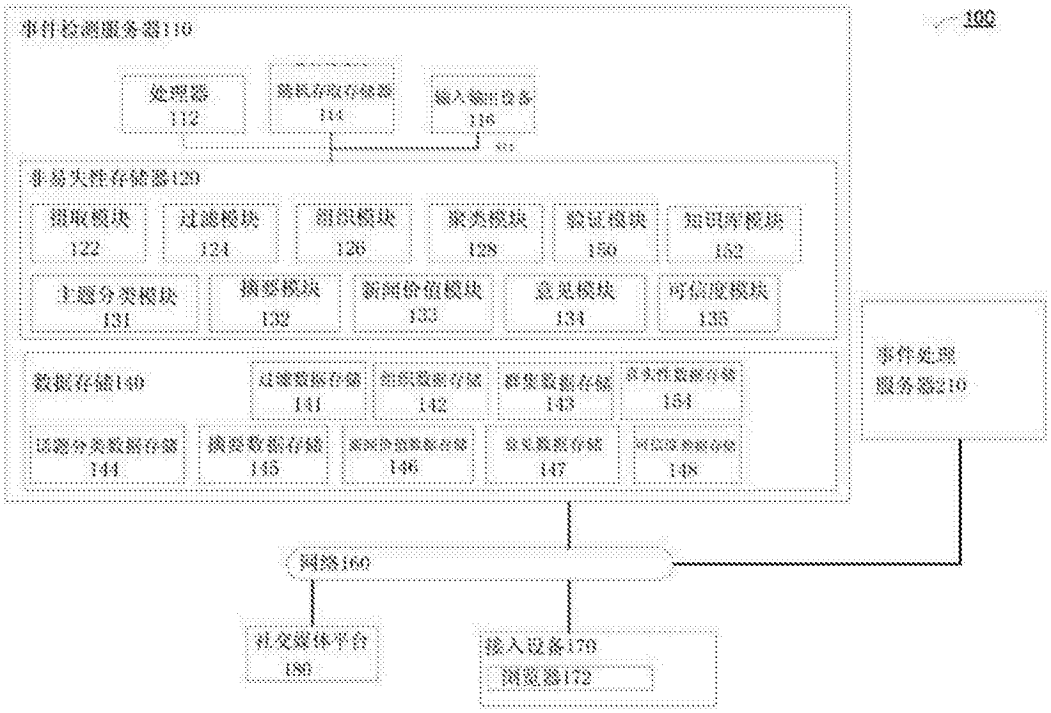

100
事件检测服务器110
处理器
112
输入输出设备
116
事件处理
服务器210
网络160
社交媒体平台
180
输入设备170
浏览器172

图1

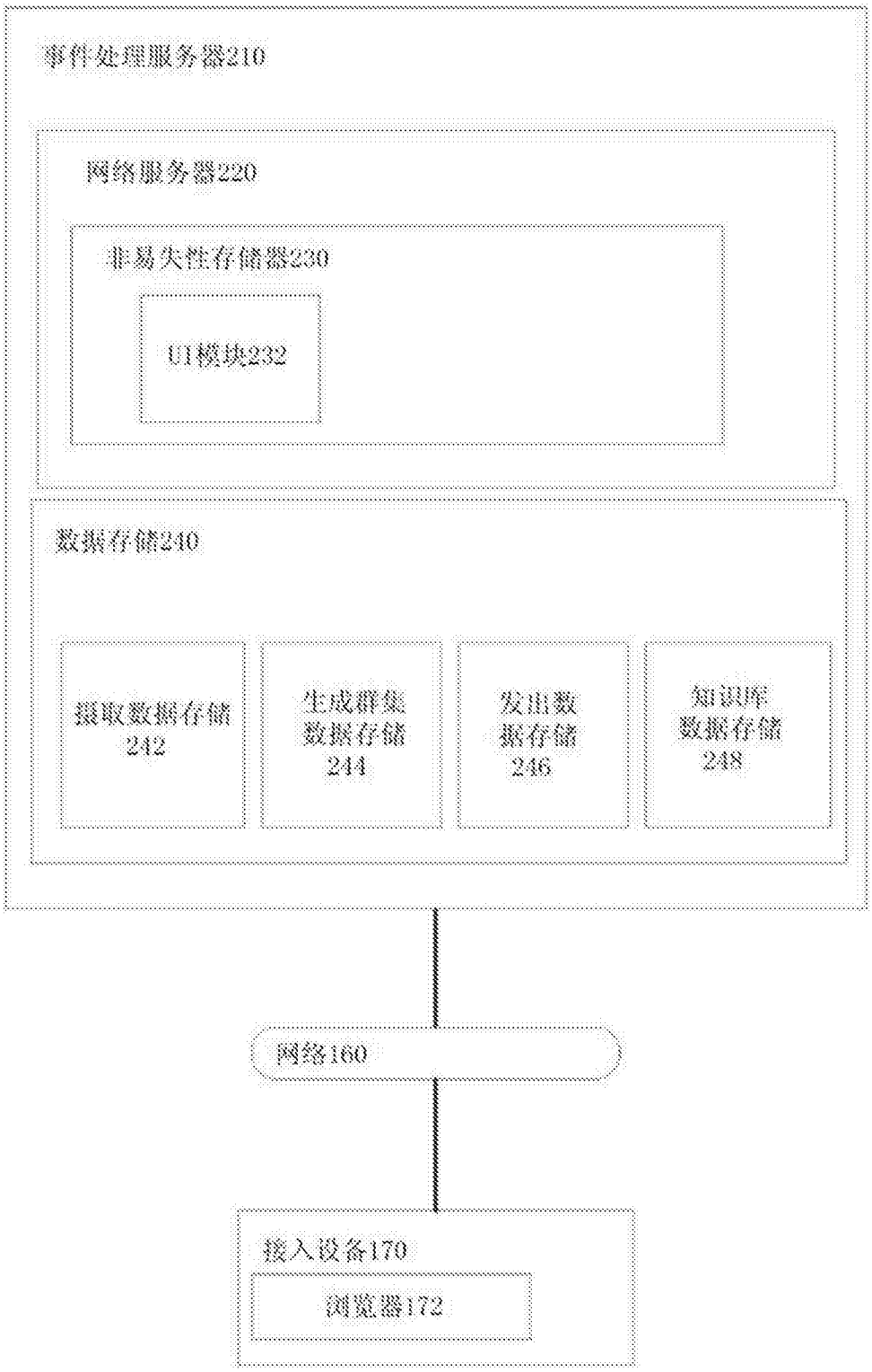

事件处理服务器210
网络服务器220
非易失性存储器230
UI模块232
数据存储240
摄取数据存储242
生成群集数据存储244
发出数据存储246
知识库数据存储248
网络160
接入设备170
浏览器172

