

(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



(10) 国际公布号

WO 2019/218452 A1

(43) 国际公布日
2019 年 11 月 21 日 (21.11.2019)

(51) 国际专利分类号:
G06F 17/00 (2019.01)

(21) 国际申请号: PCT/CN2018/096267

(22) 国际申请日: 2018 年 7 月 19 日 (19.07.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201810456973.4 2018 年 5 月 14 日 (14.05.2018) 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)。

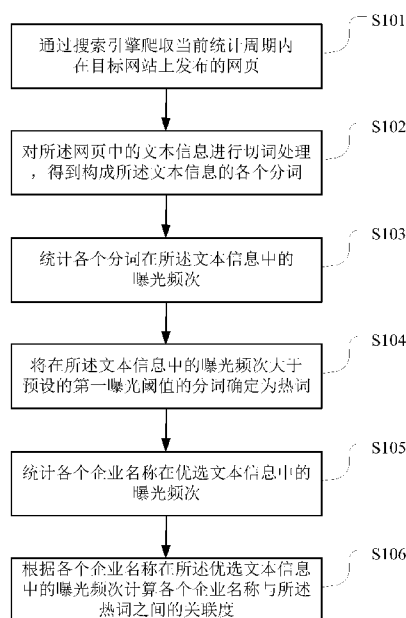
(72) 发明人: 张依(ZHANG, Yi); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 汪伟(WANG, Wei); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 肖京(XIAO, Jing); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳中一专利商标事务所(SHENZHEN ZHONGYI PATENT AND TRADEMARK OFFICE);
中国广东省深圳市福田区深南中路 1014 号老特区报社四楼 (5 号信箱), Guangdong 518028 (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,

(54) Title: METHOD, COMPUTER READABLE STORAGE MEDIUM, TERMINAL APPARATUS, AND DEVICE FOR ANALYZING TRENDING TERMS

(54) 发明名称: 热词分析方法、计算机可读存储介质、终端设备及装置



S101 Crawl a web page published on a target website in a current counting period by means of a search engine
S102 Perform term segmentation processing on text information in the web page, so as to obtain respective segmented terms forming the text information
S103 Count the frequency of each of the segmented terms in the text information
S104 Determine a segmented term having a frequency greater than a first preset threshold in the text information to be a trending term
S105 Count the frequencies of respective company names in preferred text information
S106 Calculate, according to the frequency of each of the company names in the preferred text information, a degree of association between each of the company names and the trending terms

图 1

(57) Abstract: A method, a computer readable storage medium, a terminal apparatus, and a device for analyzing trending terms, applied in the technical field of computers. The method comprises: crawling a web page published on a target website in a current counting period by means of a search engine (S101); performing term segmentation processing on text information in the web page, so as to obtain respective segmented terms forming the text information (S102); counting the frequency of each of the segmented terms in the text information (S103); determining a segmented term having a frequency greater than a first preset threshold in the text information to

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。

(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))。

be a trending term (S104); counting the frequencies of respective company names in preferred text information (S105); and calculating, according to the frequency of each of the company names in the preferred text information, a degree of association between each of the company names and the trending terms (S106). The method provides an objective evaluation criterion for trending term determination, and comprehensively considers relationships between companies and trending terms after the trending terms are obtained, and an analysis result thereof provides better guidance for companies.

(57) 摘要: 一种热词分析方法、计算机可读存储介质、终端设备及装置, 应用于计算机技术领域。所述方法通过搜索引擎爬取当前统计周期内在目标网站上发布的网页(S101); 对所述网页中的文本信息进行切词处理, 得到构成所述文本信息的各个分词(S102); 统计各个分词在所述文本信息中的曝光频次(S103); 将在所述文本信息中的曝光频次大于预设的第一曝光阈值的分词确定为热词(S104); 统计各个企业名称在优选文本信息中的曝光频次(S105); 根据各个企业名称在所述优选文本信息中的曝光频次计算各个企业名称与所述热词之间的关联度(S106)。该方法为热词的确定提供了一套客观的评估标准, 而且在得到热词之后, 综合考虑了企业与热词之间的关系, 分析结果对于企业具有更强的指导意义。

热词分析方法、计算机可读存储介质、终端设备及装置

本申请要求于 2018 年 5 月 14 日提交中国专利局、申请号为 201810456973.4、发明名称为“一种热词分析方法、计算机可读存储介质及终端设备”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

本申请属于计算机技术领域，尤其涉及一种热词分析方法、计算机可读存储介质、终端设备及装置。

背景技术

热词，即网络热门词汇，是指一种词汇现象，反映了一个国家、一个地区在一个时期人们普遍关注的问题和事物。热词具有时代特征，可以作为一个时期的热点话题及民生问题的代表。

目前对于热词的确定，主要是由网络分析人员根据自己在互联网上所浏览到信息进行处理所得到的，这种方式依赖于网络分析人员个人的判断，主观性极强，难以客观地反应真实的热词情况，而且在得到热词之后，往往仅仅是只针对热词进行片面的分析，分析维度单一，分析结果对于企业的指导意义极差。

技术问题

有鉴于此，本申请实施例提供了一种热词分析方法、计算机可读存储介质、终端设备及装置，以解决现有技术中热词的确定过程主观性极强且分析结果对于企业的指导意义极差的问题。

20 技术解决方案

本申请实施例的第一方面提供了一种热词分析方法，可以包括：

通过搜索引擎爬取当前统计周期内在目标网站上发布的网页，所述目标网站为浏览量大于预设的浏览量阈值的网站；

对所述网页中的文本信息进行切词处理，得到构成所述文本信息的各个分词；

统计各个分词在所述文本信息中的曝光频次；

将在所述文本信息中的曝光频次大于预设的第一曝光阈值的分词确定为热词；

统计各个企业名称在优选文本信息中的曝光频次，所述优选文本信息为包含所述热词的文本信息；

根据各个企业名称在所述优选文本信息中的曝光频次计算各个企业名称与所述热词之间的关联度。

本申请实施例的第二方面提供了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计算机可读指令，所述计算机可读指令被处理器执行时实现上述热词分析方法的步骤。

本申请实施例的第三方面提供了一种热词分析终端设备，包括存储器、处理器以及存储在该存储器中并可在所述处理器上运行的计算机可读指令，所述处理器执行

所述计算机可读指令时实现上述热词分析方法的步骤。

本申请实施例的第四方面提供了一种热词分析装置，可以包括用于实现上述热词分析方法的步骤的模块。

有益效果

- 5 本申请实施例与现有技术相比存在的有益效果是：本申请实施例为热词的确定提供了一套客观的评估标准，摆脱了对网络分析人员个人能力的依赖，确定出的热词能够更加难以客观地反应真实的情况，而且在得到热词之后，综合考虑了企业与热词之间的关系，分析结果对于企业具有更强的指导意义。

附图说明

- 10 图 1 为本申请实施例中一种热词分析方法的一个实施例流程图；
图 2 为第一曝光阈值和第二曝光阈值的设置过程在一种具体实现中的示意流程图；
图 3 为第一曝光阈值和第二曝光阈值的设置过程在另一种具体实现中的示意流程图；
图 4 为本申请实施例中一种热词分析装置的一个实施例结构图；
15 图 5 为本申请实施例中一种热词分析终端设备的示意框图。

本发明的实施方式

请参阅图 1，本申请实施例中一种热词分析方法的一个实施例可以包括：

步骤 S101、通过搜索引擎爬取当前统计周期内在目标网站上发布的网页。

- 20 所述目标网站为浏览量大于预设的浏览量阈值的网站，可以根据实际情况将所述浏览量阈值设置为 10 万次、50 万次、100 万次等等，所述目标网站可以是百度新闻（<http://news.baidu.com/>）、网易新闻（<http://news.163.com/>）、腾讯新闻（<http://news.qq.com/>）、凤凰新闻（<http://news.ifeng.com/>）等新闻网站或者其它的新闻网站。

统计周期可以根据实际情况设置为一天、一周、半个月或者一个月等等。

- 25 步骤 S102、对所述网页中的文本信息进行切词处理，得到构成所述文本信息的各个分词。

切词处理是指将一个语句文本切分成一个一个单独的词，也即各个所述分词，在本实施例中，可以根据通用词典对语句文本进行切分，保证分出的词语都是正常词汇，如词语不在词典内则分出单字。

- 30 步骤 S103、统计各个分词在所述文本信息中的曝光频次。

也即分别统计各个分词在所述文本信息中出现的次数。

步骤 S104、将在所述文本信息中的曝光频次大于预设的第一曝光阈值的分词确定为热词。

- 35 例如，可以设置所述第一曝光阈值为 10000，若在当前统计周期内，“十九大”的曝光频次为 20000，大于了所述第一曝光阈值，则直接将其确认为热词。

可选地，还可以将在所述文本信息中的曝光频次小于或等于所述第一曝光阈值且

大于预设的第二曝光阈值的分词确定为候选分词，然后从历史统计记录中获取各个候选分词在所述当前统计周期之前的 T 个统计周期内的曝光频次，将满足下述条件的候选分词也确定为热词：

$$\ln\left(\frac{ExpNum_{n,t+1}}{ExpNum_{n,t}}\right) > ThreshRatio$$

对于任意的 t 的取值，不等式均成立。

- 5 其中， n 为所述候选分词的序号， $1 \leq n \leq N$ ， N 为所述候选分词的总数， t 为各个统计周期按照时间先后顺序依次排列的序号， $1 \leq t \leq T$ ， T 为正整数， $ExpNum_{n,t}$ 为第 n 个候选分词在第 t 个统计周期内的曝光频次， $ExpNum_{n,t+1}$ 为第 n 个候选分词在所述当前统计周期内的曝光频次， $\ln$ 为自然对数函数， $ThreshRatio$ 为预设的比例阈值。

