

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局



(43) 国际公布日
2018 年 10 月 11 日 (11.10.2018)

WIPO | PCT

(10) 国际公布号
WO 2018/184306 A1

(51) 国际专利分类号:
G06F 17/27 (2006.01)

(21) 国际申请号: PCT/CN2017/090579

(22) 国际申请日: 2017 年 6 月 28 日 (28.06.2017)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201710225685.3 2017 年 4 月 7 日 (07.04.2017) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY(SHENZHEN)CO.,LTD.) [CN/CN];
中国广东省深圳市福田区八卦岭八卦三路平安大厦, Guangdong 518000 (CN)。

(72) 发明人: 王健宗(WANG, Jianzong); 中国广东省深圳市福田区八卦岭八卦三路平安大厦, Guangdong 518000 (CN)。 黄章成(HUANG, Zhangcheng); 中国广东省深圳市福田区八卦岭八卦三路平安大厦, Guangdong 518000 (CN)。 吴天博(WU, Tianbo); 中国广东省深圳市福田区八卦岭八卦三路平

安大厦, Guangdong 518000 (CN)。 肖京(XIAO, Jing); 中国广东省深圳市福田区八卦岭八卦三路平安大厦, Guangdong 518000 (CN)。

(74) 代理人: 广州华进联合专利商标代理有限公司(ADVANCE CHINA IP LAW OFFICE); 中国广东省广州市天河区花城大道 85 号 3901 房, Guangdong 510623 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(54) Title: TOPIC ALARM METHOD, DEVICE, COMPUTER APPARATUS, AND STORAGE MEDIUM

(54) 发明名称: 话题预警的方法、装置、计算机设备及存储介质

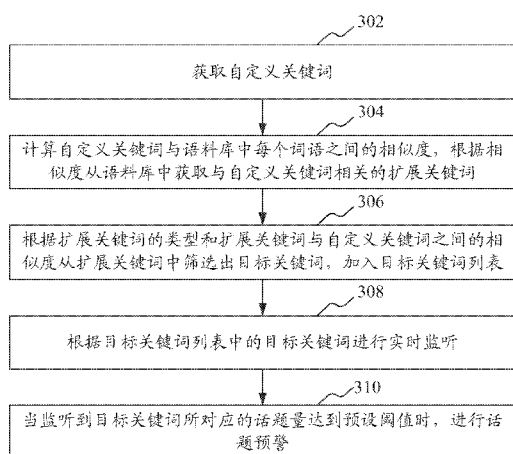


图 3

- 302 Acquire a customized keyword
304 Compute similarity levels of the customized keyword and words in a corpus, and acquire, according to the similarity levels and from the corpus, expansion keywords related to the customized keyword
306 Filter and select, according to the type of the expansion keywords and the similarity levels between the expansion keywords and the customized keyword, a target keyword from the expansion keywords, and add the same to a target keyword list
308 Perform, according to the target keyword in the target keyword list, real-time monitoring
310 When a topic number corresponding to the target keyword reaches a predetermined threshold, provide a topic alarm

(57) Abstract: A topic alarm method, comprising: acquiring a customized keyword; computing similarity levels of the customized keyword and words in a corpus, and acquiring, according to the similarity levels and from the corpus, expansion keywords related to the customized keyword; filtering and selecting, according to the type of the expansion keywords and the similarity levels between the expansion keywords and the customized keyword, a target keyword from the expansion keywords, and adding the same to a target keyword list; performing, according to the target keyword in the target keyword list, real-time monitoring; and when a topic number corresponding to the target keyword reaches a predetermined threshold, providing a topic alarm.

(57) 摘要: 一种话题预警的方法, 包括: 获取自定义关键词, 计算所述自定义关键词与语料库中每个词语之间的相似度, 根据所述相似度从语料库中获取与所述自定义关键词相关的扩展关键词, 根据所述扩展关键词的类型和所述扩展关键词与所述自定义关键词之间的相似度从所述扩展关键词中筛选出目标关键词, 加入目标关键词列表, 根据所述目标关键词列表中的目标关键词进行实时监听, 当监听到目标关键词所对应的话题量达到预设阈值时, 进行话题预警。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

说明书

发明名称：话题预警的方法、装置、计算机设备及存储介质

本申请要求于 2017 年 4 月 7 日提交中国专利局、申请号为 2017102256853、发明名称为“话题预警的方法和装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5 技术领域

本发明涉及计算机处理领域，特别是涉及一种话题预警的方法、装置、计算机设备及存储介质。

背景技术

10 随着社交媒体的发展，社交网站、在线社区、微博等已逐渐成为人们生活中不可或缺的一部分，也是当今时代信息传播的主要渠道，与此同时，社交媒体也是舆情传播的重要途径。通过对社交媒体的话题监听预警，能够为决策者提供科学化的信息支持。传统的对社交媒体话题监听预警是通过对获取到的历史数据进行分析，然后针对不同的话题进行标签分级。由于话题更
15 新速度非常快，仅仅针对历史数据进行分析得出的结果显然不够准确，且传统的话题监听是针对所有的话题进行监听，没有考虑到用户的个性化需求。

发明内容

20 根据本申请的各种实施例，提供一种话题预警的方法、装置、计算机设备及存储介质。

一种话题预警的方法，包括：

获取自定义关键词；

计算所述自定义关键词与语料库中每个词语之间的相似度，根据所述相

似度从语料库中获取与所述自定义关键词相关的扩展关键词;

根据所述扩展关键词的类型和所述扩展关键词与所述自定义关键词之间的相似度从所述扩展关键词中筛选出目标关键词, 加入目标关键词列表;

根据所述目标关键词列表中的目标关键词进行实时监听; 及

5 当监听到目标关键词所对应的话题量达到预设阈值时, 进行话题预警。

一种话题预警的装置, 包括:

自定义关键词获取模块, 用于获取自定义关键词;

扩展关键词获取模块, 用于计算所述自定义关键词与语料库中每个词语之间的相似度, 根据所述相似度从语料库中获取与所述自定义关键词相关的
10 扩展关键词;

目标关键词筛选模块, 用于根据所述扩展关键词的类型和所述扩展关键词与所述自定义关键词之间的相似度从所述扩展关键词中筛选出目标关键词, 加入目标关键词列表;

监听模块, 用于根据所述目标关键词列表中的目标关键词进行实时监听;

15 及

预警模块, 用于当监听所述目标关键词所对应的话题量达到预设阈值时, 进行话题预警。

一种计算机设备, 包括存储器和处理器, 所述存储器中存储有计算机可读指令, 所述计算机可读指令被所述处理器执行时, 使得所述处理器执行以下
20 步骤:

获取自定义关键词;

计算所述自定义关键词与语料库中每个词语之间的相似度, 根据所述相似度从语料库中获取与所述自定义关键词相关的扩展关键词;

