(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 109344316 A
(43)申请公布日 2019.02.15

(21)申请号 201810923014.9

(22)申请日 2018.08.14

(71)申请人 优视科技(中国)有限公司
地址 510627 广东省广州市天河区黄埔大道西平云路163号广电平云广场B塔13楼自编01单元

(72)发明人 熊逵 马镇新 孙连生

(74)专利代理机构 北京展翼知识产权代理事务所(特殊普通合伙) 11452
代理人 张阳

(51)Int.Cl.
G06F 16/953(2019.01)

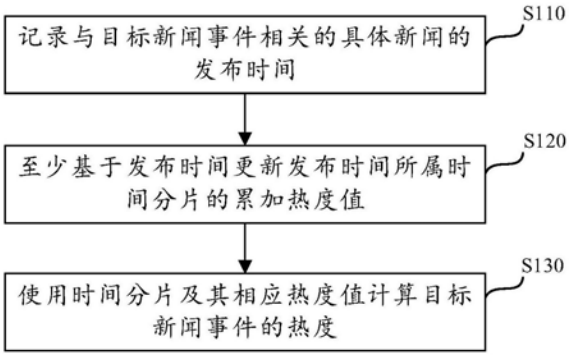
权利要求书4页 说明书11页 附图4页

(54)发明名称

新闻热度计算方法及装置

(57)摘要

公开了一种新闻热度计算方法和装置。所述方法包括:记录与目标新闻事件相关的具体新闻的发布时间;至少基于具体新闻的发布时间将该具体新闻的发布更新至所述发布时间所属时间分片的累加热度值;以及使用时间分片及其相应热度值计算所述目标新闻事件的热度。由此,能够通过引入时间分片来方便地计算目标新闻事件的热度。通过引入包括累加值的时间分片,使得本发明的热度计算方案仅需极少的存储空间就能够快速更新新闻事件的热度信息,根据新闻事件相关报道的发表时间序列,还能够快速拟合出新闻事件的长期热度信息与短期爆发热度信息。



CN 109344316 A

1.一种新闻热度计算方法,包括:

记录与目标新闻事件相关的具体新闻的发布时间;

至少基于具体新闻的发布时间将该具体新闻的发布更新至所述发布时间所属时间分片的累加热度值;以及

使用时间分片及其相应热度值计算所述目标新闻事件的热度。

2.如权利要求1所述的方法,其中,记录与目标新闻事件相关的具体新闻的发布时间包括:

实时、轮询或以预定间隔获取每个媒体发布端下与目标新闻事件相关的具体新闻的发布时间。

3.如权利要求1所述的方法,其中,至少基于具体新闻的发布时间将该具体新闻的发布更新至所述发布时间所属时间分片的累加热度值包括:

使用所述具体新闻的发布时间和热度值更新所述发布时间所属时间分片的累加热度值和最后更新时间值。

4.如权利要求3所述的方法,其中,使用所述具体新闻的发布时间和热度值更新所述发布时间所属时间分片的累加热度值和最后更新时间值包括:

求取当前最后更新时间t_0与所述具体新闻的发布时间t的时间差d;

当d=0,直接将所述具体新闻的热度值h累加至当前累加热度值h_0,以得到所述时间分片的更新的累加热度值;

当d<0,将t的值作为所述时间分片的更新的最后更新时间t_0,并且使用下式更新所述时间分片的累加热度值:

更新的累加热度值=h_0*exp(a*d)+h,

其中exp(x)表示指数函数,a为预定的取值为正的系数;

当d>0,使用下式更新所述时间分片的累加热度值:

更新的累加热度值=h_0+h*exp(-a*d)。

5.如权利要求3所述的方法,其中,

为每个具体新闻分配作为预定常数的热度值h;和/或

基于具体新闻的发布媒体,为所述具体新闻分配不同的热度值h。

6.如权利要求3所述的方法,还包括:

为所述目标新闻事件构造按时间顺序排序的时间分片序列,其中,每个时间分片对应地包括所述累加热度值和所述最后更新时间值。

7.如权利要求6所述的方法,其中,使用时间分片及其相应热度值计算所述目标新闻事件的热度包括:

直接使用所述时间分片序列生成所述目标新闻事件的热度-时间分布图。

8.如权利要求6所述的方法,其中,使用时间分片及其相应热度值计算所述目标新闻事件的热度包括:

使用所述时间分片序列求取所述目标新闻事件的长期热度信息或短期爆发热度信息。

9.如权利要求8所述的方法,其中,使用所述时间分片序列求取所述目标新闻事件的长期热度信息包括:

按照预定规则划分距当前时间或特定时间的多个热度计算时段,每个热度计算时段都

包括在前更短的热度计算时段所包括的所有时间分片；

求取每个热度计算时段的时段累加热度值；以及

基于所述时段累加热度值加权求取所述长期热度信息。

10.如权利要求9所述的方法，其中，所述多个热度计算时段还包括距当前时间或特定时间最长热度计算时段之外的在前热度计算时间。

11.如权利要求9所述的方法，其中，基于时间分片的最后更新时间与所述当前时间或特定时间的时间差确定每个热度计算时段所包括的具体时间分片。

12.如权利要求8所述的方法，其中，使用所述时间分片序列求取所述目标新闻事件的短期爆发热度信息包括：

选取距当前时间或特定时间的预定数量的连续时间分片；

基于所述连续时间分片中取值小于预定阈值的时间分片将所述连续时间分片进行分组；以及

基于每个分组的时间分片最后更新时间和累加热度值求取所述短期爆发热度信息。

13.如权利要求12所述的方法，其中，基于每个分组的时间分片最后更新时间和累加热度值求取所述短期爆发热度信息包括：

基于每个时间分片分组，得到集合C＝{(t_i,h_i)|t_i为分组中所有时间分片最后更新时间的平均值，h_i为分组中所有时间分片的热度值的和}，并且基于下式获取短期爆发热度信息：

$$\text{短期爆发热度信息}=\frac{\sum_{i=1}^{n} h_i * n}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}|t_i - t_j|}$$