- 10 例如，可以设置所述第二曝光阈值为 2000，所述比值阈值为 2，且 $T=1$ 。若在当前统计周期内，“雄安新区”的曝光频次为 9000，小于所述第一曝光阈值，但是大于所述第二曝光阈值，则获取其在当前统计周期之前的 1 个统计周期内的曝光频次，若上一统计周期内的曝光频次为 1000，不等式 $\ln\left(\frac{9000}{1000}\right) > 2$ 成立，则将其也确定为热词。

若在当前统计周期内，“人工智能”的曝光量为 1500 次，不仅小于所述第一曝光阈值，且小于所述第二曝光阈值，则直接将其确认为普通词语。

- 15 进一步地，还可以对确定出的热词进行过滤，也即从确定出的热词中将产生干扰的词语过滤掉，可以预先设置好一些常见的干扰热词，例如，像“我们”、“大家”、“这个”等等。这些干扰热词的曝光频次与新闻内容并无任何关系，也即不管新闻内容是什么，这些干扰热词的曝光频次都可能保持大于所述第一曝光阈值。在做热词统计时，如果不对这些干扰热词进行过滤处理，则会影响分析结果的准确度，因而需要从确定出的热词中滤除干扰热词，得到过滤后的热词，也即得到真正所需的热词。
- 20 具体地，在确定出热词之后，可以从数据列表中获取预先设定的干扰热词，然后，将所有确定出的热词与所有的干扰热词逐一对比，如果某个热词与某个干扰热词一致，则将其滤除，否则，如果某个热词与任何一个干扰热词都不一致，则保留该热词，最后被保留下来的热词即为过滤后的热词。

- 25 步骤 S105、统计各个企业名称在优选文本信息中的曝光频次。

所述优选文本信息为包含所述热词的文本信息。

可选地，所述当前统计周期可以包括 M 个子周期，其中， M 为正整数，则需统计各个企业名称在各个子时间段的优选文本信息中的曝光频次。

- 30 步骤 S106、根据各个企业名称在所述优选文本信息中的曝光频次计算各个企业名称与所述热词之间的关联度。

具体地，可以根据下式计算各个企业名称与所述热词之间的关联度：

$$Rel_{q,p} = \frac{\sum_{m=1}^M (k_m \times EntExpNum_{q,p,m})}{\sum_{m=1}^M [k_m \times (\sum_{q=1}^Q EntExpNum_{q,p,m})]}$$

其中, q 为企业名称的序号, $1 \leq q \leq Q$, Q 为企业名称的总数, p 为热词的序号, $1 \leq p \leq P$, P 为热词的总数, m 为各个子时间段按照时间先后顺序依次排列的序号, $1 \leq m \leq M$, $EntExpNum_{q,p,m}$ 为第 m 个子时间段内第 q 个企业名称在包含第 p 个热词的优

5 选文本信息中的曝光频次, k_m 为预设的权值系数, $k_m < k_{m+1}$ 且 $\sum_{m=1}^M k_m = 1$, $Rel_{q,p}$ 为第 q 个企业名称与第 p 个热词之间的关联度。

特殊地, 若企业名称 A 与所述热词 B 之间的关联度超过预设的关联度阈值, 则可以认为两者是唯一匹配的, 所述关联度阈值可以根据实际情况设置为 80%、90%或者 95%等等。

10 例如, 若要对热词“王者荣耀”进行企业关联分析, 则在包含该热词的文本信息中搜索企业名称, 统计各个企业名称的曝光频次, 并按照上述公式计算各个企业名称与该热词之间的关联度, 若关联度排序结果第一位为腾讯, 且其与该热词之间的关联度为 98%, 超过了所述关联度阈值, 则确定与热词“王者荣耀”唯一匹配的企业名称为腾讯。

15 以上是通过热词关联对应的企业, 另一个关联角度是通过企业关联对应的热词。具体地, 在网页中搜索出现该企业名称的文本信息, 然后在出现该企业的文本信息中搜索热词, 统计各个热词出现的频次, 并按照频次从大到小的顺序对热词进行排序, 排序越靠前的热词与该企业的关联度越高, 排序越靠后的热词与该企业的关联度越低。

20 例如, 若要对腾讯进行热词关联分析, 则在网页中搜索出现腾讯的文本信息, 然后在这些文本信息中搜索热词, 统计各个热词出现的频次, 并按照频次从大到小的顺序对热词进行排序, 若排序结果从大到小依次为: “王者荣耀”、“绝地求生”、“光荣使命”……, 则可确定当前与腾讯关联度最高的热词分别为“王者荣耀”、“绝地求生”、“光荣使命”……。

25 进一步地, 在进行企业关联分析时, 企业名称应包括其别称在内, 例如, 对腾讯进行关联分析时, 不仅要搜索“腾讯”, 还需搜索“Tencent”“鹅厂”等, 对阿里巴巴进行关联分析时, 不仅要搜索“阿里巴巴”, 还需搜索“Alibaba”“阿里”等。具体地, 可以预先设置企业名称的别称列表, 记录企业的正式名称与别称之间的对应关系, 当进行企业关联分析时, 从该列表中获取个企业对应的别称, 将该别称也计入对企业的统计过程中。

30 在本申请实施例的一种具体实现中, 所述第一曝光阈值和所述第二曝光阈值的设置过程可以包括如图 2 所示的步骤:

步骤 S201、从历史统计记录中获取各个历史热词在各个统计周期内的曝光频次。

所述历史热词为在所述当前统计周期之前已经确定的热词。

步骤 S202、构造各个历史热词的第一曝光序列。

具体地，可以根据下式构造各个历史热词的第一曝光序列：

$$ExpSeq1_{nh} = \{HsExpNum_{nh,1}, HsExpNum_{nh,2}, \dots, HsExpNum_{nh,th}, \dots, HsExpNum_{nh,TH_{nh}}\}$$

- 5 其中， nh 为所述历史热词的序号， $1 \leq nh \leq NH$ ， NH 为所述历史热词的总数， th 为各个统计周期按照时间先后顺序依次排列的序号， $1 \leq th \leq TH_{nh}$ ， TH_{nh} 为第 nh 个历史热词的统计周期的总数， $HsExpNum_{nh,th}$ 为第 nh 个历史热词在第 th 个统计周期内的曝光频次， $ExpSeq1_{nh}$ 为第 nh 个历史热词的第一曝光序列。

步骤 S203、计算各个第一曝光序列的均值。

- 10 具体地，可以根据下式计算各个第一曝光序列的均值：

$$AvExpSeq1_{nh} = \frac{\sum_{th=1}^{TH_{nh}} HsExpNum_{nh,th}}{TH_{nh}}$$

其中， $AvExpSeq1_{nh}$ 为第 nh 个第一曝光序列的均值。

步骤 S204、构造各个第一曝光序列的均值按照从大到小的顺序依次排列的序列。

- 15 具体地，可以根据下式构造各个第一曝光序列的均值按照从大到小的顺序依次排列的序列：

$$\{AvExpSeq1'_1, AvExpSeq1'_2, \dots, AvExpSeq1'_{nh1}, \dots, AvExpSeq1'_{NH}\}$$

其中， $AvExpSeq1'_{nh1}$ 为按照从大到小的顺序排列在第 $nh1$ 位上的第一曝光序列的均值， $1 \leq nh1 \leq NH$ 。

步骤 S205、计算所述第一曝光阈值和所述第二曝光阈值。

- 20 具体地，可以根据下式计算所述第一曝光阈值：

$$Threshold1 = \frac{\sum_{nh1=1}^{NMAX} AvExpSeq1'_{nh1}}{NMAX}$$

其中， $NMAX = floor(\xi_{\max} \times NH)$ ， $\xi_{\max}$ 为预设的系数，且 $0 < \xi_{\max} < 1$ ， $floor$ 为向下取值函数， $Threshold1$ 为所述第一曝光阈值；

根据下式计算所述第二曝光阈值：

$$Threshold2 = \frac{\sum_{nh1=NH-NMIN+1}^{NH} AvExpSeq1'_{nh1}}{NMIN}$$

- 25 其中， $NMIN = floor(\xi_{\min} \times NH)$ ， $\xi_{\min}$ 为预设的系数，且 $0 < \xi_{\min} < 1$ ， $Threshold2$ 为所述第二曝光阈值。

在本申请实施例的另一种具体实现中，所述第一曝光阈值和所述第二曝光阈值的设置过程可以包括如图 3 所示的步骤：

- 30 步骤 S301、从历史统计记录中获取各个历史热词在各个统计周期内的曝光频次。

步骤 S302、构造各个历史热词的第一曝光序列。

步骤 S303、计算各个第一曝光序列的均值。

其中，步骤 S301-步骤 S303 的过程与步骤 S201-步骤 S203 的过程相同，具体可参照上述说明，在此不再赘述。

5 步骤 S304、构造各个历史热词的第二曝光序列。

具体地，可以根据下式构造各个历史热词的第二曝光序列：

$$ExpSeq2_{nh} = \{HsExpNum_{nh,1}', HsExpNum_{nh,2}', \dots, HsExpNum_{nh,th1}', \dots, HsExpNum_{nh,TH_{nh}}'\}$$

其中， $HsExpNum_{nh,th1}' \in ExpSeq1_{nh}$ ， $1 \leq th1 \leq TH_{nh}$ ， $HsExpNum_{nh,th1}' \geq HsExpNum_{nh,th1+1}'$ ， $ExpSeq2_{nh}$ 为第 nh 个历史热词的第二曝光序列。

10 步骤 S305、计算各个第二曝光序列的均值。

具体地，可以根据下式计算各个第二曝光序列的均值：

$$AvExpSeq2_{nh} = \frac{\sum_{th1=1}^{TH1_{nh}} HsExpNum_{nh,th1}'}{TH1_{nh}}$$

其中， $AvExpSeq2_{nh}$ 为第 nh 个第二曝光序列的均值， $TH1_{nh}$ 满足以下条件： $HsExpNum_{nh,TH1}' \geq AvExpSeq1_{nh}$ 且 $HsExpNum_{nh,TH1+1}' < AvExpSeq1_{nh}$ 。