根据所述扩展关键词的类型和所述扩展关键词与所述自定义关键词之间的相似度从所述扩展关键词中筛选出目标关键词, 加入目标关键词列表;
25

根据所述目标关键词列表中的目标关键词进行实时监听; 及

当监听到目标关键词所对应的话题量达到预设阈值时, 进行话题预警。

一个或多个存储有计算机可读指令的非易失性可读存储介质，所述计算机可读指令被一个或多个处理器执行时，使得所述一个或多个处理器执行以下步骤：

获取自定义关键词；

5 计算所述自定义关键词与语料库中每个词语之间的相似度，根据所述相似度从语料库中获取与所述自定义关键词相关的扩展关键词；

根据所述扩展关键词的类型和所述扩展关键词与所述自定义关键词之间的相似度从所述扩展关键词中筛选出目标关键词，加入目标关键词列表；

根据所述目标关键词列表中的目标关键词进行实时监听；及

10 当监听到目标关键词所对应的话题量达到预设阈值时，进行话题预警。

本发明的一个或多个实施例的细节在下面的附图和描述中提出。本发明的其它特征、目的和优点将从说明书、附图以及权利要求书变得明显。

附图说明

15 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。

图 1 为一个实施例中终端的内部结构框图；

20 图 2 为一个实施例中服务器的内部结构框图；

图 3 为一个实施例中话题预警的方法流程图；

图 4 为一个实施例中根据扩展关键词的类型和扩展关键词与自定义关键词之间的相似度从扩展关键词中筛选出目标关键词的方法流程图；

图 5 为另一个实施例中话题预警的方法流程图；

25 图 6 为一个实施例中计算自定义关键词与语料库中每个词语之间的相似度，根据相似度从语料库中获取扩展关键词的方法流程图；

图 7 为一个实施例中话题预警的装置结构框图；

图 8 为一个实施例中目标关键词筛选模块的结构框图；

图 9 为另一个实施例中话题预警的装置结构框图。

具体实施方式

5 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

如图 1 所示，在一个实施例中，终端 102 的内部结构如图 1 所示，包括通过系统总线连接的处理器、非易失性存储介质、内存储器、网络接口、显示
10 屏和输入装置。其中，终端 102 的处理器用于提供计算和控制能力，支撑整个终端 102 的运行。非易失性存储介质存储有操作系统和计算机可读指令，该计算机可读指令可被处理器执行以实现适用于终端 102 的一种话题预警的方法。终端 102 中的内存储器为非易失性存储介质中的操作系统和计算机可读指令的运行提供环境。网络接口用于连接到网络进行通信。终端 102 的显
15 示屏可以是液晶显示屏或者电子墨水显示屏等，输入装置可以是显示屏上覆盖的触摸层，也可以是电子设备外壳上设置的按键、轨迹球或触控板，也可以是外接的键盘、触控板或鼠标等。该终端 102 可以是平板电脑、笔记本电脑、台式计算机等。本领域技术人员可以理解，图 1 中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的
20 的终端的限定，具体的终端可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

如图 2 所示，在一个实施例中，服务器 104 的内部结构如图 2 所示，包括通过系统总线连接的处理器、非易失性存储介质、内存储器和网络接口。其中，该服务器 104 的处理器用于提供计算和控制能力，支撑整个服务器的
25 运行。该非易失存储介质包括操作系统和计算机可读指令。该计算机可读指令可被处理器执行以实现适用于服务器 104 的一种话题预警的方法该服务器 104 的内存储器为非易失性存储介质中的操作系统和计算机可读指令的运行

提供环境，该服务器的网络接口用于与外部的服务器和终端通过网络连接通信。本领域技术人员可以理解，图 2 中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的服务器的限定，具体的服务器可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

如图 3 所示，在一个实施例中，提出了一种话题预警的方法，该方法可应用于计算机设备中，其中，计算机设备可以是终端，也可以是服务器，具体包括以下步骤：

步骤 302，获取自定义关键词。

在本实施例中，自定义关键词是指用户给出的符合用户监听需求的关键词。为了能够满足用户的个性化的监听需求，监听关键词的设定是根据用户自定义关键词来设定的。由于大数据时代的社交媒体信息错综复杂，主体多种多样，而不同的用户所关注的话题不尽相同，其中，话题是指谈论的主体，由于不同人关注的主体不同，所以通过自定义关键词不仅能带来友好的用户交互，更多的是能够实现用户监听需求的个性化以及多元化。

步骤 304，计算自定义关键词与语料库中每个词语之间的相似度，根据相似度从语料库中获取与自定义关键词相关的扩展关键词。

在本实施例中，由于用户给定的自定义关键词往往不够完整和全面，因此有必要对该自定义关键词进行一定的扩展。获取与该自定义关键词相关的扩展关键词，有利于保证用户对所需要监听的话题更加全面和完整，从而保证监听结果的完整性和多样性。通过计算自定义关键词与语料库中每个词语之间的相似度，从语料库中选取与自定义关键词相似度比较大的词语作为扩展关键词。相似度越大，说明该词语与自定义关键词的语义越相近。词语相似度的计算方法有多种，比如，可以采用同义词词林的方式计算词语之间的相似度，也可以采用皮尔森相关系数来计算词语之间的相似度。这里并不对词语相似度的计算方法进行限定。

在一个实施例中，相似度的计算是通过计算词向量之间的相似度得到的。

首先，采用 word2vec 模型计算自定义关键词对应的词向量，其中，word2vec 是一款将词表征为实数值向量的高效工具，其利用深度学习的思想，可以通过训练，把对文本内容的处理简化为 k 维向量空间中的向量运算，而向量空间上的相似度可以用来表示文本语义上的相似度。具体地，将自定义关键词作为 word2vec 模型的输入，输出该自定义关键词的词向量表示。获取到自定义关键词的词向量表示之后，通过计算词向量之间的相似度从语料库中筛选出自定义关键词的扩展关键词。为了能够更快的获取到与自定义关键词相关的扩展关键词，可以将语料库中的词语均以词向量的形式存储。在一个实施例中，采用皮尔森相关系数（Pearson Correlation Coefficient）来计算词向量之间的相似度。假设自定义关键词的向量表示为 $W=(w_1, w_2, \dots, w_n)$ ，语料库中任一词语的向量表示为 $X=(x_1, x_2, \dots, x_n)$ ，那么它们之间的相似度 $s(W, X)$ 为：

$$s(W, X) = \frac{n \sum_{i=1}^n w_i x_i - \sum_{i=1}^n w_i \sum_{i=1}^n x_i}{\sqrt{n \sum_{i=1}^n w_i^2 - (\sum_{i=1}^n w_i)^2} \cdot \sqrt{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}}。$$

其中，n 表示词向量的第 n 个词向量特征，i 表示词向量中的第 i 个词向量特征。通过计算自定义关键词与语料库中每个词语的相似度筛选出与自定义关键词相关的扩展关键词。具体地，可以将相似度按照从高到低的顺序进行排列，选出相似度最高的前 k 个词语作为自定义关键词的扩展关键词。将自定义关键词进行扩展，使得关键词更具多样性，保证了话题监听结果具有与相似关键词的对比性，便于为决策者提供更丰富的信息。

步骤 306，根据扩展关键词的类型和扩展关键词与自定义关键词之间的相似度从扩展关键词中筛选出目标关键词，加入目标关键词列表。

在本实施例中，如果对步骤 204 得到的扩展关键词全部监听，将会使得信息错杂冗乱。所以为了保证信息的清楚，需要对获取到的扩展关键词进行进一步的筛选。根据扩展关键词的类型和扩展关键词与自定义关键词之间的相似度从扩展关键词中筛选出目标关键词的方法有多种。在一个实施例中，首先，将获取到的全部扩展关键词进行分类，然后从每一类中选取与自定

义关键词相似度最高的前 h 个扩展关键词作为目标关键词，其中， h 为大于 0 的正整数，将每一类筛选出来的目标关键词进行聚合，生成用于监听的目标关键词列表。在另一个实施例中，首先，获取全部扩展词对应的类型，然后将相同类型的关键词分为一组。分别获取每一类扩展关键词对应的扩展词数目，以扩展词数目最少的类型为基准，假设扩展词数目最少的类型对应的数目为 X 个，那么分别从其他每一类型中也筛选出 X 个扩展关键词作为目标关键词，其中，从其他每一类型中筛选出 X 个扩展关键词是根据相似度的大小进行筛选的，分别筛选出其他每一类扩展关键词中相似度最高的前 X 个扩展关键词作为目标关键词，加入目标关键词列表。

