

(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
2017 年 2 月 9 日 (09.02.2017)



(10) 国际公布号
WO 2017/020451 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2015/095754
- (22) 国际申请日: 2015 年 11 月 27 日 (27.11.2015)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201510483126.3 2015 年 8 月 3 日 (03.08.2015) CN
- (71) 申请人: 百度在线网络技术(北京)有限公司
(BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) [CN/CN]; 中国北京市海淀区上地十街 10 号百度大厦三层, Beijing 100085 (CN)。
- (72) 发明人: 裘皓萍 (QIU, Haoping); 中国北京市海淀区上地十街 10 号百度大厦三层, Beijing 100085 (CN)。
陈炜于 (CHEN, Weiyu); 中国北京市海淀区上地十街 10 号百度大厦三层, Beijing 100085 (CN)。
- (74) 代理人: 北京英赛嘉华知识产权代理有限公司 (INSIGHT INTELLECTUAL PROPERTY LIM-

ITED); 中国北京市海淀区知春路甲 48 号盈都大厦 A 座 19A, Beijing 100098 (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, 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。
- (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)。

[见续页]

(54) Title: INFORMATION PUSH METHOD AND DEVICE

(54) 发明名称: 信息推送方法和装置

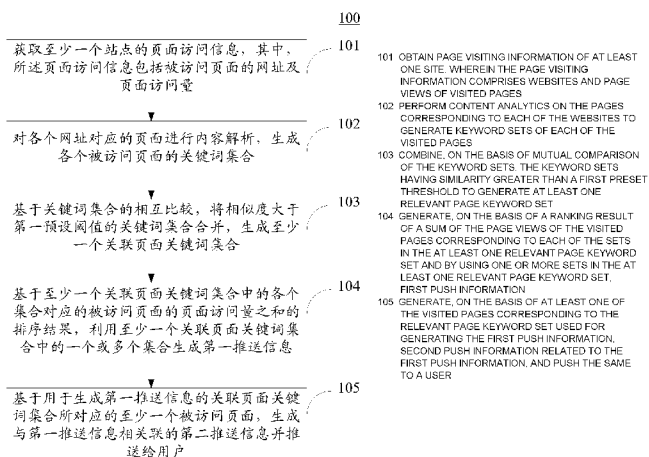


图 1

(57) Abstract: Provided are an information push method and device. A specific embodiment of the method comprises: obtaining websites and page views of visited pages of at least one site (101); performing content analytics on the pages corresponding to each of the websites to generate keyword sets of each of the visited pages (102); combining, on the basis of mutual comparison of the keyword sets, the keyword sets having similarity greater than a first preset threshold to generate at least one relevant page keyword set (103); generating, on the basis of a ranking result of a sum of the page views of the visited pages corresponding to each of the sets in the at least one relevant page keyword set and by using one or more sets in the at least one relevant page keyword set, first push information (104); and generating, on the basis of at least one of the visited pages corresponding to the relevant page keyword set used for generating the first push information, second push information related to the first push information, and pushing the same to a user (105). The embodiment can enrich content of the push information.

(57) 摘要:

[见续页]



WO 2017/020451 A1

**本国际公布:**

— 包括国际检索报告(条约第 21 条(3))。

一种信息推送方法和装置。所述方法的一具体实施方式包括：获取至少一个站点的被访问页面的网址及页面访问量(101)；对各个网址对应的页面进行内容解析，生成各个被访问页面的关键词集合(102)；基于关键词集合的相互比较，将相似度大于第一预设阈值的关键词集合合并，生成至少一个关联页面关键词集合(103)；基于至少一个关联页面关键词集合中各个集合对应的被访问页面的页面访问量之和的排序结果，利用所述至少一个关联页面关键词集合中的一个或多个集合生成第一推送信息(104)；基于用于生成所述第一推送信息的关联页面关键词集合所对应的至少一个被访问页面，生成与所述第一推送信息相关联的第二推送信息并推送给用户(105)。该实施方式可以丰富推送信息的内容。

信息推送方法和装置

相关申请的交叉引用

本申请要求于 2015 年 08 月 03 日提交的中国专利申请号为“201510483126.3”的优先权，其全部内容作为整体并入本申请中。

技术领域

本申请涉及计算机技术领域，具体涉及互联网技术领域，尤其涉及一种信息推送方法和装置。

背景技术

信息推送，又称为“网络广播”，是通过一定的技术标准或协议，在互联网上通过推送用户需要的信息来减少信息过载的一项技术。信息推送技术通过主动推送信息给用户，可以减少用户在网络上搜索所花的时间。

然而，在现有的信息推送技术中，推送给用户的信息往往是一条或多条相互独立的信息，缺乏信息之间的关联性。如果所推送信息是某一事件进展的片段，难以通过所推送的内容使用户了解所推送信息的事件背景或发展过程。因此，这种信息推送技术存在着网络信息相关数据利用不足，推送信息内容不够丰富的问题。

发明内容

本申请的目的在于提出一种改进的信息推送方法和装置，来解决以上背景技术部分提到的技术问题。

一方面，本申请提供了一种信息推送方法，所述方法包括：获取至少一个站点的页面访问信息，其中，所述页面访问信息包括被访问页面的网址及页面访问量；对各个网址对应的页面进行内容解析，生成各个被访问页面的关键词集合；基于关键词集合的相互比较，将相

似度大于第一预设阈值的关键词集合合并，生成至少一个关联页面关键词集合，其中，用于生成关联页面关键词集合的关键词集合对应的被访问页面互为关联页面；基于所述至少一个关联页面关键词集合中的各个集合对应的被访问页面的页面访问量之和的排序结果，利用所述至少一个关联页面关键词集合中的一个或多个集合生成第一推送信息；基于用于生成所述第一推送信息的关联页面关键词集合所对应的至少一个被访问页面，生成与所述第一推送信息相关联的第二推送信息并推送给用户。

在一些实施例中，所述基于用于生成所述第一推送信息的关联页面关键词集合所对应的至少一个被访问页面，生成与所述第一推送信息相关联的第二推送信息并推送给用户，包括：对用于生成所述第一推送信息的关联页面关键词集合所对应的被访问页面的发布时间按照预设的时间间隔进行聚类，划分成至少一个时间段，其中，当所述至少一个时间段包括两个以上的时间段时，分别取自任意两个时间段的发布时间之间的时间差大于所述时间间隔；对于所述至少一个时间段中的一个或多个时间段，分别从每个时间段所对应的被访问页面中提取一个页面；基于所提取的页面，生成第二推送信息并推送给用户。

在一些实施例中，所述对用于生成所述第一推送信息的关联页面关键词集合所对应的被访问页面的发布时间按照预设的时间间隔进行聚类，划分成至少一个时间段之前，还包括：对于关联页面关键词集合所对应的被访问页面，将相似度大于第二预设阈值的关键词集合所对应的被访问页面筛选至一个页面，将筛选页面后剩余的被访问页面作为关联页面关键词集合所对应的被访问页面，其中，所述第二预设阈值大于第一预设阈值。

在一些实施例中，所述对各个网址对应的页面进行内容解析，生成各个被访问页面的关键词集合包括：对所述被访问页面的内容进行统计分析和/或语义分析，提取至少一个关键词；基于所述至少一个关键词，生成关键词集合。

在一些实施例中，所述基于所述至少一个关键词，生成关键词集合包括：对于每个所述至少一个关键词中的单个关键词，进行扩展以

生成扩展关键词，其中，所述扩展关键词包括以下至少一项：所述单个关键词的同义词、所述单个关键词的近义词、所述单个关键词的关联词；基于所述至少一个关键词和所述扩展关键词，生成关键词集合。

在一些实施例中，将满足以下条件之一的关键词集合作为相似度大于第一预设阈值的关键词集合：相同关键词的个数大于个数阈值；相同关键词的个数与进行比较的关键词集合中关键词的总个数的比值大于比值阈值。

在一些实施例中，所述关键词集合中的各关键词还具有重要度系数，以及，所述基于关键词集合的相互比较，将相似度大于第一预设阈值的关键词集合合并，生成至少一个关联页面关键词集合包括：基于所述重要度系数对不同的关键词集合进行相似度计算；将相似度大于相似度阈值的关键词集合合并，生成关联页面关键词集合。

第二方面，本申请提供了一种信息推送装置，所述装置包括：信息获取模块，配置用于获取至少一个站点的页面访问信息，其中，所述页面访问信息包括被访问页面的网址及页面访问量；关键词集合生成模块，配置用于对各个网址对应的页面进行内容解析，生成各个被访问页面的关键词集合；关键词集合合并模块，配置用于基于关键词集合的相互比较，将相似度大于第一预设阈值的关键词集合合并，生成至少一个关联页面关键词集合，其中，用于生成关联页面关键词集合的关键词集合对应的被访问页面互为关联页面；第一推送信息生成模块，配置用于基于所述至少一个关联页面关键词集合中的各个集合对应的被访问页面的页面访问量之和的排序结果，利用所述至少一个关联页面关键词集合中的一个或多个集合生成第一推送信息；第二推送信息生成及推送模块，配置用于基于用于生成所述第一推送信息的关联页面关键词集合所对应的至少一个被访问页面，生成与所述第一推送信息相关联的第二推送信息并推送给用户。