15 步骤 S306、构造各个第二曝光序列的均值按照从大到小的顺序依次排列的序列。

具体地，可以根据下式构造各个第二曝光序列的均值按照从大到小的顺序依次排列的序列：

$$\{AvExpSeq2_1', AvExpSeq2_2', \dots, AvExpSeq2_{nh1}', \dots, AvExpSeq2_{NH}'\}$$

20 其中， $AvExpSeq2_{nh1}'$ 为按照从大到小的顺序排列在第 nh1 位上的第二曝光序列的均值。

步骤 S307、计算所述第一曝光阈值和所述第二曝光阈值。

具体地，可以根据下式计算所述第一曝光阈值：

$$Threshold1 = \frac{\sum_{nh1=1}^{NMAX} AvExpSeq2_{nh1}'}{NMAX};$$

根据下式计算所述第二曝光阈值：

$$Threshold2 = \frac{\sum_{nh1=NH-NMIN+1}^{NH} AvExpSeq2_{nh1}'}{NMIN}.$$

综上所述，本申请实施例为热词的确定提供了一套客观的评估标准，摆脱了对网络分析人员个人能力的依赖，确定出的热词能够更加难以客观地反应真实的情况，而且在得到热词之后，综合考虑了企业与热词之间的关系，分析结果对于企业具有更强的指导意义。

对应于上文实施例所述的一种热词分析方法，图 4 示出了本申请实施例提供的一种热词分析装置的一个实施例结构图。

本实施例中，一种热词分析装置可以包括：

网页爬取模块 401，用于通过搜索引擎爬取当前统计周期内在目标网站上发布的网页，所述目标网站为浏览量大于预设的浏览量阈值的网站；

切词处理模块 402，用于对所述网页中的文本信息进行切词处理，得到构成所述文本信息的各个分词；

第一统计模块 403，用于统计各个分词在所述文本信息中的曝光频次；

第一热词确定模块 404，用于将在所述文本信息中的曝光频次大于预设的第一曝光阈值的分词确定为热词；

第二统计模块 405，用于统计各个企业名称在优选文本信息中的曝光频次，所述优选文本信息为包含所述热词的文本信息；

关联度计算模块 406，用于根据各个企业名称在所述优选文本信息中的曝光频次计算各个企业名称与所述热词之间的关联度。

进一步地，所述热词分析装置还可以包括：

候选分词确定模块，用于将在所述文本信息中的曝光频次小于或等于所述第一曝光阈值且大于预设的第二曝光阈值的分词确定为候选分词；

第三统计模块，用于从历史统计记录中获取各个候选分词在所述当前统计周期之前的 T 个统计周期内的曝光频次，其中，T 为正整数；

第二热词确定模块，用于将满足下述条件的候选分词确定为热词：

对于任意的 t 的取值，不等式 $\ln\left(\frac{ExpNum_{n,t+1}}{ExpNum_{n,t}}\right) > ThreshRatio$ 均成立，其中，n 为所述候选分词的序号， $1 \leq n \leq N$ ，N 为所述候选分词的总数，t 为各个统计周期按照时间先后顺序依次排列的序号， $1 \leq t \leq T$ ， $ExpNum_{n,t}$ 为第 n 个候选分词在第 t 个统计周期内的曝光频次， $ExpNum_{n,t+1}$ 为第 n 个候选分词在所述当前统计周期内的曝光频次，ln 为自然对数函数，ThreshRatio 为预设的比例阈值；

进一步地，所述热词分析装置还可以包括：

第四统计模块，用于从历史统计记录中获取各个历史热词在各个统计周期内的曝光频次，所述历史热词为在所述当前统计周期之前已经确定的热词；

第一曝光序列构造模块，用于根据下式构造各个历史热词的第一曝光序列：

$ExpSeq1_{nh} = \{HsExpNum_{nh,1}, HsExpNum_{nh,2}, \dots, HsExpNum_{nh,th}, \dots, HsExpNum_{nh,TH_{nh}}\}$

其中，nh 为所述历史热词的序号， $1 \leq nh \leq NH$ ，NH 为所述历史热词的总数，th 为各个统计周期按照时间先后顺序依次排列的序号， $1 \leq th \leq TH_{nh}$ ， TH_{nh} 为第 nh 个历史热词的统计周期的总数， $HsExpNum_{nh,th}$ 为第 nh 个历史热词在第 th 个统计周期内的曝光频次， $ExpSeq1_{nh}$ 为第 nh 个历史热词的第一曝光序列；

第一曝光序列均值计算模块，用于根据下式计算各个第一曝光序列的均值：

$$AvExpSeq1_{nh} = \frac{\sum_{th=1}^{TH_{nh}} HsExpNum_{nh,th}}{TH_{nh}}$$

其中， $AvExpSeq1_{nh}$ 为第 nh 个第一曝光序列的均值；

第一均值序列构造模块，用于根据下式构造各个第一曝光序列的均值按照从大到小的顺序依次排列的序列：

$$\{AvExpSeq1'_1, AvExpSeq1'_2, \dots, AvExpSeq1'_{nh1}, \dots, AvExpSeq1'_{NH}\}$$

其中， $AvExpSeq1'_{nh1}$ 为按照从大到小的顺序排列在第 $nh1$ 位上的第一曝光序列的均值， $1 \leq nh1 \leq NH$ ；

第一曝光阈值计算模块，用于根据下式计算所述第一曝光阈值：

$$Threshold1 = \frac{\sum_{nh1=1}^{NMAX} AvExpSeq1'_{nh1}}{NMAX}$$

其中， $NMAX = floor(\xi_{max} \times NH)$ ， ξ_{max} 为预设的系数，且 $0 < \xi_{max} < 1$ ， $floor$ 为向下取值函数， $Threshold1$ 为所述第一曝光阈值；

第二曝光阈值计算模块，用于根据下式计算所述第二曝光阈值：

$$Threshold2 = \frac{\sum_{nh1=NH-NMIN+1}^{NH} AvExpSeq1'_{nh1}}{NMIN}$$

其中， $NMIN = floor(\xi_{min} \times NH)$ ， ξ_{min} 为预设的系数，且 $0 < \xi_{min} < 1$ ， $Threshold2$ 为所述第二曝光阈值。

进一步地，所述热词分析装置还可以包括：

第二曝光序列构造模块，用于根据下式构造各个历史热词的第二曝光序列：

$$ExpSeq2_{nh} = \{HsExpNum'_{nh,1}, HsExpNum'_{nh,2}, \dots, HsExpNum'_{nh,th1}, \dots, HsExpNum'_{nh,TH_{nh}}\}$$

其中， $HsExpNum'_{nh,th1} \in ExpSeq1_{nh}$ ， $1 \leq th1 \leq TH_{nh}$ ， $HsExpNum'_{nh,th1} \geq HsExpNum'_{nh,th1+1}$ ， $ExpSeq2_{nh}$ 为第 nh 个历史热词的第二曝光序列；

第二曝光序列均值计算模块，用于根据下式计算各个第二曝光序列的均值：

$$AvExpSeq2_{nh} = \frac{\sum_{th=1}^{TH1_{nh}} HsExpNum'_{nh,th1}}{TH1_{nh}}$$

其中， $AvExpSeq2_{nh}$ 为第 nh 个第二曝光序列的均值， $TH1_{nh}$ 满足以下条件：

$$HsExpNum'_{nh,TH1} \geq AvExpSeq1_{nh} \text{ 且 } HsExpNum'_{nh,TH1+1} < AvExpSeq1_{nh} ;$$

第二均值序列构造模块，用于根据下式构造各个第二曝光序列的均值按照从大到小的顺序依次排列的序列：

$$\{AvExpSeq2'_1, AvExpSeq2'_2, \dots, AvExpSeq2'_{nh1}, \dots, AvExpSeq2'_{NH}\}$$

其中, $AvExpSeq2_{nh1}'$ 为按照从大到小的顺序排列在第 $nh1$ 位上的第二曝光序列的均值;

第一曝光阈值计算模块, 用于根据下式计算所述第一曝光阈值:

$$Threshold1 = \frac{\sum_{nh1=1}^{NMAX} AvExpSeq2_{nh1}'}{NMAX};$$

5 第二曝光阈值计算模块, 用于根据下式计算所述第二曝光阈值:

$$Threshold2 = \frac{\sum_{nh1=NH-NMIN+1}^{NH} AvExpSeq2_{nh1}'}{NMIN}。$$

进一步地, 所述当前统计周期包括 M 个子周期, 其中, M 为正整数, 所述第二统计模块可以包括:

10 子时间段统计单元, 用于统计各个企业名称在各个子时间段的优选文本信息中的曝光频次;

所述关联度计算模块可以包括:

第一计算单元, 用于根据下式计算各个企业名称与所述热词之间的关联度:

$$Rel_{q,p} = \frac{\sum_{m=1}^M (k_m \times EntExpNum_{q,p,m})}{\sum_{m=1}^M [k_m \times (\sum_{q=1}^Q EntExpNum_{q,p,m})]}$$

15 其中, q 为企业名称的序号, $1 \leq q \leq Q$, Q 为企业名称的总数, p 为热词的序号, $1 \leq p \leq P$, P 为热词的总数, m 为各个子时间段按照时间先后顺序依次排列的序号, $1 \leq m \leq M$, $EntExpNum_{q,p,m}$ 为第 m 个子时间段内第 q 个企业名称在包含第 p 个热词的优选文本信息中的曝光频次, k_m 为预设的权值系数, $k_m < k_{m+1}$ 且 $\sum_{m=1}^M k_m = 1$, $Rel_{q,p}$ 为第 q 个企业名称与第 p 个热词之间的关联度。

20 图 5 示出了本申请实施例提供的一种热词分析终端设备的示意框图, 为了便于说明, 仅示出了与本申请实施例相关的部分。

在本实施例中, 所述热词分析终端设备 5 可以是手机、平板电脑、桌上型计算机、笔记本及云端服务器等计算设备。该热词分析终端设备 5 可包括: 处理器 50、存储器 51 以及存储在所述存储器 51 中并可在所述处理器 50 上运行的计算机可读指令 52, 例如执行上述的热词分析方法的计算机可读指令。所述处理器 50 执行所述计算机可读指令 52 时实现上述各个热词分析方法实施例中的步骤。