步骤 308，根据目标关键词列表中的目标关键词进行实时监听。

在本实施例中，当确定了目标关键词列表后，根据目标关键词列表中的目标关键词进行实时监听。由于社交媒体数据每时每刻都在产生，迅速而规模庞大，形成了庞大的网络数据流。为了更好的对话题进行监听，可以采用基于滑动窗口的时序管理框架。基于滑动窗口的时序管理框架的主要思想是：对于目标监听列表中的每一个目标关键词，以滑动窗口的形式对话题数据流进行管理，每个目标关键词维护一个一定大小的缓存，每过一个时间片(为了实时监听，时间片的设置通常很小，比如 5 分钟)，数据窗口进行滑动，然后对缓存中的数据进行处理。

步骤 310，当监听到目标关键词所对应的话题量达到预设阈值时，进行话题预警。

在本实施例中，良好的监听必定需要预警，通过监听目标关键词所对应的话题量是否达到预设阈值，对话题进行预警。预警可以从两个方面来进行考虑，第一，对预设的时间片内的话题量进行监听预警。由于时间片是一个较短的时间，所以通过对短时间内的话题监听，能够对短时间内的突发事件进行预警。第二，对于一段时间段的话题进行预警，很多时候事件的发生或舆情的走势并不一定是急剧的，因此，考察一段时间内话题的热点能够帮助决策者发现事件的兴起或舆情的逐渐走势。具体地，采用两种评价策略进行

关键词的实时预警，一种是采用话题热度进行预警，通过分析大量的关键词的热度变化趋势及其生命周期，以经验的方式确定热度临界阈值，当监听的目标关键词在一个滑动窗口的时间片内出现的频率大于该热度临界阈值时，进行预警响应。一种是采用情感极性比率进行预警，对监听的目标关键词列表相关的社会网络文本进行情感极性分析，主要包括正面、中性和负面三个方面的情感极性，当负面情感在所有该目标关键词对应的话题量中占的比率大于情感极性阈值时，进行预警。该话题预警的方法可以应用于很多领域，尤其是可以应用于金融领域。以应用于金融产品为例，说明一下该话题预警的益处。首先，互联网与金融产业息息相关，根据对互联网数据的监控可以为金融产品避免诸多损失。其次，与金融相关的关键词比较有规律，而且相对比较固定，通过对金融产品相关的话题进行监听预警，可以实现快速响应而不失准确率。

在本实施例中，通过获取用户自定义关键词，然后在语料库中根据相似度对该自定义关键词进行扩展，获取相关的扩展关键词，再根据扩展关键词的类型和相似度进行筛选，筛选出最终用于监听的目标关键词，之后在社交媒体上根据该目标关键词进行实时监听，当监听到目标关键词的话题量达到预设阈值时，进行话题预警。该方法不仅能够实时对话题进行监听，而且可以基于用户自定义的关键词有针对性的进行监控，满足了用户的个性化监听预警的需求。通过对用户所要监控的自定义关键词进行扩展和筛选，保证了监听的多样性和全面性。

如图 4 所示，在一个实施例中，根据扩展关键词的类型和扩展关键词与自定义关键词之间的相似度从扩展关键词中筛选出目标关键词，加入目标关键词列表的步骤包括：

步骤 306A，将扩展关键词按照预设的类型进行分类。

在本实施例中，为了对基于自定义关键词的监听能够监听的更加全面和平衡化。首先，需要对扩展关键词按照预设的类型进行分类，比如，将扩展关键词按照“品牌”、“产品”、“竞品”分为三类。这样，便于后续针对每一

类挑选出相同个数的目标关键词进行监听，有利于保证监听信息的清楚全面和平衡。

步骤 306B，从每一类的扩展关键词中筛选出与自定义关键词相似度最高的前 h 个扩展关键词作为目标关键词，其中，h 为大于 0 的正整数。

5 在本实施例中，将扩展关键词按照预设的类型进行分类后，采用众包策略从每一类的扩展关键词中筛选出与自定义关键词相似度最高的前 h 个扩展关键词作为目标关键词。例如，从每一类中挑选出与自定义关键词相似度最高的前 5 个词语，最后将挑选出的每一类的目标关键词进行聚合。

10 步骤 306C，将每一类筛选出来的目标关键词进行聚合，生成用于监听的目标关键词列表。

 在本实施例中，通过从每一类的扩展关键词中筛选出与自定义关键词相似度最高的前 h 个扩展关键词作为目标关键词后，将每一类筛选出来的目标关键词聚集起来，放在同一张列表中，即生成目标关键词列表，后续便于根据该目标关键词列表中的目标关键词进行实时监听。比如，若将扩展关键词
15 按照“品牌”、“产品”、“竞品”分为三类。若每一类都挑选出 5 个目标关键词，那么将总共挑选出 15 个目标关键词进行监听。通过将扩展关键词进行分类，然后再针对每一类进行筛选有利于监听的内容更加清晰和全面，不会出现偏激化的结果。

 如图 5 所示，在一个实施例中，提出了一种话题预警的方法，该方法包
20 括：

 步骤 502，获取自定义关键词。

 步骤 504，计算自定义关键词对应的词向量。

 步骤 506，计算自定义关键词的词向量与语料库中每个词语的词向量之间的相似度，根据词向量之间的相似度从语料库中获取与自定义关键词相关的
25 扩展关键词。

 步骤 508，根据扩展关键词的类型和扩展关键词与自定义关键词之间的相似度从扩展关键词中筛选出目标关键词，加入目标关键词列表。

步骤 510, 根据目标关键词列表中的目标关键词进行实时监听。

步骤 512, 当监听到目标关键词所对应的话题量达到预设阈值时, 进行话题预警。

在本实施例中, 当获取到自定义关键词后, 为了后续计算词向量之间的
5 相似度, 首先需要计算该自定义关键词对应的词向量, 通过将自定义关键词
作为 word2vec 模型的输入, 生成与该自定义关键词对应的词向量并输出。为
了监听的更加全面, 需要对自定义关键词进行扩展, 即找出相关的与该自定义
关键词语义相近的词语表示。通过计算自定义关键词与语料库中的每个词
语之间的相似度来获取与自定义关键词相关的扩展关键词, 其中, 相似度越
10 高, 说明与自定义关键词的语义越相近。具体地, 可以采用皮尔森相关系数
(Pearson Correlation Coefficient) 方法计算自定义关键词的词向量与语料库中
每个词语的词向量之间的相似度, 从中挑选出与自定义关键词相似度最高的
前 K 个 (比如, 设 $K=50$) 词语作为扩展关键词。如果对挑选出来的扩展关
键词全部进行监听, 将会使得信息显得冗杂, 为了解决这一问题, 还需要对
15 挑选出来的扩展关键词进行进一步的筛选。基于众包策略对扩展关键词进行
进一步的筛选, 首先对挑选出来的扩展关键词进行分类, 比如, 按照 “品牌”、
“产品”、“竞品” 分为三类。分类完成后, 针对每一类, 根据之前计算得到
的每个扩展关键词与自定义关键词之间的相似度, 每一类选出与自定义关键词
相似度最高的前 h 个词语作为目标关键词, 然后将每一类筛选出来的目标
20 关键词进行汇总, 放在同一个列表中, 即都加入目标关键词列表。之后根据
该目标关键词列表进行监听, 并进行相应的预警。该方法通过对用户自定义
关键词进行扩展, 保证了监听的多样性和全面性, 结合众包技术对扩展关键词
进行进一步甄选保证了监听结果不具有偏激化。

如图 6 所示, 在一个实施例中, 计算自定义关键词与语料库中每个词语
25 之间的相似度, 根据相似度从语料库中获取与自定义关键词相关的扩展关键词
的步骤包括:

步骤 304A, 采用皮尔森相关系数方法计算自定义关键词与语料库中每个

词语之间的相似度。

在本实施例中，为了对自定义关键词进行扩展，找出与自定义关键词语义相近的扩展关键词，通过采用皮尔森相关系数方法来计算自定义关键词与语料库中每个词语之间的相似度。相似度越大，语义越相近。具体地，首先，
5 获取自定义关键词的词向量表示，可以通过 word2vec 方法计算得到。然后计算自定义关键词的词向量与语料库中词语的词向量之间的相似度。为了能够更加快捷的计算自定义关键词与语料库中词语之间的相似度，在语料库中，词语是以词向量的形式存在的。假设自定义关键词的词向量表示为 $W=(w_1, w_2, \dots, w_n)$ ，语料库中任一词语的词向量表示为 $X=(x_1, x_2, \dots, x_n)$ ，那么它们之间的相似度 $s(W, X)$ 为：

$$s(W, X) = \frac{n \sum_{i=1}^n w_i x_i - \sum_{i=1}^n w_i \sum_{i=1}^n x_i}{\sqrt{n \sum_{i=1}^n w_i^2 - (\sum_{i=1}^n w_i)^2} \cdot \sqrt{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}}。$$

步骤 304B，获取与自定义关键词相似度最高的前 K 个词语作为自定义关键词的扩展关键词，其中，K 为大于 0 的正整数。

在本实施例中，显然，对自定义关键词进行无限扩展是不切实际的，所以
15 需要从语料库中筛选出相似度比较大的词语作为扩展关键词。具体地，采用贪心策略选择与自定义关键词相似度最高的前 K 个词语作为自定义关键词的扩展，设扩展关键词集合为 $ES(W)$ ，那么 $ES(W) = \{X | s(W, X) \geq s(W, X_k)\}$ ，其中， W 表示自定义关键词， X_k 表示与自定义关键词相似度第 K 大的词汇，比如，可以设置 $K=50$ ，即选取与自定义关键词相似度最高的前 50 个词汇作为其扩展
20 关键词集合。

在一个实施例中，根据目标关键词列表中的目标关键词进行实时监听的步骤包括：采用滑动窗口的形式对目标关键词列表中的每一个目标关键词进行实时监听。

在本实施例中由于社交媒体数据每时每刻都在产生，且迅速而规模庞大，
25 为了达到对话题进行实时监听，需要解决如何在数据流的环境下进行话题的

实时监听。在该实施例中，通过采用基于滑动窗口的形式对目标关键词列中的每一个目标关键词进行实时监听。即以滑动窗口的形式对话题数据流进行管理，每个目标关键词维护一个一定大小的缓存，每过一个时间片，数据窗口进行滑动，然后对缓存中的数据进行处理，从而实现了对每个目标关键词进行实时监听。

如图 7 所示，在一个实施例中，提出了一种话题预警的装置 700，该装置包括：

自定义关键词获取模块 702，用于获取自定义关键词。

扩展关键词获取模块 704，用于计算自定义关键词与语料库中每个词语之间的相似度，根据相似度从语料库中获取与自定义关键词相关的扩展关键词。

目标关键词筛选模块 706，用于根据扩展关键词的类型和扩展关键词与自定义关键词之间的相似度从扩展关键词中筛选出目标关键词，加入目标关键词列表。

监听模块 708，用于根据目标关键词列表中的目标关键词进行实时监听。

预警模块 710，用于当监听目标关键词所对应的话题量达到预设阈值时，进行话题预警。

如图 8 所示，在一个实施例中，目标关键词筛选模块 706 包括：

分类模块 706A，用于将扩展关键词按照预设的类型进行分类。

筛选模块 706B，用于从每一类的扩展关键词中筛选出与自定义关键词相似度最高的前 h 个扩展关键词作为目标关键词，其中，h 为大于 0 的正整数。

聚合模块 706C，用于将每一类筛选出来的目标关键词进行聚合，生成用于监听的目标关键词列表。

如图 9 所示，在一个实施例中，提出了一种话题预警的装置 900，除了包括上述模块 702-710，还包括：

计算模块 703，用于计算自定义关键词对应的词向量。

扩展关键词获取模块 704 还用于计算自定义关键词的词向量与语料库中

每个词语的词向量之间的相似度，根据词向量之间的相似度从语料库中获取与自定义关键词相关的扩展关键词。

在一个实施例中，扩展词获取模块还用于采用皮尔森相关系数方法计算自定义关键词与语料库中每个词语之间的相似度，获取与自定义关键词相似
5 度最高的前 K 个词语作为自定义关键词的扩展关键词，其中，K 为大于 0 的正整数。

在一个实施例中，预警模块还用于采用滑动窗口的形式对目标关键词列表中的每一个目标关键词进行实时监听。

上述话题预警的装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。其中，网络接口可以是以太网卡或无线网卡等。上述各模块可以
10 硬件形式内嵌于或独立于服务器中的处理器中，也可以以软件形式存储于服务器中的存储器中，以便于处理器调用执行以上各个模块对应的操作。该处理器可以为中央处理单元（CPU）、微处理器、单片机等。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，该计算机程序可存储于一计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，前述的存储介质可为磁碟、光盘、只读存储记忆体（Read-Only Memory, ROM）等非易失性存储介质，或随机存储记忆体（Random Access Memory, RAM）等。
15

以上所述实施例的各技术特征可以进行任意的组合，为使描述简洁，未
20 对上述实施例中的各个技术特征所有可能的组合都进行描述，然而，只要这些技术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

以上所述实施例仅表达了本发明的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对发明专利范围的限制。应当指出的是，对于本
25 领域的普通技术人员来说，在不脱离本发明构思的前提下，还可以做出若干变形和改进，这些都属于本发明的保护范围。因此，本发明的保护范围应以所附权利要求为准。

权利要求书

1、一种话题预警的方法，包括：

获取自定义关键词；

计算所述自定义关键词与语料库中每个词语之间的相似度，根据所述相似度从语料库中获取与所述自定义关键词相关的扩展关键词；

5 根据所述扩展关键词的类型和所述扩展关键词与所述自定义关键词之间的相似度从所述扩展关键词中筛选出目标关键词，加入目标关键词列表；

根据所述目标关键词列表中的目标关键词进行实时监听；及

当监听到目标关键词所对应的话题量达到预设阈值时，进行话题预警。

2、根据权利要求1所述的方法，其特征在于，所述根据所述扩展关键词
10 的类型和所述扩展关键词与所述自定义关键词之间的相似度从所述扩展关键词中筛选出目标关键词，加入目标关键词列表包括：