图2

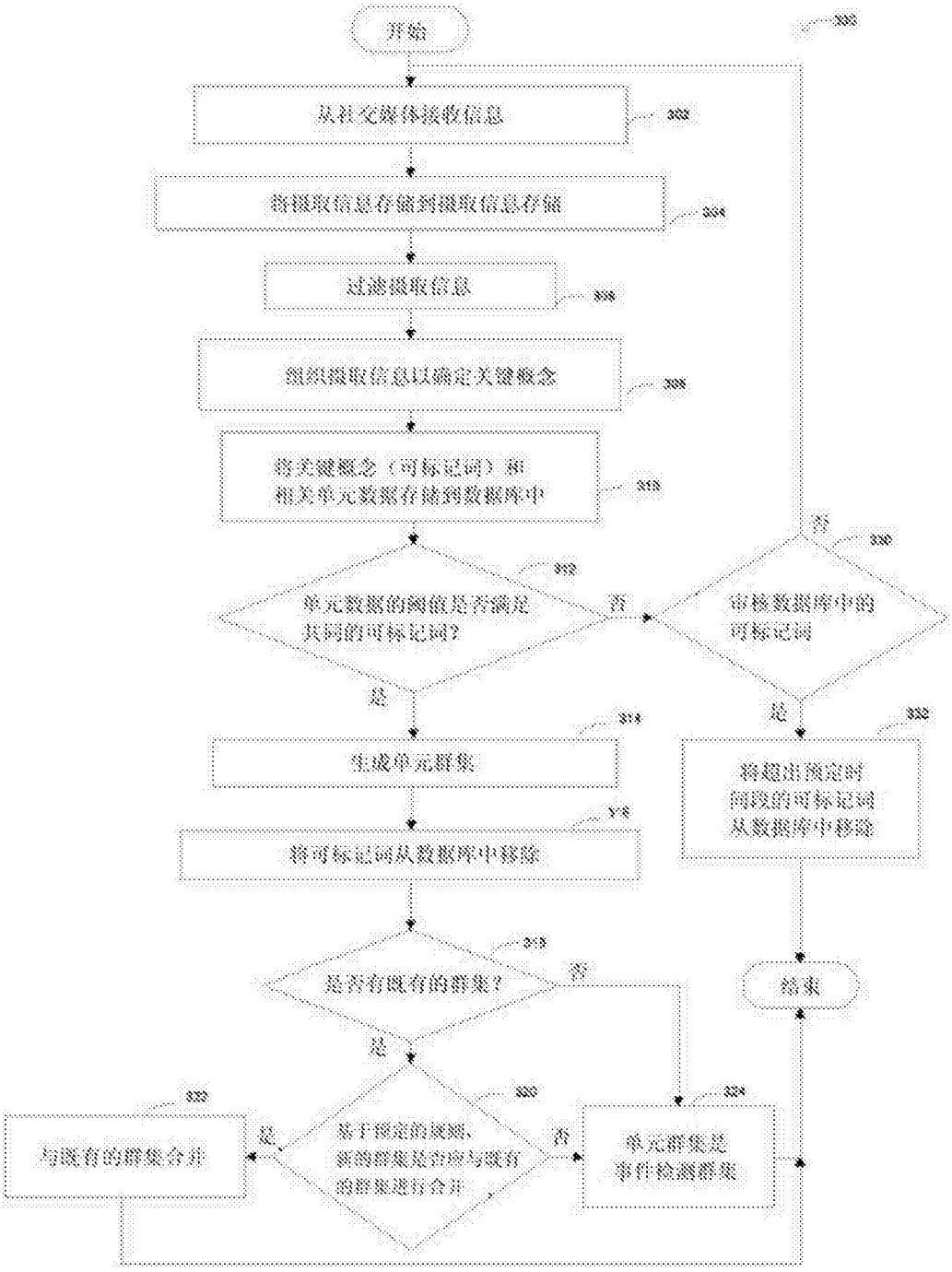

开始
300
从社交媒体接收信息
302
将提取信息存储到提取信息存储
304
过滤提取信息
306
组织提取信息以确定关键概念
308
将关键概念（可标记词）和
相关单元数据存储到数据库中
310
单元数据的阈值是否满足
共同的可标记词?
312
否
是
审核数据库中的
可标记词
320
是
否
生成单元群集
314
将超出预定时
间段的可标记词
从数据库中移除
322
将可标记词从数据库中移除
316
是否有既有的群集?
318
否
是
结束
与既有的群集合并
322
基于预定的规则，
新的群集是否应与既有
的群集进行合并
320
是
否
单元群集是
事件检测群集
324

