

(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
2021 年 2 月 18 日 (18.02.2021)



(10) 国际公布号
WO 2021/027116 A1

- (51) 国际专利分类号:
G06F 16/35 (2019.01)
- (21) 国际申请号: PCT/CN2019/116550
- (22) 国际申请日: 2019 年 11 月 8 日 (08.11.2019)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201910768143.X 2019年8月15日 (15.08.2019) 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福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。
- (72) 发明人: 苏智辉(SU, ZhiHui); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 侯丽(HOU, Li); 中国广东省深圳市福田区福田街道福安社

区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 姚飞(YAO, Fei); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

- (74) 代理人: 深圳市沃德知识产权代理事务所(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,

(54) **Title:** METHOD AND APPARATUS FOR DISCOVERING TEXT HOTSPOT AND COMPUTER-READABLE STORAGE MEDIUM

(54) 发明名称: 文本热点的发现方法、装置及计算机可读存储介质

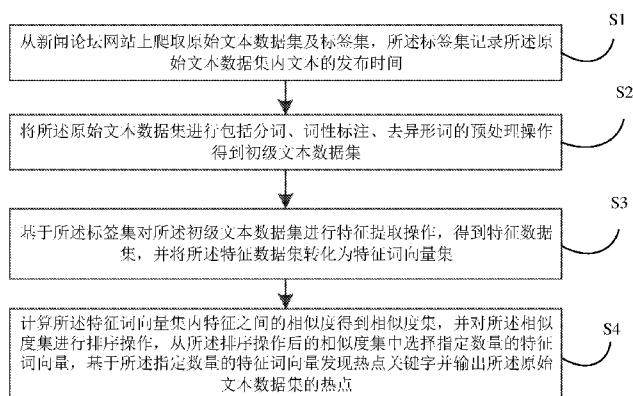


图 1

- S1 Crawl an original text data set and a tag set from a news forum website, the tag set recording the publication time of text in the original text data set
- S2 Perform a preprocessing operation comprising word segmentation, part-of-speech tagging, and heteromorphic word removal on the original text data set to obtain a primary text data set
- S3 Perform a feature extraction operation on the primary text data set on the basis of the tag set to obtain a feature data set, and convert the feature data set into a feature word vector set
- S4 Calculate the similarity between features in the feature word vector set to obtain a similarity set, perform a ranking operation on the similarity set, select a specified quantity of feature word vectors from the ranked similarity set, discover hotspot keywords on the basis of the specified quantity of feature word vectors and output a hotspot of the original text data set

(57) **Abstract:** The present application relates to artificial intelligence technology, and disclosed is a method for discovering a text hotspot, comprising: receiving an original text data set and a tag set, the tag set recording the publication time of text in the original text data set; performing a preprocessing operation comprising word segmentation, part-of-speech tagging, and heteromorphic word removal on the original text data set to obtain a primary text data set; performing a feature extraction operation on the primary text data set on the basis of the tag set to obtain a feature data set, and converting the feature data set into a feature word vector set; calculating the similarity between features in the feature word vector set to obtain a similarity set; selecting a specified quantity of feature word vectors from within the similarity set; and discovering hotspot keywords on the basis of the specified quantity of feature word vectors and outputting a hotspot. Further proposed by the present application are an apparatus for discovering a text hotspot and a computer-readable storage medium. The present application may achieve the function of accurately and efficiently discovering a text hotspot.

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)。

本国际公布:

- 包括国际检索报告(条约第21条(3))。

(57) 摘要: 本申请涉及一种人工智能技术, 揭露了一种文本热点的发现方法, 包括: 接收原始文本数据集及标签集, 所述标签集记录所述原始文本数据集内文本的发布时间, 将所述原始文本数据集进行包括分词、词性标注、去异形词的预处理操作得到初级文本数据集, 基于所述标签集对所述初级文本数据集进行特征提取操作得到特征数据集, 并将所述特征数据集转化为特征词向量集, 计算所述特征词向量集内特征之间的相似度得到相似度集, 从所述相似度集中选择指定数量的特征词向量, 基于所述指定数量的特征词向量发现热点关键字并输出所述热点。本申请还提出一种文本热点的发现装置以及一种计算机可读存储介质。本申请可以实现精准高效的文本热点的发现功能。

文本热点的发现方法、装置及计算机可读存储介质

5 本申请基于巴黎公约申明享有 2019 年 8 月 15 日递交的申请号为 CN 201910768143.X、名称为“文本热点的发现方法、装置及计算机可读存储介质”的中国专利申请的优先权，该中国专利申请的整体内容以参考的方式结合在本申请中。

技术领域

10 本申请涉及人工智能技术领域，尤其涉及一种提取文本数据集中关键字进而发现文本热点的方法、装置及计算机可读存储介质。

背景技术

15 随着互联网技术的飞速发展，各大门户网站应运而生，且多数门户网站也成为了人们获取信息的主要渠道。但由于网络的复杂性、冗余性、更新和传播的快速性等，都给人们快速、准确地获取自己所需的关键信息带来了困难，也不利于网络舆情的监控。因此及时发现网络热点关键字已成为当下研究的侧重点。目前有基于single-pass的文本聚类算法，由于算法简单易实现，时空复杂度低，聚类效果优异等特点被广泛用作发现网络热点关键字。但
20 single-pass算法存在局限性，如关键字相似匹配只根据经验阈值归类，不仅造成网络中的每条文本数据的话题分析效率慢，且由于数据量和时间是指数级正相关关系，进一步影响了准确性。

发明内容

25 本申请提供一种文本热点的发现方法、装置及计算机可读存储介质，其主要目的在于通过提取文本数据集中关键字进而发现文本热点。

为实现上述目的，本申请提供的一种文本热点的发现方法，包括：

从新闻论坛网站上爬取原始文本数据集及标签集，所述标签集记录所述原始文本数据集内文本的发布时间；

30 将所述原始文本数据集进行包括分词、词性标注、去异形词的预处理操

作得到初级文本数据集；

基于所述标签集对所述初级文本数据集进行特征提取操作，得到特征数据集，并将所述特征数据集转化为特征词向量集；

5 计算所述特征词向量集内特征之间的相似度得到相似度集，并对所述相似度集进行排序操作，从所述排序操作后的相似度集中选择指定数量的特征词向量，基于所述指定数量的特征词向量发现热点关键字并输出所述原始文本数据集的热点。

此外，为实现上述目的，本申请还提供一种文本热点的发现装置，该装置包括存储器和处理器，所述存储器中存储有可在所述处理器上运行的文本热点的发现程序，所述文本热点的发现程序被所述处理器执行时实现如下步骤：
10