25

在本申请各个实施例中的各功能单元如果以软件功能单元的形式实现并作为独立的产品销售或使用, 可以存储在一个计算机可读存储介质中。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或者该技术方案的全部或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中，包括若干计算机可读指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本申请各个实施例所述方法的全部或部分步骤。

权利要求书

1. 一种热词分析方法，其特征在于，包括：

通过搜索引擎爬取当前统计周期内在目标网站上发布的网页，所述目标网站为浏览量大于预设的浏览量阈值的网站；

5 对所述网页中的文本信息进行切词处理，得到构成所述文本信息的各个分词；

统计各个分词在所述文本信息中的曝光频次；

将在所述文本信息中的曝光频次大于预设的第一曝光阈值的分词确定为热词；

统计各个企业名称在优选文本信息中的曝光频次，所述优选文本信息为包含所述热词的文本信息；

10 根据各个企业名称在所述优选文本信息中的曝光频次计算各个企业名称与所述热词之间的关联度。

2. 根据权利要求 1 所述的热词分析方法，其特征在于，在所述统计各个分词在所述文本信息中的曝光频次之后，还包括：

将在所述文本信息中的曝光频次小于或等于所述第一曝光阈值且大于预设的第二

15 曝光阈值的分词确定为候选分词；

从历史统计记录中获取各个候选分词在所述当前统计周期之前的 T 个统计周期内的曝光频次，其中，T 为正整数；

将满足下述条件的候选分词确定为热词：

$$\ln\left(\frac{ExpNum_{n,t+1}}{ExpNum_{n,t}}\right) > ThreshRatio$$

对于任意的 t 的取值，不等式 均成立，其中，n 为所述

20 候选分词的序号， $1 \leq n \leq N$ ，N 为所述候选分词的总数，t 为各个统计周期按照时间先后顺序依次排列的序号， $1 \leq t \leq T$ ， $ExpNum_{n,t}$ 为第 n 个候选分词在第 t 个统计周期内的曝光频次， $ExpNum_{n,t+1}$ 为第 n 个候选分词在所述当前统计周期内的曝光频次，ln 为自然对数函数，ThreshRatio 为预设的比例阈值；

3. 根据权利要求 2 所述的热词分析方法，其特征在于，所述第一曝光阈值和所述
25 第二曝光阈值的设置过程包括：

从历史统计记录中获取各个历史热词在各个统计周期内的曝光频次，所述历史热词为在所述当前统计周期之前已经确定的热词；

根据下式构造各个历史热词的第一曝光序列：

$$ExpSeq1_{nh} = \{HsExpNum_{nh,1}, HsExpNum_{nh,2}, \dots, HsExpNum_{nh,th}, \dots, HsExpNum_{nh,TH_{nh}}\}$$

30 其中，nh 为所述历史热词的序号， $1 \leq nh \leq NH$ ，NH 为所述历史热词的总数，th 为各个统计周期按照时间先后顺序依次排列的序号， $1 \leq th \leq TH_{nh}$ ， TH_{nh} 为第 nh 个历史热词的统计周期的总数， $HsExpNum_{nh,th}$ 为第 nh 个历史热词在第 th 个统计周期内的曝光频次， $ExpSeq1_{nh}$ 为第 nh 个历史热词的第一曝光序列；

根据下式计算各个第一曝光序列的均值：

$$AvExpSeq1_{nh} = \frac{\sum_{th=1}^{TH_{nh}} HsExpNum_{nh,th}}{TH_{nh}}$$

其中, $AvExpSeq1_{nh}$ 为第 nh 个第一曝光序列的均值;

根据下式构造各个第一曝光序列的均值按照从大到小的顺序依次排列的序列:

$$\{AvExpSeq1'_1, AvExpSeq1'_2, \dots, AvExpSeq1'_{nh1}, \dots, AvExpSeq1'_{NH}\}$$

- 5 其中, $AvExpSeq1'_{nh1}$ 为按照从大到小的顺序排列在第 $nh1$ 位上的第一曝光序列的均值, $1 \leq nh1 \leq NH$;

根据下式计算所述第一曝光阈值:

$$Threshold1 = \frac{\sum_{nh1=1}^{NMAX} AvExpSeq1'_{nh1}}{NMAX}$$

- 10 其中, $NMAX = floor(\xi_{\max} \times NH)$, $\xi_{\max}$ 为预设的系数, 且 $0 < \xi_{\max} < 1$, $floor$ 为向下取值函数, $Threshold1$ 为所述第一曝光阈值;

根据下式计算所述第二曝光阈值:

$$Threshold2 = \frac{\sum_{nh1=NH-NMIN+1}^{NH} AvExpSeq1'_{nh1}}{NMIN}$$

其中, $NMIN = floor(\xi_{\min} \times NH)$, $\xi_{\min}$ 为预设的系数, 且 $0 < \xi_{\min} < 1$, $Threshold2$ 为所述第二曝光阈值。

- 15 4. 根据权利要求 3 所述的热词分析方法, 其特征在于, 在计算各个第一曝光序列的均值之后, 还包括:

根据下式构造各个历史热词的第二曝光序列:

$$ExpSeq2_{nh} = \{HsExpNum'_{nh,1}, HsExpNum'_{nh,2}, \dots, HsExpNum'_{nh,th1}, \dots, HsExpNum'_{nh,TH_{nh}}\}$$

其中, $HsExpNum'_{nh,th1} \in ExpSeq1_{nh}$, $1 \leq th1 \leq TH_{nh}$, $HsExpNum'_{nh,th1} \geq HsExpNum'_{nh,th1+1}$,

- 20 $ExpSeq2_{nh}$ 为第 nh 个历史热词的第二曝光序列;

根据下式计算各个第二曝光序列的均值:

$$AvExpSeq2_{nh} = \frac{\sum_{th1=1}^{TH1_{nh}} HsExpNum'_{nh,th1}}{TH1_{nh}}$$

其中, $AvExpSeq2_{nh}$ 为第 nh 个第二曝光序列的均值, $TH1_{nh}$ 满足以下条件:
 $HsExpNum'_{nh,TH1} \geq AvExpSeq1_{nh}$ 且 $HsExpNum'_{nh,TH1+1} < AvExpSeq1_{nh}$;

- 25 根据下式构造各个第二曝光序列的均值按照从大到小的顺序依次排列的序列:

$$\{AvExpSeq2'_1, AvExpSeq2'_2, \dots, AvExpSeq2'_{nh1}, \dots, AvExpSeq2'_{NH}\}$$

其中, $AvExpSeq2'_{nh1}$ 为按照从大到小的顺序排列在第 $nh1$ 位上的第二曝光序列的均值;

根据下式计算所述第一曝光阈值:

$$Threshold1 = \frac{\sum_{nh1=1}^{NMAX} AvExpSeq2_{nh1}'}{NMAX};$$

根据下式计算所述第二曝光阈值:

$$Threshold2 = \frac{\sum_{nh1=NH-NMIN+1}^{NH} AvExpSeq2_{nh1}'}{NMIN}。$$

- 5 5. 根据权利要求 1 至 4 中任一项所述的热词分析方法, 其特征在于, 所述当前统计周期包括 M 个子周期, 其中, M 为正整数, 所述统计各个企业名称在优选文本信息中的曝光频次包括:

统计各个企业名称在各个子时间段的优选文本信息中的曝光频次;

- 10 所述根据各个企业名称在所述优选文本信息中的曝光频次计算各个企业名称与所述热词之间的关联度包括:

根据下式计算各个企业名称与所述热词之间的关联度:

$$Rel_{q,p} = \frac{\sum_{m=1}^M (k_m \times EntExpNum_{q,p,m})}{\sum_{m=1}^M [k_m \times (\sum_{q=1}^Q EntExpNum_{q,p,m})]}$$

- 15 其中, q 为企业名称的序号, $1 \leq q \leq Q$, Q 为企业名称的总数, p 为热词的序号, $1 \leq p \leq P$, P 为热词的总数, m 为各个子时间段按照时间先后顺序依次排列的序号, $1 \leq m \leq M$, $EntExpNum_{q,p,m}$ 为第 m 个子时间段内第 q 个企业名称在包含第 p 个热词的优选文本信息中的曝光频次, k_m 为预设的权值系数, $k_m < k_{m+1}$ 且 $\sum_{m=1}^M k_m = 1$, $Rel_{q,p}$ 为第 q 个企业名称与第 p 个热词之间的关联度。

6. 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计算机可读指令, 其特征在于, 所述计算机可读指令被处理器执行时实现如下步骤:

- 20 通过搜索引擎爬取当前统计周期内在目标网站上发布的网页, 所述目标网站为浏览量大于预设的浏览量阈值的网站;

对所述网页中的文本信息进行切词处理, 得到构成所述文本信息的各个分词;

统计各个分词在所述文本信息中的曝光频次;

将在所述文本信息中的曝光频次大于预设的第一曝光阈值的分词确定为热词;

- 25 统计各个企业名称在优选文本信息中的曝光频次, 所述优选文本信息为包含所述热词的文本信息;

根据各个企业名称在所述优选文本信息中的曝光频次计算各个企业名称与所述热词之间的关联度。

7. 根据权利要求 6 所述的计算机可读存储介质, 其特征在于, 在所述统计各个分

词在所述文本信息中的曝光频次之后，还包括：

将在所述文本信息中的曝光频次小于或等于所述第一曝光阈值且大于预设的第二曝光阈值的分词确定为候选分词；

从历史统计记录中获取各个候选分词在所述当前统计周期之前的 T 个统计周期内的曝光频次，其中， T 为正整数；

将满足下述条件的候选分词确定为热词：

$$\ln\left(\frac{ExpNum_{n,t+1}}{ExpNum_{n,t}}\right) > ThreshRatio$$

对于任意的 t 的取值，不等式均成立，其中， n 为所述候选分词的序号， $1 \leq n \leq N$ ， N 为所述候选分词的总数， t 为各个统计周期按照时间先后顺序依次排列的序号， $1 \leq t \leq T$ ， $ExpNum_{n,t}$ 为第 n 个候选分词在第 t 个统计周期内的曝光频次， $ExpNum_{n,T+1}$ 为第 n 个候选分词在所述当前统计周期内的曝光频次， $\ln$ 为自然对数函数， $ThreshRatio$ 为预设的比例阈值；

