

(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
2019 年 5 月 9 日 (09.05.2019)



(10) 国际公布号
WO 2019/085355 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01)

(21) 国际申请号: PCT/CN2018/077638

(22) 国际申请日: 2018 年 2 月 28 日 (28.02.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201711060246.8 2017 年 11 月 1 日 (01.11.2017) 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工业区平安大厦六楼, Guangdong 518000 (CN)。

(72) 发明人: 张师琲(ZHANG, ShiBei); 中国广东省深圳市福田区八卦岭工业区平安大厦六楼, Guangdong 518000 (CN)。 侯丽(HOU, Li); 中国

广东省深圳市福田区八卦岭工业区平安大厦六楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市沃德知识产权代理有限公司(普通合伙)(SHENZHEN WORLD INTELLECTUAL PROPERTY AGENCY (GENERAL PARTNERSHIP)); 中国广东省深圳市福田区园岭街道八卦四路 10 号中浩大厦 1528-1530 室于志光, Guangdong 518000 (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: PUBLIC SENTIMENT CLUSTERING ANALYSIS METHOD FOR INTERNET NEWS, APPLICATION SERVER, AND COMPUTER-READABLE STORAGE MEDIUM

(54) 发明名称: 互联网新闻的舆情聚类分析方法、应用服务器及计算机可读存储介质

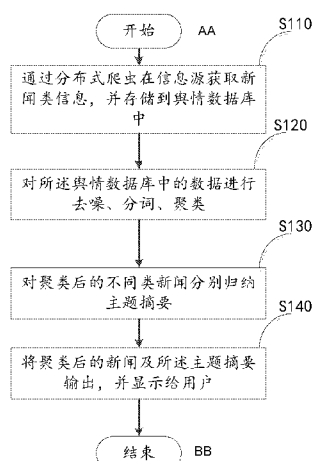


图 3

S110 ACQUIRE NEWS-TYPE INFORMATION FROM AN INFORMATION SOURCE BY MEANS OF A DISTRIBUTED CRAWLER AND STORE SAME IN A PUBLIC SENTIMENT DATABASE
S120 CARRY OUT DENOISING, WORD SEGMENTATION AND CLUSTERING ON DATA IN THE PUBLIC SENTIMENT DATABASE
S130 RESPECTIVELY SUMMARIZE A TOPIC DIGEST FOR DIFFERENT TYPES OF CLUSTERED NEWS
S140 OUTPUT THE CLUSTERED NEWS AND THE TOPIC DIGEST, AND DISPLAY SAME TO A USER
AA START
BB END

(57) Abstract: A public sentiment clustering analysis method for Internet news. The method comprises: acquiring news-type information from an information source by means of a distributed crawler, and storing same in a public sentiment database (S110); carrying out denoising, word segmentation and clustering on data in the public sentiment database (S120); respectively summarizing a topic digest for different types of clustered news (S130); and outputting the clustered news and the topic digest, and displaying same to a user (S140). Further provided are an application server and a computer readable storage medium. According to the public sentiment clustering analysis method for Internet news, the application server and the computer readable storage medium provided therein, news on a network can be obtained quickly, the acquired news is clustered to acquire hot news, and keywords and a digest can be automatically extracted from the acquired news, wherein same is more convenient, faster and more accurate compared with the prior art.

(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))。

(57) 摘要: 一种互联网新闻的舆情聚类分析方法, 该方法包括: 通过分布式爬虫在信息源获取新闻类信息, 并存储到舆情数据库中 (S110); 对所述舆情数据库中的数据进行去噪、分词、聚类 (S120); 对聚类后的不同类新闻分别归纳主题摘要 (S130); 及将聚类后的新闻及所述主题摘要输出, 并显示给用户 (S140)。还提供一种应用服务器及计算机可读存储介质。提供的互联网新闻的舆情聚类分析方法、应用服务器及计算机可读存储介质能够快速获得网络上的新闻, 对获取的新闻进行聚类获取热点新闻, 并且可以对获取的新闻进行自动关键词及摘要, 相较于现有技术, 更加方便、快捷、准确。

互联网新闻的舆情聚类分析方法、应用服务器及计算机可读存储介质

本申请要求于2017年11月01日提交中国专利局、申请号为201711060246.8、发明名称为“互联网新闻的舆情聚类分析方法、应用服务器及计算机可读存储介质”的中国专利申请的优先权，其全部内容通过引用结合在5 合在申请中。

技术领域

本申请涉及数据分析技术领域，尤其涉及一种互联网新闻的舆情聚类分10 析方法、应用服务器及计算机可读存储介质。

背景技术

随着 Internet 的迅猛发展，网络信息已经成为人们生活中必不可少的一部分，目前中国网民数量已经超过2亿，中国网页数量也超过了80亿。网络媒15 体已被公认为继报纸、广播和电视之后的“第四媒体”，网络成为反应社会舆情的主要载体之一。网络舆情与社会舆情相互作用、相互影响，网络舆情与社会舆情在内容表现形态方面具有一致性，网络舆情一定程度上会影响社会舆情的发展趋势，因此网络舆情热点话题的发现具有十分重要的意义。

而各大互联网站的新闻热点层出不穷，抓取并找出这些新闻舆情热点，20 对互联网信息发布、舆情监督等都有重要意义。

因此，如何从互联网的海量信息中发现热点并呈现给用户，成为当下亟需解决的一大问题。

发明内容

有鉴于此，本申请提出一种互联网新闻的舆情聚类分析方法、应用服务25 器及计算机可读存储介质，以解决如何从互联网的海量信息中发现热点并呈

现给用户的问题。

首先，为实现上述目的，本申请提出一种互联网新闻的舆情聚类分析方法，该方法包括步骤：

通过分布式爬虫在信息源获取新闻类信息，并存储到舆情数据库中；

5 对所述舆情数据库中的数据进行去噪、分词、聚类；

对聚类后的不同类新闻分别归纳主题摘要；及

将聚类后的新闻及所述主题摘要输出，并显示给用户。

此外，为实现上述目的，本申请还提供一种应用服务器，包括存储器、
处理器，所述存储器上存储有可在所述处理器上运行的互联网新闻的舆情聚
10 类分析系统，所述互联网新闻的舆情聚类分析系统被所述处理器执行时实现
如上述的互联网新闻的舆情聚类分析方法的步骤。

进一步地，为实现上述目的，本申请还提供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互联网新闻的舆情聚类分析系统，所述互联
网新闻的舆情聚类分析系统可被至少一个处理器执行，以使所述至少一个处
15 理器执行如上述的互联网新闻的舆情聚类分析方法的步骤。