在一些实施例中，所述第二推送信息生成及推送模块包括：聚类单元，配置用于对用于生成所述第一推送信息的关联页面关键词集合所对应的被访问页面的发布时间按照预设的时间间隔进行聚类，划分成至少一个时间段，其中，当所述至少一个时间段包括两个以上的时间

间段时，分别取自任意两个时间段的发布时间之间的时间差大于所述时间间隔；提取单元，配置用于对于所述至少一个时间段中的一个或多个时间段，分别从每个时间段所对应的被访问页面中提取一个页面；生成单元，配置用于基于所提取的页面，生成第二推送信息并推送给用户。

在一些实施例中，所述第二推送信息生成及推送模块还包括：筛选单元，配置用于对于关联页面关键词集合所对应的被访问页面，将相似度大于第二预设阈值的关键词集合所对应的被访问页面筛选至一个页面，将筛选页面后剩余的被访问页面作为关联页面关键词集合所对应的被访问页面，其中，所述第二预设阈值大于第一预设阈值。

在一些实施例中，所述关键词集合生成模块包括：关键词提取单元，配置用于对所述被访问页面的内容进行统计分析和/或语义分析，提取至少一个关键词；关键词集合生成单元，配置用于基于所述至少一个关键词，生成关键词集合。

在一些实施例中，所述关键词集合生成单元包括：扩展子单元，配置用于对于每个所述至少一个关键词中的单个关键词，进行扩展以生成扩展关键词，其中，所述扩展关键词包括以下至少一项：所述单个关键词的同义词、所述单个关键词的近义词、所述单个关键词的关联词；关键词集合生成子单元，配置用于基于所述至少一个关键词和所述扩展关键词，生成关键词集合。

在一些实施例中，所述关键词集合合并模块进一步配置用于：将满足以下条件之一的关键词集合作为相似度大于第一预设阈值的关键词集合：相同关键词的个数大于个数阈值；相同关键词的个数与进行比较的关键词集合中关键词的总个数的比值大于比值阈值。

在一些实施例中，所述关键词集合中的各关键词还具有重要度系数，以及，所述关键词集合合并模块包括：计算单元，配置用于基于所述重要度系数对不同的关键词集合进行相似度计算；合并及生成单元，配置用于将相似度大于相似度阈值的关键词集合合并，生成关联页面关键词集合。

本申请提供的信息推送方法和装置，通过获取至少一个站点的页

面访问信息，接着对各个网址对应的页面进行内容解析，生成各个被访问页面的关键词集合，然后基于关键词集合的相互比较，将相似度大于第一预设阈值的关键词集合合并，生成至少一个关联页面关键词集合，接着基于至少一个关联页面关键词集合各自对应的被访问页面的页面访问量之和的排序结果，利用至少一个关联页面关键词集合中的一个或多个集合生成第一推送信息，并且，基于用于生成第一推送信息的关联页面关键词集合所对应的至少一个被访问页面，生成与第一推送信息相关联的第二推送信息并推送给用户。这种信息推送方法和装置在向用户推送第一推送信息之后，还可以进一步向用户推送与第一推送信息相关联的第二推送信息，从而丰富了推送信息的内容。

附图说明

通过阅读参照以下附图所作的对非限制性实施例的详细描述，本申请的其它特征、目的和优点将会变得更明显：

图 1 是根据本申请的信息推送方法的一个实施例的流程图；

图 2 是根据本申请的信息推送方法的一个应用示例的示意图；

图 3 是根据本申请的信息推送方法的又一个实施例的流程图；

图 4 是图 3 所示的信息推送方法的实施例的一个应用场景的效果图；

图 5 是根据本申请的信息推送装置的一个实施例的结构示意图；

图 6 是适于用来实现本申请实施例的电子设备的计算机系统的结构示意图。

具体实施方式

下面结合附图和实施例对本申请作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释相关发明，而非对该发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与有关发明相关的部分。

需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本

申请。

请参考图 1，其示出了信息推送的方法的一个实施例的流程 100。本实施例主要以该方法应用于有一定运算能力的电子设备中来举例说明，该电子设备可以包括但不限于智能手机、平板电脑、电子书阅读器、MP3 播放器（Moving Picture Experts Group Audio Layer III，动态影像专家压缩标准音频层面 3）、MP4（Moving Picture Experts Group Audio Layer IV，动态影像专家压缩标准音频层面 4）播放器、膝上型便携计算机和台式计算机等等。该信息推送方法，包括以下步骤：

步骤 101，获取至少一个站点的页面访问信息，其中，页面访问信息包括被访问页面的网址及页面访问量。

在本实施例中，电子设备（例如可以是包含信息推送的应用运行于其上的电子终端或为包含信息推送的应用提供支持的后台服务器）可以从本地或远程地获取至少一个站点的页面访问信息。其中，当上述电子设备是为至少一个站点提供支持的网站服务器时，其可以直接从本地获取上述页面访问信息；而当上述电子设备不是为站点提供支持的网站服务器时，其可以通过有线连接方式或者无线连接方式从网站服务器获取上述页面访问信息。上述无线连接方式包括但不限于 3G/4G 连接、WiFi 连接、蓝牙连接、WiMAX 连接、Zigbee 连接、UWB（ultra wideband）连接、以及其他现在已知或将来开发的无线连接方式。

在这里，页面访问信息可以包括被访问页面的网址及页面访问量。被访问页面可以是被用户访问过的页面。通常，用户访问的每个页面都对应一个网址，该网址可以用统一资源定位器（Uniform Resource Locator，URL）来表示。电子设备可以从一个或多个站点（例如论坛网站）中获取被用户访问过的页面的 URL。可选地，电子设备也可以获取被访问页面的页面内容。

对于电子设备获取的每个页面，电子设备获取页面的 URL 的同时，还可以获取页面访问量。其中页面访问量可以是页面的总被访问次数，也可以是页面在一定时间段（例如 24 个小时）内的被访问次数。电子设备获取的被访问页面可以是被用户访问过的所有页面，也可以

是访问量大于一定阈值（如 50 次）的页面，还可以是访问量由高到低排列靠前的预设个数（如 10 万个）的页面。

步骤 102，对各个网址对应的页面进行内容解析，生成各个被访问页面的关键词集合。

在本实施例中，电子设备可以对上述的各个网址对应的页面的内容通过各种方法进行解析，从中提取出一个或多个关键词，生成关键词集合。

在本实施例的可选实现方式中，电子设备对上述页面的内容的分析方法可以是统计分析方法。例如电子设备可以通过隐含狄利克雷分布（Latent Dirichlet Allocation, LDA）模型提取上述页面的关键词。具体的，电子设备可以将每个页面视为一个词频向量（例如包括各个词语及其出现频率的向量），从而将文本信息转化成易于建模的数字信息，并根据词、主题和文档（可以将每个页面的页面内容作为一个文档）三层结构建立三层贝叶斯概率模型。其中，文档到主题服从多项式分布，主题到词服从多项式分布。这样，每一个页面代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多词所构成的一个概率分布。电子设备可以根据词的概率分布，将分布概率大于一定阈值（例如大于 1%）的词作为页面的关键词，也可以从每个页面中按照分布概率由高到低选择一定数量（例如 20 个）的词作为页面的关键词。

在本实施例的可选实现方式中，电子设备对上述页面的内容的分析方法也可以是语义分析方法。例如，电子设备可以对被访问页面的内容进行全切分方法等处理，把内容分割成词；再对所得到的词进行重要性计算（例如采用词频-逆向文件频率方法（Term Frequency-Inverse Document Frequency, TF-IDF）），基于重要性计算的结果过滤掉一些常用的虚词（对于中文而言，如“了”、“的”）等不产生实际语义的词汇，进而得到关键词。

具体地，电子设备可以首先利用全切分方法切分出与语言词库匹配的所有可能的词，再运用统计语言模型确定最优的切分结果。以页面内容的主题为“本季度居民收入”为例，可以首先进行语言词库匹

配，找到匹配的所有词——本，季度，居民，收入，本季，本季度，度，居民收入，民；这些词以词网格（word lattices）形式表示，接着基于词网格做路径搜索，再基于统计语言模型（例如 N-Gram 模型，）找到最优路径。如果结果显示“本季度 居民 收入”的语言模型得分最高，则“本季度 居民 收入”即为“本季度居民收入”的最优切分。在这里所述的 N-Gram 模型是常用的一种语言模型，对中文而言，可以称之为汉语语言模型（Chinese Language Model, CLM）。该 N-Gram 模型基于这样一种假设，第 N 个词的出现只与前面 N-1 个词相关，而与其它任何词都不相关，整句的概率就是各个词出现概率的乘积，而这些概率可以通过直接从语料中统计 N 个词同时出现的次数得到。