从新闻论坛网站上爬取原始文本数据集及标签集，所述标签集记录所述原始文本数据集内文本的发布时间；

15 将所述原始文本数据集进行包括分词、词性标注、去异形词的预处理操作得到初级文本数据集；

基于所述标签集对所述初级文本数据集进行特征提取操作，得到特征数据集，并将所述特征数据集转化为特征词向量集；

20 计算所述特征词向量集内特征之间的相似度得到相似度集，并对所述相似度集进行排序操作，从所述排序操作后的相似度集中选择指定数量的特征词向量，基于所述指定数量的特征词向量发现热点关键字并输出所述原始文本数据集的热点。

此外，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上存储有文本热点的发现程序，所述文本热点的发现程序可被一个或者多个处理器执行，以实现如上所述的文本热点的发现方法的步骤。
25

本申请首先爬取新闻论坛的实时文本数据，通过前期较精准的分词、词性标准的预处理，可以有效提取出可能属于热点关键字的词语，进一步地，通过词向量的转换，在不损失特征精准的同时，可高效的让计算机进行分析，最后基于特征相似度的计算遍历出热点关键字，从而得到当前的文本热点。
30

因此本申请提出的文本热点的发现方法、装置及计算机可读存储介质可以实现精准高效的文本热点发现功能。

附图说明

- 5 图1为本申请一实施例提供的文本热点的发现方法的流程示意图；
图2为本申请一实施例提供的文本热点的发现装置的内部结构示意图；
图3为本申请一实施例提供的文本热点的发现装置中文本热点的发现程序的模块示意图。
- 本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步
10 说明。

具体实施方式

应当理解，此处所描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

- 15 本申请提供一种文本热点的发现方法。参照图 1 所示，为本申请一实施例提供的文本热点的发现方法的流程示意图。该方法可以由一个装置执行，该装置可以由软件和/或硬件实现。

在本实施例中，文本热点的发现方法包括：

- 20 S1、从新闻论坛网站上爬取原始文本数据集及标签集，所述标签集记录所述原始文本数据集内文本的发布时间。

较佳地，所述爬取可采用爬虫技术，所述爬虫技术是先创建 URL 队列，其中所述 URL 队列包括若干个 URL，然后依次读取所述 URL 队列内的 URL 并解析为 IP 地址，最后基于 HTTP 通信协议下载所述 IP 地址指定的网页数据，并对所述网页数据进行分析得到原始文本数据集及标签集。

- 25 较佳地，所述 URL 称为统一资源定位符，是对所述新闻论坛网站内的各种资源位置和访问方法的一种简洁表示，又称为所述新闻论坛网站内的各种资源的地址。所述 URL 由协议（protocol）、主机名（hostname）、端口（port）、路径（path）、查询字符串、散列符元素等组成。所述协议（protocol）表示访问资源和服务的协议，例如 http、ftp、mailto 和 file 等；所述主机名（hostname）
30 表示资源所在主机的完全限定域名，例如 www.baidu.com。所述端口（port）

表示协议使用的 TCP 端口号，其中所述 HTTP 通信协议的常用端口为 80，一般采用默认省略的模式；所述路径（path）表示资源的目录/文件路径名；所述查询字符串表示 URL 中传递的查询字符串；所述散列符元素表示所述 URL 所指定的文件偏移量，包括散列符（#）加上所述文件偏移量相关的位置。

5 进一步地，所述解析为 IP 地址即提取所述协议（protocol）、主机名（hostname）、端口（port）、路径（path）等得到所述 IP 地址。

优选地，所述 URL 一般为指定的新闻、微博等 URL，因为所述新闻、微博等 URL 的网页数据有文本数据及发布时间，将所述文本数据组建成原始文本数据集，所述原始文本数据集内文本数据的发布时间在标签集中。

10 S2、将所述原始文本数据集进行包括分词、词性标注、去停用词的预处理操作得到初级文本数据集。

较佳地，因为在汉语表示中，词和词之间没有明确的分隔标识，因此要对所述原始文本数据集进行分词处理。所述分词处理使用基于 Python、JAVA 等编程语言的 jieba 分词进行处理，如所述原始文本数据集中有文本数据为：

15 “杨宇彬是一位有名的创业青年，靠着扎实的知识 and 勤劳实干在当地开始了自己的事业”。基于所述 jieba 分词进行处理后得到为：[杨宇彬][是][一位][有名的][创业][青年][，][靠着][扎实的][知识][和][勤劳实干][在][当地][开始][了][自己的][事业]。

进一步地，所述词性标注是基于预先构建的词性标记模板标注出所述分
20 词完成的原始文本数据集中的名词、动词。其中，所述词性标记模板是指名词、动词特征的识别器，所述词性标记模板可以通过识别词语的特征来确定名词、动词。如上述[杨宇彬][是][一位][有名的][创业][青年][，][靠着][扎实的][知识][和][勤劳实干][在][当地][开始][了][自己的][事业]，根据所述词性标记模板标注出为名词为[青年]、[知识]、[当地]、[事业]，动词为[创业]、[靠着]、
25 [开始]；

搜索所述原始文本数据集内长度大于预设长度，如两个字符并含有“的”或“地”的词，并判断所述长度大于两个字符并含有“的”或“地”的词在所述文本数据中的前后词是否是名词或动词。若所述前后词是名词或动词，则所述长度大于预设长度字符并含有“的”或“地”的词即为形容词或副词，如[杨宇彬]
30 [是][一位][有名的][创业][青年][，][靠着][扎实的][知识][和][勤劳实干][在][当

地][开始][了][自己的][事业], 先根据所述词性标记模板识别出名词为[青年]、[知识]、[当地]、[事业], 动词为[创业]、[靠着]、[开始], 同时识别出长度大于两个字符并含有“的”或“地”的词为[有名的]、[扎实的]、[自己的], 判断出所述词前后都有名词或动词如[创业]、[靠着]、[知识]等, 因此为形容词或副词并标注。较佳地, 所述标注方式可采用包括标注符号的形式, 如[杨宇彬是一位有名的^{adj}创业^v青年ⁿ, 靠着^v扎实的^{adj}知识ⁿ和勤劳实干在当地ⁿ开始^v了自己的^{adj}事业ⁿ]。

进一步地, 所述异形词如所有英文字母、阿拉伯数字、中文数字、标点符号、停用词等, 所述停用词包括“了”“于”等用词, 如上述经过去除所述异形词后得到为[有名的^{adj}创业^v青年ⁿ靠着^v扎实的^{adj}知识ⁿ当地ⁿ开始^v自己的^{adj}事业ⁿ]。