相较于现有技术，本申请所提出的互联网新闻的舆情聚类分析方法、应
用服务器及计算机可读存储介质，首先通过分布式爬虫在信息源获取新闻类
信息，并存储到舆情数据库中；其次，对所述舆情数据库中的数据进行去噪、
分词、聚类；然后，对聚类后的不同类新闻分别归纳主题摘要；最后，将聚
20 类后的新闻及所述主题摘要输出，并显示给用户。采用本申请所提出的互联
网新闻的舆情聚类分析方法、应用服务器及计算机可读存储介质可以快速获
得网络上的新闻，对获取的新闻进行聚类获取热点新闻，并且可以对获取的
新闻进行自动关键词及摘要，相较于现有技术，更加方便、快捷、准确。

25 附图说明

图1是本申请应用服务器一可选的硬件架构的示意图；

图 2 是本申请互联网新闻的舆情聚类分析系统实施方式的程序模块示意图；

图 3 是本申请互联网新闻的舆情聚类分析方法第一实施方式的流程图；

图 4 是本申请互联网新闻的舆情聚类分析方法第二实施方式的流程图；

5 图 5 是本申请互联网新闻的舆情聚类分析方法第三实施方式的流程图；

图 6 是本申请互联网新闻的舆情聚类分析方法第四实施方式的流程图；

图 7 是本申请互联网新闻的舆情聚类分析方法第五实施方式的流程图；

图 8 是本申请互联网新闻的舆情聚类分析方法第六实施方式的流程图；

图 9 是本申请互联网新闻的舆情聚类分析方法第七实施方式的流程图。

10 本申请目的的实现、功能特点及优点将结合实施方式，参照附图做进一步说明。

具体实施方式

为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及
15 实施方式，对本申请进行进一步详细说明。应当理解，此处所描述的具体实施方式仅用以解释本申请，并不用于限定本申请。基于本申请中的实施方式，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方式，都属于本申请保护的范围。

需要说明的是，在本申请中涉及“第一”、“第二”等的描述仅用于描述目的，
20 而不能理解为指示或暗示其相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括至少一个该特征。另外，各个实施方式之间的技术方案可以相互结合，但是必须是以本领域普通技术人员能够实现为基础，当技术方案的结合出现相互矛盾或无法实现时应当认为这种技术方案的结合不存在，也不在本申请要求的保护范围
25 之内。

参阅图 1 所示，是本申请应用服务器 1 一可选的硬件架构的示意图。

本实施方式中，所述应用服务器 1 可包括，但不仅限于，可通过系统总线相互通信连接存储器 11、处理器 12、网络接口 13。需要指出的是，图 1 仅示出了具有组件 11-13 的应用服务器 1，但是应理解的是，并不要求实施所有示出的组件，可以替代的实施更多或者更少的组件。

5 其中，所述应用服务器 1 可以是机架式服务器、刀片式服务器、塔式服务器或机柜式服务器等计算设备，该应用服务器 1 可以是独立的服务器，也可以是多个服务器所组成的服务器集群。

所述存储器 11 至少包括一种类型的可读存储介质，所述可读存储介质包括闪存、硬盘、多媒体卡、卡型存储器（例如，SD 或 DX 存储器等）、随机访问存储器（RAM）、静态随机访问存储器（SRAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、可编程只读存储器（PROM）、磁性存储器、磁盘、光盘等。在一些实施方式中，所述存储器 11 可以是所述应用服务器 1 的内部存储单元，例如该应用服务器 1 的硬盘或内存。在另一些实施方式中，所述存储器 11 也可以是所述应用服务器 1 的外部存储设备，例如该应用服务器 1 上配备的插接式硬盘，智能存储卡（Smart Media Card, SMC），安全数字（Secure Digital, SD）卡，闪存卡（Flash Card）等。当然，所述存储器 11 还可以既包括所述应用服务器 1 的内部存储单元也包括其外部存储设备。本实施方式中，所述存储器 11 通常用于存储安装于所述应用服务器 1 的操作系统和各类应用软件，例如互联网新闻的舆情聚类分析系统 200 的程序代码等。此外，所述存储器 11 还可以用于暂时地存储已经输出或者将要输出的各类数据。

所述处理器 12 在一些实施方式中可以是中央处理器（Central Processing Unit, CPU）、控制器、微控制器、微处理器、或其他数据处理芯片。该处理器 12 通常用于控制所述应用服务器 1 的总体操作。本实施方式中，所述处理器 12 用于运行所述存储器 11 中存储的程序代码或者处理数据，例如运行所述的互联网新闻的舆情聚类分析系统 200 等。

所述网络接口 13 可包括无线网络接口或有线网络接口，该网络接口 13 通常用于在所述应用服务器 1 与其他电子设备之间建立通信连接。

至此，已经详细介绍了本申请相关设备的硬件结构和功能。下面，将基于上述介绍提出本申请的各个实施方式。

5 首先，本申请提出一种互联网新闻的舆情聚类分析系统 200。

参阅图 2 所示，是本申请互联网新闻的舆情聚类分析系统 200 第一实施方式的程序模块图。

在一实施方式中，所述互联网新闻的舆情聚类分析系统 200 包括一系列的存储于存储器 11 上的计算机程序指令，当该计算机程序指令被处理器 12
10 执行时，可以实现本申请各实施方式的互联网新闻的舆情聚类分析操作。在一些实施方式中，基于该计算机程序指令各部分所实现的特定的操作，互联网新闻的舆情聚类分析系统 200 可以被划分为一个或多个模块。例如，在图 2 中，所述互联网新闻的舆情聚类分析系统 200 可以被分割成获取模块 21、处理模块 22、归纳模块 23 及输出显示模块 24。其中：

15 所述获取模块 21，用于通过分布式爬虫在信息源获取新闻类信息，并存储到舆情数据库中。

具体地，所述分布式爬虫可对分布于不同服务器上的网页进行多任务抓取，提高信息的抓取效率。所述分布式爬虫框架主要分成两部分：下载器和解析器。下载器负责抓取网页，解析器负责解析网页并入库。两者之间依靠
20 消息队列进行通信，两者可以分布在不同机器，也可分布在同一台机器。两者的数量也是灵活可变的，例如可能有五台机在做下载、两台机在做解析，这都是可以根据爬虫系统的状态及时调整的。

具体地，所述消息队列有两个管道：HTML/JS 文件和待爬种子。下载器从待爬种子里拿到一条种子，根据种子信息调用相应的抓取模块进行网页抓
25 取，然后存入 HTML/JS 文件这个通道；解析器从 HTML/JS 文件里拿到一条网页内容，根据里面的信息调用相应的解析模块进行解析，将目标字段入库，