利用全切分方法将内容分割成词之后，电子设备可以采用词频-逆向文件频率（term frequency-inverse document frequency, TF-IDF）方法对这些词进行重要性计算。词频-逆向文件频率方法的主要思想是，如果某个词或短语在一个文档或页面中出现较多，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。其中，频率（Term Frequency, TF）可以衡量某个词或短语对于一个文档或页面的重要性，如果某个词或短语在一个文档或页面中出现的次数多，则 TF 越大，反之，TF 越小；逆向文档频率（inverse document frequency, IDF）可以衡量一个词或短语的普遍重要性，词语在文档集或语料库出现的频率越高，该词语的普遍重要性越高，IDF 越小，反之 IDF 越大。电子设备可以根据 TF 与 IDF 的乘积来衡量某个词或短语在某个页面里面的重要性，从而提取出页面的一个或多个关键词。

需要说明的是，上述语义分析方式的各种方法是目前广泛研究和应用的公知技术，在此不再赘述。

在本实施例的一些可选实现方式中，电子设备还可以对上述一个或多个关键词中的单个关键词进行扩展生成扩展关键词，并将扩展关键词和所提取的关键词一起生成关键词集合。实践中，每个词语可以有同义词，例如“爸爸”可以具有同义词“父亲”，每个词语也可以有近义词，例如“出席”可以具有近义词“参加”，每个词语还可以有

联词，例如“工程图”可以具有关联词“绘制”，等等。电子设备可以将上述一个或多个关键词中单个关键词的同义词、近义词、关联词汇总，作为单个关键词的扩展关键词，并将这些扩展关键词加入上述关键词集合。其中，经常一起出现的词或短语可以作为关联词。可选地，单个关键词的关联词可以根据预先抓取的大量文档或页面数据通过机器学习预先训练的关联词模型获取。例如，该关联词模型可以是根据预先抓取的大量文档或页面内容，经过全切分方法等处理，把内容分割成词，再统计至少两个词同时出现的概率的模型。其中，同时出现概率大于一定阈值的词可以互为关联词。

在本实施例的一些可选实现方式中，关键词集合中的每个关键词还可以具有重要度系数。其中，重要度系数是衡量一个关键词相对于其所在的页面的重要度的数值。例如，可以将从页面中提取的关键词的重要度系数设为 1，将该关键词的同义词的重要度系数设为 0.8，将该关键词的近义词或关联词的重要度系数设为 0.5，等等。值得说明的是，重要度系数是为了区分关键词的重要程度，以上具体数值是对重要度系数的示例性说明，并不构成对重要度系数的限定。可选地，从页面提取的关键词的重要度系数还可以与关键词在页面中出现的次数相关联，出现次数越多，重要度系数越大；扩展关键词的重要度系数还可以和扩展关键词与从页面提取的关键词之间的关联度有关，例如，从页面提取的关键词的同义词可以具有与该关键词相同的重要度系数。实践中，预设的关联词模型中也可以包括关联词的关联度，该关联度可以与词语同时出现的概率成正比，从页面提取的关键词的关联词的重要度系数可以为该关键词的重要度系数与关联度的乘积。

步骤 103，基于关键词集合的相互比较，将相似度大于第一预设阈值的关键词集合合并，生成至少一个关联页面关键词集合。

在本实施例中，电子设备可以进一步对不同的关键词集合的相互比较，计算各个关键词集合之间的相似度，并将相似度大于第一预设阈值的关键词集合合并，生成关联页面关键词集合。其中，用于生成关联页面关键词集合的关键词集合对应的被访问页面可以互为关联页面。

在这里，关键词集合之间的相似度可以表征不同的关键词集合之间的相似程度。在本实施例中，电子设备可以用两个集合之间的相同关键词的个数来表征关键词集合之间的相似程度。电子设备也可以采用余弦相似度（cosine similarity）算法、Jaccard 系数之类的公知的文本相似度计算方法来进行相似度计算。以 Jaccard 系数方法为例，电子设备可以采用如下的公式计算两个关键词集合 A 和 B 之间的相似度：关键词集合 A 与关键词集合 B 之间的相似度=关键词集合 A 与关键词集合 B 之间共有的词的数目/关键词集合 A 与关键词集合 B 一起包括的词的数目。

在一些实现中，关键词集合中的词还可以具有重要度系数。此时，以余弦相似度算法为例，电子设备可以采用如下方法计算关键词集合 A 与关键词集合 B 之间的相似度：关键词集合 A 与关键词集合 B 之间共有的词的重要度系数的乘积之和/关键词集合 A 中各词的重要度系数的平方和与关键词集合 B 中各词的重要度系数的平方和分别开平方后的乘积。例如，关键词集合 A 包括（日本 1，造岛 0.8，填海 0.5），其中，1、0.8 和 0.5 分别是关键词集合 A 中的关键词“日本”、“造岛”和“填海”具有的重要度系数，关键词集合 B 包括（日本 0.7，造岛 1，主权 0.6），其中，0.7、1 和 0.6 分别是关键词集合 B 中的关键词“日本”、“造岛”和“主权”具有的重要度系数，则关键词集合 A 与关键词集合 B 之间的相似度可以为：

$$(1 \times 0.7 + 0.8 \times 1) / (\sqrt{1^2 + 0.8^2 + 0.5^2} \times \sqrt{0.7^2 + 1^2 + 0.6^2})。$$

值得说明的是，第一预设阈值可以是根据经验人为设定的阈值（例如 0.5），也可以是根据预先获取的页面样本进行训练获得分类模型，并通过验证样本对该分类模型进行验证，在该分类模型具有一定的分类准确率（如 99%）时的阈值。

其中，电子设备可以仅将不同的关键词集合中的各词去重后放入一个集合进行合并，电子设备也可以将不同的关键词集合中的各词去重后放入一个集合，同时将相同关键词的重要度系数相加以进行合并。

通过该步骤，电子设备可以将步骤 101 中获取的被访问页面划分为多个分类。其中，每个分类由至少一个被访问页面组成，这些被访

问页面的页面内容相似或相关联，互为关联页面。同时，这些关联页面对应的关键词集合被合并生成关联页面关键词集合。

在一些实现中，该步骤中电子设备也可以通过文本聚类（如 K-means）的方法获取关联页面，并生成关联页面关键词集合。以 K-means 聚类方法为例，电子设备可以首先选取页面访问量最高的 K 个页面作为聚类的质心，然后测量其他页面到每个质心的距离，并把它归到最近的质心的类，接着重新计算已经得到的各个类的质心，循环执行步骤“测量其他页面到每个质心的距离，并把它归到最近的质心的类”直至新的质心与原质心等于或小于指定阈值，此时，页面被划分为 K 个分类。这 K 个分类中，每个分类对应的被访问页面可以互为关联页面。将互为关联页面的被访问页面的关键词集合按照上述的方法合并，可以得到关联页面关键词集合。

步骤 104，基于至少一个关联页面关键词集合中的各个集合对应的被访问页面的页面访问量之和的排序结果，利用至少一个关联页面关键词集合中的一个或多个集合生成第一推送信息。

在本实施例，电子设备可以首先获取上述至少一个关联页面关键词集合所对应的被访问页面的页面访问量的总和，并将这些页面访问量的总和进行排序（例如排序顺序为页面访问量的总和从高到低），然后基于排序结果，利用至少一个关联页面关键词集合中的一个或多个集合生成第一推送信息。

例如，当按照上述页面访问量的总和从高到低的顺序排序时，电子设备可以获取排列靠前的预设个数（例如 10 个）的关联页面关键词集合，然后根据这些关联页面关键词集合或者这些关联页面关键词集合所对应的被访问页面，生成第一推送信息。在这里，电子设备可以选取关联页面关键词集合对应的被访问页面中发布时间最近的页面，将该页面的主题或关键词作为第一推送信息。电子设备也可以将关联页面关键词集合中的各词按照所对应的被访问页面的页面数量或页面访问量由大到小进行排序，选取排在最前的预设个数的关键词作为第一推送信息。电子设备还可以将关联页面关键词集合对应的关联页面中页面访问量最高的页面的主题作为第一推送信息。电子设备还可以

以其他方式，如将关联页面关键词集合对应的被访问页面中页面访问量最高的页面的关键词作为第一推送信息。本申请对此不做限定。可选地，第一推送信息还可以包括关联页面关键词集合对应的关联页面的页面访问量的总和，或者用于生成第一推送信息的被访问页面的页面访问量。

在一些实现中，电子设备可以将该第一推送信息推送给用户。电子设备还可以将第一推送信息直接呈现给用户，还可以将第一推送信息以超链接形式推送给用户，该超链接可以是包括关键词或主题名称的文本，用于链接到该第一推送信息对应的被访问页面或生成该第一推送信息的关联页面关键词集合所对应的关联页面中页面访问量最高的被访问页面。

通过本步骤，电子设备可以获取上述页面对应的分类中访问量最高的前 N (N 为正整数) 个分类，并将这 N 个分类生成 N 条第一推送信息。

步骤 105，基于用于生成第一推送信息的关联页面关键词集合所对应的至少一个被访问页面，生成与第一推送信息相关联的第二推送信息并推送给用户。

在本实施例中，对于每条第一推送信息，电子设备可以获取用于生成第一推送信息的关联页面关键词集合所对应的被访问页面，并从中选取至少一个被访问页面，根据该至少一个被访问页面生成与前述第一推送信息相关联的第二推送信息。