S3、基于所述标签集对所述初级文本数据集进行特征提取得到特征数据集, 并将所述特征数据集转化为特征词向量集。

较佳地, 所述特征提取为:

$$DF(t, c) = \frac{\lg(DF_t)}{\lg(N_c)}$$

其中, DF_t 表示所述初级文本数据集内特征词 t 在所述初级文本数据集中出现的文本数, N_c 为所述初级文本数据集的数据总数, c 为所述初级文本数据集, $\lg$ 表示以10为底数的 $\log$ 函数。例如在上述[有名的^{adj}创业^v青年ⁿ靠着^v扎实的^{adj}知识ⁿ当地ⁿ开始^v自己的^{adj}事业ⁿ]中, [有名的]、[创业]等都为特征词。

进一步地, 所述特征数据集转化为特征词向量集的转化过程包括假设出所述特征数据集内的特征与所述特征词向量集内的特征词向量之间的权重关系、基于所述权重关系计算所述权重, 完成所述转化过程。

具体地, 所述权重关系为:

$$d = \{(t_1, w_1), (t_2, w_2), \dots, (t_i, w_i), \dots, (t_n, w_n)\}$$

其中, d 为所述特征词向量集, t_1 、 t_2 、 $\dots$ 、 t_n 为所述特征数据集内的特征, 如上述[有名的]、[创业]等, w_1 、 w_2 、 $\dots$ 、 w_n 为所述对应特征的权重。

进一步地, 所述权重的计算方法为:

$$w_i = \frac{\frac{f_i}{N_i} \times \log \frac{N+1}{N_j}}{\sqrt{\sum_{k=1}^i \left(\frac{f_{ij}}{N_i} \times \log \frac{N+1}{N_j} \right)^2}} \times F_m$$

其中， f_i 表示特征词在所述初级文本数据集中出现的次数， N 为文档合集中文档的总数， N_j 表示所述初级文本数据集中特征词总数， N_i 表示特征词 i 在所述初级文本数据集的出现次数， F_m 为加权因子，一般取值为小于1。

5 S4、计算所述特征词向量集内特征之间的相似度得到相似度集，并对所述相似度集进行排序操作，从所述排序操作后的相似度集中选择指定数量的特征词向量，基于所述指定数量的特征词向量发现热点关键字并输出所述原始文本数据集的热点。

较佳地，所述相似度的计算方法为：

$$\text{sim}(d, t) = \alpha \times \frac{\sum_{k=1}^n w_{dk} \times w_{tk}}{\sqrt{\sum_{k=1}^n w_{dk}^2 \times \sum_{k=1}^n w_{tk}^2}} + \beta \times T$$

10 其中， $\text{sim}(d, t)$ 表示特征词向量 d 、 t 之间的相似度， w 表示所述特征词向量 d 、 t 与所述特征词向量集其他特征词向量 k 的权重系数， n 为所述特征词向量集中的数据总数， α 、 β 为偏置系数，其中 $\alpha + \beta = 1$ ， T 为时间距离函数。

进一步地，所述时间距离函数 T 为：

$$T = \frac{2t_d - t_{Ts} - t_{Te}}{2}$$

其中， t_d 表示所述标签集内所述特征词向量 d 所在文本的发布时间， t_{Ts} 是所述标签集中文本数据最早的发布时间， t_{Te} 是所述标签集中最新的发布时间。

15 较佳地，遍历并按照相似度从大到小的排序方法排序所述相似度集，选取相似度高所对应的特征词向量并最终得到特征词，如上述原始文本数据集中包括了如“杨宇彬是一位有名的创业青年，靠着扎实的知识和勤劳实干在当地开始了自己的事业”的文本数据，经过本申请所述方法进行分析后得到所述原始文本数据集的文本热点关键词为“创业”。

20

发明还提供一种文本热点的发现装置。参照图2所示，为本申请一实施例提供的文本热点的发现装置的内部结构示意图。

在本实施例中，所述文本热点的发现装置1可以是PC(Personal Computer，个人电脑)，或者是智能手机、平板电脑、便携计算机等终端设备，也可以是一种服务器等。该文本热点的发现装置1至少包括存储器11、处理器12，通

25

信总线 13，以及网络接口 14。

其中，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）、磁性存储器、磁盘、光盘等。存储器 11 在一些实施例中可以是文本热点的发现装置 1 的内部存储单元，例如该文本热点的发现装置 1 的硬盘。存储器 11 在另一些实施例中也可以是文本热点的发现装置 1 的外部存储设备，例如文本热点的发现装置 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。进一步地，存储器 11 还可以既包括文本热点的发现装置 1 的内部存储单元也包括外部存储设备。存储器 11 不仅可以用于存储安装于文本热点的发现装置 1 的应用软件及各类数据，例如文本热点的发现程序 01 的代码等，还可以用于暂时地存储已经输出或者将要输出的数据。

处理器 12 在一些实施例中可以是一中央处理器（Central Processing Unit, CPU）、控制器、微控制器、微处理器或其他数据处理芯片，用于运行存储器 11 中存储的程序代码或处理数据，例如执行文本热点的发现程序 01 等。

通信总线 13 用于实现这些组件之间的连接通信。

网络接口 14 可选的可以包括标准的有线接口、无线接口（如 WI-FI 接口），通常用于在该装置 1 与其他电子设备之间建立通信连接。

可选地，该装置 1 还可以包括用户接口，用户接口可以包括显示器（Display）、输入单元比如键盘（Keyboard），可选的用户接口还可以包括标准的有线接口、无线接口。可选地，在一些实施例中，显示器可以是 LED 显示器、液晶显示器、触控式液晶显示器以及 OLED（Organic Light-Emitting Diode，有机发光二极管）触摸器等。其中，显示器也可以适当的称为显示屏或显示单元，用于显示在文本热点的发现装置 1 中处理的信息以及用于显示可视化的用户界面。

图 2 仅示出了具有组件 11-14 以及文本热点的发现程序 01 的文本热点的发现装置 1，本领域技术人员可以理解的是，图 1 示出的结构并不构成对文本热点的发现装置 1 的限定，可以包括比图示更少或者更多的部件，或者组合某些部件，或者不同的部件布置。

在图 2 所示的装置 1 实施例中，存储器 11 中存储有文本热点的发现程序

01; 处理器 12 执行存储器 11 中存储的文本热点的发现程序 01 时实现如下步骤:

步骤一、从新闻论坛网站上爬取原始文本数据集及标签集, 所述标签集记录所述原始文本数据集内文本的发布时间。

5 较佳地, 所述爬取可采用爬虫技术, 所述爬虫技术是先创建 URL 队列, 其中所述 URL 队列包括若干个 URL, 然后依次读取所述 URL 队列内的 URL 并解析为 IP 地址, 最后基于 HTTP 通信协议下载所述 IP 地址指定的网页数据, 并对所述网页数据进行分析得到原始文本数据集及标签集。

较佳地, 所述 URL 称为统一资源定位符, 是对所述新闻论坛网站内的各种资源位置和访问方法的一种简洁表示, 又称为所述新闻论坛网站内的各种资源的地址。所述 URL 由协议(protocol)、主机名(hostname)、端口(port)、路径(path)、查询字符串、散列符元素等组成。所述协议(protocol)表示访问资源和服务的协议, 例如 http、ftp、mailto 和 file 等; 所述主机名(hostname)表示资源所在主机的完全限定域名, 例如 www.baidu.com。所述端口(port)表示协议使用的 TCP 端口号, 其中所述 HTTP 通信协议的常用端口为 80, 一般采用默认省略的模式; 所述路径(path)表示资源的目录/文件路径名; 所述查询字符串表示 URL 中传递的查询字符串; 所述散列符元素表示所述 URL 所指定的文件偏移量, 包括散列符(#)加上所述文件偏移量相关的位置。

进一步地, 所述解析为 IP 地址即提取所述协议(protocol)、主机名(hostname)、端口(port)、路径(path)等得到所述 IP 地址。

优选地, 所述 URL 一般为指定的新闻、微博等 URL, 因为所述新闻、微博等 URL 的网页数据有文本数据及发布时间, 将所述文本数据组建成原始文本数据集, 所述原始文本数据集内文本数据的发布时间在标签集中。

步骤二、将所述原始文本数据集进行包括分词、词性标注、去停用词的预处理操作得到初级文本数据集。

较佳地, 因为在汉语表示中, 词和词之间没有明确的分隔标识, 因此要对所述原始文本数据集进行分词处理。所述分词处理使用基于 Python、JAVA 等编程语言的 jieba 分词进行处理, 如所述原始文本数据集中有文本数据为:

“杨宇彬是一位有名的创业青年, 靠着扎实的知识和勤劳实干在当地开始了自己的事业”。基于所述 jieba 分词进行处理后得到为: [杨宇彬][是][一位][有

名的[[创业]]青年]], [[靠着]]扎实的[[知识]]和[[勤劳实干]]在[[当地]]开始]]了]]自己的]]事业]]。

进一步地, 所述词性标注是基于预先构建的词性标记模板标注出所述分词完成的原始文本数据集中的名词、动词。其中, 所述词性标记模板是指名词、动词特征的识别器, 所述词性标记模板可以通过识别词语的特征来确定名词、动词。如上述[杨宇彬] [是]]一位]]有名的]]创业]]青年]], [[靠着]]扎实的[[知识]]和[[勤劳实干]]在[[当地]]开始]]了]]自己的]]事业]], 根据所述词性标记模板标注出为名词为[青年]、[知识]、[当地]、[事业], 动词为[创业]、[靠着]、[开始];

搜索所述原始文本数据集内长度大于预设长度, 如两个字符并含有“的”或“地”的词, 并判断所述长度大于两个字符并含有“的”或“地”的词在所述文本数据中的前后词是否是名词或动词。若所述前后词是名词或动词, 则所述长度大于预设长度字符并含有“的”或“地”的词即为形容词或副词, 如[杨宇彬] [是]]一位]]有名的]]创业]]青年]], [[靠着]]扎实的[[知识]]和[[勤劳实干]]在[[当地]]开始]]了]]自己的]]事业]], 先根据所述词性标记模板识别出名词为[青年]、[知识]、[当地]、[事业], 动词为[创业]、[靠着]、[开始], 同时识别出长度大于两个字符并含有“的”或“地”的词为[有名的]、[扎实的]、[自己的], 判断出所述词前后都有名词或动词如[创业]、[靠着]、[知识]等, 因此为形容词或副词并标注。较佳地, 所述标注方式可采用包括标注符号的形式, 如[杨宇彬是一位有名的^{adj}创业^v青年ⁿ, 靠着^v扎实的^{adj}知识ⁿ和勤劳实干在当地ⁿ开始^v了自己的^{adj}事业ⁿ]]。

进一步地, 所述异形词如所有英文字母、阿拉伯数字、中文数字、标点符号、停用词等, 所述停用词包括“了”“于”等用词, 如上述经过去除所述异形词后得到为[有名的^{adj}创业^v青年ⁿ靠着^v扎实的^{adj}知识ⁿ当地ⁿ开始^v自己的^{adj}事业ⁿ]]。

步骤三、基于所述标签集对所述初级文本数据集进行特征提取得到特征数据集, 并将所述特征数据集转化为特征词向量集。

较佳地, 所述特征提取为:

$$DF(t, c) = \frac{\lg(DF_t)}{\lg(N_c)}$$

其中, DF_t 表示所述初级文本数据集内特征词 t 在所述初级文本数据集中

出现的文本数, N_c 为所述初级文本数据集的数据总数, c 为所述初级文本数据集, $\lg$ 表示以 10 为底数的 $\log$ 函数。例如在上述[有名的^{adj}创业^v青年ⁿ靠着^v扎实的^{adj}知识ⁿ当地ⁿ开始^v自己的^{adj}事业ⁿ]中, [有名的]、[创业]等都为特征词。

- 5 进一步地, 所述特征数据集转化为特征词向量集的转化过程包括假设出所述特征数据集内的特征与所述特征词向量集内的特征词向量之间的权重关系、基于所述权重关系计算所述权重, 完成所述转化过程。

具体地, 所述权重关系为:

$$d = \{(t_1, w_1), (t_2, w_2), \dots, (t_i, w_i), \dots, (t_n, w_n)\}$$

- 10 其中, d 为所述特征词向量集, t_1 、 t_2 、 $\dots$ 、 t_n 为所述特征数据集内的特征, 如上述[有名的]、[创业]等, w_1 、 w_2 、 $\dots$ 、 w_n 为所述对应特征的权重。

进一步地, 所述权重的计算方法为:

$$w_i = \frac{\frac{f_i}{N_i} \times \log \frac{N+1}{N_j}}{\sqrt{\sum_{k=1}^i \left(\frac{f_{ij}}{N_i} \times \log \frac{N+1}{N_j} \right)^2}} \times F_m$$