将所述扩展关键词按照预设的类型进行分类；

从每一类的扩展关键词中筛选出与所述自定义关键词相似度最高的前 h
个扩展关键词作为目标关键词，其中，h 为大于 0 的正整数；

15 将每一类筛选出来的目标关键词进行聚合，生成用于监听的目标关键词列表。

3、根据权利要求1所述的方法，其特征在于，在获取自定义关键词之后还包括：

计算所述自定义关键词对应的词向量；

20 所述计算所述自定义关键词与语料库中每个词语之间的相似度，根据所述相似度从语料库中获取与所述自定义关键词相关的扩展关键词包括：

计算自定义关键词的词向量与所述语料库中每个词语的词向量之间的相似度；

25 根据词向量之间的相似度从语料库中获取与所述自定义关键词相关的扩展关键词。

4、根据权利要求1所述的方法，其特征在于，所述计算所述自定义关键词

词与语料库中每个词语之间的相似度，根据相似度从语料库中获取与所述自定义关键词相关的扩展关键词包括：

采用皮尔森相关系数方法计算所述自定义关键词与语料库中每个词语之间的相似度；

5 获取与所述自定义关键词相似度最高的前 K 个词语作为所述自定义关键词的扩展关键词，其中，K 为大于 0 的正整数。

5、根据权利要求 1 所述的方法，其特征在于，所述根据所述目标关键词列表中的目标关键词进行实时监听包括：

10 采用滑动窗口的形式对所述目标关键词列表中的每一个目标关键词进行实时监听。

6、一种话题预警的装置，包括：

自定义关键词获取模块，用于获取自定义关键词；

15 扩展关键词获取模块，用于计算所述自定义关键词与语料库中每个词语之间的相似度，根据所述相似度从语料库中获取与所述自定义关键词相关的扩展关键词；

目标关键词筛选模块，用于根据所述扩展关键词的类型和所述扩展关键词与所述自定义关键词之间的相似度从所述扩展关键词中筛选出目标关键词，加入目标关键词列表；

20 监听模块，用于根据所述目标关键词列表中的目标关键词进行实时监听；及

预警模块，用于当监听所述目标关键词所对应的话题量达到预设阈值时，进行话题预警。

7、根据权利要求 6 述的装置，其特征在于，所述目标关键词筛选模块包括：

25 分类模块，用于将所述扩展关键词按照预设的类型进行分类；

筛选模块，用于从每一类的扩展关键词中筛选出与所述自定义关键词相似度最高的前 h 个扩展关键词作为目标关键词，其中，h 为大于 0 的正整数；

聚合模块，用于将每一类筛选出来的目标关键词进行聚合，生成用于监听的目标关键词列表。

8、根据权利要求6所述的装置，其特征在于，所述装置还包括：

计算模块，用于计算所述自定义关键词对应的词向量；

5 扩展关键词获取模块还用于计算自定义关键词的词向量与所述语料库中每个词语的词向量之间的相似度，根据词向量之间的相似度从语料库中获取与所述自定义关键词相关的扩展关键词。

9、根据权利要求6所述的装置，其特征在于，所述扩展词获取模块还用于采用皮尔森相关系数方法计算所述自定义关键词与语料库中每个词语之间的相似度，获取与所述自定义关键词相似度最高的前K个词语作为所述自定义关键词的扩展关键词，其中，K为大于0的正整数。

10、根据权利要求6所述的装置，其特征在于，所述预警模块还用于采用滑动窗口的形式对所述目标关键词列表中的每一个目标关键词进行实时监听。

11、一种计算机设备，包括存储器和处理器，所述存储器中存储有计算机可读指令，所述计算机可读指令被所述处理器执行时，使得所述处理器执行以下步骤：

获取自定义关键词；

计算所述自定义关键词与语料库中每个词语之间的相似度，根据所述相似度从语料库中获取与所述自定义关键词相关的扩展关键词；

20 根据所述扩展关键词的类型和所述扩展关键词与所述自定义关键词之间的相似度从所述扩展关键词中筛选出目标关键词，加入目标关键词列表；

根据所述目标关键词列表中的目标关键词进行实时监听；及

当监听到目标关键词所对应的话题量达到预设阈值时，进行话题预警。

12、根据权利要求11所述的计算机设备，其特征在于，所述处理器所执行的所述根据所述扩展关键词的类型和所述扩展关键词与所述自定义关键词之间的相似度从所述扩展关键词中筛选出目标关键词，加入目标关键词列表的步骤包括：

将所述扩展关键词按照预设的类型进行分类;

从每一类的扩展关键词中筛选出与所述自定义关键词相似度最高的前 h 个扩展关键词作为目标关键词, 其中, h 为大于 0 的正整数;

5 将每一类筛选出来的目标关键词进行聚合, 生成用于监听的目标关键词列表。

13、根据权利要求 11 所述的计算机设备, 其特征在于, 在获取自定义关键词的步骤之后, 所述处理器还用于执行以下步骤:

计算所述自定义关键词对应的词向量;

10 所述计算所述自定义关键词与语料库中每个词语之间的相似度, 根据所述相似度从语料库中获取与所述自定义关键词相关的扩展关键词包括:

计算自定义关键词的词向量与所述语料库中每个词语的词向量之间的相似度;

根据词向量之间的相似度从语料库中获取与所述自定义关键词相关的扩展关键词。

15 14、根据权利要求 11 所述的计算机设备, 其特征在于, 所述处理器所执行的所述计算所述自定义关键词与语料库中每个词语之间的相似度, 根据相似度从语料库中获取与所述自定义关键词相关的扩展关键词的步骤包括:

采用皮尔森相关系数方法计算所述自定义关键词与语料库中每个词语之间的相似度;

20 获取与所述自定义关键词相似度最高的前 K 个词语作为所述自定义关键词的扩展关键词, 其中, K 为大于 0 的正整数。

15 15、根据权利要求 11 所述的计算机设备, 其特征在于, 所述处理器所执行的所述根据所述目标关键词列表中的目标关键词进行实时监听的步骤包括: 采用滑动窗口的形式对所述目标关键词列表中的每一个目标关键词进行
25 实时监听。

16、一个或多个存储有计算机可读指令的非易失性可读存储介质, 所述计算机可读指令被一个或多个处理器执行时, 使得所述一个或多个处理器执

行以下步骤:

获取自定义关键词;

计算所述自定义关键词与语料库中每个词语之间的相似度, 根据所述相似度从语料库中获取与所述自定义关键词相关的扩展关键词;

5 根据所述扩展关键词的类型和所述扩展关键词与所述自定义关键词之间的相似度从所述扩展关键词中筛选出目标关键词, 加入目标关键词列表;

根据所述目标关键词列表中的目标关键词进行实时监听; 及

当监听到目标关键词所对应的话题量达到预设阈值时, 进行话题预警。

17、根据权利要求 16 所述的非易失性可读存储介质, 其特征在于, 所述
10 处理器所执行的所述根据所述扩展关键词的类型和所述扩展关键词与所述自定义关键词之间的相似度从所述扩展关键词中筛选出目标关键词, 加入目标关键词列表的步骤包括:

将所述扩展关键词按照预设的类型进行分类;

从每一类的扩展关键词中筛选出与所述自定义关键词相似度最高的前 h
15 个扩展关键词作为目标关键词, 其中, h 为大于 0 的正整数;

将每一类筛选出来的目标关键词进行聚合, 生成用于监听的目标关键词列表。

18、根据权利要求 16 所述的非易失性可读存储介质, 其特征在于, 在获取自定义关键词的步骤之后, 所述处理器还用于执行以下步骤:

20 计算所述自定义关键词对应的词向量;

所述计算所述自定义关键词与语料库中每个词语之间的相似度, 根据所述相似度从语料库中获取与所述自定义关键词相关的扩展关键词包括:

计算自定义关键词的词向量与所述语料库中每个词语的词向量之间的相似度;

25 根据词向量之间的相似度从语料库中获取与所述自定义关键词相关的扩展关键词。

19、根据权利要求 16 所述的非易失性可读存储介质, 其特征在于, 所述

处理器所执行的所述计算所述自定义关键词与语料库中每个词语之间的相似度，根据相似度从语料库中获取与所述自定义关键词相关的扩展关键词的步骤包括：

5 采用皮尔森相关系数方法计算所述自定义关键词与语料库中每个词语之间的相似度；

获取与所述自定义关键词相似度最高的前 K 个词语作为所述自定义关键词的扩展关键词，其中，K 为大于 0 的正整数。

20、根据权利要求 16 所述的非易失性可读存储介质，其特征在于，所述处理器所执行的所述根据所述目标关键词列表中的目标关键词进行实时监听的步骤包括：采用滑动窗口的形式对所述目标关键词列表中的每一个目标关键词进行实时监听。