在这里，第二推送信息可以根据与第一推送信息相关联的页面生成。例如，如果第一推送信息是关联页面关键词集合中各词按照所对应的被访问页面的页面数量或页面访问量由大到小进行排序而选取的排在最前的预设个数的关键词，第二推送信息可以是包含这预设个数的关键词中词的个数最多的 M (M 为正整数) 个页面的主题；如果第一推送信息是关联页面关键词集合对应的关联页面中页面访问量最高的被访问页面的主题，第二推送信息可以是关联页面关键词集合对应的关联页面中页面访问量最高的前 M (M 为正整数) 个页面（可以包括用于生成第一推送信息的页面，也可以不包括用于生成第一推送信

息的页面)的主题。

其中,电子设备可以将第二推送信息和第一推送信息一起呈现给用户,也可以在向用户呈现第一推送信息后,检测用户的预定操作,响应于检测到预定操作,将第二推送信息展示给用户。例如,第二推送信息可以在用户点击第一推送信息时呈现,也可以在用户点击第一推送信息对应的按钮时呈现,还可以响应于鼠标悬停而呈现,等等。可选地,第二推送信息可以以超链接的形式推送给用户,该超链接可以关联到第二推送信息对应的页面。

如图2所示,给出了本实施例在具体应用时的一个示例。在图2的示例中,电子设备首先从至少一个站点获取被访问页面的网址及页面访问量,接着对各个被访问页面进行内容解析,生成各个被访问页面的关键词集合,然后基于关键词集合的相互比较,将相似度大于第一预设阈值的关键词集合合并,生成关联页面关键词集合,接着从关联页面关键词集合中选取对应的关联页面的页面访问量之和最高的3个关联页面关键词集合,并将这3个关联页面关键词集合分别对应的页面访问量最高的被访问页面的主题生成第一推送信息201(如网络中的热点新闻);然后,从用于生成第一推送信息201的关联页面关键词集合所对应的被访问页面中获取至少一个(如3个)被访问页面,生成与第一推送信息相关联的第二推送信息202(如热点新闻的背景新闻),并推送给用户。

在图2中,第一推送信息201可以包括主题2011、关联页面关键词集合对应的关联页面的页面访问量之和2012和按钮2013,当按钮2013被用户点击时,电子设备显示第二推送信息202包含的主题2021。其中主题2011和主题2021都可以是超链接形式的文本,分别用以链接到主题2011和主题2021对应的被访问页面。该示例的应用场景例如可以是电子设备向网站的编辑人员推送网络上较受关注的新闻事件,以及这些新闻事件的背景资料,以便编辑人员对新闻事件进行编辑并更新网站内容。

本申请的上述实施例通过向用户推送与第一关联信息相关联的第二推送信息,从而可以向用户展示更丰富的推送信息的内容。

进一步参考图 3，其示出了本申请的信息推送的方法的又一个实施例的流程 300。该信息推送方法 300，包括以下步骤：

步骤 301，获取至少一个站点的页面访问信息，其中，页面访问信息包括被访问页面的网址及页面访问量。

在本实施例中，电子设备（例如可以是包含信息推送的应用运行于其上的电子终端或为包含信息推送的应用提供支持的后台服务器）可以从本地或远程地获取至少一个站点的页面访问信息。在这里，页面访问信息可以包括被访问页面的网址（例如 URL）及页面访问量。

步骤 302，对各个网址对应的页面进行内容解析，生成各个被访问页面的关键词集合。

在本实施例中，电子设备可以对上述的各个网址对应的页面的内容通过各种方法（例如统计分析方法或语义分析方法）进行解析，从中提取出一个或多个关键词，生成关键词集合。在一些实现中，电子设备还可以对上述一个或多个关键词中的单个关键词进行扩展生成扩展关键词，并将所提取的关键词和扩展关键词一起生成关键词集合。其中，扩展关键词可以包括所提取的单个关键词的同义词、近义词和关联词。可选地，关键词集合中的每个关键词还可以具有重要度系数。

步骤 303，基于关键词集合的相互比较，将相似度大于第一预设阈值的关键词集合合并，生成至少一个关联页面关键词集合。

在本实施例中，电子设备可以进一步对不同的关键词集合的相互比较，计算各个关键词集合之间的相似度，并将相似度大于第一预设阈值的关键词集合合并，生成关联页面关键词集合。其中，用于生成关联页面关键词集合的关键词集合对应的被访问页面可以互为关联页面。

在这里，关键词集合之间的相似度可以表征不同的关键词集合之间的相似程度。在本实施例中，电子设备可以用两个集合之间的相同关键词的个数来表征关键词集合之间的相似程度。电子设备可以采用余弦相似度（cosine similarity）算法、Jaccard 系数之类的公知的文本相似度计算方法来进行相似度计算。在一些实现中，关键词集合中的词还可以具有重要度系数。此时，电子设备可以基于重要度系数计算

关键词集合之间的相似度。

步骤 304，基于至少一个关联页面关键词集合中的各个集合对应的被访问页面的页面访问量之和的排序结果，利用至少一个关联页面关键词集合中的一个或多个集合生成第一推送信息。

在本实施例中，电子设备可以首先获取上述至少一个关联页面关键词集合所对应的被访问页面的页面访问量的总和，并将这些页面访问量的总和进行排序（例如排序顺序为页面访问量的总和从高到低），然后基于排序结果，利用至少一个关联页面关键词集合中的一个或多个集合生成第一推送信息。

步骤 305，对用于生成第一推送信息的关联页面关键词集合所对应的被访问页面的发布时间按照预设的时间间隔进行聚类，划分成至少一个时间段。

在本实施例中，电子设备可以对用于生成第一推送信息的关联页面关键词集合对应的被访问页面的发布时间按照预设的时间间隔进行聚类，划分成至少一个时间段。这里，当上述至少一个时间段包括两个以上的时间段时，聚类的结果可以是：分别取自任意两个时间段的发布时间之间的时间差大于上述预设的时间间隔。

聚类是将物理或抽象对象的集合分成由类似的对象组成的多个类的过程。在这里，电子设备将被访问页面的发布时间按照预设的时间间隔聚类的目的是：将被访问页面的发布时间划分成至少一个时间段，从而将被访问页面分成发布时间相近的多个类。

在本实施例中，按照发布时间的聚类可以使用各种公知的聚类算法。例如，电子设备可以基于层次聚类算法，每次合并时间间隔最小的两个发布时间，直到时间间隔最小的两个发布时间之间的时间差大于或等于预设的时间间隔，从而，将关联页面关键词集合对应的被访问页面按照发布时间划分成在不同时间段内发布的页面。不同时间段内发布的任意两个被访问页面，他们的发布时间都大于预设的时间间隔。

在本实施例的可选实现方式中，电子设备还可以按照一天的不同时间段确定聚类的预设时间间隔。例如，电子设备可以预先获取多天

的页面发布量，根据页面发布量的分布划分时间间隔。例如，假设每天 0:00 到 6:00 的网页发布量比较少，则可以将发布时间为 0:00 到 6:00 的预设时间间隔设置为一个较长时间段，如 2 小时；同样，假设在每天 9:00 到 11:00 之间的网页发布量比较多，则可以将发布时间为 9:00 到 11:00 的预设时间间隔设置为一个较短的时间段，如 20 分钟。

通过本步骤，电子设备可以将一个关联页面关键词集合对应的被访问页面按时间划分开来，不同时间段的被访问页面可能记录了不同发展阶段的事件内容。

步骤 306，对于上述至少一个时间段中的一个或多个时间段，分别从每个时间段所对应的被访问页面中提取一个页面。

在本实施例中，电子设备可以针对上述至少一个时间段中的一个或多个时间段，分别从每个时间段所对应的被访问页面中提取出一个页面。

在这里，电子设备所提取的页面，可以是所对应的时间段内发布的任意页面，也可以是按一定规则获取的页面。当电子设备按一定规则获取页面时，可以获取所对应的时间段内页面访问量最高的页面，也可以获取所对应的时间段内发布时间最早的页面，还可以按照预先设定的发布页面的站点的优先级别获取页面，等等，本申请对此不做限定。

步骤 307，基于所提取的页面，生成第二推送信息并推送给用户。

在本实施例中，电子设备可以基于步骤 306 中所提取的页面，根据一定的规则，生成第二推送信息，并可以将第二推送信息推送给用户。电子设备基于所提取的页面生成第二推送信息的方式有很多，例如，电子设备可以将所提取的页面的主题或关键字作为第二推送信息，电子设备也可以从所提取的页面中按照发布时间由近及远的顺序选取前预设个数的页面，将这些页面的主题或关键字作为第二推送信息，等等。本申请对此不做限定。

在本实施例的可选实现方式中，在步骤 304 和步骤 305 之间，还可以包括页面的去重步骤。电子设备可以将关联页面关键词集合所对应的被访问页面做以下处理：对于关联页面关键词集合所对应的被访