其中, f_i 表示特征词在所述初级文本数据集中出现的次数, N 为文档合集中文档的总数, N_j 表示所述初级文本数据集中特征词总数, N_i 表示特征词 i 在所述初级文本数据集的出现次数, F_m 为加权因子, 一般取值为小于 1。

- 15 步骤四、计算所述特征词向量集内特征之间的相似度得到相似度集, 并对所述相似度集进行排序操作, 从所述排序操作后的相似度集中选择指定数量的特征词向量, 基于所述指定数量的特征词向量发现热点关键字并输出所述原始文本数据集的热点。

较佳地, 所述相似度的计算方法为:

$$\text{sim}(d, t) = \alpha \times \frac{\sum_{k=1}^n w_{dk} \times w_{tk}}{\sqrt{\sum_{k=1}^n w_{dk}^2 \times \sum_{k=1}^n w_{tk}^2}} + \beta \times T$$

- 20 其中, $\text{sim}(d, t)$ 表示特征词向量 d 、 t 之间的相似度, w 表示所述特征词向量 d 、 t 与所述特征词向量集其他特征词向量 k 的权重系数, n 为所述特征词向量集中的数据总数, α 、 β 为偏置系数, 其中 $\alpha + \beta = 1$, T 为时间距离函数。

进一步地, 所述时间距离函数 T 为:

$$T = \frac{2t_d - t_{Ts} - t_{Te}}{2}$$

其中, t_d 表示所述标签集内所述特征词向量 d 所在文本的发布时间, t_{Ts} 是

所述标签集中文本数据最早的发布时间, t_{Te} 是所述标签集中最新的发布时间。

较佳地, 遍历并按照相似度从大到小的排序方法排序所述相似度集, 选取相似度高所对应的特征词向量并最终得到特征词, 如上述原始文本数据集中包括了如“杨宇彬是一位有名的创业青年, 靠着扎实的知识和勤劳实干在当地开始了自己的事业”的文本数据, 经过本申请所述方法进行分析后得到所述原始文本数据集的文本热点关键词为“创业”。

可选地, 在其他实施例中, 文本热点的发现程序还可以被分割为一个或者多个模块, 一个或者多个模块被存储于存储器 11 中, 并由一个或多个处理器 (本实施例为处理器 12) 所执行以完成本申请, 本申请所称的模块是指能够完成特定功能的一系列计算机程序指令段, 用于描述文本热点的发现程序在文本热点的发现装置中的执行过程。

例如, 参照图 3 所示, 为本申请文本热点的发现装置一实施例中的文本热点的发现程序的程序模块示意图, 该实施例中, 所述文本热点的发现程序可以被分割为数据接收模块 10、数据处理模块 20、词向量转化模块 30、文本热点输出模块 40 示例性地:

所述数据接收模块 10 用于: 从新闻论坛网站上爬取原始文本数据集及标签集, 所述标签集记录所述原始文本数据集内文本的发布时间。

所述数据处理模块 20 用于: 将所述原始文本数据集进行包括分词、词性标注、去异形词的预处理操作得到初级文本数据集;

所述词向量转化模块 30 用于: 基于所述标签集对所述初级文本数据集进行特征提取操作, 得到特征数据集, 并将所述特征数据集转化为特征词向量集。

所述文本热点输出模块 40 用于: 计算所述特征词向量集内特征之间的相似度得到相似度集, 并对所述相似度集进行排序操作, 从所述排序操作后的相似度集中选择指定数量的特征词向量, 基于所述指定数量的特征词向量发现热点关键字并输出所述原始文本数据集的热点。

上述数据接收模块 10、数据处理模块 20、词向量转化模块 30、文本热点输出模块 40 等程序模块被执行时所实现的功能或操作步骤与上述实施例大体相同, 在此不再赘述。

此外，本申请实施例还提出一种计算机可读存储介质，所述计算机可读存储介质上存储有文本热点的发现程序，所述文本热点的发现程序可被一个或多个处理器执行，以实现如下操作：

5 从新闻论坛网站上爬取原始文本数据集及标签集，所述标签集记录所述原始文本数据集内文本的发布时间；

将所述原始文本数据集进行包括分词、词性标注、去异形词的预处理操作得到初级文本数据集；

10 基于所述标签集对所述初级文本数据集进行特征提取操作，得到特征数据集，并将所述特征数据集转化为特征词向量集；

计算所述特征词向量集内特征之间的相似度得到相似度集，并对所述相似度集进行排序操作，从所述排序操作后的相似度集中选择指定数量的特征词向量，基于所述指定数量的特征词向量发现热点关键字并输出所述原始文本数据集的热点。

15 需要说明的是，上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。并且本文中的术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、装置、物品或者方法不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、装置、物品或者方法所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、装置、物品或者方法中还存在另外的相同要素。

25 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例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如上所述的一个存储介质(如 ROM/RAM、磁碟、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，或者网络设备等)执行本申请各个实施例所述的方法。

30 以上仅为本申请的优选实施例，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

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。

权利要求书

1、一种文本热点的发现方法，其特征在于，所述方法包括：

从新闻论坛网站上爬取原始文本数据集及标签集，所述标签集记录所述
5 原始文本数据集内文本的发布时间；

将所述原始文本数据集进行包括分词、词性标注、去异形词的预处理操作得到初级文本数据集；

基于所述标签集对所述初级文本数据集进行特征提取操作，得到特征数据集，并将所述特征数据集转化为特征词向量集；

10 计算所述特征词向量集内特征之间的相似度得到相似度集，并对所述相似度集进行排序操作，从所述排序操作后的相似度集中选择指定数量的特征词向量，基于所述指定数量的特征词向量发现热点关键字并输出所述原始文本数据集的热点。

2、如权利要求 1 所述的文本热点的发现方法，其特征在于，从新闻论坛
15 网站上爬取原始文本数据集及标签集，包括：

创建 URL 队列，其中，所述 URL 队列包括若干个 URL；

依次读取所述 URL 队列内的 URL 并解析为 IP 地址；

基于 HTTP 通信协议下载所述 IP 地址指定的网页数据，并对所述网页数据进行分析得到所述原始文本数据集和标签集。

20 3、如权利要求 2 所述的文本热点的发现方法，其特征在于，所述特征提取操作为：

$$DF(t, c) = \frac{\lg(DF_t)}{\lg(N_c)}$$

其中， $DF(t, c)$ 为所述特征数据集， DF_t 表示所述初级文本数据集内特征词 t 在所述初级文本数据集中出现的文本数， N_c 为所述初级文本数据集的数据总数， c 为所述初级文本数据集。