8. 根据权利要求 7 所述的计算机可读存储介质，其特征在于，所述第一曝光阈值和所述第二曝光阈值的设置过程包括：

从历史统计记录中获取各个历史热词在各个统计周期内的曝光频次，所述历史热词为在所述当前统计周期之前已经确定的热词；

根据下式构造各个历史热词的第一曝光序列：

$$ExpSeq1_{nh} = \{HsExpNum_{nh,1}, HsExpNum_{nh,2}, \dots, HsExpNum_{nh,th}, \dots, HsExpNum_{nh,TH_{nh}}\}$$

其中， nh 为所述历史热词的序号， $1 \leq nh \leq NH$ ， NH 为所述历史热词的总数， th 为各个统计周期按照时间先后顺序依次排列的序号， $1 \leq th \leq TH_{nh}$ ， TH_{nh} 为第 nh 个历史热词的统计周期的总数， $HsExpNum_{nh,th}$ 为第 nh 个历史热词在第 th 个统计周期内的曝光频次， $ExpSeq1_{nh}$ 为第 nh 个历史热词的第一曝光序列；

根据下式计算各个第一曝光序列的均值：

$$AvExpSeq1_{nh} = \frac{\sum_{th=1}^{TH_{nh}} HsExpNum_{nh,th}}{TH_{nh}}$$

其中， $AvExpSeq1_{nh}$ 为第 nh 个第一曝光序列的均值；

根据下式构造各个第一曝光序列的均值按照从大到小的顺序依次排列的序列：

$$\{AvExpSeq1'_1, AvExpSeq1'_2, \dots, AvExpSeq1'_{nh1}, \dots, AvExpSeq1'_{NH}\}$$

其中， $AvExpSeq1'_{nh1}$ 为按照从大到小的顺序排列在第 $nh1$ 位上的第一曝光序列的均值， $1 \leq nh1 \leq NH$ ；

根据下式计算所述第一曝光阈值：

$$Threshold1 = \frac{\sum_{nh1=1}^{NMAX} AvExpSeq1'_{nh1}}{NMAX}$$

其中， $NMAX = floor(\xi_{max} \times NH)$ ， ξ_{max} 为预设的系数，且 $0 < \xi_{max} < 1$ ， $floor$ 为向下

取值函数, $Threshold1$ 为所述第一曝光阈值;

根据下式计算所述第二曝光阈值:

$$Threshold2 = \frac{\sum_{nh1=NH-NMIN+1}^{NH} AvExpSeq1_{nh1}'}{NMIN}$$

其中, $NMIN = floor(\xi_{min} \times NH)$, ξ_{min} 为预设的系数, 且 $0 < \xi_{min} < 1$, $Threshold2$ 为所述第二曝光阈值。

9. 根据权利要求 8 所述的计算机可读存储介质, 其特征在于, 在计算各个第一曝光序列的均值之后, 还包括:

根据下式构造各个历史热词的第二曝光序列:

$$ExpSeq2_{nh} = \{HsExpNum_{nh,1}', HsExpNum_{nh,2}', \dots, HsExpNum_{nh,th1}', \dots, HsExpNum_{nh,TH_{nh}}'\}$$

其中, $HsExpNum_{nh,th1}' \in ExpSeq1_{nh}$, $1 \leq th1 \leq TH_{nh}$, $HsExpNum_{nh,th1}' \geq HsExpNum_{nh,th1+1}'$, $ExpSeq2_{nh}$ 为第 nh 个历史热词的第二曝光序列;

根据下式计算各个第二曝光序列的均值:

$$AvExpSeq2_{nh} = \frac{\sum_{th1=1}^{TH1_{nh}} HsExpNum_{nh,th1}'}{TH1_{nh}}$$

其中, $AvExpSeq2_{nh}$ 为第 nh 个第二曝光序列的均值, $TH1_{nh}$ 满足以下条件:

$HsExpNum_{nh,TH1}' \geq AvExpSeq1_{nh}$ 且 $HsExpNum_{nh,TH1+1}' < AvExpSeq1_{nh}$;

根据下式构造各个第二曝光序列的均值按照从大到小的顺序依次排列的序列:

$$\{AvExpSeq2_1', AvExpSeq2_2', \dots, AvExpSeq2_{nh1}', \dots, AvExpSeq2_{NH}'\}$$

其中, $AvExpSeq2_{nh1}'$ 为按照从大到小的顺序排列在第 $nh1$ 位上的第二曝光序列的均值;

根据下式计算所述第一曝光阈值:

$$Threshold1 = \frac{\sum_{nh1=1}^{NMAX} AvExpSeq2_{nh1}'}{NMAX};$$

根据下式计算所述第二曝光阈值:

$$Threshold2 = \frac{\sum_{nh1=NH-NMIN+1}^{NH} AvExpSeq2_{nh1}'}{NMIN}。$$

10. 根据权利要求 6 至 9 中任一项所述的计算机可读存储介质, 其特征在于, 所述当前统计周期包括 M 个子周期, 其中, M 为正整数, 所述统计各个企业名称在优选文本信息中的曝光频次包括:

统计各个企业名称在各个子时间段的优选文本信息中的曝光频次;

所述根据各个企业名称在所述优选文本信息中的曝光频次计算各个企业名称与所

述热词之间的关联度包括：

根据下式计算各个企业名称与所述热词之间的关联度：

$$Rel_{q,p} = \frac{\sum_{m=1}^M (k_m \times EntExpNum_{q,p,m})}{\sum_{m=1}^M [k_m \times (\sum_{q=1}^Q EntExpNum_{q,p,m})]}$$

其中， q 为企业名称的序号， $1 \leq q \leq Q$ ， Q 为企业名称的总数， p 为热词的序号， $1 \leq p \leq P$ ， P 为热词的总数， m 为各个子时间段按照时间先后顺序依次排列的序号， $1 \leq m \leq M$ ， $EntExpNum_{q,p,m}$ 为第 m 个子时间段内第 q 个企业名称在包含第 p 个热词的优选文本信息中的曝光频次， k_m 为预设的权值系数， $k_m < k_{m+1}$ 且 $\sum_{m=1}^M k_m = 1$ ， $Rel_{q,p}$ 为第 q 个企业名称与第 p 个热词之间的关联度。

11. 一种热词分析终端设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机可读指令，其特征在于，所述处理器执行所述计算机可读指令时实现如下步骤：

通过搜索引擎爬取当前统计周期内在目标网站上发布的网页，所述目标网站为浏览量大于预设的浏览量阈值的网站；

对所述网页中的文本信息进行切词处理，得到构成所述文本信息的各个分词；

统计各个分词在所述文本信息中的曝光频次；

将在所述文本信息中的曝光频次大于预设的第一曝光阈值的分词确定为热词；

统计各个企业名称在优选文本信息中的曝光频次，所述优选文本信息为包含所述热词的文本信息；

根据各个企业名称在所述优选文本信息中的曝光频次计算各个企业名称与所述热词之间的关联度。

12. 根据权利要求 11 所述的热词分析终端设备，其特征在于，在所述统计各个分词在所述文本信息中的曝光频次之后，还包括：

将在所述文本信息中的曝光频次小于或等于所述第一曝光阈值且大于预设的第二曝光阈值的分词确定为候选分词；

从历史统计记录中获取各个候选分词在所述当前统计周期之前的 T 个统计周期内的曝光频次，其中， T 为正整数；

将满足下述条件的候选分词确定为热词：

$$\ln\left(\frac{ExpNum_{n,t+1}}{ExpNum_{n,t}}\right) > ThreshRatio$$

对于任意的 t 的取值，不等式均成立，其中， n 为所述候选分词的序号， $1 \leq n \leq N$ ， N 为所述候选分词的总数， t 为各个统计周期按照时间先后顺序依次排列的序号， $1 \leq t \leq T$ ， $ExpNum_{n,t}$ 为第 n 个候选分词在第 t 个统计周期内的曝光频次， $ExpNum_{n,t+1}$ 为第 n 个候选分词在所述当前统计周期内的曝光频次， $\ln$ 为自

然对数函数，*ThreshRatio* 为预设的比例阈值；

13. 根据权利要求 12 所述的热词分析终端设备，其特征在于，所述第一曝光阈值和所述第二曝光阈值的设置过程包括：

从历史统计记录中获取各个历史热词在各个统计周期内的曝光频次，所述历史热词为在所述当前统计周期之前已经确定的热词；

根据下式构造各个历史热词的第一曝光序列：

$$ExpSeq1_{nh} = \{HsExpNum_{nh,1}, HsExpNum_{nh,2}, \dots, HsExpNum_{nh,th}, \dots, HsExpNum_{nh,TH_{nh}}\}$$

其中，*nh* 为所述历史热词的序号， $1 \leq nh \leq NH$ ，*NH* 为所述历史热词的总数，*th* 为各个统计周期按照时间先后顺序依次排列的序号， $1 \leq th \leq TH_{nh}$ ，*TH_{nh}* 为第 *nh* 个历史热词的统计周期的总数，*HsExpNum_{nh,th}* 为第 *nh* 个历史热词在第 *th* 个统计周期内的曝光频次，*ExpSeq1_{nh}* 为第 *nh* 个历史热词的第一曝光序列；