10

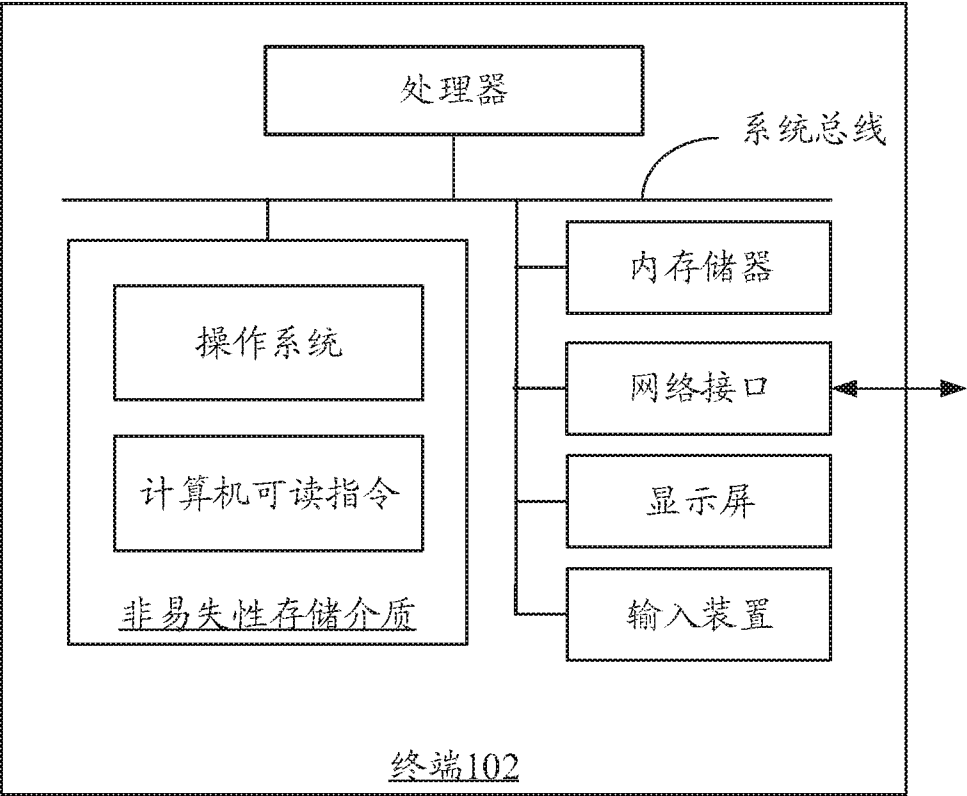


图 1

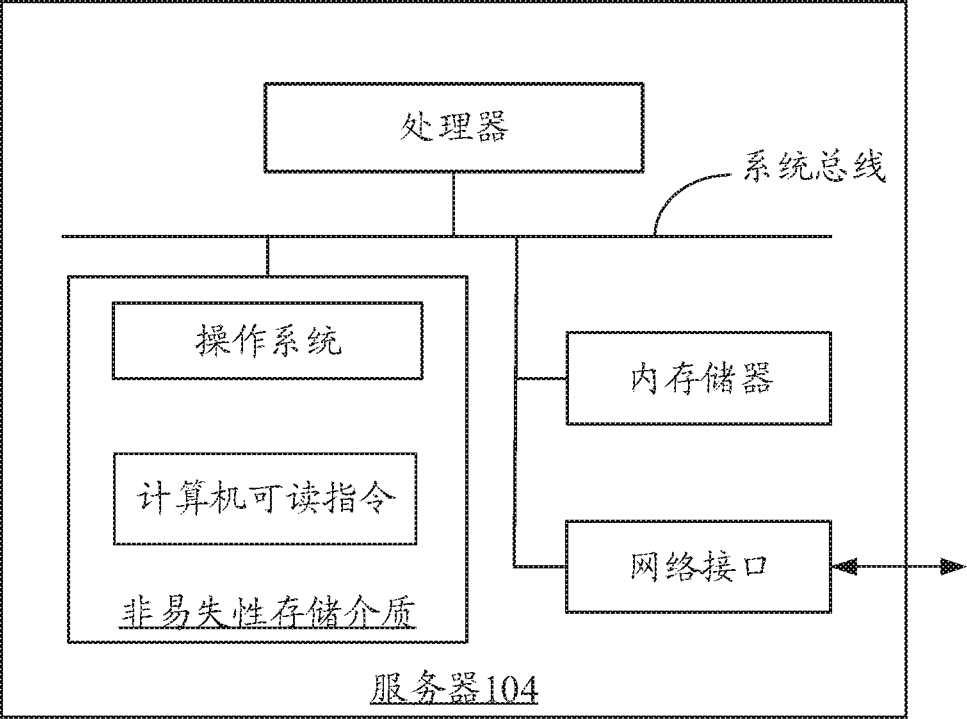


图 2

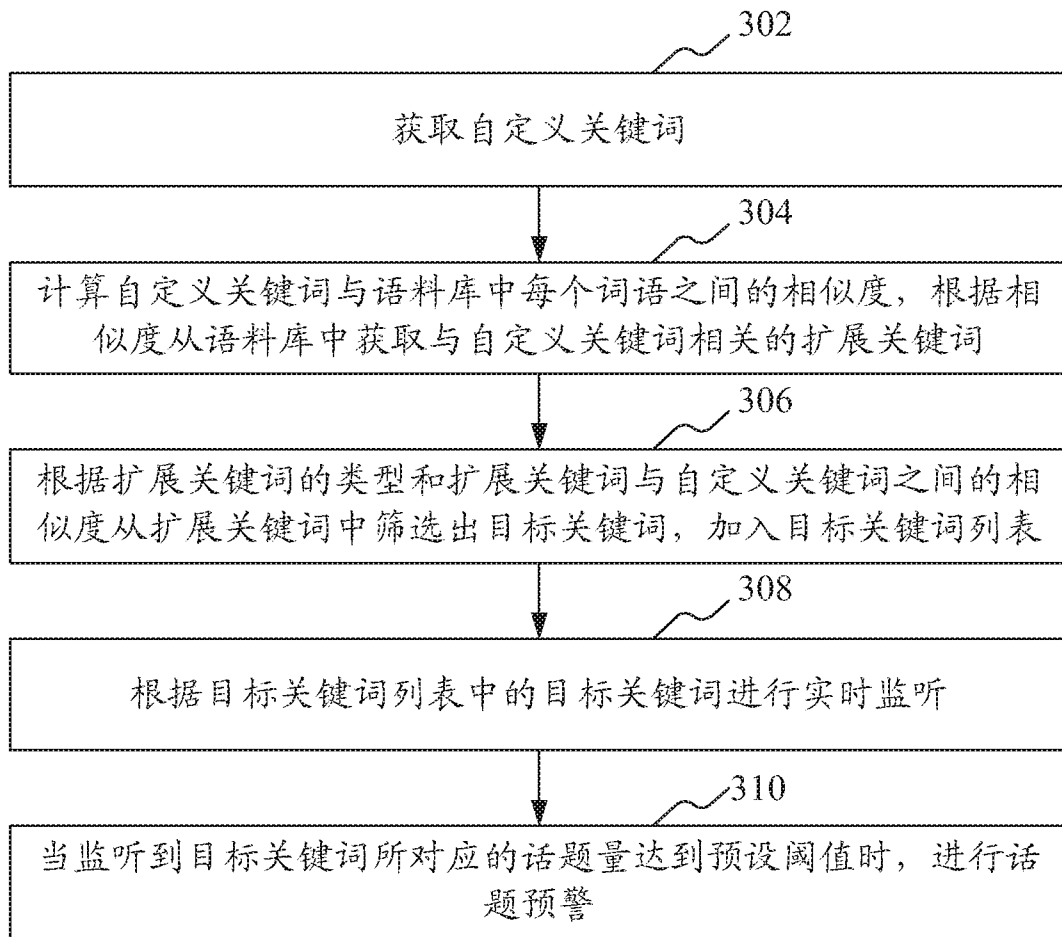


图 3

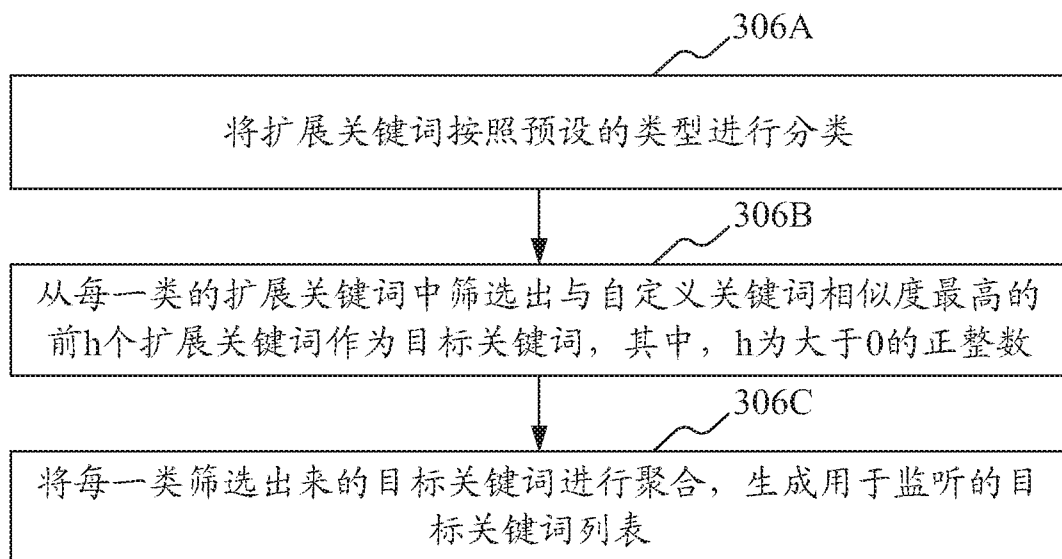


图 4

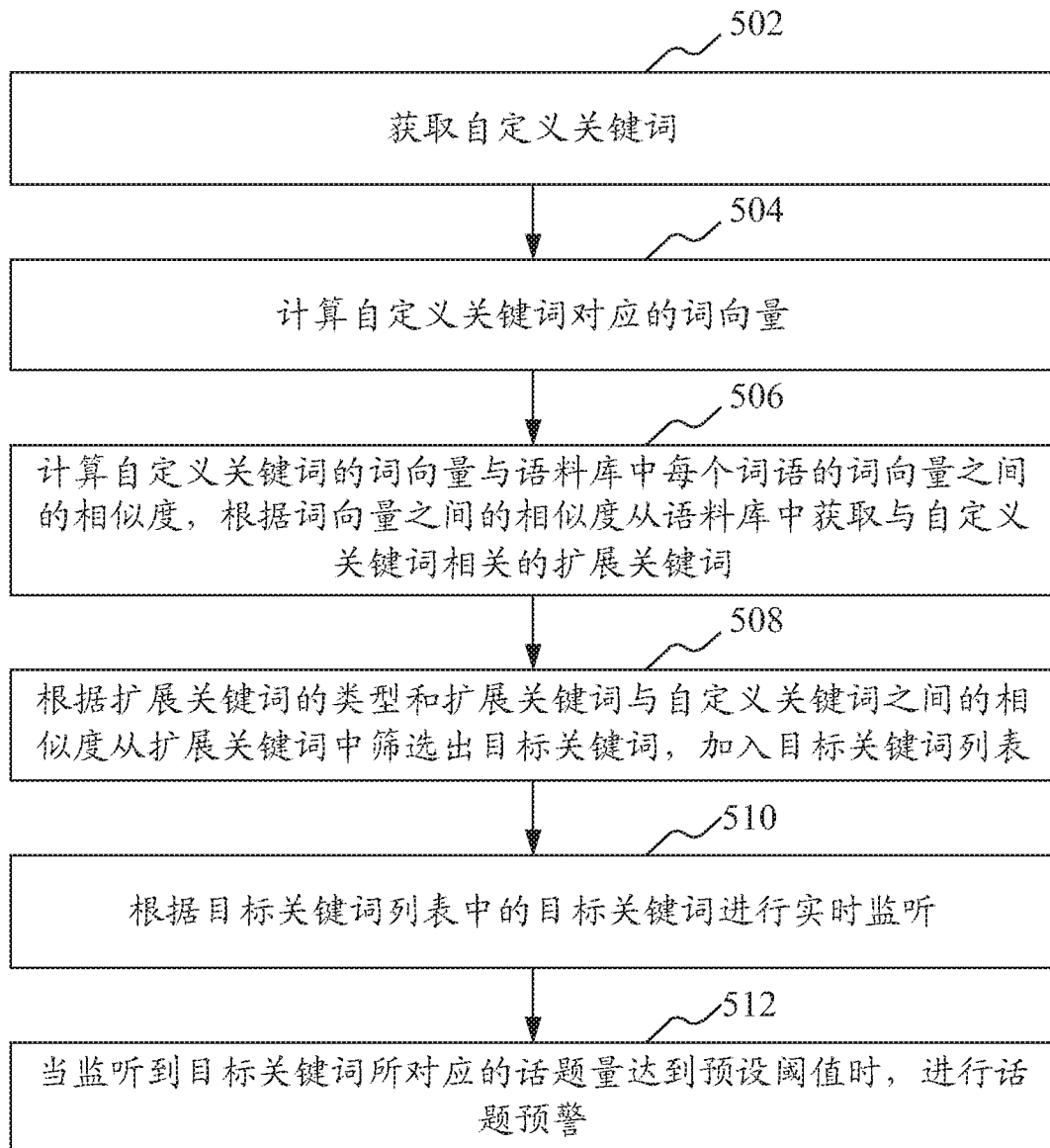


图 5

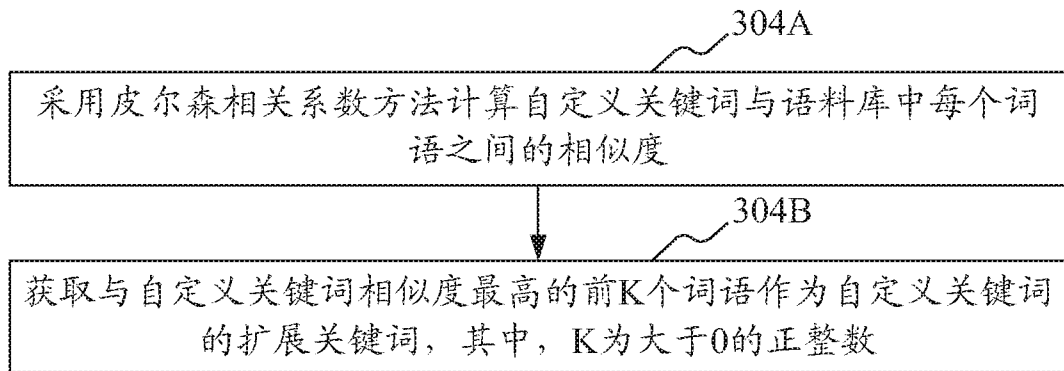


图 6

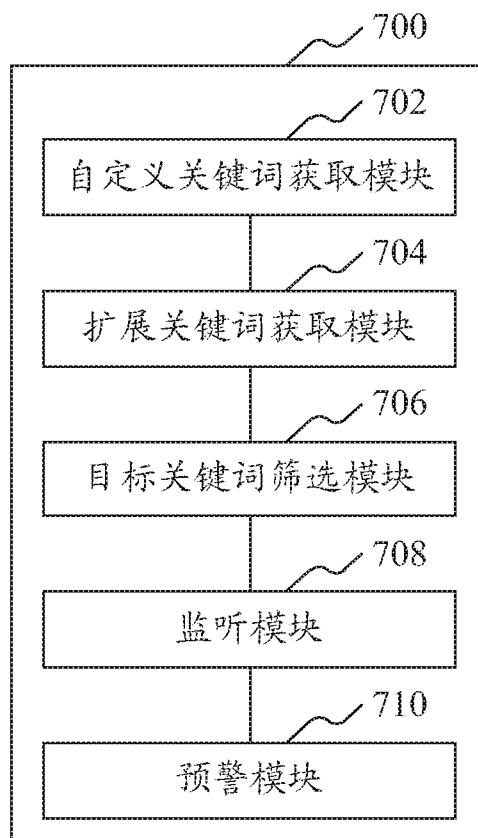


图 7

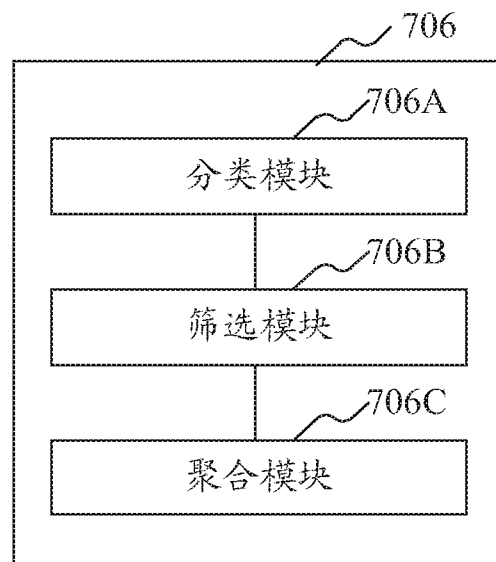


图 8

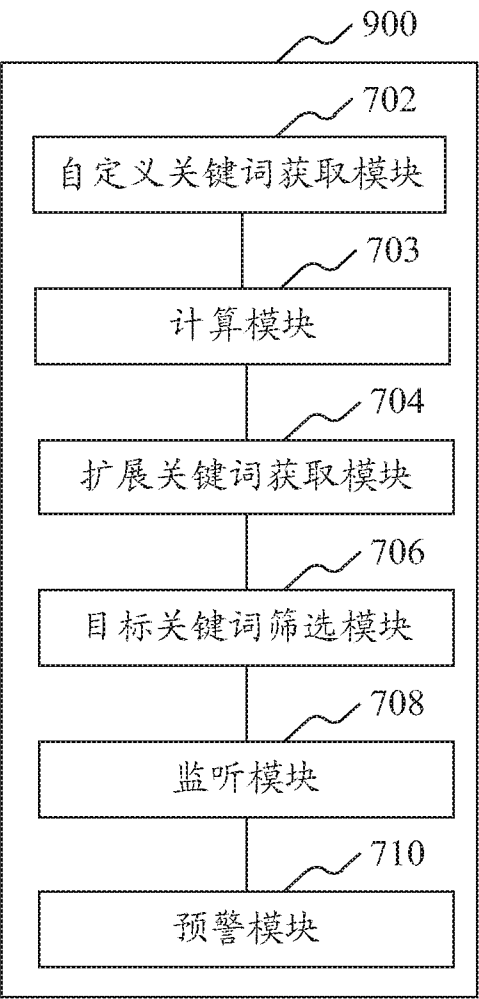


图 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/090579

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/27 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CPRSABS, DWPI, CNKI: 关键词, 话题, 相似度, 库, 阈值, 扩展, 预警, keyword, topic, similarity, database, threshold, expanded, warning

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 103853720 A (SUZHOU XINYI SYSTEM INTEGRATION CO., LTD.) 11 June 2014 (11.06.2014), entire document	1-20
A	CN 104516903 A (PEKING UNIVERSITY FOUNDER GROUP CO., LTD. et al.) 15 April 2015 (15.4.2015) entire document	1-20
A	CN 102194001 A (HANGZHOU DIANZI UNIVERSITY) 21 September 2011 (21.09.2011), entire document	1-20
A	CN 104408157 A (SICHUAN ESLITE E-COMMERCE CO., LTD.) 11 March 2015 (11.03.2015), entire document	1-20

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&”document member of the same patent family

Date of the actual completion of the international search
03 January 2018

Date of mailing of the international search report
11 January 2018

Name and mailing address of the ISA