其中n为分组个数。

14.如权利要求1所述的方法，其中，所述新闻热度计算方法在执行在线新闻聚类实时抓取的同时实时进行。

15.一种新闻热度计算装置，包括：

记录装置，用于记录与目标新闻事件相关的具体新闻的发布时间；

更新装置，用于至少基于具体新闻的发布时间将该具体新闻的发布更新至所述发布时间所属时间分片的累加热度值；以及

计算装置，使用时间分片及其相应热度值计算所述目标新闻事件的热度。

16.如权利要求15所述的装置，其中，所述记录装置进一步用于实时、轮询或以预定间隔获取每个媒体发布端下与目标新闻事件相关的具体新闻的发布时间。

17.如权利要求15所述的装置，其中，所述更新装置进一步用于使用所述具体新闻的发布时间和热度值更新所述发布时间所属时间分片的累加热度值和最后更新时间值。

18.如权利要求17所述的装置，其中，所述更新装置进一步包括：

求取当前最后更新时间t_0与所述具体新闻的发布时间t的时间差d；

当d＝0，直接将所述具体新闻的热度值h累加至当前累加热度值h_0，以得到所述时间分片的更新的累加热度值；

当d<0，将t的值作为所述时间分片的更新的最后更新时间t_0，并且使用下式更新所述时间分片的累加热度值：

更新的累加热度值＝h_0*exp(a*d)+h，

其中exp(x)表示指数函数，a为预定的取值为正的系数；

当d>0，使用下式更新所述时间分片的累加热度值：

更新的累加热度值＝h_0+h*exp(-a*d)。

19.如权利要求17所述的装置，其中，

为每个具体新闻分配作为预定常数的热度值h；和/或

基于具体新闻的发布媒体，为所述具体新闻分配不同的热度值h。

20.如权利要求17所述的装置，还包括：

构造装置，用于为所述目标新闻事件构造按时间顺序排序的时间分片序列，其中，每个时间分片对应地包括所述累加热度值和所述最后更新时间值。

21.如权利要求20所述的装置，其中，所述计算装置直接使用所述构造装置构造的时间分片序列生成所述目标新闻事件的热度-时间分布图。

22.如权利要求20所述的装置，其中，所述计算装置用于使用所述时间分片序列求取所述目标新闻事件的长期热度信息或短期爆发热度信息。

23.如权利要求22所述的装置，其中，所述计算装置求取所述目标新闻事件的长期热度信息包括：

按照预定规则划分距当前时间或特定时间的多个热度计算时段，每个热度计算时段都包括在前更短的热度计算时段所包括的所有时间分片；

求取每个热度计算时段的时段累加热度值；以及

基于所述时段累加热度值加权求取所述长期热度信息。

24.如权利要求23所述的装置，其中，所述多个热度计算时段还包括距当前时间或特定时间最长热度计算时段之外的在前热度计算时间。

25.如权利要求23所述的装置，其中，基于时间分片的最后更新时间与所述当前时间或特定时间的时间差确定每个热度计算时段所包括的具体时间分片。

26.如权利要求22所述的装置，其中，所述计算装置求取所述目标新闻事件的短期爆发热度信息包括：

选取距当前时间或特定时间的预定数量的连续时间分片；

基于所述连续时间分片中取值小于预定阈值的时间分片将所述连续时间分片进行分组；以及

基于每个分组的时间分片最后更新时间和累加热度值求取所述短期爆发热度信息。

27.如权利要求26所述的装置，其中，基于每个分组的时间分片最后更新时间和累加热度值求取所述短期爆发热度信息包括：

基于每个时间分片分组，得到集合C＝{(t_i,h_i)|t_i为分组中所有时间分片最后更新时间的平均值，h_i为分组中所有时间分片的热度值的和}，并且基于下式获取短期爆发热度信息：

$$\text{短期爆发热度信息}=\frac{\sum_{i=1}^{n} h_i * n}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}|t_i - t_j|}$$

其中n为分组个数。

28.一种在线新闻聚类服务器,包括:

抓取装置,用于实时抓取新闻文章并进行聚合分类;

如权利要求15-27中任一项所述的新闻热度计算装置,用于获取所述抓取装置抓取的新闻文章的发布时间,并将其更新至对应目标新闻事件的时间分片中,并基于所述时间分片计算所述目标新闻事件的热度。

29.一种计算设备,包括:

处理器;以及

存储器,其上存储有可执行代码,当所述可执行代码被所述处理器执行时,使所述处理器执行如权利要求1-14中任一项所述的方法。

30.一种非暂时性机器可读存储介质,其上存储有可执行代码,当所述可执行代码被电子设备的处理器执行时,使所述处理器执行如权利要求1-14中任一项所述的方法。

新闻热度计算方法及装置

技术领域

[0001] 本发明涉及互联网领域，尤其涉及一种新闻热度计算方法及装置。

背景技术

[0002] 获取新闻信息是人们进行互联网浏览的一大目的，新闻聚类技术可以将相关的新闻报道聚合在一起，让人们对新闻事件的了解更加全面、准确。在这其中，热度信息是新闻事件的重要属性之一。一个正在发生的重大事件，需要实时地更新其热度信息。虽然在线的新闻聚类能够将实时抓取到的新闻文章进行聚合，但由于存储容量、计算速度以及更新时延的限制，无法方便地实时地存储并获取一个新闻事件的全部历史文章序列以进行热度拟合。

[0003] 为此，需要一种更为快捷简便的新闻热度计算方案。

发明内容

[0004] 为了解决如上至少一个问题，本发明提出了一种仅需极少的存储空间就能够快速更新新闻事件热度信息的方案，根据新闻事件相关报道的发表时间序列，能够快速拟合出新闻事件的长期热度信息与短期爆发热度信息并实现实时更新。

[0005] 根据本发明的一个方面，提出了一种新闻热度计算方法，包括：记录与目标新闻事件相关的具体新闻的发布时间；至少基于具体新闻的发布时间将该具体新闻的发布更新至所述发布时间所属时间分片的累加热度值；以及使用时间分片及其相应热度值计算所述目标新闻事件的热度。由此，能够通过引入时间分片来方便地计算目标新闻事件的热度。

[0006] 所述新闻热度计算方法可以在执行在线新闻聚类实时抓取的同时实时进行。优选地，记录多个媒体发布端下与目标新闻事件相关的具体新闻的发布时间可以包括：实时、轮询或以预定间隔获取每个媒体发布端下与目标新闻事件相关的具体新闻的发布时间。由此，能够以极小的开销实现对新闻热度的计算，优选为实时计算。

[0007] 至少基于具体新闻的发布时间将该具体新闻的发布更新至所述发布时间所属时间分片的累加热度值可以包括：使用所述具体新闻的发布时间和热度值更新所述发布时间所属时间分片的累加热度值和最后更新时间值。由此，通过引入最后更新时间值，能够更为准确地对每个时间片的属性进行描述。

[0008] 使用所述具体新闻的发布时间和热度值更新所述发布时间所属时间分片的累加热度值和最后更新时间值可以包括：求取所述当前最后更新时间t_0与所述具体新闻的发布时间t的时间差d；当d＝0，直接将所述具体新闻的热度值h累加至当前累加热度值h_0，以得到所述时间分片的更新的累加热度值；当d<0，将t的值作为所述时间分片的更新的最后更新时间t_0，并且使用下式更新所述时间分片的累加热度值：