根据下式计算各个第一曝光序列的均值：

$$AvExpSeq1_{nh} = \frac{\sum_{th=1}^{TH_{nh}} HsExpNum_{nh,th}}{TH_{nh}}$$

其中，*AvExpSeq1_{nh}* 为第 *nh* 个第一曝光序列的均值；

根据下式构造各个第一曝光序列的均值按照从大到小的顺序依次排列的序列：

$$\{AvExpSeq1'_1, AvExpSeq1'_2, \dots, AvExpSeq1'_{nh1}, \dots, AvExpSeq1'_{NH}\}$$

其中，*AvExpSeq1'_{nh1}* 为按照从大到小的顺序排列在第 *nh1* 位上的第一曝光序列的均值， $1 \leq nh1 \leq NH$ ；

根据下式计算所述第一曝光阈值：

$$Threshold1 = \frac{\sum_{nh1=1}^{NMAX} AvExpSeq1'_{nh1}}{NMAX}$$

其中， $NMAX = \text{floor}(\xi_{\max} \times NH)$ ， $\xi_{\max}$ 为预设的系数，且 $0 < \xi_{\max} < 1$ ，*floor* 为向下取值函数，*Threshold1* 为所述第一曝光阈值；

根据下式计算所述第二曝光阈值：

$$Threshold2 = \frac{\sum_{nh1=NH-NMIN+1}^{NH} AvExpSeq1'_{nh1}}{NMIN}$$

其中， $NMIN = \text{floor}(\xi_{\min} \times NH)$ ， $\xi_{\min}$ 为预设的系数，且 $0 < \xi_{\min} < 1$ ，*Threshold2* 为所述第二曝光阈值。

14. 根据权利要求 13 所述的热词分析终端设备，其特征在于，在计算各个第一曝光序列的均值之后，还包括：

根据下式构造各个历史热词的第二曝光序列：

$$ExpSeq2_{nh} = \{HsExpNum'_{nh,1}, HsExpNum'_{nh,2}, \dots, HsExpNum'_{nh,th1}, \dots, HsExpNum'_{nh,TH_{nh}}\}$$

其中, $HsExpNum_{nh,th1}' \in ExpSeq1_{nh}$, $1 \leq th1 \leq TH_{nh}$, $HsExpNum_{nh,th1}' \geq HsExpNum_{nh,th1+1}'$, $ExpSeq2_{nh}$ 为第 nh 个历史热词的第二曝光序列;

根据下式计算各个第二曝光序列的均值:

$$AvExpSeq2_{nh} = \frac{\sum_{th=1}^{TH1_{nh}} HsExpNum_{nh,th1}'}{TH1_{nh}}$$

- 5 其中, $AvExpSeq2_{nh}$ 为第 nh 个第二曝光序列的均值, $TH1_{nh}$ 满足以下条件:
 $HsExpNum_{nh,TH1}' \geq AvExpSeq1_{nh}$ 且 $HsExpNum_{nh,TH1+1}' < AvExpSeq1_{nh}$;

根据下式构造各个第二曝光序列的均值按照从大到小的顺序依次排列的序列:

$$\{AvExpSeq2_1', AvExpSeq2_2', \dots, AvExpSeq2_{nh1}', \dots, AvExpSeq2_{NH}'\}$$

- 10 其中, $AvExpSeq2_{nh1}'$ 为按照从大到小的顺序排列在第 $nh1$ 位上的第二曝光序列的均值;

根据下式计算所述第一曝光阈值:

$$Threshold1 = \frac{\sum_{nh1=1}^{NMAX} AvExpSeq2_{nh1}'}{NMAX};$$

根据下式计算所述第二曝光阈值:

$$Threshold2 = \frac{\sum_{nh1=NH-NMIN+1}^{NH} AvExpSeq2_{nh1}'}{NMIN}.$$

- 15 15. 根据权利要求 11 至 14 中任一项所述的热词分析终端设备, 其特征在于, 所述当前统计周期包括 M 个子周期, 其中, M 为正整数, 所述统计各个企业名称在优选文本信息中的曝光频次包括:

统计各个企业名称在各个子时间段的优选文本信息中的曝光频次;

- 20 所述根据各个企业名称在所述优选文本信息中的曝光频次计算各个企业名称与所述热词之间的关联度包括:

根据下式计算各个企业名称与所述热词之间的关联度:

$$Rel_{q,p} = \frac{\sum_{m=1}^M (k_m \times EntExpNum_{q,p,m})}{\sum_{m=1}^M [k_m \times (\sum_{q=1}^Q EntExpNum_{q,p,m})]}$$

- 25 其中, q 为企业名称的序号, $1 \leq q \leq Q$, Q 为企业名称的总数, p 为热词的序号, $1 \leq p \leq P$, P 为热词的总数, m 为各个子时间段按照时间先后顺序依次排列的序号, $1 \leq m \leq M$, $EntExpNum_{q,p,m}$ 为第 m 个子时间段内第 q 个企业名称在包含第 p 个热词的优

选文本信息中的曝光频次, k_m 为预设的权值系数, $k_m < k_{m+1}$ 且 $\sum_{m=1}^M k_m = 1$, $Rel_{q,p}$ 为第 q

个企业名称与第 p 个热词之间的关联度。

16. 一种热词分析装置，其特征在于，包括：

网页爬取模块，用于通过搜索引擎爬取当前统计周期内在目标网站上发布的网页，所述目标网站为浏览量大于预设的浏览量阈值的网站；

5 切词处理模块，用于对所述网页中的文本信息进行切词处理，得到构成所述文本信息的各个分词；

第一统计模块，用于统计各个分词在所述文本信息中的曝光频次；

第一热词确定模块，用于将在所述文本信息中的曝光频次大于预设的第一曝光阈值的分词确定为热词；

10 第二统计模块，用于统计各个企业名称在优选文本信息中的曝光频次，所述优选文本信息为包含所述热词的文本信息；

关联度计算模块，用于根据各个企业名称在所述优选文本信息中的曝光频次计算各个企业名称与所述热词之间的关联度。

17. 根据权利要求 16 所述的热词分析装置，其特征在于，还包括：

15 候选分词确定模块，用于将在所述文本信息中的曝光频次小于或等于所述第一曝光阈值且大于预设的第二曝光阈值的分词确定为候选分词；

第三统计模块，用于从历史统计记录中获取各个候选分词在所述当前统计周期之前的 T 个统计周期内的曝光频次，其中， T 为正整数；

第二热词确定模块，用于将满足下述条件的候选分词确定为热词：

20 对于任意的 t 的取值，不等式 $\ln\left(\frac{ExpNum_{n,t+1}}{ExpNum_{n,t}}\right) > ThreshRatio$ 均成立，其中， n 为所述候选分词的序号， $1 \leq n \leq N$ ， N 为所述候选分词的总数， t 为各个统计周期按照时间先后顺序依次排列的序号， $1 \leq t \leq T$ ， $ExpNum_{n,t}$ 为第 n 个候选分词在第 t 个统计周期内的曝光频次， $ExpNum_{n,T+1}$ 为第 n 个候选分词在所述当前统计周期内的曝光频次， $\ln$ 为自然对数函数， $ThreshRatio$ 为预设的比例阈值；

25 18. 根据权利要求 17 所述的热词分析装置，其特征在于，还包括：

第四统计模块，用于从历史统计记录中获取各个历史热词在各个统计周期内的曝光频次，所述历史热词为在所述当前统计周期之前已经确定的热词；

第一曝光序列构造模块，用于根据下式构造各个历史热词的第一曝光序列：

$$ExpSeq1_{nh} = \{HsExpNum_{nh,1}, HsExpNum_{nh,2}, \dots, HsExpNum_{nh,th}, \dots, HsExpNum_{nh,TH_{nh}}\}$$

30 其中， nh 为所述历史热词的序号， $1 \leq nh \leq NH$ ， NH 为所述历史热词的总数， th 为各个统计周期按照时间先后顺序依次排列的序号， $1 \leq th \leq TH_{nh}$ ， TH_{nh} 为第 nh 个历史热词的统计周期的总数， $HsExpNum_{nh,th}$ 为第 nh 个历史热词在第 th 个统计周期内的曝光频次， $ExpSeq1_{nh}$ 为第 nh 个历史热词的第一曝光序列；

第一曝光序列均值计算模块，用于根据下式计算各个第一曝光序列的均值：

$$AvExpSeq1_{nh} = \frac{\sum_{th=1}^{TH_{nh}} HsExpNum_{nh,th}}{TH_{nh}}$$

其中, $AvExpSeq1_{nh}$ 为第 nh 个第一曝光序列的均值;

第一均值序列构造模块, 用于根据下式构造各个第一曝光序列的均值按照从大到小的顺序依次排列的序列:

$$\{AvExpSeq1'_1, AvExpSeq1'_2, \dots, AvExpSeq1'_{nh1}, \dots, AvExpSeq1'_{NH}\}$$

其中, $AvExpSeq1'_{nh1}$ 为按照从大到小的顺序排列在第 $nh1$ 位上的第一曝光序列的均值, $1 \leq nh1 \leq NH$;

第一曝光阈值计算模块, 用于根据下式计算所述第一曝光阈值:

$$Threshold1 = \frac{\sum_{nh1=1}^{NMAX} AvExpSeq1'_{nh1}}{NMAX}$$

其中, $NMAX = floor(\xi_{\max} \times NH)$, $\xi_{\max}$ 为预设的系数, 且 $0 < \xi_{\max} < 1$, $floor$ 为向下取值函数, $Threshold1$ 为所述第一曝光阈值;

第二曝光阈值计算模块, 用于根据下式计算所述第二曝光阈值:

$$Threshold2 = \frac{\sum_{nh1=NH-NMIN+1}^{NH} AvExpSeq1'_{nh1}}{NMIN}$$

其中, $NMIN = floor(\xi_{\min} \times NH)$, $\xi_{\min}$ 为预设的系数, 且 $0 < \xi_{\min} < 1$, $Threshold2$ 为所述第二曝光阈值。

19. 根据权利要求 18 所述的热词分析装置, 其特征在于, 还包括:

第二曝光序列构造模块, 用于根据下式构造各个历史热词的第二曝光序列:

$$ExpSeq2_{nh} = \{HsExpNum'_{nh,1}, HsExpNum'_{nh,2}, \dots, HsExpNum'_{nh,th1}, \dots, HsExpNum'_{nh,TH_{nh}}\}$$

其中, $HsExpNum'_{nh,th1} \in ExpSeq1_{nh}$, $1 \leq th1 \leq TH_{nh}$, $HsExpNum'_{nh,th1} \geq HsExpNum'_{nh,th1+1}$,

$ExpSeq2_{nh}$ 为第 nh 个历史热词的第二曝光序列;

第二曝光序列均值计算模块, 用于根据下式计算各个第二曝光序列的均值:

$$AvExpSeq2_{nh} = \frac{\sum_{th=1}^{TH1_{nh}} HsExpNum'_{nh,th1}}{TH1_{nh}}$$

其中, $AvExpSeq2_{nh}$ 为第 nh 个第二曝光序列的均值, $TH1_{nh}$ 满足以下条件:

$HsExpNum'_{nh,TH1} \geq AvExpSeq1_{nh}$ 且 $HsExpNum'_{nh,TH1+1} < AvExpSeq1_{nh}$;

第二均值序列构造模块, 用于根据下式构造各个第二曝光序列的均值按照从大到小的顺序依次排列的序列:

$$\{AvExpSeq2'_1, AvExpSeq2'_2, \dots, AvExpSeq2'_{nh1}, \dots, AvExpSeq2'_{NH}\}$$

其中, $AvExpSeq2'_{nh1}$ 为按照从大到小的顺序排列在第 $nh1$ 位上的第二曝光序列的均

值；

第一曝光阈值计算模块，用于根据下式计算所述第一曝光阈值：

$$Threshold1 = \frac{\sum_{nh1=1}^{NMAX} AvExpSeq2_{nh1}'}{NMAX};$$

第二曝光阈值计算模块，用于根据下式计算所述第二曝光阈值：

$$Threshold2 = \frac{\sum_{nh1=NH-NMIN+1}^{NH} AvExpSeq2_{nh1}'}{NMIN}。$$

5

20. 根据权利要求 16 至 19 中任一项所述的热词分析装置，其特征在于，所述当前统计周期包括 M 个子周期，其中，M 为正整数，所述第二统计模块包括：

子时间段统计单元，用于统计各个企业名称在各个子时间段的优选文本信息中的曝光频次；

10 所述关联度计算模块可以包括：

第一计算单元，用于根据下式计算各个企业名称与所述热词之间的关联度：

$$Rel_{q,p} = \frac{\sum_{m=1}^M (k_m \times EntExpNum_{q,p,m})}{\sum_{m=1}^M [k_m \times (\sum_{q=1}^Q EntExpNum_{q,p,m})]}$$

其中，q 为企业名称的序号， $1 \leq q \leq Q$ ，Q 为企业名称的总数，p 为热词的序号， $1 \leq p \leq P$ ，P 为热词的总数，m 为各个子时间段按照时间先后顺序依次排列的序号，

15 $1 \leq m \leq M$ ， $EntExpNum_{q,p,m}$ 为第 m 个子时间段内第 q 个企业名称在包含第 p 个热词的优

选文本信息中的曝光频次， k_m 为预设的权值系数， $k_m < k_{m+1}$ 且 $\sum_{m=1}^M k_m = 1$ ， $Rel_{q,p}$ 为第 q 个企业名称与第 p 个热词之间的关联度。