图3a

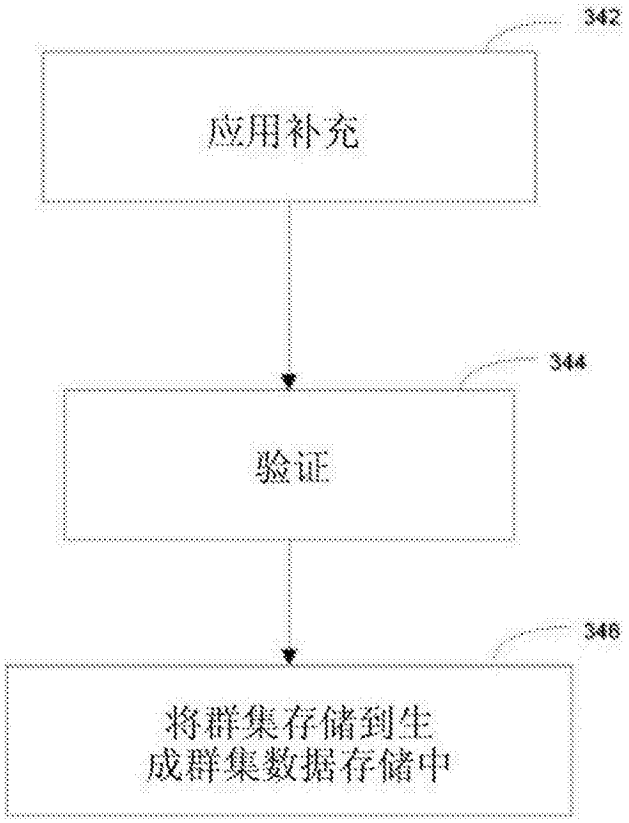


图3b

表1：特征描述

图4a

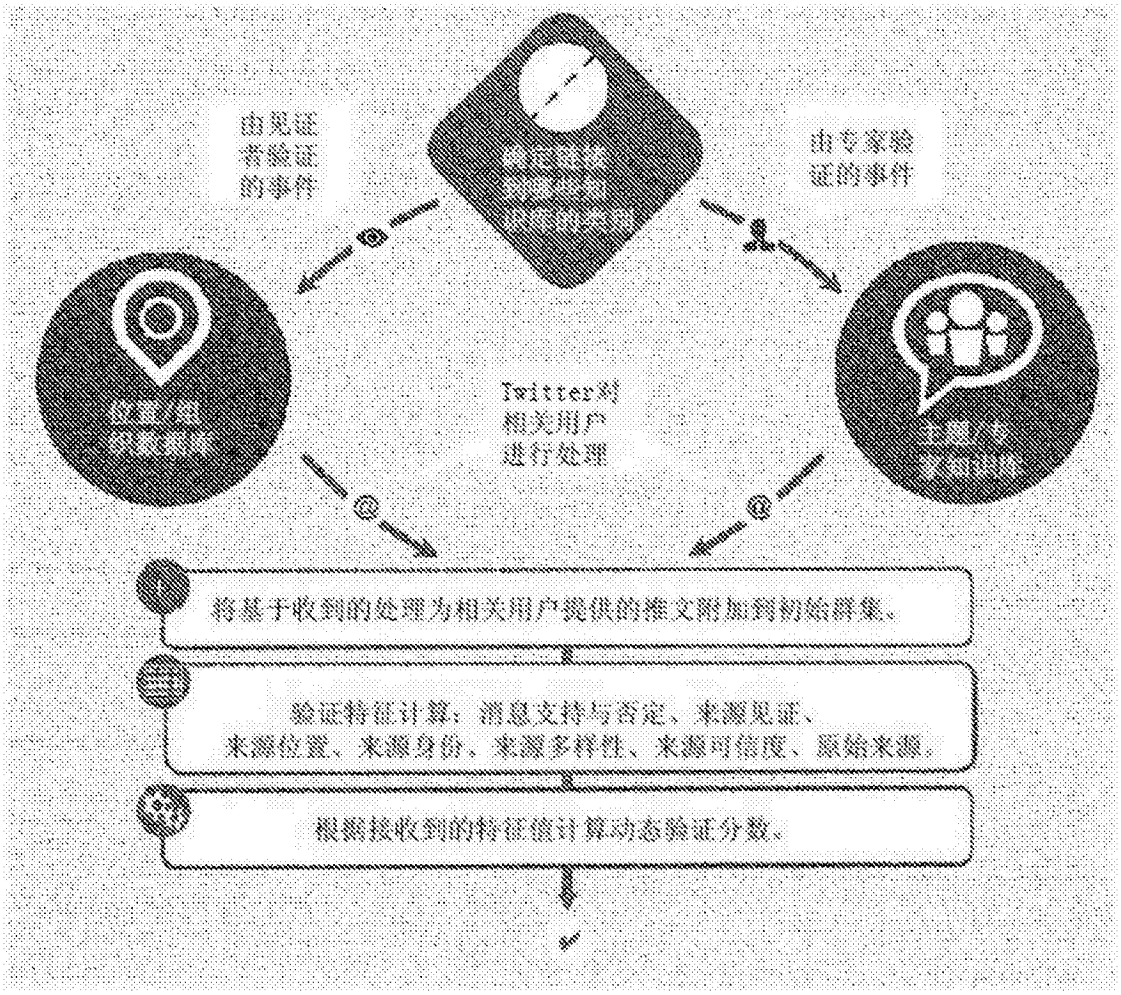

由见证者验证的事件
由专家验证的事件
Twitter对相关用户进行处理
将基于收到的处理为相关用户提供的推文附加到初始群集。
验证特征计算：消息支持与否定、来源见证、来源位置、来源身份、来源多样性、来源可信度、原始来源。
根据接收到的特征值计算动态验证分数。

图4b

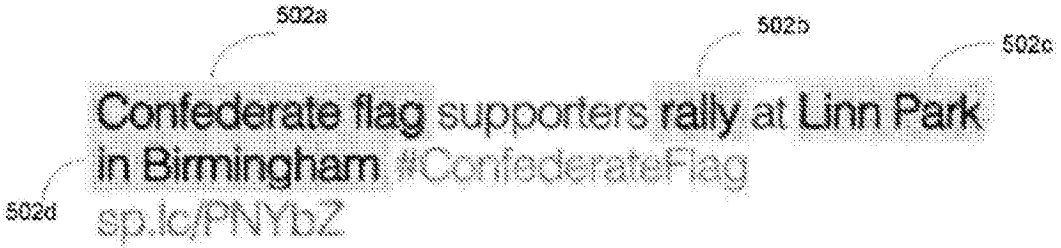

502a
502b
502c
502d
Confederate flag supporters rally at Linn Park in Birmingham #ConfederateFlag
sp.lc/PNYbZ

图5a

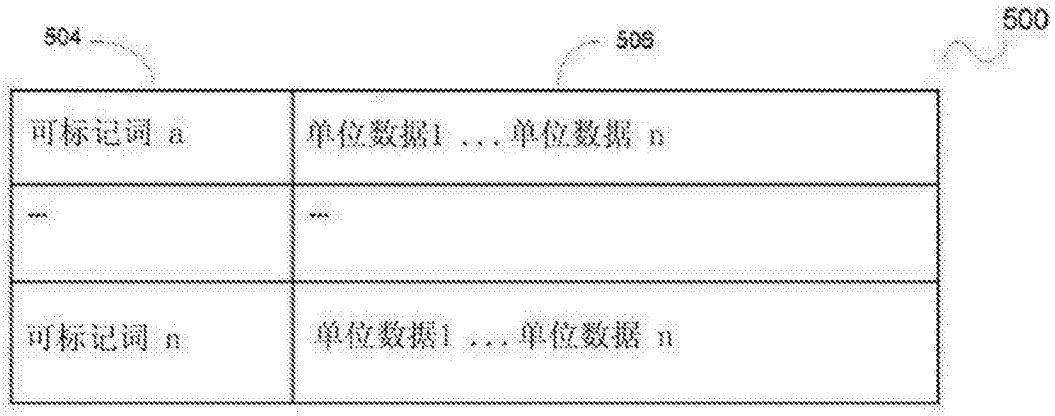

504
506
500
可标记词 a
单位数据1 ...单位数据 n
...
...
可标记词 n
单位数据1 ...单位数据 n

图5b

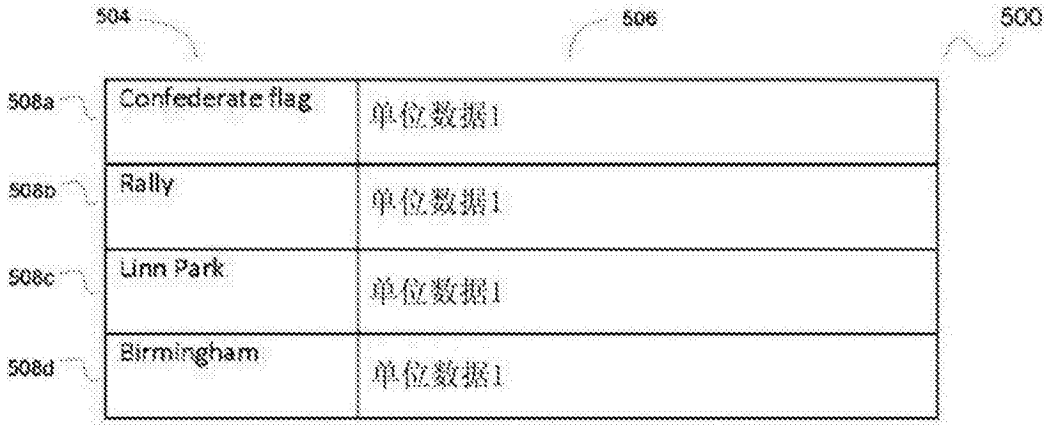

504
506
500
508a
508b
508c
508d
Confederate flag
单位数据1
Rally
单位数据1
Linn Park
单位数据1
Birmingham
单位数据1