需要的话还会解析出新的待爬种子。

具体地，所述下载器包含 User-Agent 池、Proxy 池、Cookie 池的，可以适应复杂网站的抓取。

具体地，所述信息源包括但不限于新闻网站、微博、微信、贴吧、论坛等 5 等平台。其中，新闻网站包括但不限于新浪、网易、凤凰新闻、新华网、人民网、腾讯网等国内主要新闻网站及各地地方网络报纸等。

所述处理模块 22，用于对所述舆情数据库中的数据进行去噪、分词、聚类。

所述归纳模块 23，用于对聚类后的不同类新闻分别归纳主题摘要。

10 具体地，举例而言，对同一事件不同新闻网站、媒体会有不同角度、形式的报道，但各种报道的主题都是大同小异的。新闻事件中的新闻摘要为该新闻内容的浓缩，目的是在用户阅读了新闻标题后，进一步了解新闻相关的重要信息，以便决定是否进一步阅读新闻的详细内容。用户阅读新闻大多利用手机，由于手机屏幕小，为了使有限的文字传递给用户的信息最大化的同时，15 尽可能减少重复信息，对聚类后的不同类新闻的主题摘要进行智能、自动归纳，从而呈现给用户，可节省用户的事件及增强用户的体验度。

所述输出显示模块 24，用于将聚类后的新闻及所述主题摘要输出，并显示给用户。

具体地，所述输出显示模块 24 包括 LCD、LED 等显示屏。

20 此外，本申请还提出一种互联网新闻的舆情聚类分析方法。

参阅图 3 所示，是本申请互联网新闻的舆情聚类分析方法第一实施方式的流程图示意图。在本实施方式中，根据不同的需求，图 3 所示的流程图中的步骤的执行顺序可以改变，某些步骤可以省略。

步骤 S110，通过分布式爬虫在信息源获取新闻类信息，并存储到舆情数 25 据库中。

步骤 S120，对所述舆情数据库中的数据进行去噪、分词、聚类。

步骤 S130, 对聚类后的不同类新闻分别归纳主题摘要。

具体地, 所述归纳方法包括步骤:

步骤一: 对该新闻的正文进行分句, 并保留句子长度在预设长度范围内的句子, 记为保留句子。

5 具体地, 通过该步骤可以限定句子的长度, 从而限定了标题的长度, 同时, 选择预设长度范围内的句子做保留句子也有方便处理的好处。

步骤二: 分别计算每个保留句子与标题的相似度 $S(s)$, 以及每个保留句子的权重 $Q(s)$ 。其中, 引入保留句子与标题的相似度是为了使最后选取的摘要与标题的相似度低, 而句子的权重则表明该句子在该新闻中的价值, 通常是
10 句子包含的关键词越多, 则其价值越大。

其中, 计算保留句子与标题的相似度 $S(s)$ 的步骤如下: 步骤 1, 基于同义词词库对保留句子和标题进行同义词转换; 步骤 2, 针对同义词转换后的保留句子和标题采用 Jaccard 距离计算保留句子和标题的相似度 $S(s)$ 。即将保留句子和标题中的词组的交集除以词组的并集得到相似度 $S(s)$ 。

15 步骤三, 根据公式 $R(s) = Q(s)/S(S)$ 计算每个保留句子的排序分, 其中, $R(s)$ 为保留句子的排序分。通过上述公式, 排序分越高, 则对应的句子越可能成为摘要。

步骤四, 选取排序分最高的保留句子作为同类新闻的摘要。

具体地, 排序分最高的保留句子代表这些保留句子越能代表新闻的主要内容, 也可以设定一个阈值, 当排序分大于所述阈值, 就将该排序分所对应的句子作为保留句子, 最后经过分析处理后聚合形成摘要。

步骤 S140, 将聚类后的新闻及所述主题摘要输出, 并显示给用户。

如图 4 所示, 是本申请互联网新闻的舆情聚类分析方法的第二实施方式
25 的流程示意图。本实施方式中, 所述步骤“对所述舆情数据库中的数据进行去噪、分词、聚类”中去噪具体包括以下步骤:

步骤 S210, 过滤图片、版权说明、广告, 获得文档信息。

具体地，采集的网页信息含有广告、导航信息、图片、版权说明等噪声数据，对舆情信息分析来说真正需要的是正文部分的元信息，清除掉这些无关内容，保留采集网页中的文档信息。

步骤 S220，过滤停用词。

- 5 具体地，在文本信息中，还包括很多无意义的词语、符号等，这些词统称为停用词。设置一个无意词表，在无意词表中添加无意义的词语、符号，比如“了”，“的”，“和”，“或”，符号等，将无意词表中的词语、符号从上一步骤获得的文档信息中去除。

10 如图 5 所示，是本申请互联网新闻的舆情聚类分析方法的第三实施方式
的流程示意图。本实施方式中，所述步骤“对所述舆情数据库中的数据进行
去噪、分词、聚类”中分词具体包括以下步骤：

步骤 S310，运用中文分词技术对采集到的网页文本数据进行分词。

- 具体地，所述的中文分词技术包括两类，第一类方法应用词典匹配、汉语词法或其它汉语语言知识进行分词，如：最大匹配法、最小分词方法等。
15 第二类基于统计的分词方法则基于字和词的统计信息，如把相邻字间的信息、
词频及相应的共现信息等应用于分词。

具体地，所述的中文分词技术包括以下方法：

- 最大正向匹配法，其基本思想为：假定分词词典中的最长词有 i 个汉字字符，则用被处理文档的当前字串中的前 i 个字作为匹配字段，查找字典。若字典中存在这样的一个 i 字词，则匹配成功，匹配字段被作为一个词切分出来。
20 如果词典中找不到这样的一个 i 字词，则匹配失败，将匹配字段中的最后一个字去掉，对剩下的字串重新进行匹配处理……如此进行下去，直到匹配成功，
即切分出一个词或剩余字串的长度为零为止。这样就完成了一轮匹配，然后
取下一个 i 字字串进行匹配处理，直到文档被扫描完为止。

- 25 逆向最大匹配法，逆向最大匹配法从被处理文档的末端开始匹配扫描，
每次取最末端的 $2i$ 个字符（ i 字字串）作为匹配字段，若匹配失败，则去掉匹

配字段最前面的一个字，继续匹配。相应地，它使用的分词词典是逆序词典，其中的每个词条都将按逆序方式存放。在实际处理时，先将文档进行倒排处理，生成逆序文档。然后，根据逆序词典，对逆序文档用正向最大匹配法处理即可。