问页面，将相似度大于第二预设阈值的关键词集合所对应的被访问页面筛除至一个页面，将筛除页面后剩余的被访问页面作为关联页面关键词集合所对应的被访问页面。

这里，相似度的算法与前述实施例的步骤 103 中的计算方法相同，在此不再赘述。其中，第二预设阈值可以大于第一预设阈值。电子设备通过该步骤对关联页面关键词集合所对应的被访问页面去重的原理是：

例如，第二预设阈值取值为 98%，则当两个关键词集合的相似度大于 98% 时，电子设备可以认为这两个关键词集合对应的被访问页面为相同内容的页面，即重复的页面。电子设备可以从重复的页面中保留任意一个页面，也可以从重复的页面中按一定的规则选取一个页面进行保留，如选取发布时间最早的页面进行保留等，同时筛除重复的页面中的其他页面，将筛除页面后剩余的被访问页面作为关联页面关键词集合所对应的被访问页面。假设关联页面关键词集合所对应的被访问页面有 1000 个，其中有 30 组重复的页面，每组重复的页面都包括 2 个页面，则电子设备从这 30 组的每一组页面中筛除 1 个页面，保留 1 个页面，则剩余 970 个页面作为关联页面关键词集合所对应的被访问页面。对于重复的页面中不被保留的页面，电子设备可以将该页面的页面信息删除。可选地，对于重复的页面，电子设备可以将不被保留的页面的页面访问量累加在保留的页面的页面访问量上。

在本实施例中，上述实现流程中的步骤 301、步骤 302、步骤 303 和步骤 304 分别与前述实施例中的步骤 101、步骤 102、步骤 103 和步骤 104 基本相同，在此不再赘述。

从图 3 中可以看出，与图 1 对应的实施例不同的是，本实施例中的信息推送方法的流程 300 用步骤 305、306、307 代替了步骤 105。通过步骤 305、306、307，本实施例可以将第一推送信息相对应的关联页面关键词集合对应的被访问页面按照时间段进行抽取，从而生成与第一推送信息相关联的第二推送信息。当这些页面属于同一个事件时，每个时间段内的页面的页面内容可以给出事件的一个发展状态，从每个时间段内提取一个页面生成第二推送信息，可以使用户通过第

二推送信息对整个事件的发展过程进行了解。如图 4 所示，为本实施例的信息推送方法的一应用场景的效果图。图 4 示出的应用场景为热点新闻信息的推送场景，其中，401 指示第一推送信息，402 指示第二推送信息。本实施例有助于向用户推送第一推送信息的各个时间段内的发展信息。可选地，在将被访问页面的发布时间进行聚类前，可以先对页面去重，以避免在不同时间段内获取具有相同内容的页面从而降低信息推送的有效性。

进一步参考图 5，作为对上述各图所示方法的实现，本申请提供了一种信息推送的装置的一个实施例，该装置实施例与图 1 所示的方法实施例相对应，该装置具体可以应用于电子设备中。

如图 5 所示，本实施例所述的信息推送的装置 500 包括：信息获取模块 501、关键词集合生成模块 502、关键词集合合并模块 503、第一推送信息生成模块 504 及第二推送信息生成及推送模块 505。其中，信息获取模块 501 配置用于获取至少一个站点的页面访问信息，其中，页面访问信息包括被访问页面的网址及页面访问量；关键词集合生成模块 502 配置用于对各个网址对应的页面进行内容解析，生成各个被访问页面的关键词集合；关键词集合合并模块 503 配置用于基于关键词集合的相互比较，将相似度大于第一预设阈值的关键词集合合并，生成至少一个关联页面关键词集合，其中，用于生成关联页面关键词集合的关键词集合对应的被访问页面互为关联页面；第一推送信息生成模块 504 配置用于基于至少一个关联页面关键词集中的各个集合对应的被访问页面的页面访问量之和的排序结果，利用上述至少一个关联页面关键词集合中的一个或多个集合生成第一推送信息；第二推送信息生成及推送模块 505 配置用于基于用于生成第一推送信息的关联页面关键词集合所对应的至少一个被访问页面，生成与第一推送信息相关联的第二推送信息并推送给用户。

在本实施例中，信息推送装置 500 可以首先通过信息获取模块 501 从本地或远程地获取至少一个站点的页面访问信息。在这里，页面访问信息可以包括被访问页面的网址（例如 URL）及页面访问量。

在本实施例中，关键词集合生成模块 502 可以接着对上述的各个

网址对应的页面的内容通过各种方法（例如统计分析方法或语义分析方法）进行解析，从中提取出一个或多个关键词，生成关键词集合。在一些实现中，关键词集合生成模块 502 还可以对上述一个或多个关键词中的单个关键词进行扩展生成扩展关键词，并将所提取的关键词和扩展关键词一起生成关键词集合。其中，扩展关键词可以包括所提取的单个关键词的同义词、近义词和关联词。可选地，关键词集合中的每个关键词还可以具有重要度系数。

在本实施例中，关键词集合合并模块 503 接着可以对关键词集合生成模块 502 生成的关键词集合相互比较，并将相似度大于第一预设阈值的关键词集合合并，生成至少一个关联页面关键词集合。其中，用于生成关联页面关键词集合的关键词集合对应的被访问页面互为关联页面。这里，关键词集合之间的相似度可以通过多种方法计算。

在本实施例中，第一推送信息生成模块 504 接着可以获取上述至少一个关联页面关键词集合所对应的被访问页面的页面访问量的总和，并将这些页面访问量的总和进行排序（例如排序顺序为页面访问量的总和从高到低），然后基于排序结果，利用至少一个关联页面关键词集合中的一个或多个集合生成第一推送信息。

在本实施例中，第二推送信息生成及推送模块 505 接着可以对于每条第一推送信息，获取用于生成第一推送信息的关联页面关键词集合所对应的被访问页面，并从中选取至少一个被访问页面，然后根据该至少一个被访问页面生成与前述第一推送信息相关联的第二推送信息并推送给第二用户。

在本实施例的一些可选实现方式中，第二推送信息生成及推送模块 505 可以包括：聚类单元（未示出），配置用于对用于生成第一推送信息的关联页面关键词集合所对应的被访问页面的发布时间按照预设的时间间隔进行聚类，划分成至少一个时间段；提取单元（未示出），配置用于对于至少一个时间段中的一个或多个时间段，分别从每个时间段所对应的被访问页面中提取一个页面；生成单元（未示出），配置用于基于所提取的页面，生成第二推送信息并推送给用户。这里，当上述至少一个时间段包括两个以上的时间段时，聚类的结果可以是：

分别取自任意两个时间段的发布时间之间的时间差大于上述预设的时间间隔。

在本实施例的一些可选实现方式中，第二推送信息生成及推送模块 505 还可以包括：筛除单元（未示出），配置用于对于关联页面关键词集合所对应的被访问页面，将相似度大于第二预设阈值的关键词集合所对应的被访问页面筛除至一个页面，将筛除页面后剩余的被访问页面作为关联页面关键词集合所对应的被访问页面。其中，第二预设阈值大于第一预设阈值。筛除单元的作用是对关联页面关键词集合所对应的被访问页面去重。

值得说明的是，信息推送装置 500 中记载的诸模块或单元与参考图 1 描述的方法中的各个步骤相对应。由此，上文针对方法描述的操作和特征同样适用于信息推送装置 500 及其中包含的模块或单元，在此不再赘述。

本领域技术人员可以理解，上述信息推送装置 500 还包括一些其他公知结构，例如处理器、存储器等，为了不必要地模糊本公开的实施例，这些公知的结构在图 5 中未示出。

下面参考图 6，其示出了适于用来实现本申请实施例的电子设备的计算机系统 600 的结构示意图。

如图 6 所示，计算机系统 600 包括中央处理单元（CPU）601，其可以根据存储在只读存储器（ROM）602 中的程序或者从存储部分 608 加载到随机访问存储器（RAM）603 中的程序而执行各种适当的动作和处理。在 RAM 603 中，还存储有系统 600 操作所需的各种程序和数据。CPU 601、ROM 602 以及 RAM 603 通过总线 604 彼此相连。输入/输出（I/O）接口 605 也连接至总线 604。

以下部件连接至 I/O 接口 605：包括键盘、鼠标等的输入部分 606；包括诸如阴极射线管（CRT）、液晶显示器（LCD）等以及扬声器等的输出部分 607；包括硬盘等的存储部分 608；以及包括诸如 LAN 卡、调制解调器等网络接口卡的通信部分 609。通信部分 609 经由诸如因特网的网络执行通信处理。驱动器 610 也根据需要连接至 I/O 接口 605。可拆卸介质 611，诸如磁盘、光盘、磁光盘、半导体存储器等等，

根据需要安装在驱动器 610 上，以便于从其上读出的计算机程序根据需要被安装入存储部分 608。

特别地，根据本申请的实施例，上文参考流程图描述的过程可以被实现为计算机软件程序。例如，本申请的实施例包括一种计算机程序产品，其包括有形地包含在机器可读介质上的计算机程序，所述计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中，该计算机程序可以通过通信部分 609 从网络上被下载和安装，和/或从可拆卸介质 611 被安装。