图5c

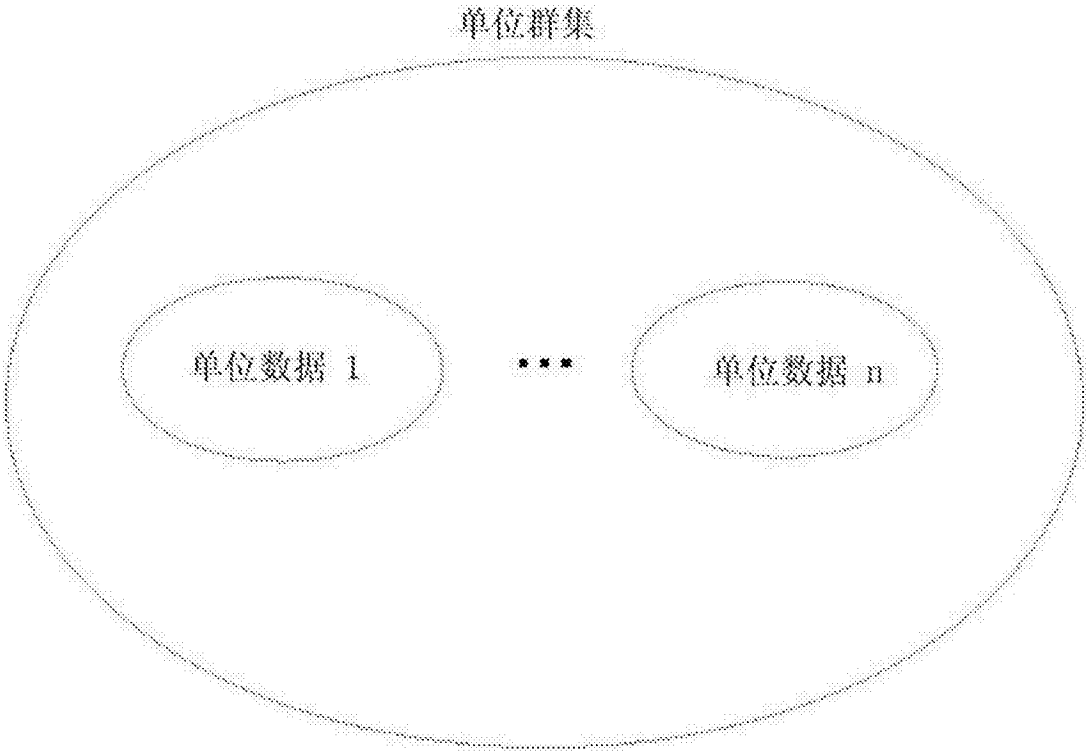

单位群集
单位数据 1
单位数据 n

图5d

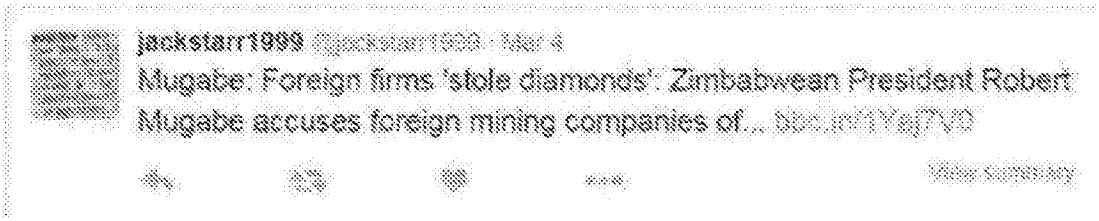

jackstarr1999
Mugabe: Foreign firms 'stole diamonds'. Zimbabwean President Robert Mugabe accuses foreign mining companies of...

图5e

```
{
"_index": "blip2016.03.04",
"_type": "twitter",
"_id": "72ea1b70e23411e5a6f5fa163ee42fcd",
"_score": null,
"_source": {
"user_handle": "jackstarr1999",
"friends_count": 9976,
"source": "<a href=\"http://twitterfeed.com\" rel=\"nofollow\">twitterfeed</a>",
"favorite_count": 0,
"content_type": "Twitter",
"urls": [
{
"expanded_url": "http://bbc.in/1Yaj7V0",
"indices": [
"first": 114,
"second": 137
],
"url": "https://t.co/ZIsuqo6Nn",
"domain": "bbc.in",
"display_url": "bbc.in/1Yaj7V0",
"is_short_url": false
}
],
"followers_count": 13262,
"is_retweet": false,
"text": "Mugabe: Foreign firms 'stole diamonds': Zimbabwean President Robert
Mugabe
accuses foreign mining companies of... https://t.co/ZIsuqo6Nn",
"retweet_count": 0,
"user_id": "268036007",
"id": "72ea1b70e23411e5a6f5fa163ee42fcd",
"language": "en",
"raw": {
"coordinates": null,
"retweeted": false,
"source": "<a href=\"http://twitterfeed.com\" rel=\"nofollow\">twitterfeed</a>",
"entities": {
"hashtags": [],
"urls": [
{
"url": "https://t.co/ZIsuqo6Nn",
```


图5f

```
"expanded_url": "http://bbc.in/1Yaj7V0",
"display_url": "bbc.in/1Yaj7V0",
"indices": [
114,
137
]
}
],
"user_mentions": [],
"symbols": []
},
"favorite_count": 0,
"in_reply_to_status_id_str": null,
"geo": null,
"id_str": "705817833658142720",
"in_reply_to_user_id": null,
"timestamp_ms": "1457115059662",
"truncated": false,
"text": "Mugabe: Foreign firms 'stole diamonds': Zimbabwean President Robert Mugabe
accuses foreign mining companies of... https://t.co/ZisuqoGNn",
"retweet_count": 0,
"id": 705817833658142700,
"in_reply_to_status_id": null,
"possibly_sensitive": false,
"filter_level": "low",
"created_at": "Fri Mar 04 18:10:59 +0000 2016",
"place": null,
"favorited": false,
"lang": "en",
"contributors": null,
"in_reply_to_screen_name": null,
"is_quote_status": false,
"in_reply_to_user_id_str": null,
"user": {
"utc_offset": 28800,
"name": "jackstarr1999",
"friends_count": 9976,
"screen_name": "jackstarr1999",
"location": "Looking Ahead Maryland",
"protected": false,
"url": null,
```


图5g

"profile_image_url":
"http://pbs.twimg.com/profile_images/3419216884/5fb506f5bfea5ed0b99c2cfa24849c81_normal
.jpeg",
"profile_background_color": "C0DEED",
"profile_use_background_image": true,
"is_translator": false,
"geo_enabled": false,
"description": "buzzing about all kinds of stuff do
say hello trying
to stay positive here",
"profile_link_color": "9EB300",
"id_str": "268036007",
"listed_count": 107,
"default_profile_image": false,
"followers_count": 13262,
"profile_image_url_https":
"https://pbs.twimg.com/profile_images/3419216884/5fb506f5bfea5ed0b99c2cfa24849c81_norm
al.jpeg",
"profile_sidebar_border_color": "EBC1C1",
"profile_background_image_url":
"http://pbs.twimg.com/profile_background_images/21916995 2/2.jpg",
"favourites_count": 494,
"following": null,
"default_profile": false,
"id": 268036007,
"profile_background_tile": true,
"contributors_enabled": false,
"follow_request_sent": null,
"created_at": "Fri Mar 18 01:18:56 +0000 2011",
"profile_sidebar_fill_color": "DDEEF6",
"lang": "en",
"profile_text_color": "333333",
"notifications": null,
"verified": false,
"time_zone": "Tijuana",
"profile_banner_url": "https://pbs.twimg.com/profile_banners/268036007/1363413391",
"statuses_count": 94782,
"profile_background_image_url_https":
"https://pbs.twimg.com/profile_background_images/21916995 2/2.jpg"
}
},
"date": "20160304T18:
10:59.000Z",
"possibly_sensitive": false,
"source_id": "705817833658142720",