- 5 如图 6 所示，是本申请互联网新闻的舆情聚类分析方法的第四实施方式的流程示意图。本实施方式中，所述步骤“对所述舆情数据库中的数据进行去噪、分词、聚类”中聚类具体包括以下步骤：

步骤 S410，设置包括敏感词、情感词的参照表。

- 具体地，预设的情感词包括具有强烈情感的副词、连词以及观点词。例如，连词包括不过、但是、于是、此外等等；副词包括相当、完美、几乎、绝对等等；观点词包括察觉、发现、认为、主张、猜想、表示、以为等等。而敏感词，可为禁止词，侵权词，不雅词，政治性、煽动性的词语。

步骤 S420，获得关键词，并根据获得的关键词设置关键词参照表。

具体地，该步骤可包括如下步骤：

- 15 步骤一：对分词之后的文档进行分析，统计词语出现的频率，出现的位置（标题，摘要，正文，备注），历史平均频率（若存在）；

步骤二：根据公式获得词语的重要度 D；

$$D=a*F_n+\sum b_i*W_i+c*F_h, i=1,2,3\dots n,$$

- 其中，a, b, c 为词语当时出现的频率，位置，历史平均频率对应权重值，
20 F_n, W_i, F_h 分别对应词语出现的频率，出现的位置，历史平均频率，对于词语出现的位置也设置不同的权重值；

步骤三：根据重要度 D 对各词语进行排序，将重要度 D 大于预设值的词语作为关键词并生成关键词参照表。

- 25 步骤 S430，对照所述敏感词、情感词及关键词参照表分析出网页文本中的关键词、敏感词和带有情感倾向的词语。

进一步地，对比近义词、同义词数据库对所述敏感词、情感词及关键词

进行分析，将关键词及其同义词、近义词统一列为关键词，将敏感词、情感词的同义词也进行对比分析。

进一步地，位于数据库服务器词库中的敏感词、关键词和情感词，也可以根据社会舆情发展变化的需要由管理员添加新的词汇到此词库，以实现词库的实时更新。

进一步地，依据处理结果提供的关键词以及情感词实现对网页信息的过滤、预警（即对网页文本信息进行标记如依照关键词标记或依照情感倾向词标记），处理结果提交给敏感词处理模块进行敏感词处理；敏感词处理模块，用于根据国家相关法律法规，对经关键词处理模块处理的文本数据中的敏感词进行过滤、屏蔽，处理结果交由聚类分析模块进行处理，其中敏感词来自分词处理模块提供的处理结果。

步骤 S440，根据关键词、情感词、敏感词将从获得的网页数据按照网页所属类别自动聚类。

具体地，根据获得的关键词对数据库中的新闻网页进行分类，例如可为政治、经济、军事、社会、科技、游戏、时尚、体育及电影等；若关键词为地名，也可根据不同地名分类，例如可为北京、上海、广州、深圳，香港、台湾等等；

具体地，根据情感词可对数据库中的新闻网页进行分类，例如可分为爱情、亲情、友情、失恋、治愈等等。

具体地，根据敏感词可将数据库中的新闻网页进行分类，例如对敏感政策、敏感地域、敏感人物的分类。

进一步地，聚类之后的各类包含多个新闻的分组。

步骤 S450，对聚类后的新闻按照热度进行排序。

具体地，还可获取新闻的评论数、点击数、转发数，根据以上数据进行排序。

具体地，还可根据同一或者相似新闻的数目计算热度，对不同聚类出的

不同新闻分别计算热度后根据热度排序。

如图 7 所示,是本申请互联网新闻的舆情聚类分析方法的第五实施方式
的流程示意图。本实施方式中,所述互联网新闻的舆情聚类分析方法的步骤
“获得所述关键词,并根据获得的所述关键词设置关键词参照表”具体包括:

5 步骤 S510,对分词之后的文档进行分析,统计词语出现的频率,出现的
位置及历史平均频率。

步骤 S520,根据如下公式获得词语的重要度 D:

$$D=a*F_n+\sum b_i*W_i+c*F_h, i=1,2,3\dots n。$$

步骤 S530,根据所述重要度 D 对各词语进行排序,将所述重要度 D 大于
10 预设值的词语作为所述关键词并生成所述关键词参照表。

其中,a,b,c 为词语当时出现的频率,位置及历史平均频率对应的权重值;
F_n,W_i,F_h 分别对应词语出现的频率,出现的位置,所述历史平均频率。

如图 8 所示,是本申请互联网新闻的舆情聚类分析方法的第六实施方式
的流程示意图。本实施方式中,所述互联网新闻的舆情聚类分析方法的步骤
15 “对聚类后的不同类新闻分别归纳主题摘要”具体包括:

步骤 S610,对该新闻的正文进行分句,并保留句子长度在预设长度范围
内的句子,记为保留句子。

具体地,通过该步骤可以限定句子的长度,从而限定了标题的长度。

步骤 S620,分别计算所述保留句子与标题的相似度 S(s),以及所述保留
20 句子的权重 Q(s)。

具体地,引入保留句子与标题的相似度是为了使最后选取的摘要与标题
的相似度低,而句子的权重则表明该句子在该新闻中的价值,通常是句子包
含的关键词越多,则其价值越大。

步骤 S630,根据公式 $R(s) = Q(s)/S(S)$ 计算所述保留句子的排序分。

25 步骤 S640,选取排序分最高的所述保留句子作为同类新闻的摘要。

如图 9 所示,是本申请互联网新闻的舆情聚类分析方法的第七实施方式

的流程示意图。本实施方式中，所述互联网新闻的舆情聚类分析方法的步骤“分别计算所述保留句子与标题的相似度 $S(s)$ ，以及所述保留句子的权重 $Q(s)$ ”具体包括：

步骤 S710，基于同义词词库对所述保留句子和标题进行同义词转换。

5 步骤 S720，针对同义词转换后的所述保留句子和标题采用 Jaccard 距离计算保留句子和标题的相似度 $S(s)$ 。

具体地，针对同义词转换后的所述保留句子和标题采用 Jaccard 距离计算保留句子和标题的相似度 $S(s)$ ，即将保留句子和标题中的词组的交集除以词组的并集得到相似度 $S(s)$ 。

10 上述本申请实施方式序号仅仅为了描述，不代表实施方式的优劣。

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施方式方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质（如 ROM/RAM、磁碟、
15 光盘）中，包括若干指令用以使得一台终端设备（可以是手机，计算机，服务器，空调器，或者网络设备）执行本申请各个实施方式所述的方法。

以上仅为本申请的优选实施方式，并非因此限制本申请的专利范围，凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本申请的专利保护范围内。
20