25 4、如权利要求 3 所述的文本热点的发现方法，其特征在于，计算所述特征词向量集内特征之间的相似度的计算方法为：

$$\text{sim}(d, t) = \alpha \times \frac{\sum_{k=1}^n w_{dk} \times w_{tk}}{\sqrt{\sum_{k=1}^n w_{dk}^2 \times \sum_{k=1}^n w_{tk}^2}} + \beta \times T$$

其中， $\text{sim}(d, t)$ 表示特征词向量 d 、 t 之间的相似度， w 表示所述特征词向

量 d 、 t 与所述特征词向量集其他特征词向量的权重系数， n 为所述特征词向量集中的数据总数， α 、 β 为偏置系数，其中 $\alpha + \beta = 1$ ， T 为时间距离函数。

5、如权利要求4所述的文本热点的发现方法，其特征在于，所述时间距离函数 T 为：

$$T = \frac{2t_d - t_{Ts} - t_{Te}}{2}$$

5 其中， t_d 表示所述标签集内所述特征词向量 d 所在文本的发布时间， t_{Ts} 是所述标签集中最早的发布时间， t_{Te} 是所述标签集中最新的发布时间。

6、如权利要求1所述的文本热点的发现方法，其特征在于，所述将所述特征数据集转化为特征词向量集包括：

10 设定所述特征数据集内的特征与所述特征词向量集内的特征词向量之间的权重关系、基于所述权重关系计算所述权重，完成所述转化过程。

7、如权利要求6所述的文本热点的发现方法，其特征在于，所述权重的计算方法包括：

$$w_i = \frac{\frac{f_i}{N_i} \times \log \frac{N+1}{N_j}}{\sqrt{\sum_{k=1}^i \left(\frac{f_{ik}}{N_i} \times \log \frac{N+1}{N_j} \right)^2}} \times F_m$$

15 其中， f_i 表示特征词在所述初级文本数据集中出现的次数， N 表示文档合集中文档的总数， N_j 表示所述初级文本数据集中特征词总数， N_i 表示特征词 i 在所述初级文本数据集的出现次数， F_m 表示加权因子， F_m 取值小于1。

8、一种文本热点的发现装置，其特征在于，所述装置包括存储器和处理器，所述存储器上存储有可在所述处理器上运行的文本热点的发现程序，所述文本热点的发现程序被所述处理器执行时实现如下步骤：

20 从新闻论坛网站上爬取原始文本数据集及标签集，所述标签集记录所述原始文本数据集内文本的发布时间；

将所述原始文本数据集进行包括分词、词性标注、去异形词的预处理操作得到初级文本数据集；

基于所述标签集对所述初级文本数据集进行特征提取操作，得到特征数据集，并将所述特征数据集转化为特征词向量集；

25 计算所述特征词向量集内特征之间的相似度得到相似度集，并对所述相似度集进行排序操作，从所述排序操作后的相似度集中选择指定数量的特征

词向量，基于所述指定数量的特征词向量发现热点关键字并输出所述原始文本数据集的热点。

9、如权利要求 8 所述的文本热点的发现装置，其特征在于，从新闻论坛网站上爬取原始文本数据集及标签集，包括：

5 创建 URL 队列，其中，所述 URL 队列包括若干个 URL；

依次读取所述 URL 队列内的 URL 并解析为 IP 地址；

基于 HTTP 通信协议下载所述 IP 地址指定的网页数据，并对所述网页数据进行分析得到所述原始文本数据集和标签集。

10 10、如权利要求 9 所述的文本热点的发现装置，其特征在于，所述特征提取操作为：

$$DF(t, c) = \frac{\lg(DF_t)}{\lg(N_c)}$$

其中， $DF(t, c)$ 为所述特征数据集， DF_t 表示所述初级文本数据集内特征词 t 在所述初级文本数据集中出现的文本数， N_c 为所述初级文本数据集的数据总数， c 为所述初级文本数据集。

15 11、如权利要求 10 所述的文本热点的发现装置，其特征在于，计算所述特征词向量集内特征之间的相似度的计算方法为：

$$\text{sim}(d, t) = \alpha \times \frac{\sum_{k=1}^n w_{dk} \times w_{tk}}{\sqrt{\sum_{k=1}^n w_{dk}^2 \times \sum_{k=1}^n w_{tk}^2}} + \beta \times T$$

其中， $\text{sim}(d, t)$ 表示特征词向量 d 、 t 之间的相似度， w 表示所述特征词向量 d 、 t 与所述特征词向量集其他特征词向量的权重系数， n 为所述特征词向量集中的数据总数， α 、 β 为偏置系数，其中 $\alpha + \beta = 1$ ， T 为时间距离函数。

20 12、如权利要求 11 所述的文本热点的发现装置，其特征在于，所述时间距离函数 T 为：

$$T = \frac{2t_d - t_{Ts} - t_{Te}}{2}$$

其中， t_d 表示所述标签集内所述特征词向量 d 所在文本的发布时间， t_{Ts} 是所述标签集中最早的发布时间， t_{Te} 是所述标签集中最新的发布时间。

13、如权利要求 8 所述的文本热点的发现装置，其特征在于，所述将所述特征数据集转化为特征词向量集包括：

25 设定所述特征数据集内的特征与所述特征词向量集内的特征词向量之间的权重关系、基于所述权重关系计算所述权重，完成所述转化过程。

14、如权利要求 13 所述的文本热点的发现装置，其特征在于，所述权重的计算方法包括：

$$w_i = \frac{\frac{f_i}{N_i} \times \log \frac{N+1}{N_j}}{\sqrt{\sum_{k=1}^i \left(\frac{f_{ij}}{N_i} \times \log \frac{N+1}{N_j} \right)^2}} \times F_m$$

其中， f_i 表示特征词在所述初级文本数据集中出现的次数， N 表示文档合集中文档的总数， N_j 表示所述初级文本数据集中特征词总数， N_i 表示特征词 i 在所述初级文本数据集的出现次数， F_m 表示加权因子， F_m 取值小于 1。

15、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质上存储有文本热点的发现程序，所述文本热点的发现程序可被一个或者多个处理器执行时，实现如下步骤：

10 从新闻论坛网站上爬取原始文本数据集及标签集，所述标签集记录所述原始文本数据集内文本的发布时间；

将所述原始文本数据集进行包括分词、词性标注、去异形词的预处理操作得到初级文本数据集；

15 基于所述标签集对所述初级文本数据集进行特征提取操作，得到特征数据集，并将所述特征数据集转化为特征词向量集；

计算所述特征词向量集内特征之间的相似度得到相似度集，并对所述相似度集进行排序操作，从所述排序操作后的相似度集中选择指定数量的特征词向量，基于所述指定数量的特征词向量发现热点关键字并输出所述原始文本数据集的热点。