本申请实施例中所涉及到的单元可以通过软件的方式实现，也可以通过硬件的方式来实现。所描述的模块也可以设置在处理器中，例如，可以描述为：一种处理器包括。其中信息获取模块、关键词集合生成模块、关键词集合合并模块、第一推送信息生成模块及第二推送信息生成及推送模块，这些模块的名称在某种情况下并不构成对该模块本身的限定，例如，信息获取模块还可以被描述为“配置用于获取至少一个站点的页面访问信息的模块”。

作为另一方面，本申请还提供了一种计算机可读存储介质，该计算机可读存储介质可以是上述实施例中所述装置中所包含的计算机可读存储介质；也可以是单独存在，未装配入终端中的计算机可读存储介质。所述计算机可读存储介质存储有一个或者一个以上程序，所述程序被一个或者一个以上的处理器用来执行描述于本申请的信息推送的方法。

以上描述仅为本申请的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解，本申请中所涉及的发明范围，并不限于上述技术特征的特定组合而成的技术方案，同时也应涵盖在不脱离所述发明构思的情况下，由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本申请中公开的（但不限于）具有类似功能的技术特征进行互相替换而形成的技术方案。

权 利 要 求 书

1、一种信息推送方法，其特征在于，所述方法包括：

获取至少一个站点的页面访问信息，其中，所述页面访问信息包括被访问页面的网址及页面访问量；

对各个网址对应的页面进行内容解析，生成各个被访问页面的关键词集合；

基于关键词集合的相互比较，将相似度大于第一预设阈值的关键词集合合并，生成至少一个关联页面关键词集合，其中，用于生成关联页面关键词集合的关键词集合对应的被访问页面互为关联页面；

基于所述至少一个关联页面关键词集合中的各个集合对应的被访问页面的页面访问量之和的排序结果，利用所述至少一个关联页面关键词集合中的一个或多个集合生成第一推送信息；

基于用于生成所述第一推送信息的关联页面关键词集合所对应的至少一个被访问页面，生成与所述第一推送信息相关联的第二推送信息并推送给用户。

2、根据权利要求1所述的方法，其特征在于，所述基于用于生成所述第一推送信息的关联页面关键词集合所对应的至少一个被访问页面，生成与所述第一推送信息相关联的第二推送信息并推送给用户，包括：

对用于生成所述第一推送信息的关联页面关键词集合所对应的被访问页面的发布时间按照预设的时间间隔进行聚类，划分成至少一个时间段，其中，当所述至少一个时间段包括两个以上的时间段时，分别取自任意两个时间段的发布时间之间的时间差大于所述时间间隔；

对于所述至少一个时间段中的一个或多个时间段，分别从每个时间段所对应的被访问页面中提取一个页面；

基于所提取的页面，生成第二推送信息并推送给用户。

3、根据权利要求2所述的方法，其特征在于，所述对用于生成所

述第一推送信息的关联页面关键词集合所对应的被访问页面的发布时间按照预设的时间间隔进行聚类，划分成至少一个时间段之前，还包括：

对于关联页面关键词集合所对应的被访问页面，将相似度大于第二预设阈值的关键词集合所对应的被访问页面筛选至一个页面，将筛选页面后剩余的被访问页面作为关联页面关键词集合所对应的被访问页面，其中，所述第二预设阈值大于第一预设阈值。

4、根据权利要求1所述的方法，其特征在于，所述对各个网址对应的页面进行内容解析，生成各个被访问页面的关键词集合包括：

对所述被访问页面的内容进行统计分析和/或语义分析，提取至少一个关键词；

基于所述至少一个关键词，生成关键词集合。

5、根据权利要求4所述的方法，其特征在于，所述基于所述至少一个关键词，生成关键词集合包括：

对于每个所述至少一个关键词中的单个关键词，进行扩展以生成扩展关键词，其中，所述扩展关键词包括以下至少一项：所述单个关键词的同义词、所述单个关键词的近义词、所述单个关键词的关联词；

基于所述至少一个关键词和所述扩展关键词，生成关键词集合。

6、根据权利要求1-5中任一所述的方法，其特征在于，将满足以下条件之一的关键词集合作为相似度大于第一预设阈值的关键词集合：

相同关键词的个数大于个数阈值；

相同关键词的个数与进行比较的关键词集合中关键词的总个数的比值大于比值阈值。

7、根据权利要求1-5中任一所述的方法，其特征在于，所述关键词集合中的各关键词还具有重要度系数，以及

所述基于关键词集合的相互比较，将相似度大于第一预设阈值的关键词集合合并，生成至少一个关联页面关键词集合包括：

基于所述重要度系数对不同的关键词集合进行相似度计算；

将相似度大于相似度阈值的关键词集合合并，生成关联页面关键词集合。

8、一种信息推送装置，其特征在于，所述装置包括：

信息获取模块，配置用于获取至少一个站点的页面访问信息，其中，所述页面访问信息包括被访问页面的网址及页面访问量；

关键词集合生成模块，配置用于对各个网址对应的页面进行内容解析，生成各个被访问页面的关键词集合；

关键词集合合并模块，配置用于基于关键词集合的相互比较，将相似度大于第一预设阈值的关键词集合合并，生成至少一个关联页面关键词集合，其中，用于生成关联页面关键词集合的关键词集合对应的被访问页面互为关联页面；

第一推送信息生成模块，配置用于基于所述至少一个关联页面关键词集合中的各个集合对应的被访问页面的页面访问量之和的排序结果，利用所述至少一个关联页面关键词集合中的一个或多个集合生成第一推送信息；

第二推送信息生成及推送模块，配置用于基于用于生成所述第一推送信息的关联页面关键词集合所对应的至少一个被访问页面，生成与所述第一推送信息相关联的第二推送信息并推送给用户。

9、根据权利要求8所述的装置，其特征在于，所述第二推送信息生成及推送模块包括：

聚类单元，配置用于对用于生成所述第一推送信息的关联页面关键词集合所对应的被访问页面的发布时间按照预设的时间间隔进行聚类，划分成至少一个时间段，其中，当所述至少一个时间段包括两个以上的时间段时，分别取自任意两个时间段的发布时间之间的时间差大于所述时间间隔；

提取单元，配置用于对于所述至少一个时间段中的一个或多个时间段，分别从每个时间段所对应的被访问页面中提取一个页面；

生成单元，配置用于基于所提取的页面，生成第二推送信息并推送给用户。

10、根据权利要求 9 所述的装置，其特征在于，所述第二推送信息生成及推送模块还包括：

筛选单元，配置用于对于关联页面关键词集合所对应的被访问页面，将相似度大于第二预设阈值的关键词集合所对应的被访问页面筛选至一个页面，将筛选页面后剩余的被访问页面作为关联页面关键词集合所对应的被访问页面，其中，所述第二预设阈值大于第一预设阈值。

11、根据权利要求 8 所述的装置，其特征在于，所述关键词集合生成模块包括：

关键词提取单元，配置用于对所述被访问页面的内容进行统计分析和/或语义分析，提取至少一个关键词；

关键词集合生成单元，配置用于基于所述至少一个关键词，生成关键词集合。

12、根据权利要求 11 所述的装置，其特征在于，所述关键词集合生成单元包括：

扩展子单元，配置用于对于每个所述至少一个关键词中的单个关键词，进行扩展以生成扩展关键词，其中，所述扩展关键词包括以下至少一项：所述单个关键词的同义词、所述单个关键词的近义词、所述单个关键词的关联词；

关键词集合生成子单元，配置用于基于所述至少一个关键词和所述扩展关键词，生成关键词集合。

13、根据权利要求 8-12 中任一项所述的装置，其特征在于，所述

关键词集合合并模块进一步配置用于：

将满足以下条件之一的关键词集合作为相似度大于第一预设阈值的关键词集合：

相同关键词的个数大于个数阈值；

相同关键词的个数与进行比较的关键词集合中关键词的总个数的比值大于比值阈值。

14、根据权利要求 8-12 中任一项所述的装置，其特征在于，所述关键词集合中的各关键词还具有重要度系数，以及

所述关键词集合合并模块包括：

计算单元，配置用于基于所述重要度系数对不同的关键词集合进行相似度计算；

合并及生成单元，配置用于将相似度大于相似度阈值的关键词集合合并，生成关联页面关键词集合。

15、一种设备，包括：

处理器；和

存储器，

所述存储器中存储有能够被所述处理器执行的计算机可读指令，在所述计算机可读指令被执行时，所述处理器执行权利要求 1 至 7 中任一项所述的方法。

16、一种非易失性计算机存储介质，所述计算机存储介质存储有能够被处理器执行的计算机可读指令，当所述计算机可读指令被处理器执行时，所述处理器执行权利要求 1 至 7 中任一项所述的方法。