权利要求书

1. 一种互联网新闻的舆情聚类分析方法，应用于应用服务器，其特征在于，所述方法包括步骤：

通过分布式爬虫在信息源获取新闻类信息，并存储到舆情数据库中；

5 对所述舆情数据库中的数据进行去噪、分词、聚类；

对聚类后的不同类新闻分别归纳主题摘要；及

将聚类后的新闻及所述主题摘要输出，并显示给用户。

2. 如权利要求 1 所述的互联网新闻的舆情聚类分析方法，其特征在于，所述信息源包括新闻网站、微博、微信、贴吧及论坛。

10 3. 如权利要求 1 所述的互联网新闻的舆情聚类分析方法，其特征在于，所述去噪步骤包括：

过滤图片、版权说明、广告，获得文档信息；及

过滤停用词。

15 4. 如权利要求 1 所述的互联网新闻的舆情聚类分析方法，其特征在于，所述的分词步骤包括：

运用中文分词技术对采集到的所述新闻类信息进行分词。

5. 如权利要求 1 所述的互联网新闻的舆情聚类分析方法，其特征在于，所述聚类步骤包括：

设置包括敏感词、情感词的参照表；

20 获得关键词，并根据获得的关键词设置关键词参照表；

对照所述敏感词、情感词及关键词参照表分析出所述新闻类信息中的关键词、敏感词和带有情感倾向的词语；

根据关键词、情感词、敏感词将所述新闻类信息按照网页所属类别自动聚类；及

25 对聚类后的新闻按照热度进行排序。

6. 如权利要求 5 所述的互联网新闻的舆情聚类分析方法，其特征在于，

所述获得所述关键词，并根据获得的所述关键词设置关键词参照表的步骤还包括：

对分词之后的所述新闻类信息进行分析，统计词语出现的频率，出现的位置及历史平均频率；

5 根据如下公式获得词语的重要度 D：

$$D=a*F_n+\sum b_i*W_i+c*F_h, i=1,2,3\dots n; \text{ 及}$$

根据所述重要度 D 对各词语进行排序，将所述重要度 D 大于预设值的词语作为所述关键词并生成所述关键词参照表；

其中，a, b, c 为词语当时出现的频率，位置及历史平均频率对应的权重值；

10 F_n, W_i, F_h 分别对应词语出现的频率，出现的位置，所述历史平均频率。

7. 如权利要求 1 所述的互联网新闻的舆情聚类分析方法，其特征在于，所述对聚类后的不同类新闻分别归纳主题摘要的步骤还包括：

对该新闻的正文进行分句，并保留句子长度在预设长度范围内的句子，记为保留句子；

15 分别计算所述保留句子与标题的相似度 $S(s)$ ，以及所述保留句子的权重 $Q(s)$ ；

根据公式 $R(s) = Q(s)/S(S)$ 计算所述保留句子的排序分；及

选取排序分最高的所述保留句子作为同类新闻的摘要；

其中， $R(s)$ 为所述保留句子的排序分。

20 8. 如权利要求 7 所述的互联网新闻的舆情聚类分析方法，其特征在于，计算所述相似度 $S(s)$ 的步骤如下：

基于同义词词库对所述保留句子和标题进行同义词转换；及

针对同义词转换后的所述保留句子和标题采用 Jaccard 距离计算保留句子和标题的相似度 $S(s)$ 。

25 9. 一种应用服务器，其特征在于，所述应用服务器包括存储器、处理器，所述存储器上存储有可在所述处理器上运行的互联网新闻的舆情聚类分析系

统, 所述互联网新闻的舆情聚类分析系统被所述处理器执行时实现如下步骤:

通过分布式爬虫在信息源获取新闻类信息, 并存储到舆情数据库中;

对所述舆情数据库中的数据进行去噪、分词、聚类;

对聚类后的不同类新闻分别归纳主题摘要; 及

5 将聚类后的新闻及所述主题摘要输出, 并显示给用户。

10. 如权利要求 9 所述的应用服务器, 其特征在于, 所述去噪步骤包括:

过滤图片、版权说明、广告, 获得文档信息; 及

过滤停用词。

11. 如权利要求 9 所述的应用服务器, 其特征在于, 所述的分词步骤包括:

10 运用中文分词技术对采集到的所述新闻类信息进行分词。

12. 如权利要求 9 所述的应用服务器, 其特征在于, 所述聚类步骤包括:

设置包括敏感词、情感词的参照表;

获得关键词, 并根据获得的关键词设置关键词参照表;

15 对照所述敏感词、情感词及关键词参照表分析出所述新闻类信息中的关键词、敏感词和带有情感倾向的词语;

根据关键词、情感词、敏感词将所述新闻类信息按照网页所属类别自动聚类; 及

对聚类后的新闻按照热度进行排序。

20 13. 如权利要求 12 所述的应用服务器, 其特征在于, 所述获得所述关键词, 并根据获得的所述关键词设置关键词参照表的步骤还包括:

对分词之后的所述新闻类信息进行分析, 统计词语出现的频率, 出现的位置及历史平均频率;

根据如下公式获得词语的重要度 D:

$$D=a*Fn+\sum b_i*W_i+c*F_h, i=1,2,3\dots n;$$
 及

25 根据所述重要度 D 对各词语进行排序, 将所述重要度 D 大于预设值的词语作为所述关键词并生成所述关键词参照表;

其中, a, b, c 为词语当时出现的频率, 位置及历史平均频率对应的权重值;
 F_n, W_i, F_h 分别对应词语出现的频率, 出现的位置, 所述历史平均频率。

14. 如权利要求 9 所述的应用服务器, 其特征在于, 所述对聚类后的不同类新闻分别归纳主题摘要的步骤还包括:

5 对该新闻的正文进行分句, 并保留句子长度在预设长度范围内的句子, 记为保留句子;

分别计算所述保留句子与标题的相似度 $S(s)$, 以及所述保留句子的权重 $Q(s)$;

根据公式 $R(s) = Q(s)/S(s)$ 计算所述保留句子的排序分; 及

10 选取排序分最高的所述保留句子作为同类新闻的摘要;

其中, $R(s)$ 为所述保留句子的排序分。15. 如权利要求 14 所述的应用服务器, 其特征在于, 计算所述相似度 $S(s)$ 的步骤如下:

基于同义词词库对所述保留句子和标题进行同义词转换; 及

15 针对同义词转换后的所述保留句子和标题采用 Jaccard 距离计算保留句子和标题的相似度 $S(s)$ 。

16. 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互联网新闻的舆情聚类分析系统, 所述互联网新闻的舆情聚类分析系统可被至少一个处理器执行, 以使所述至少一个处理器执行如下步骤:

通过分布式爬虫在信息源获取新闻类信息, 并存储到舆情数据库中;