图5h

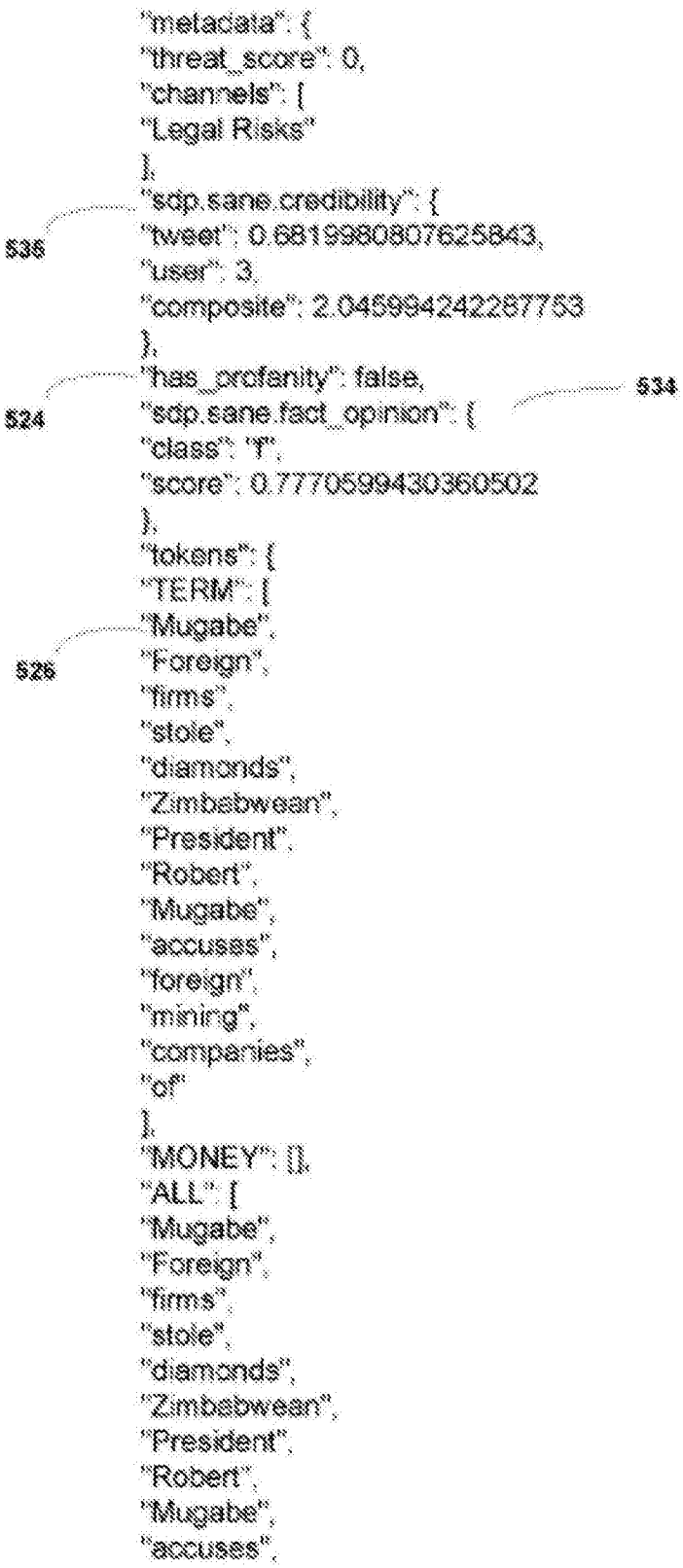

"metadata": {
"threat_score": 0,
"channels": [
"Legal Risks"
],
"sdp.sane.credibility": {
536
"tweet": 0.6819980807625843,
"user": 3,
"composite": 2.045994242267753
},
"has_profanity": false,
534
524
"sdp.sane.fact_opinion": {
"class": "f",
"score": 0.7770599430360502
},
"tokens": {
"TERM": [
526
"Mugabe",
"Foreign",
"firms",
"stole",
"diamonds",
"Zimbabwean",
"President",
"Robert",
"Mugabe",
"accusss",
"foreign",
"mining",
"companies",
"of"
],
"MONEY": [],
"ALL": [
"Mugabe",
"Foreign",
"firms",
"stole",
"diamonds",
"Zimbabwean",
"President",
"Robert",
"Mugabe",
"accuses",

图5i

"foreign",
"mining",
"companies",
"of",
",",
"""",
"""",
",",
"…",
"https://t.co/Zlsuqo6Nri",
",",
"""",
"""",
",",
"…"
],
"ENTITY": [
"https://t.co/Zlsuqo6Nri"
],
"JUNK": [
",",
"""",
"""",
",",
"…"
],
"EMOTICON": [],
"PUNC": [
",",
"""",
"""",
",",
"…"
],
"EMOJI": [],
"CRISIS": []
},
"origins": [
"onepctpublic",
"onepctfilter"
],
"sdp.sane.spam_category": "news
like"
},

图5j

```
"user": {
"name": "jackstarr1999",
"friends_count": 9976,
"screen_name": "jackstarr1999",
"location": "Looking Ahead Maryland",
"description": "buzzing about all kinds of stuff do
say hello trying
to stay positive here",
"listed_count": 107,
"user_avatar":
"http://pbs.twimg.com/profile_images/3419216884/
5fb506f5bfea5ed0b99c2cfa24849c81_normal
.jpeg",
"followers_count": 13262,
"is_verified": false,
"favourites_count": 494,
"is_default_profile": false,
"has_profile_image": true,
"id": "26803607",
"language": "en",
"created_at": "20110318T01:
18:56.000Z",
"statuses_count": 94782,
"is_protected": false
}
},
"fields": {
"raw.user.created_at": [
1300411136000
],
"raw.timestamp_ms": [
1457115059662
],
"date": [
1457115059000
],
"user.created_at": [
1300411136000
],
"raw.created_at": [
1457115059000
]
},
"sort": [
1457115059000
]
}
```