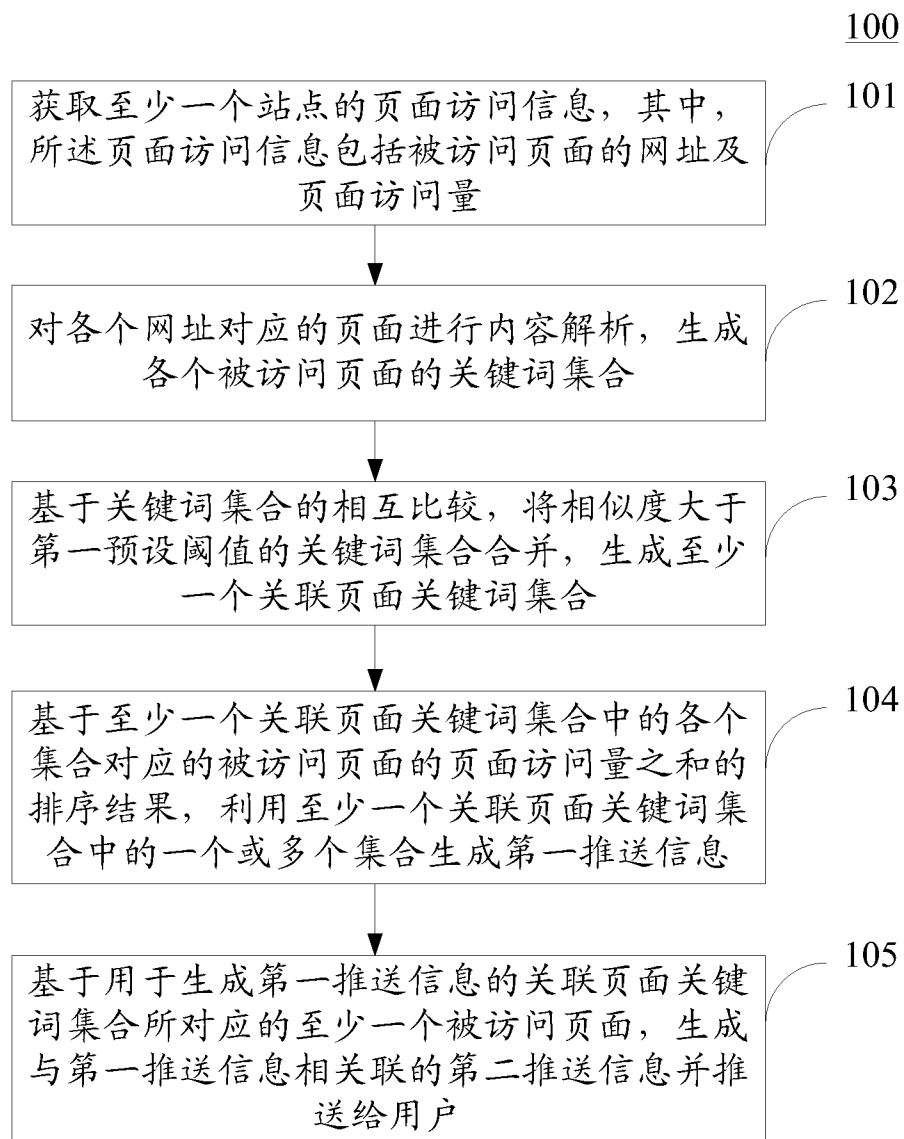


图 1

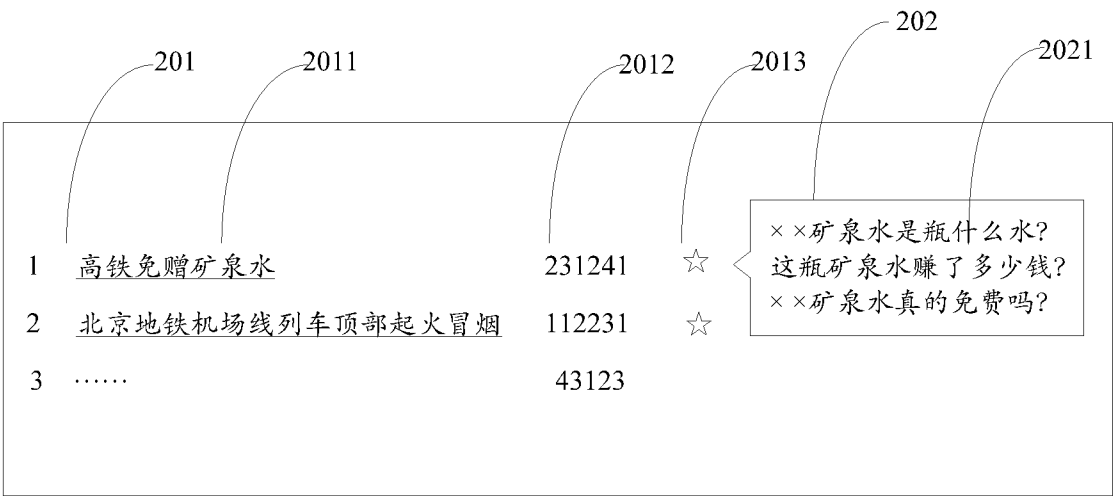


图 2

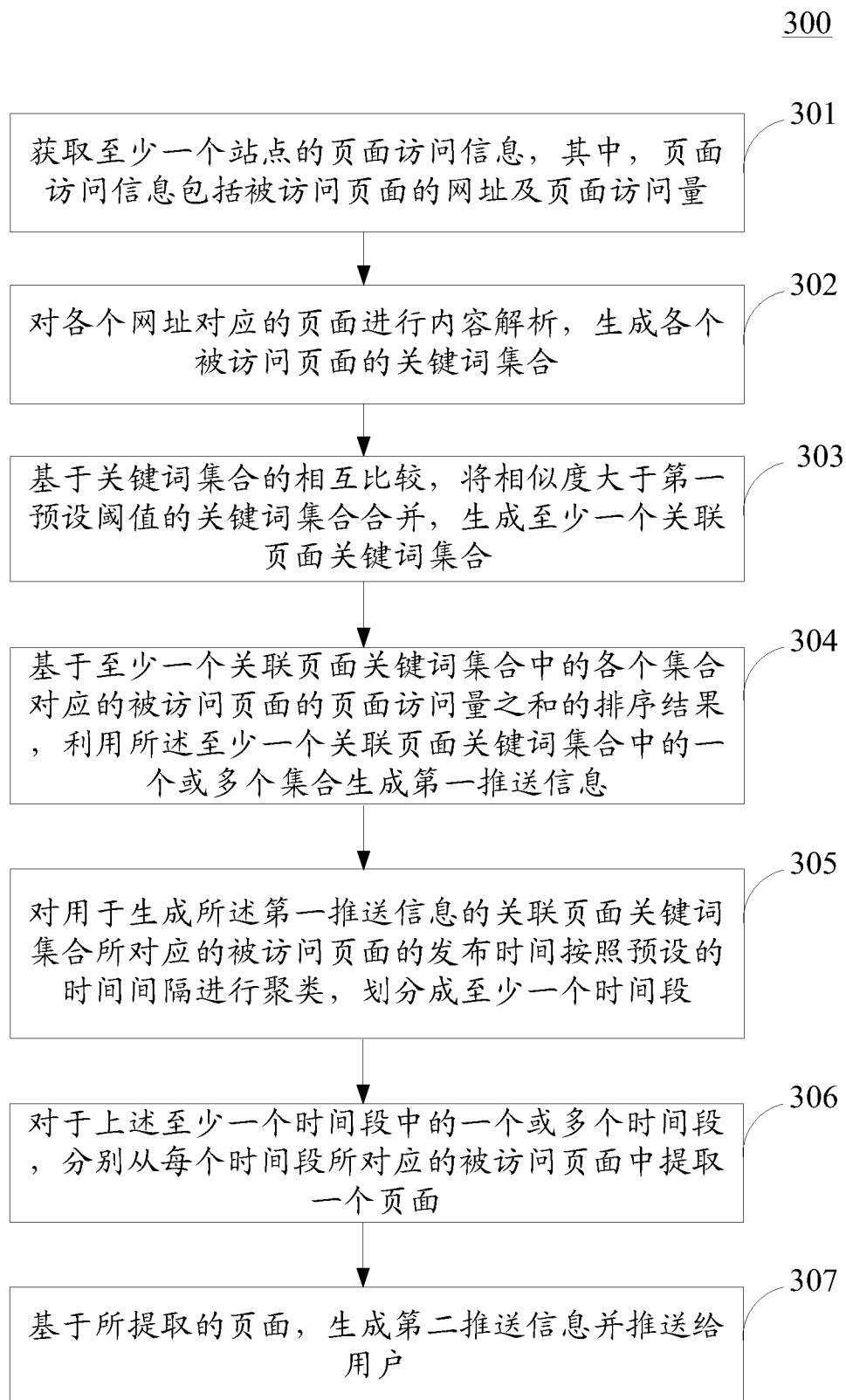


图 3

400

401

全网热点

- 1 链接：中国可能空袭南苏丹首都喀土穆
- 2 张百生：阿基诺三世是亲华 菲海基想买中国舰机
- 3 江西人大副团王任被实名一带反团12年 屡被网上曝光
- 4 阿基诺三世水基已给海基回函
- 5 三个华人被富商“砍”死呢？
- 6 意大利强盗京城 大陆数度派兵
- 7 日本手机短信被拦截，第一台手机不例外
- 8 小米为台湾首富雷军2012年发100条信息最多至
- 9 索罗斯：中国经济不会不行 或许比美国强过世界大敌
- 10 中国版在索罗斯五个字：这次不一样