20 对所述舆情数据库中的数据进行去噪、分词、聚类;

对聚类后的不同类新闻分别归纳主题摘要; 及

将聚类后的新闻及所述主题摘要输出, 并显示给用户。

17. 如权利要求 16 所述的计算机可读存储介质, 其特征在于, 所述聚类步骤包括:

25 设置包括敏感词、情感词的参照表;

获得关键词, 并根据获得的关键词设置关键词参照表;

对照所述敏感词、情感词及关键词参照表分析出所述新闻类信息中的关键词、敏感词和带有情感倾向的词语；

根据关键词、情感词、敏感词将所述新闻类信息按照网页所属类别自动聚类；及

5 对聚类后的新闻按照热度进行排序。

18. 如权利要求 17 所述的计算机可读存储介质，其特征在于，所述获得所述关键词，并根据获得的所述关键词设置关键词参照表的步骤还包括：

对分词之后的所述新闻类信息进行分析，统计词语出现的频率，出现的位置及历史平均频率；

10 根据如下公式获得词语的重要度 D：

$$D=a*F_n+\sum b_i*W_i+c*F_h, i=1,2,3\dots n; \text{ 及}$$

根据所述重要度 D 对各词语进行排序，将所述重要度 D 大于预设值的词语作为所述关键词并生成所述关键词参照表；

其中，a, b, c 为词语当时出现的频率，位置及历史平均频率对应的权重值；

15 F_n, W_i, F_h 分别对应词语出现的频率，出现的位置，所述历史平均频率。

19. 如权利要求 16 所述的计算机可读存储介质，其特征在于，所述对聚类后的不同类新闻分别归纳主题摘要的步骤还包括：

对该新闻的正文进行分句，并保留句子长度在预设长度范围内的句子，记为保留句子；

20 分别计算所述保留句子与标题的相似度 $S(s)$ ，以及所述保留句子的权重 $Q(s)$ ；

根据公式 $R(s) = Q(s)/S(S)$ 计算所述保留句子的排序分；及

选取排序分最高的所述保留句子作为同类新闻的摘要；

其中， $R(s)$ 为所述保留句子的排序分。

25 20. 如权利要求 19 所述的计算机可读存储介质，其特征在于，计算所述相似度 $S(s)$ 的步骤如下：

基于同义词词库对所述保留句子和标题进行同义词转换；及

针对同义词转换后的所述保留句子和标题采用 Jaccard 距离计算保留句子和标题的相似度 $S(s)$ 。

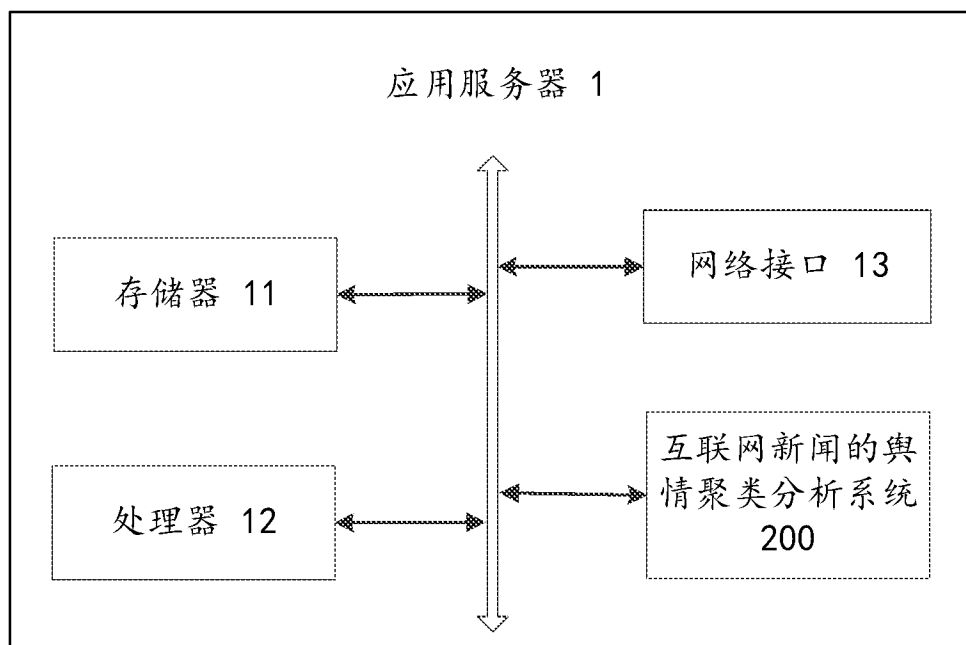


图 1

2

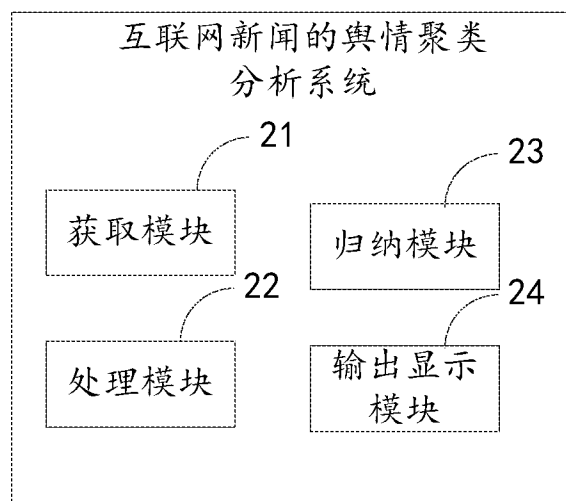


图 2

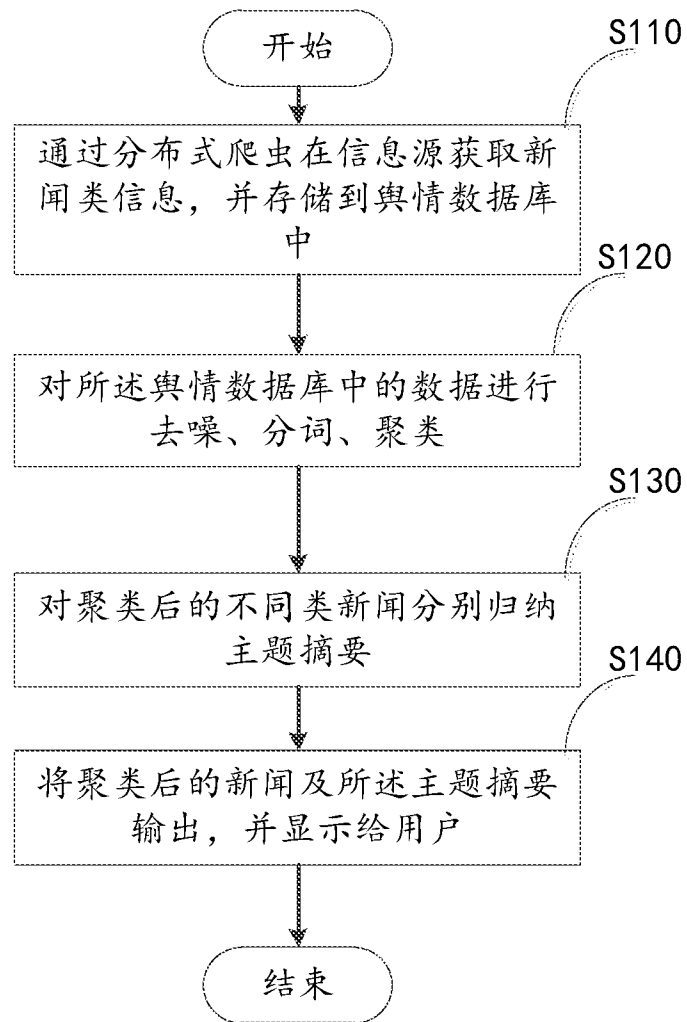


图 3

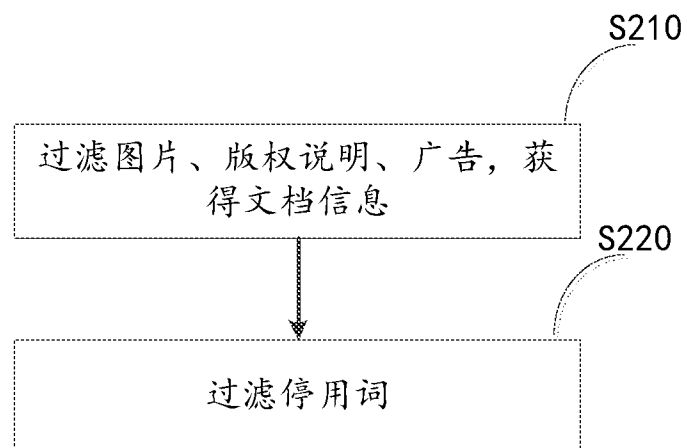


图 4

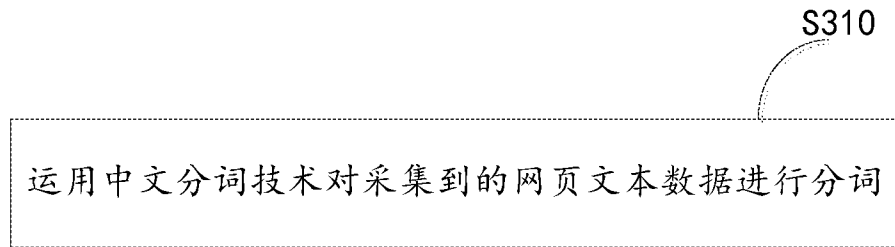


图 5

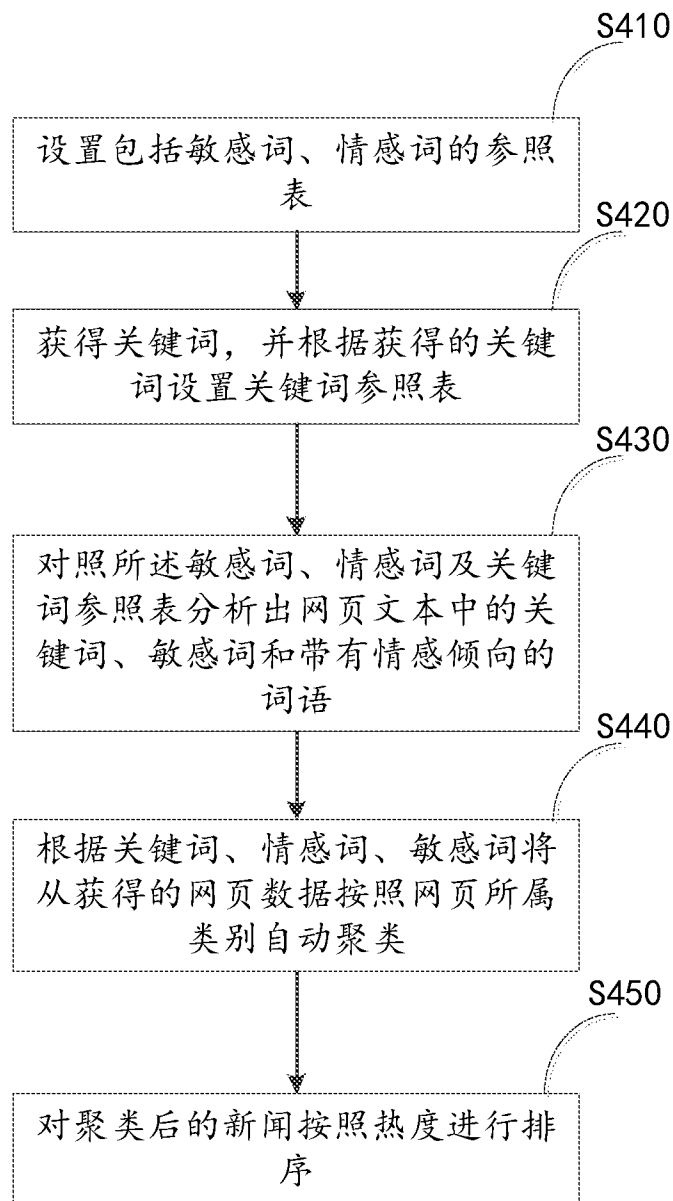


图 6

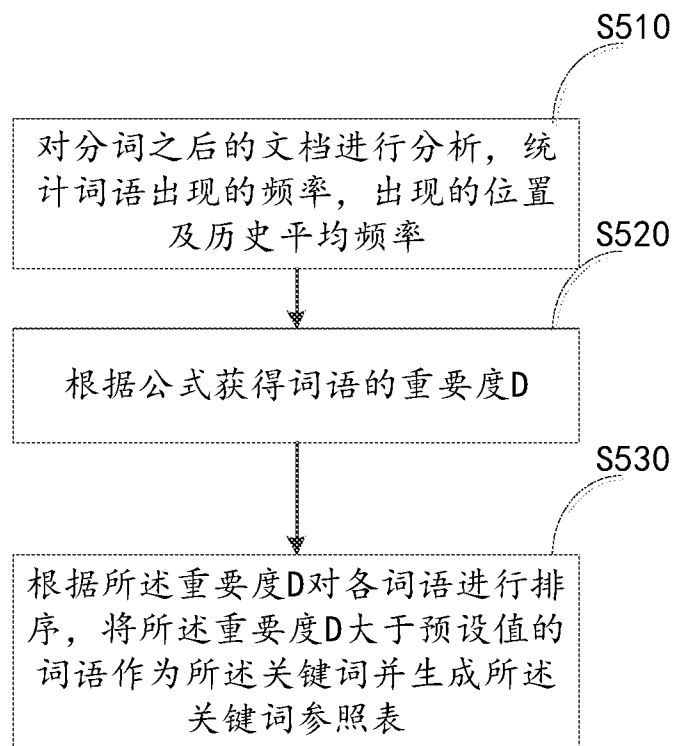


图 7

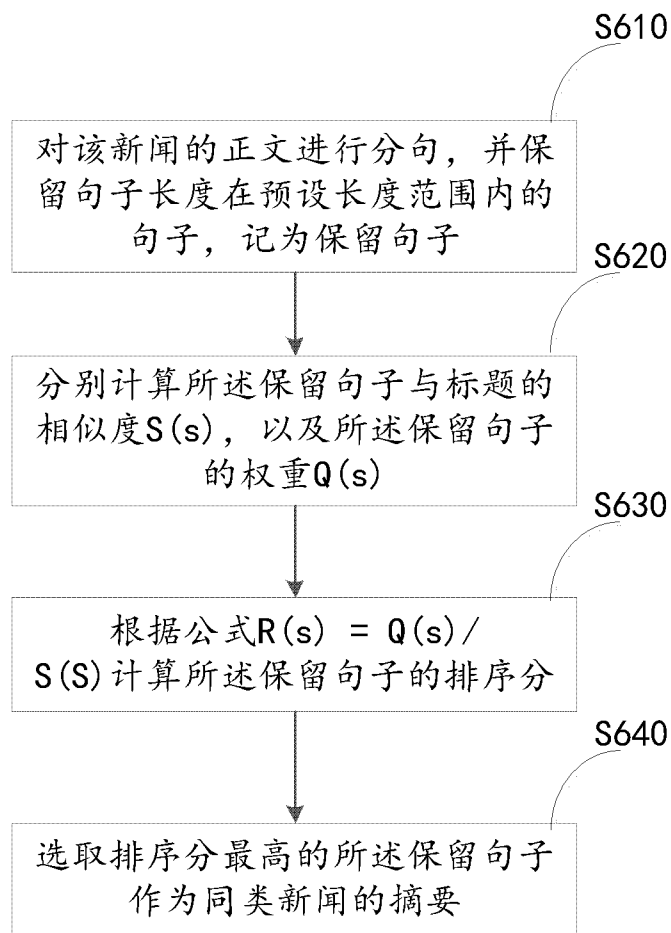


图 8

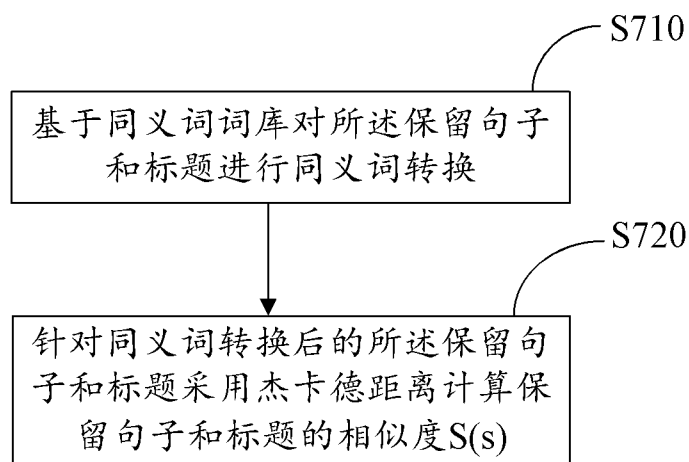


图 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/077638