图5k

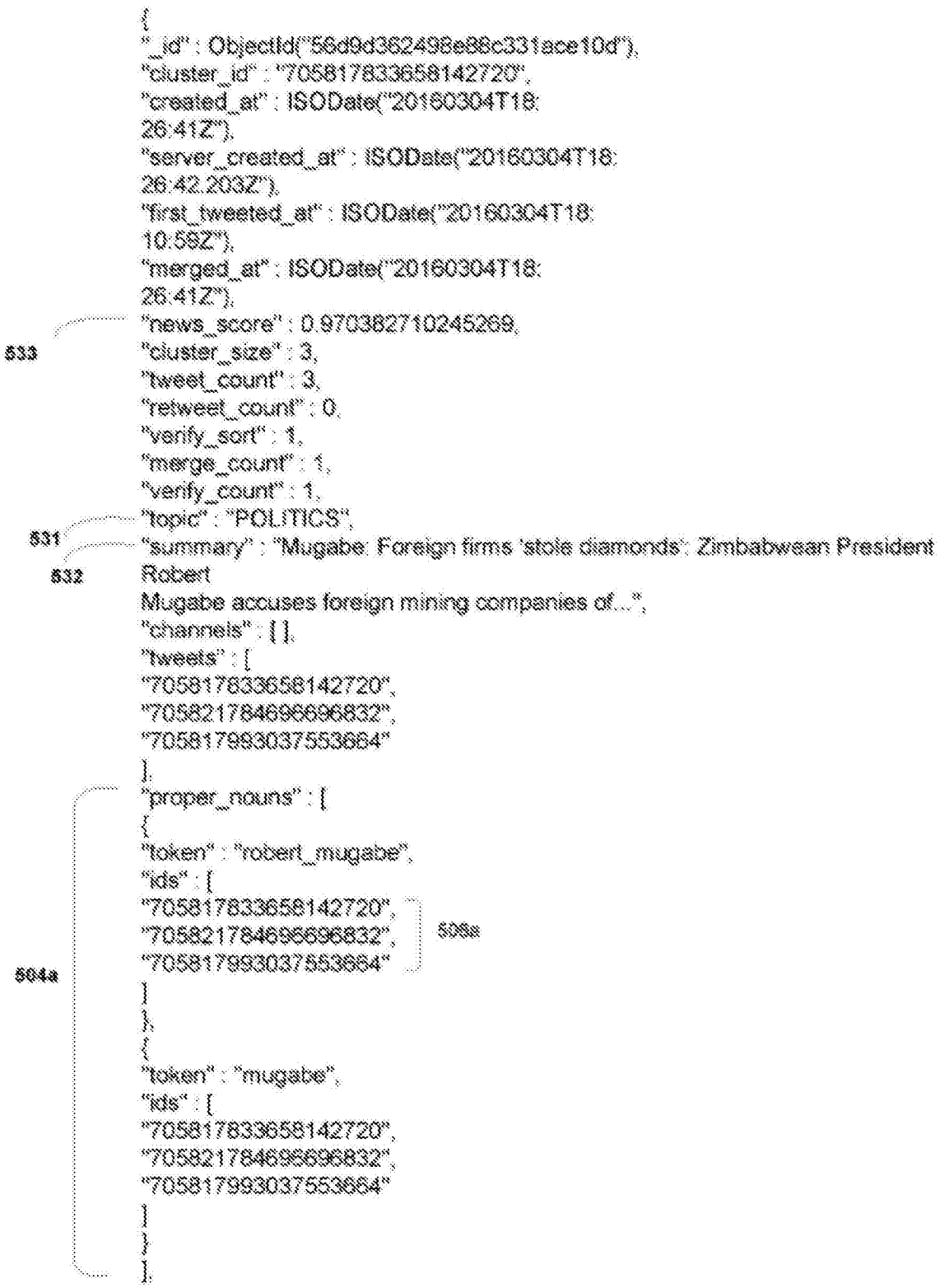

{
"_id" : ObjectId("56d9d362498e88c331ace10d"),
"cluster_id" : "7058178336581427 20",
"created_at" : ISODate("20160304T18:
26:41Z"),
"server_created_at" : ISODate("20160304T18:
26:42.203Z"),
"first_tweeted_at" : ISODate("20160304T18:
10:59Z"),
"merged_at" : ISODate("20160304T18:
26:41Z"),
"news_score" : 0.970382710245269,
533
"cluster_size" : 3,
"tweet_count" : 3,
"retweet_count" : 0,
"verify_sort" : 1,
"merge_count" : 1,
"verify_count" : 1,
"topic" : "POLITICS",
531
"summary" : "Mugabe: Foreign firms 'stole diamonds': Zimbabwean President
532
Robert
Mugabe accuses foreign mining companies of...",
"channels" : [ ],
"tweets" : [
"705817833658142720",
"705821784696696832",
"705817993037553664"
],
"proper_nouns" : [
{
"token" : "robert_mugabe",
"ids" : [
"705817833658142720",
"705821784696696832",
504a
"705817993037553664"
]
},
{
"token" : "mugabe",
"ids" : [
"705817833658142720",
"705821784696696832",
"705817993037553664"
]
}
],
508a

图51

```
"hashtags" : [
{
"token" : "news",
"ids" : [
"705821784696696832"
]
}
],
"ticker_symbols" : [ ],
"additional_tokens" : [
{
"token" : "companies",
"tag_type" : "COMMON_NOUN",
"ids" : [
"7058178336658142720",
"7058179930375536640"
]
},
{
"token" : "mining",
"tag_type" : "COMMON_NOUN",
"ids" : [
"7058178336658142720",
"705821784696696832",
"7058179930375536640"
]
},
{
"token" : "president",
"tag_type" : "COMMON_NOUN",
"ids" : [
"7058178336658142720",
"705821784696696832",
"7058179930375536640"
]
},
{
"token" : "diamonds",
"tag_type" : "COMMON_NOUN",
"ids" : [
"7058178336658142720",
"705821784696696832",
"7058179930375536640"
]
},
{
"token" : "firms",
"tag_type" : "COMMON_NOUN",
"ids" : [
"7058178336658142720",
"705821784696696832",
"7058179930375536640"
]
}
],
```


图5m


"urls" : [
"http://bit.ly/bbcNews",
"http://bbc.in/1Ya7V0"
],
"unit_clusters" : [
{
"tweets" : [
"705817833658142720",
"705817993037553664",
"705821784696696832"
],
"merged_at" :
ISODate("20160304T18:
26:41Z")
}
],
"locations" : [ ],
"merge_history" : [
{
"date" : ISODate("20160304T18:
26:41Z"),
"merge_id" : 1
}
],
"retweet_hierarchy" : [
{
"parent" : "705817833658142720",
"size" : 0,
"children" : [ ]
},
{
"parent" : "705821784696696832",
"size" : 0,
"children" : [ ]
},
{
"parent" : "705817993037553664",
"size" : 0,
"children" : [ ]
}
],
"verify_history" : [
{
"date" : ISODate("20160304T18:
26:41Z"),
"score" : 1
}
],
"verifiability" : true
}
506b
550