402

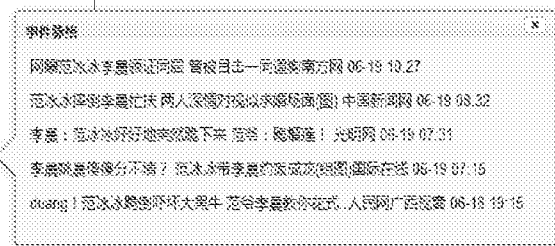


图 4

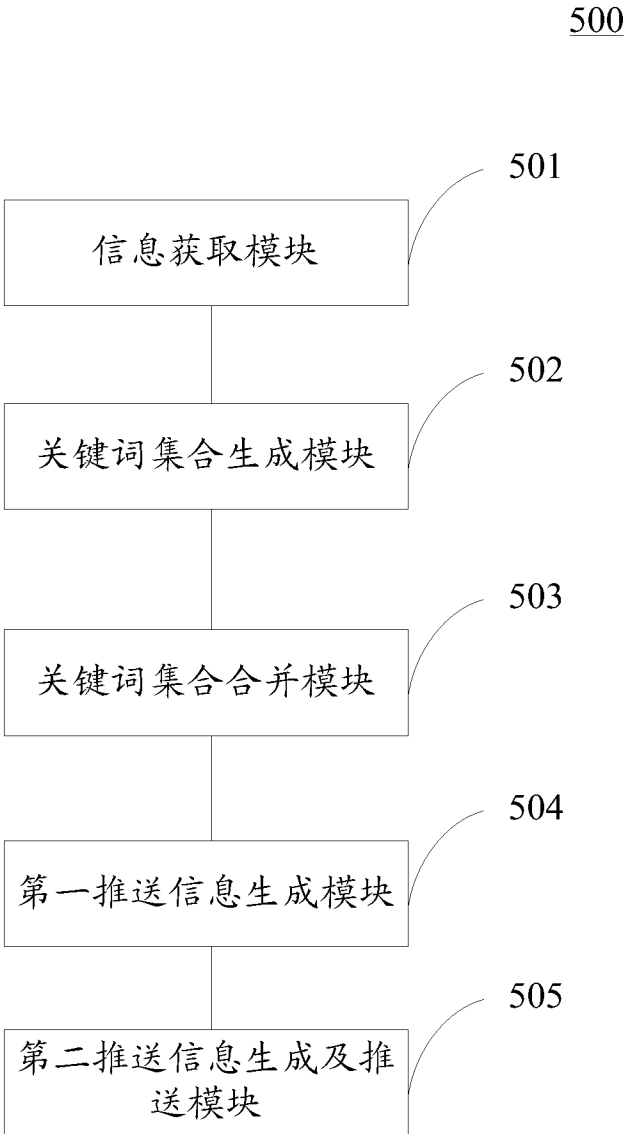


图 5

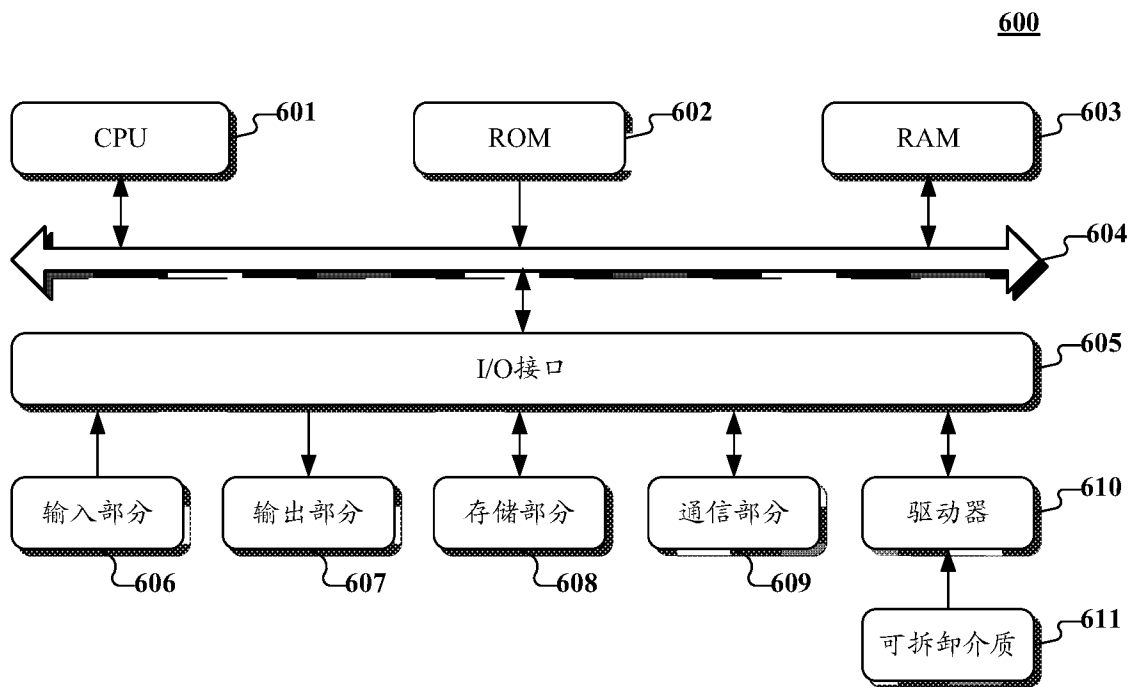


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2015/095754**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

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; IEEE: similarity; associate; generate; site; universal resource locator; analysis; word set; set; preset; threshold value; sort, accord, user, website, content, keyword, relevant, page, information, push, receive, page, access, station, obtain, first, threshold, address

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 105069102 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.), 18 November 2015 (18.11.2015), claims 1-14, and description, paragraphs [0104]-[0106]	1-16
A	CN 103164521 A (IZP (BEIJING) TECHNOLOGIES CO., LTD.), 19 June 2013 (19.06.2013), description, paragraphs [0050]-[0091]	1-16
A	CN 101984423 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.), 09 March 2011 (09.03.2011), the whole document	1-16
A	US 2007260597 A1 (CRAMER, M.), 08 November 2007 (08.11.2007), the whole document	1-16

☐ 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 11 March 2016 (11.03.2016)	Date of mailing of the international search report 27 April 2016 (27.04.2016)
Name and mailing address of the ISA/CN: 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 LI, Meng Telephone No.: (86-10) 010-62413008

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2015/095754

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 105069102 A	18 November 2015	None	
CN 103164521 A	19 June 2013	None	
CN 101984423 A	09 March 2011	None	
US 2007260597 A1	08 November 2007	US 2012078710 A1	29 March 2012
		US 2010106703 A1	29 April 2010
		WO 2007130914 A2	15 November 2007
		US 2008114751 A1	15 May 2008
		US 2013262455 A1	03 October 2013
		US 2014032311 A1	30 January 2014
		US 2015032717 A1	29 January 2015
		US 2015081691 A1	19 March 2015

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

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;EPDOC;CNPAT;CNKI;IEEE: 关键词;相似度;关联;生成;推送;站点;访问;网址;解析;词集;集合;预设;阈值;排序;accord, user, website, content, keyword, relevant, page, information, push, receive, page, access, station, obtain, first, threshold, address

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
PX	CN 105069102 A (百度在线网络技术北京有限公司) 2015年 11月 18日 (2015 - 11 - 18) 权利要求1-14, 说明书第[0104]-[0106]段	1-16
A	CN 103164521 A (亿赞普北京科技有限公司) 2013年 6月 19日 (2013 - 06 - 19) 说明书第[0050]-[0091]段	1-16
A	CN 101984423 A (百度在线网络技术北京有限公司) 2011年 3月 9日 (2011 - 03 - 09) 全文	1-16
A	US 2007260597 A1 (CRAMER, MARK) 2007年 11月 8日 (2007 - 11 - 08) 全文	1-16

☐ 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

2016年 3月 11日

国际检索报告邮寄日期

2016年 4月 27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

李萌

电话号码 (86-10)010-62413008

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2015/095754

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	105069102	A	2015年 11月 18日	无			
CN	103164521	A	2013年 6月 19日	无			
CN	101984423	A	2011年 3月 9日	无			
US	2007260597	A1	2007年 11月 8日	US	2012078710	A1	2012年 3月 29日
				US	2010106703	A1	2010年 4月 29日
				WO	2007130914	A2	2007年 11月 15日
				US	2008114751	A1	2008年 5月 15日
				US	2013262455	A1	2013年 10月 3日
				US	2014032311	A1	2014年 1月 30日
				US	2015032717	A1	2015年 1月 29日
				US	2015081691	A1	2015年 3月 19日

表 PCT/ISA/210 (同族专利附件) (2009年7月)