20 16、如权利要求 15 所述的计算机可读存储介质，其特征在于，从新闻论坛网站上爬取原始文本数据集及标签集，包括：

创建 URL 队列，其中，所述 URL 队列包括若干个 URL；

依次读取所述 URL 队列内的 URL 并解析为 IP 地址；

25 基于 HTTP 通信协议下载所述 IP 地址指定的网页数据，并对所述网页数据进行分析得到所述原始文本数据集和标签集。

17、如权利要求 16 所述的计算机可读存储介质，其特征在于，所述特征提取操作为：

$$DF(t, c) = \frac{\lg(DF_t)}{\lg(N_c)}$$

其中, $DF(t, c)$ 为所述特征数据集, DF_t 表示所述初级文本数据集内特征词 t 在所述初级文本数据集中出现的文本数, N_c 为所述初级文本数据集的数据总数, c 为所述初级文本数据集。

18、如权利要求 17 所述的计算机可读存储介质, 其特征在于, 计算所述
5 特征词向量集内特征之间的相似度的计算方法为:

$$\text{sim}(d, t) = \alpha \times \frac{\sum_{k=1}^n w_{dk} \times w_{tk}}{\sqrt{\sum_{k=1}^n w_{dk}^2 \times \sum_{k=1}^n w_{tk}^2}} + \beta \times T$$

其中, $\text{sim}(d, t)$ 表示特征词向量 d 、 t 之间的相似度, w 表示所述特征词向量 d 、 t 与所述特征词向量集其他特征词向量的权重系数, n 为所述特征词向量集中的数据总数, α 、 β 为偏置系数, 其中 $\alpha + \beta = 1$, T 为时间距离函数。

19、如权利要求 18 所述的计算机可读存储介质, 其特征在于, 所述时间
10 距离函数 T 为:

$$T = \frac{2t_d - t_{Ts} - t_{Te}}{2}$$

其中, t_d 表示所述标签集内所述特征词向量 d 所在文本的发布时间, t_{Ts} 是所述标签集中最早的发布时间, t_{Te} 是所述标签集中最新的发布时间。

20、如权利要求 15 所述的计算机可读存储介质, 其特征在于, 所述将所述特征数据集转化为特征词向量集包括:

15 设定所述特征数据集内的特征与所述特征词向量集内的特征词向量之间的权重关系、基于所述权重关系计算所述权重, 完成所述转化过程。

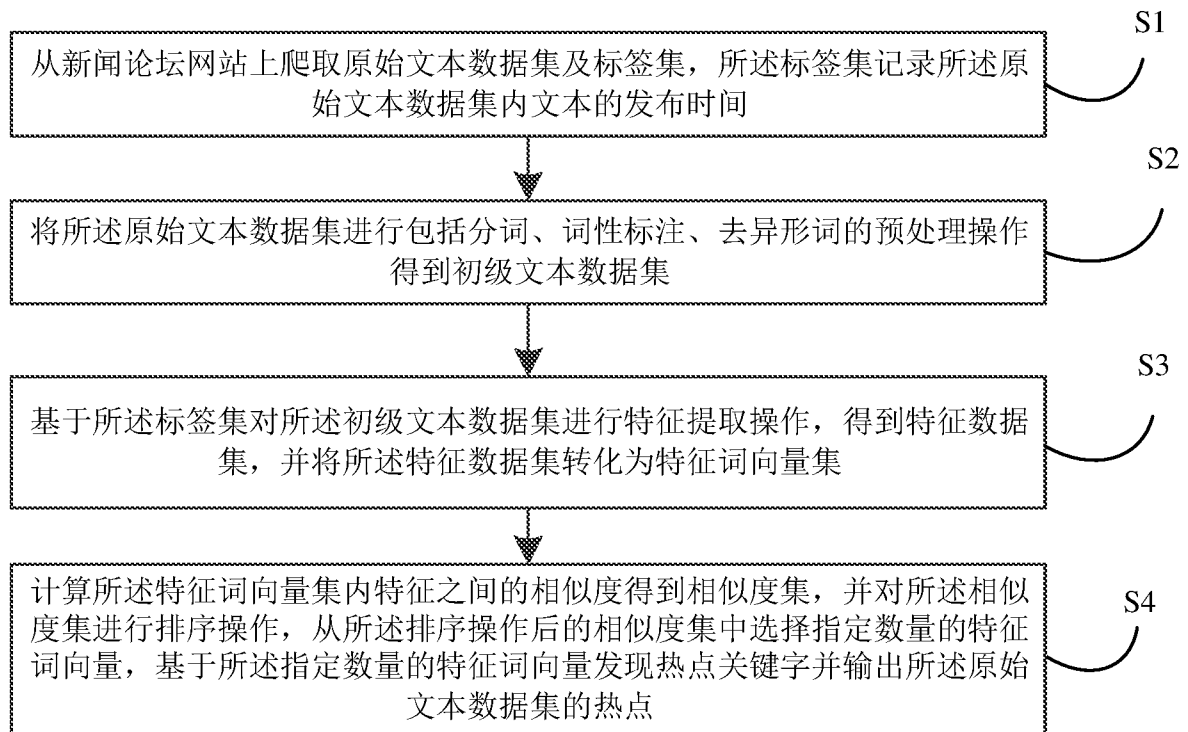


图 1

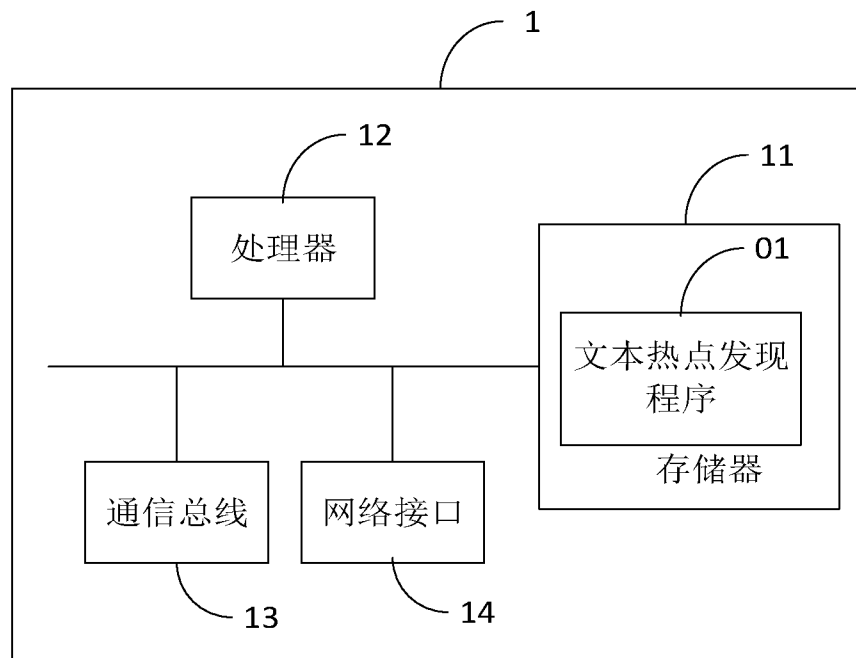


图 2

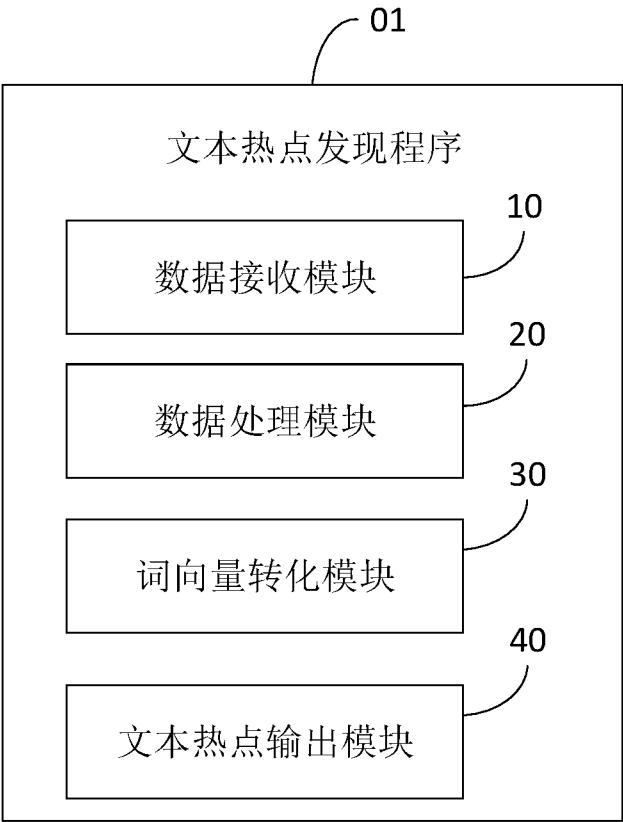


图 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/116550

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/35(2019.01)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)

CNABS, CNTXT, DWPI, CNKI, ACM, IEEE: 话题, 识别, 新闻, 热点, 发现, 微博, 推特, 文本, 相似, 向量, 权重, 时间, 词频, 倒排, topic, recognize, news, hot, detect, twitter, text, document, similar, vector, weight, time, DF, IDF

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 102937960 A (BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 20 February 2013 (2013-02-20) description, paragraphs [0068]-[0085]	1-2, 6, 8-9, 13, 15-16, 20
Y	CN 102937960 A (BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 20 February 2013 (2013-02-20) description, paragraphs [0068]-[0085]	3-5, 7, 10-12, 14, 17-19
Y	张婷 (ZHANG, Ting). "新闻报道的交互式时间线系统研究 (Research on Interactive Timeline System of News Report)" 中国优秀硕士学位论文全文数据库信息科技辑 (Information & Technology, China Master's Theses Full-Text Database), 15 January 2019 (2019-01-15), sections 3.2.1, 3.2.2	3-5, 10-12, 17-19
Y	CN 101488150 A (HARBIN ENGINEERING UNIVERSITY) 22 July 2009 (2009-07-22) claim 5	7, 14
A	CN 106599181 A (ZHEJIANG INSIGMA HENGtian SOFTWARE LTD.) 26 April 2017 (2017-04-26) 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

29 February 2020

Date of mailing of the international search report

06 March 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/116550

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 107895053 A (FUZHOU UNIVERSITY) 10 April 2018 (2018-04-10) entire document	1-20

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/116550

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)		Publication date (day/month/year)
CN	102937960	A	20 February 2013	CN	102937960	B 17 June 2015
CN	101488150	A	22 July 2009	CN	101488150	B 05 January 2011
CN	106599181	A	26 April 2017	CN	106599181	B 18 June 2019
CN	107895053	A	10 April 2018	None		

国际检索报告

国际申请号

PCT/CN2019/116550

A. 主题的分类

G06F 16/35 (2019.01) 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))

CNABS, CNTXT, DWPI, CNKI, ACM, IEEE: 话题, 识别, 新闻, 热点, 发现, 微博, 推特, 文本, 相似, 向量, 权重, 时间, 词频, 倒排, topic, recognize, news, hot, detect, twitter, text, document, similar, vector, weight, time, DF, IDF

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 102937960 A (北京邮电大学) 2013年 2月 20日 (2013 - 02 - 20) 说明书[0068]-[0085]段	1-2, 6, 8-9, 13, 15-16, 20
Y	CN 102937960 A (北京邮电大学) 2013年 2月 20日 (2013 - 02 - 20) 说明书[0068]-[0085]段	3-5, 7, 10- 12, 14, 17-19
Y	张婷. “新闻报道的交互式时间线系统研究” 中国优秀硕士学位论文全文数据库信息科技辑, 2019年 1月 15日 (2019 - 01 - 15), 3.2.1、3.2.2节	3-5, 10-12, 17-19
Y	CN 101488150 A (哈尔滨工程大学) 2009年 7月 22日 (2009 - 07 - 22) 权利要求5	7, 14
A	CN 106599181 A (浙江网新恒天软件有限公司) 2017年 4月 26日 (2017 - 04 - 26) 全文	1-20
A	CN 107895053 A (福州大学) 2018年 4月 10日 (2018 - 04 - 10) 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

2020年 2月 29日

国际检索报告邮寄日期

2020年 3月 6日

ISA/CN的名称和邮寄地址

中国国家知识产权局(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

慈丽雁

电话号码 62411748

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/116550

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	102937960	A	2013年 2月 20日	CN 102937960 B	2015年 6月 17日
CN	101488150	A	2009年 7月 22日	CN 101488150 B	2011年 1月 5日
CN	106599181	A	2017年 4月 26日	CN 106599181 B	2019年 6月 18日
CN	107895053	A	2018年 4月 10日	无	