State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimenqiao
Haidian District, Beijing 100088, China
Facsimile No. (86-10) 62019451

Authorized officer
MING, Mei
Telephone No. (86-10) 62411851

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/090579

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 103853720 A	11 June 2014	CN 103853720 B	26 April 2017
CN 104516903 A	15 April 2015	JP 6231668 B2	15 November 2017
		WO 2015043066 A1	02 April 2015
		EP 3051431 A4	03 May 2017
		EP 3051431 A1	03 August 2016
		US 2016232211 A1	11 August 2016
		JP 2016532175 A	13 October 2016
CN 102194001 A	21 September 2011	None	
CN 104408157 A	11 March 2015	None	

国际检索报告

国际申请号

PCT/CN2017/090579

A. 主题的分类 G06F 17/27 (2006.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类																	
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) CPRSABS, DWPI, CNKI: 关键词, 话题, 相似度, 库, 阈值, 扩展, 预警, keyword, topic, similarity, database, threshold, expanded, warning																	
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>A</td> <td>CN 103853720 A (苏州信颐系统集成有限公司) 2014年 6月 11日 (2014 - 06 - 11) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 104516903 A (北大方正集团有限公司等) 2015年 4月 15日 (2015 - 04 - 15) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 102194001 A (杭州电子科技大学) 2011年 9月 21日 (2011 - 09 - 21) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 104408157 A (四川诚品电子商务有限公司) 2015年 3月 11日 (2015 - 03 - 11) 