图5n

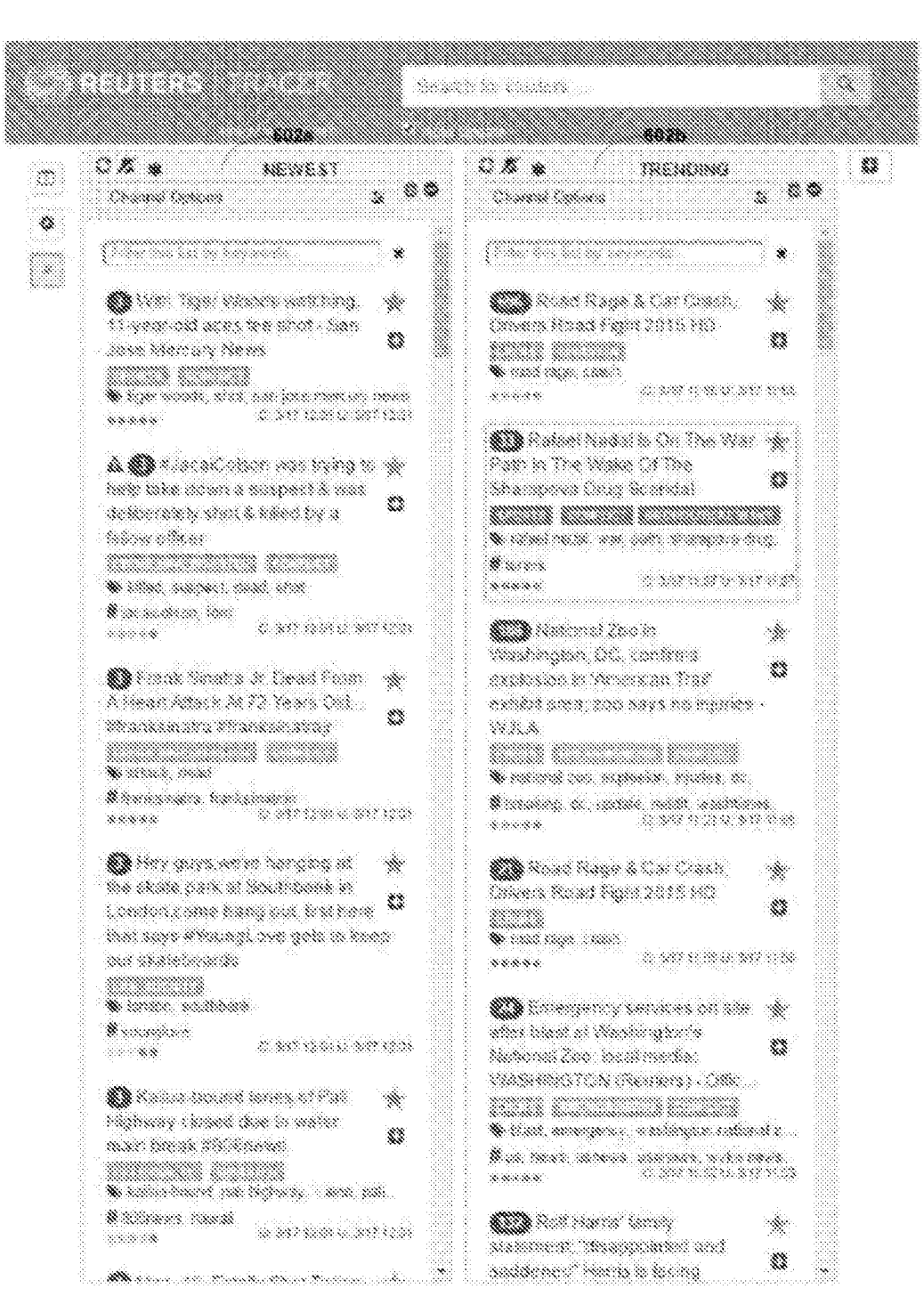


图6a

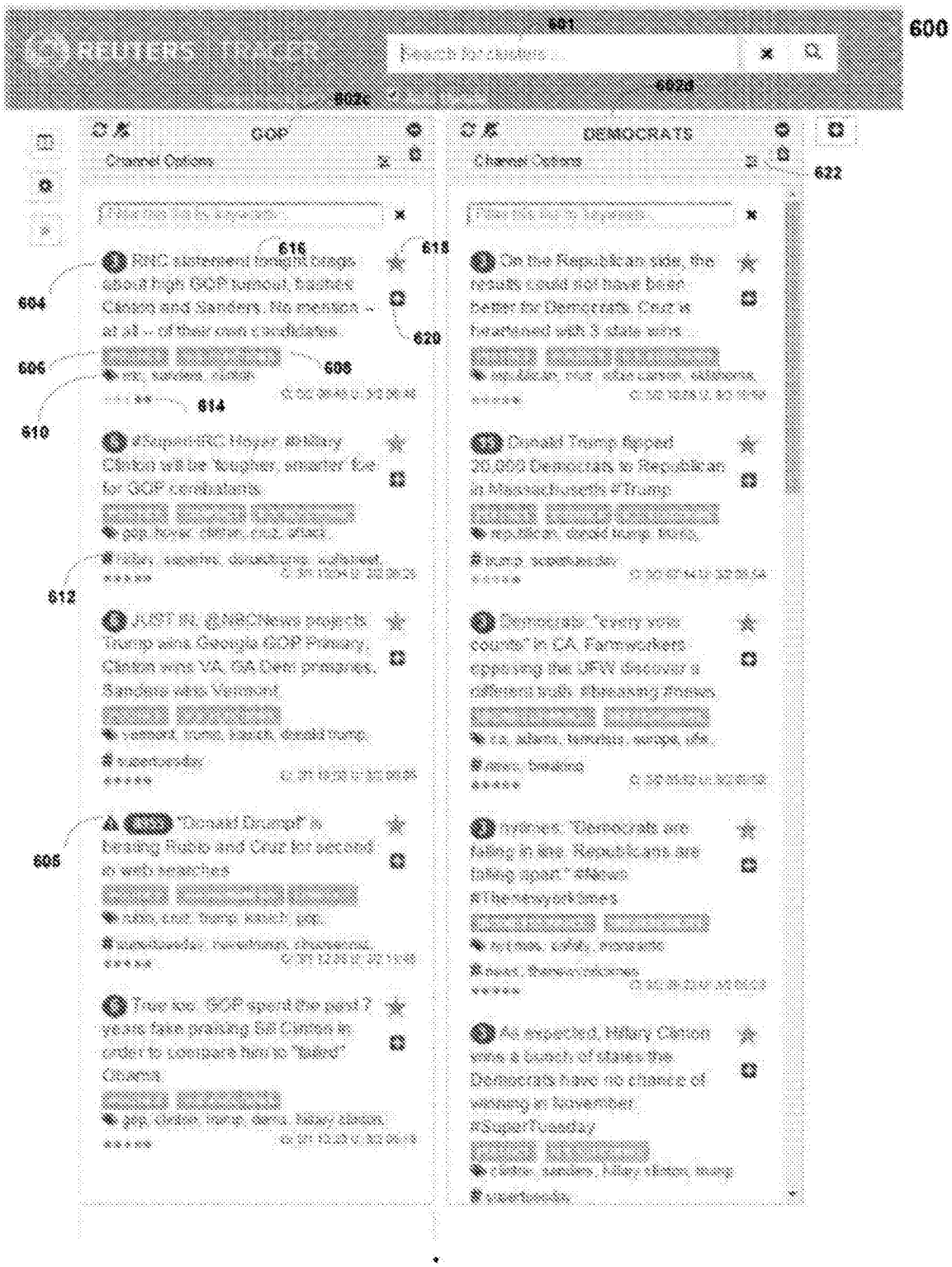

600
601
602a
602b
602c
GOP
DEMOCRATS
Channel Options
622
604
606
608
610
612
614
615
618
620
605

图6b

图6c

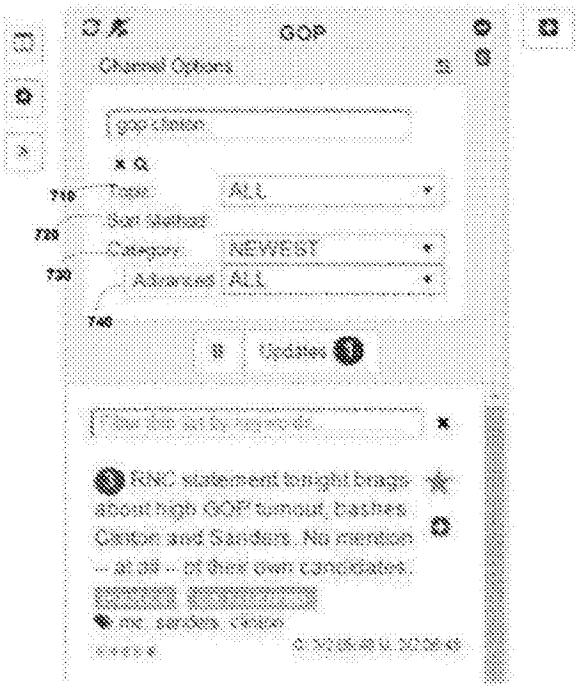

GOP
Channel Options
Topic
ALL
Sort Method
Category
NEWEST
Advanced
ALL
Update

图7a

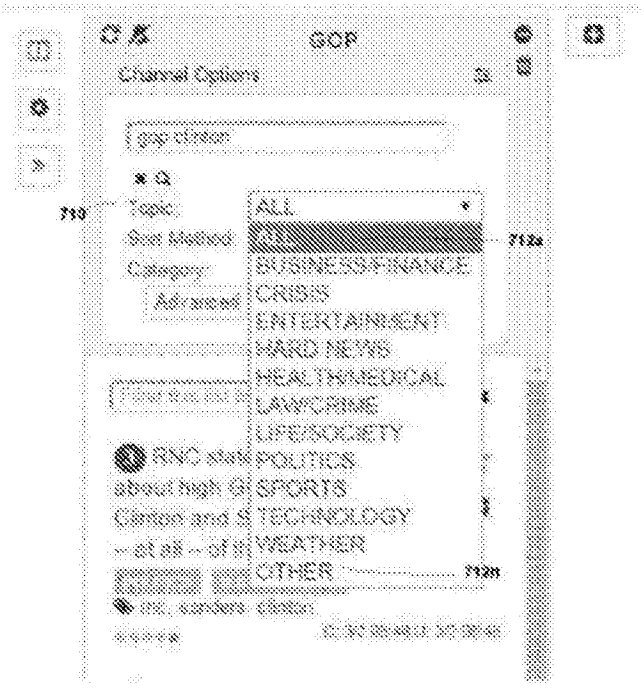

GOP
Channel Options
Topic
ALL
Sort Method
Category
Advanced
BUSINESS/FINANCE
CRISIS
ENTERTAINMENT
HARD NEWS
HEALTH/MEDICAL
LAW/CRIME
LIFE/SOCIETY
POLITICS
SPORTS
TECHNOLOGY
WEATHER
OTHER