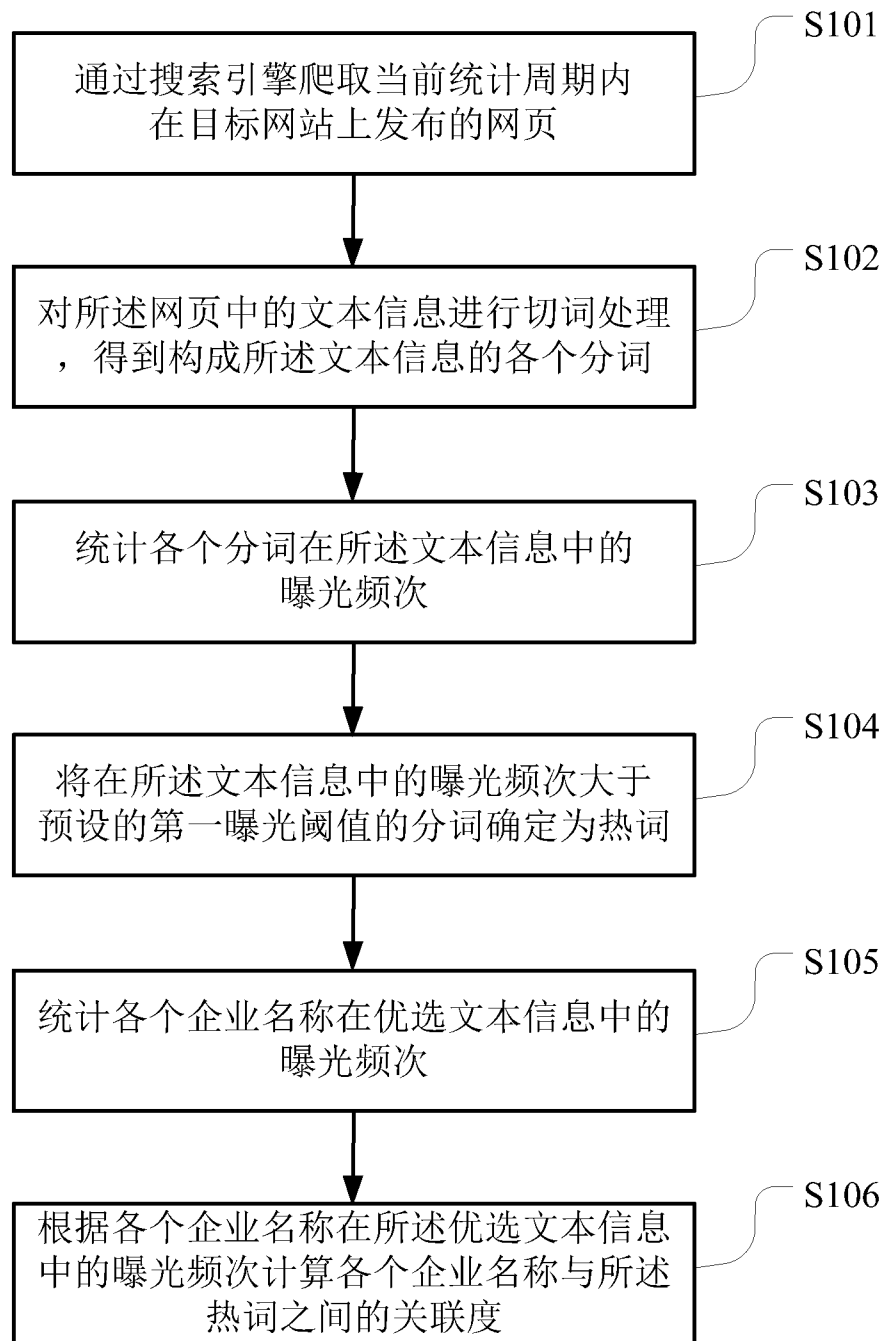


图 1

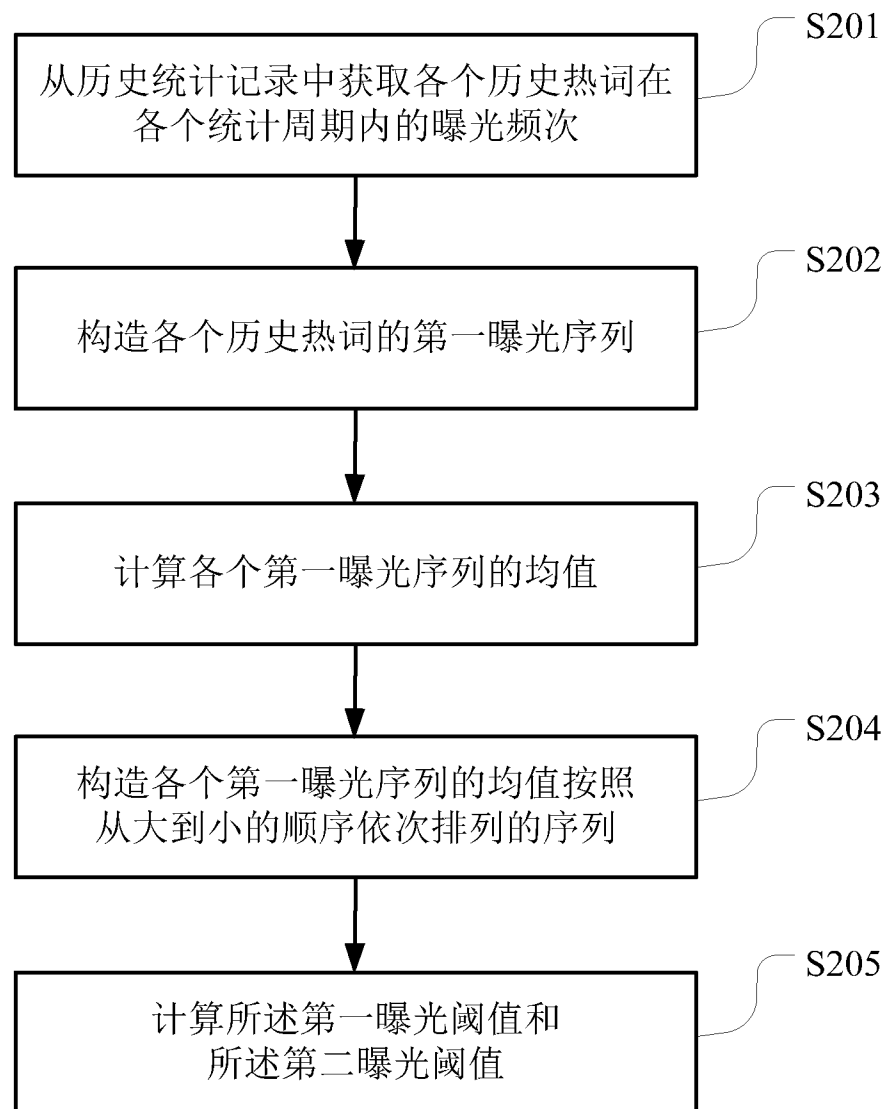


图 2

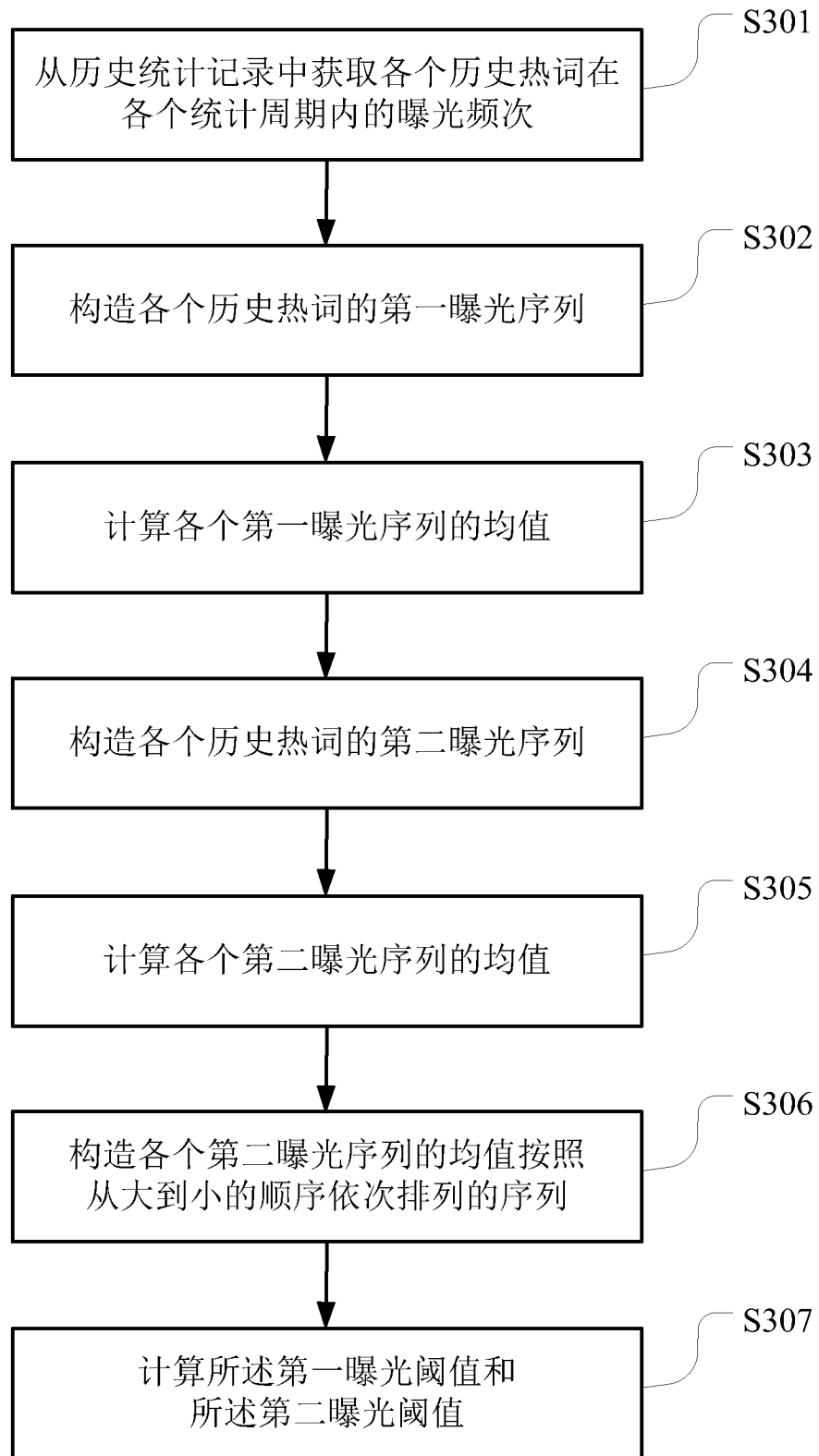


图 3

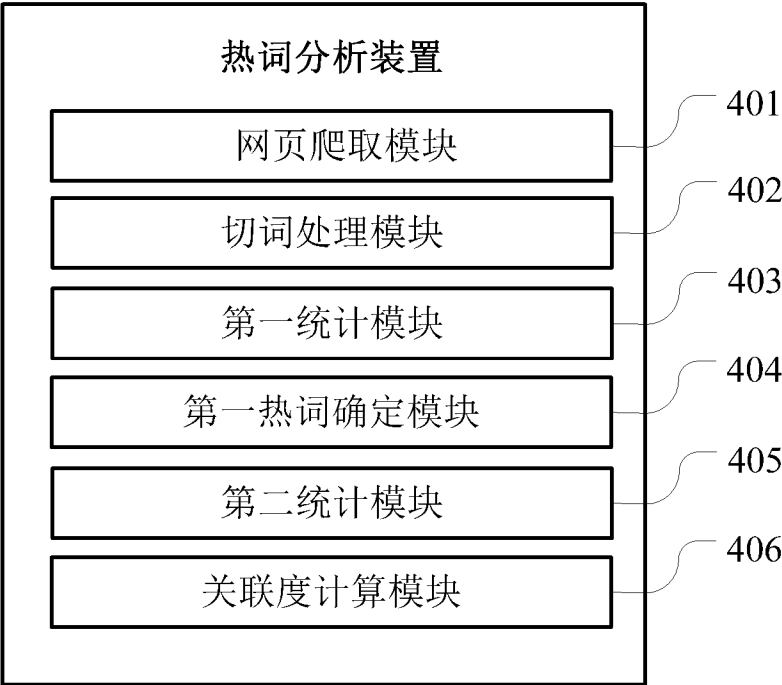


图 4

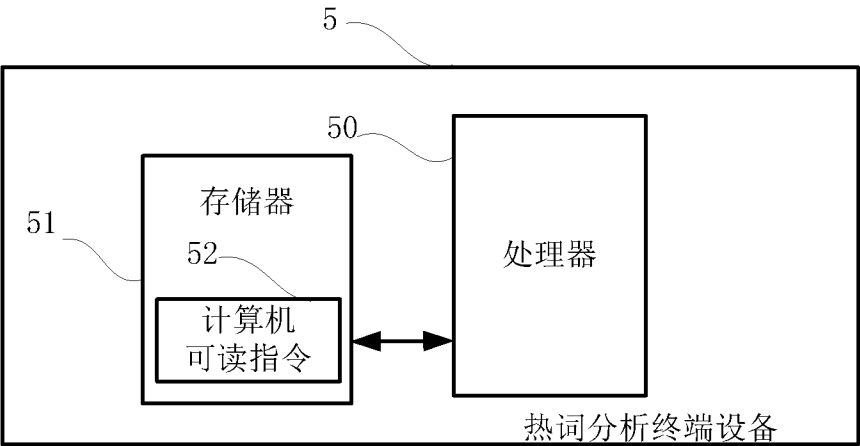


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/096267

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/00(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)

CNPAT, CNKI, WPI, EPODOC: 热词, 分词, 分析, 搜索, 爬取, 网站, 网页, 文本, 切词, 曝光, 频次, 频率, 阈值, 关联度, hot word, analysis, frequency, exposure, enterprise, text, web page, segmentation, count, relevancy

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 103186675 A (NANJING AXON TECHNOLOGY CO., LTD.) 03 July 2013 (2013-07-03) description, paragraphs [0005]-[0011]	1, 6, 11, 16
A	CN 101246499 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.) 20 August 2008 (2008-08-20) entire document	1-20
A	CN 105045882 A (WUXI TVMINING MEDIA SCIENCE & TECHNOLOGY CO., LTD.) 11 November 2015 (2015-11-11) entire document	1-20
A	CN 103106227 A (PEOPLE SEARCH NETWORK CO., LTD.) 15 May 2013 (2013-05-15) 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 January 2019

Date of mailing of the international search report

19 February 2019

Name and mailing address of the ISA/CN

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

Authorized officer

Facsimile No. (86-10)62019451

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/096267

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	103186675	A	03 July 2013	None	
CN	101246499	A	20 August 2008	None	
CN	105045882	A	11 November 2015	None	
CN	103106227	A	15 May 2013	None	

国际检索报告

国际申请号

PCT/CN2018/096267

A. 主题的分类 G06F 17/00 (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)) CNPAT, CNKI, WPI, EPDOC: 热词, 分词, 分析, 搜索, 爬取, 网站, 网页, 文本, 切词, 曝光, 频次, 频率, 阈值, 关联度, hot word, analysis, frequency, exposure, enterprise, text, web page, segmentation, count, relevancy		
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 103186675 A (南京安讯科技有限责任公司) 2013年 7月 3日 (2013 - 07 - 03) 说明书第[0005]-[0011]段	1, 6, 11, 16
A	CN 101246499 A (腾讯科技深圳有限公司) 2008年 8月 20日 (2008 - 08 - 20) 全文	1-20
A	CN 105045882 A (无锡天脉聚源传媒科技有限公司) 2015年 11月 11日 (2015 - 11 - 11) 全文	1-20
A	CN 103106227 A (人民搜索网络股份公司) 2013年 5月 15日 (2013 - 05 - 15) 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 2019年 1月 29日		国际检索报告邮寄日期 2019年 2月 19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		授权官员 白露霜 电话号码 86-(10)53961390

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/096267

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	103186675	A	2013年 7月 3日	无	
CN	101246499	A	2008年 8月 20日	无	
CN	105045882	A	2015年 11月 11日	无	
CN	103106227	A	2013年 5月 15日	无	