全文</td> <td>1-20</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	A	CN 103853720 A (苏州信颐系统集成有限公司) 2014年 6月 11日 (2014 - 06 - 11) 全文	1-20	A	CN 104516903 A (北大方正集团有限公司等) 2015年 4月 15日 (2015 - 04 - 15) 全文	1-20	A	CN 102194001 A (杭州电子科技大学) 2011年 9月 21日 (2011 - 09 - 21) 全文	1-20	A	CN 104408157 A (四川诚品电子商务有限公司) 2015年 3月 11日 (2015 - 03 - 11) 全文	1-20
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
A	CN 103853720 A (苏州信颐系统集成有限公司) 2014年 6月 11日 (2014 - 06 - 11) 全文	1-20															
A	CN 104516903 A (北大方正集团有限公司等) 2015年 4月 15日 (2015 - 04 - 15) 全文	1-20															
A	CN 102194001 A (杭州电子科技大学) 2011年 9月 21日 (2011 - 09 - 21) 全文	1-20															
A	CN 104408157 A (四川诚品电子商务有限公司) 2015年 3月 11日 (2015 - 03 - 11) 全文	1-20															
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																	
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																	
国际检索实际完成的日期 2018年 1月 3日		国际检索报告邮寄日期 2018年 1月 11日															
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		授权官员 明媚 电话号码 (86-10)62411851															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2017/090579

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	103853720	A	2014年 6月 11日	CN 103853720 B	2017年 4月 26日
CN	104516903	A	2015年 4月 15日	JP 6231668 B2	2017年 11月 15日
				WO 2015043066 A1	2015年 4月 2日
				EP 3051431 A4	2017年 5月 3日
				EP 3051431 A1	2016年 8月 3日
				US 2016232211 A1	2016年 8月 11日
				JP 2016532175 A	2016年 10月 13日
CN	102194001	A	2011年 9月 21日	无	
CN	104408157	A	2015年 3月 11日	无	