图7b

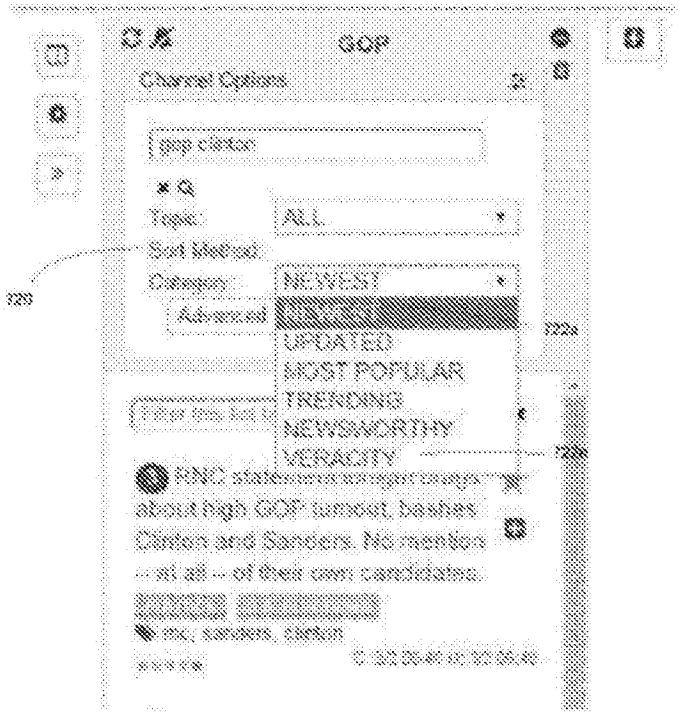

GOP
Channel Options
gop clinton
Topic
ALL
Sort Method
Category
NEWEST
Advanced
UPDATED
MOST POPULAR
TRENDING
NEWSWORTHY
VERACITY
about high GOP turnout, bashes
Clinton and Sanders. No mention
-- at all -- of their own candidates.

图7c

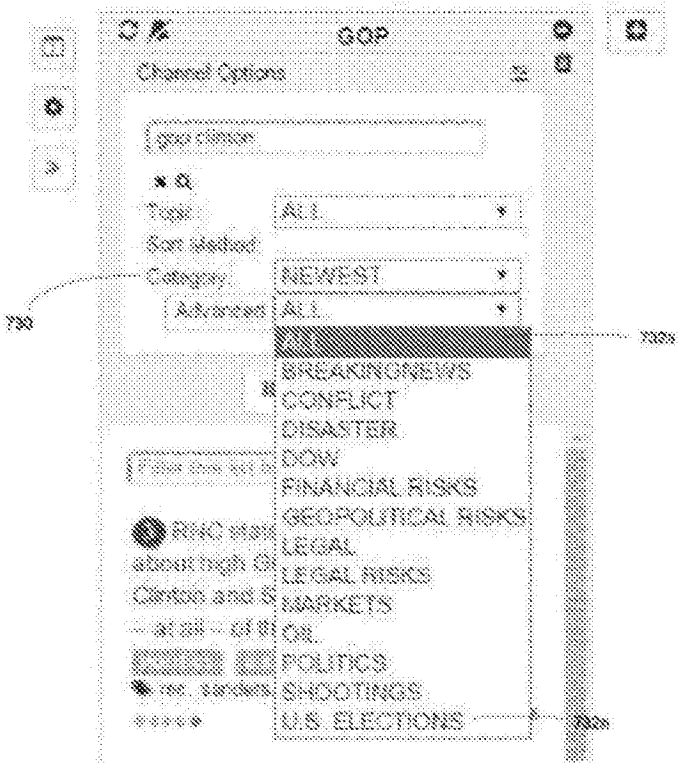

GOP
Channel Options
gop clinton
Topic
ALL
Sort Method
Category
NEWEST
Advanced
ALL
BREAKINGNEWS
CONFLICT
DISASTER
DOW
FINANCIAL RISKS
GEOPOLITICAL RISKS
LEGAL
LEGAL RISKS
MARKETS
OIL
POLITICS
SHOOTINGS
U.S. ELECTIONS

图7d

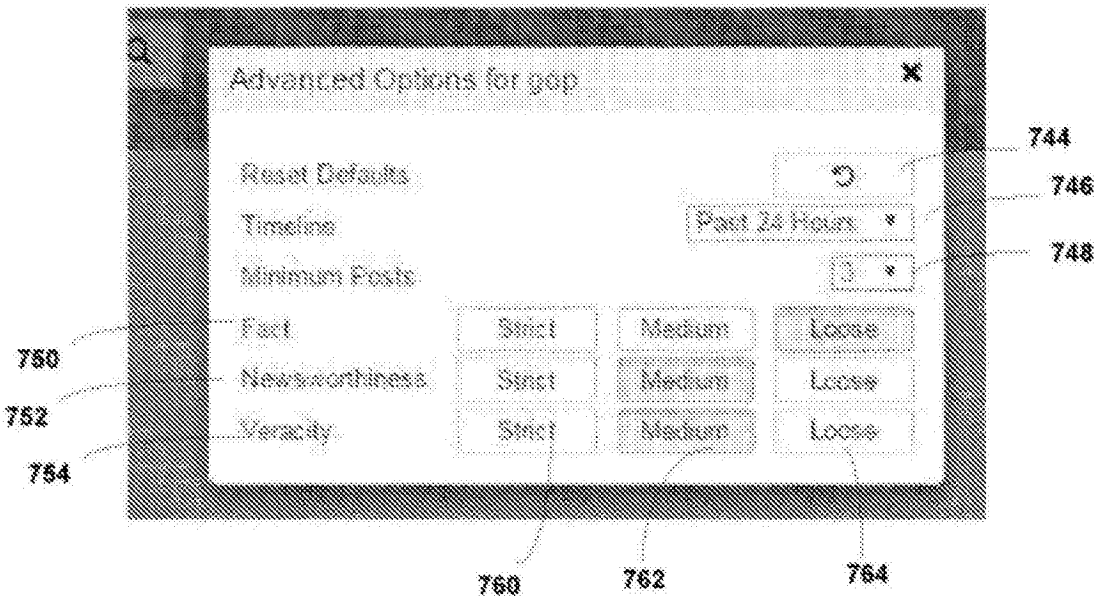

Advanced Options for gap
Reset Defaults
Timeline
Past 24 Hours
Minimum Posts
3
Fact
Newsworthiness
Veracity
Strict
Medium
Loose
744
746
748
750
752
754
760
762
764

图7e