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (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, G06Q

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)

WPI, EPODOC, CNPAT, CNKI: 爬取, 新闻, 敏感, 情感, 摘要, 聚类, 分词, 去噪, 爬虫, 舆情, 舆论, public sentiment, news, emotional, abstract, cluster, crawler

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 105824959 A (CAPINFO CO., LTD.) 03 August 2016 (03.08.2016), description, paragraphs [0024]-[0063]	1-20
X	CN 103544255 A (CHANGZHOU UNIVERSITY) 29 January 2014 (29.01.2014), description, paragraphs [0021]-[0088]	1-20
A	CN 102708096 A (DAI, Song) 03 October 2012 (03.10.2012), entire document	1-20
A	CN 104933093 A (WUHAN STORMOR WISDOM TECHNOLOGY CO., LTD.) 23 September 2015 (23.09.2015), entire document	1-20

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

☒ See patent family annex.

* Special categories of cited documents:	“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
“A” document defining the general state of the art which is not considered to be of particular relevance	“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
“E” earlier application or patent but published on or after the international filing date	“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
“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)	“&” document member of the same patent family
“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	

Date of the actual completion of the international search 18 July 2018	Date of mailing of the international search report 03 August 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 DING, Wenjing Telephone No. (86-10) 53961330

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/077638

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 105824959 A	03 August 2016	None	
CN 103544255 A	29 January 2014	None	
CN 102708096 A	03 October 2012	None	
CN 104933093 A	23 September 2015	None	

国际检索报告

国际申请号

PCT/CN2018/077638

A. 主题的分类

G06F 17/30(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, G06Q

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))

WPI, EPODOC, CNPAT, CNKI: 爬取, 新闻, 敏感, 情感, 摘要, 聚类, 分词, 去噪, 爬虫, 舆情, 舆论, public sentiment, news, emotional, abstract, cluster, crawler

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 105824959 A (首都信息发展股份有限公司) 2016年 8月 3日 (2016 - 08 - 03) 说明书第[0024]-[0063]段	1-20
X	CN 103544255 A (常州大学) 2014年 1月 29日 (2014 - 01 - 29) 说明书第[0021]-[0088]段	1-20
A	CN 102708096 A (代松) 2012年 10月 3日 (2012 - 10 - 03) 全文	1-20
A	CN 104933093 A (武汉泰迪智慧科技有限公司) 2015年 9月 23日 (2015 - 09 - 23) 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 7月 18日

国际检索报告邮寄日期

2018年 8月 3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

受权官员

丁文勃

电话号码 86-(10)-53961330

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/077638

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	105824959	A	2016年 8月 3日	无	
CN	103544255	A	2014年 1月 29日	无	
CN	102708096	A	2012年 10月 3日	无	
CN	104933093	A	2015年 9月 23日	无	