[0009] 更新的累加热度值＝h_0*exp(a*d)+h，

[0010] 其中exp(x)表示指数函数，a为预定的取值为正的系数；

[0011] 当d>0，使用下式更新所述时间分片的累加热度值：

[0012] 更新的累加热度值＝h_0+h*exp(-a*d)。

[0013] 由此，通过指数求取叠加时间衰减，从而能够更为贴切地反映新闻事件的真实热度。

[0014] 根据不同的实现，可以为每个具体新闻分配作为预定常数的热度值h；和/或基于具体新闻的发布媒体，为所述具体新闻分配不同的热度值h。

[0015] 本发明的新闻热度计算方法还可以包括：为所述目标新闻事件构造按时间顺序排序的时间分片序列，其中，每个时间分片对应地包括所述累加热度值和所述最后更新时间值。由此，通过构造并维护时间分片序列，能够进一步方便对新闻热度的计算，尤其是实时计算。

[0016] 使用时间分片及其相应热度值计算所述目标新闻事件的热度可以包括：直接使用所述时间分片序列生成所述目标新闻事件的热度-时间分布图。

[0017] 使用时间分片及其相应热度值计算所述目标新闻事件的热度还包括：使用所述时间分片序列求取所述目标新闻事件的长期热度信息或短期爆发热度信息。

[0018] 使用所述时间分片序列求取所述目标新闻事件的长期热度信息可以包括：按照预定规则划分距当前时间或特定时间的多个热度计算时段，每个热度计算时段都包括在前更短的热度计算时段所包括的所有时间分片；求取每个热度计算时段的时段累加热度值；以及基于所述时段累加热度值加权求取所述长期热度信息。

[0019] 多个热度计算时段还可以包括距当前时间或特定时间最长热度计算时段之外的在前热度计算时间。

[0020] 优选地，基于时间分片的最后更新时间与所述当前时间或特定时间的时间差确定每个热度计算时段所包括的具体时间分片。

[0021] 使用所述时间分片序列求取所述目标新闻事件的短期爆发热度信息可以包括：选取距当前时间或特定时间的预定数量的连续时间分片；基于所述连续时间分片中取值小于预定阈值的时间分片将所述连续时间分片进行分组；以及基于每个分组的时间分片最后更新时间和累加热度值求取所述短期爆发热度信息。

[0022] 基于每个分组的时间分片最后更新时间和累加热度值求取所述短期爆发热度信息可以包括：基于每个时间分片分组，得到集合C＝{(t_i,h_i)|t_i为分组中所有时间分片最后更新时间的平均值，h_i为分组中所有时间分片的热度值的和}，并且基于下式获取短期爆发热度信息：

[0023] 短期爆发热度信息= $$\frac{\sum_{i=1}^{n} h_i * n}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}|t_i - t_j|}$$

[0024] 其中n为分组个数。

[0025] 由此，能够方便地求取能够合理描述新闻事件热度趋势的长短期热度指标。

[0026] 根据本发明的另一个方面，提出了一种新闻热度计算装置，包括：记录装置，用于记录与目标新闻事件相关的具体新闻的发布时间；更新装置，用于至少基于具体新闻的发布时间将该具体新闻的发布更新至所述发布时间所属时间分片的累加热度值；以及计算装置，使用时间分片及其相应热度值计算所述目标新闻事件的热度。

[0027] 优选地，记录装置可以进一步用于实时、轮询或以预定间隔获取每个媒体发布端

下与目标新闻事件相关的具体新闻的发布时间。

[0028] 优选地，更新装置可以进一步用于使用所述具体新闻的发布时间和热度值更新所述发布时间所属时间分片的累加热度值和最后更新时间值。

[0029] 优选地，更新装置可以进一步用于：求取当前最后更新时间t_0与所述具体新闻的发布时间t的时间差d；当d＝0，直接将所述具体新闻的热度值h累加至当前累加热度值h_0，以得到所述时间分片的更新的累加热度值；当d<0，将t的值作为所述时间分片的更新的最后更新时间t_0，并且使用下式更新所述时间分片的累加热度值：

[0030] 更新的累加热度值＝h_0*exp(a*d)+h，

[0031] 其中exp(x)表示指数函数，a为预定的取值为正的系数；

[0032] 当d>0，使用下式更新所述时间分片的累加热度值：

[0033] 更新的累加热度值＝h_0+h*exp(-a*d)。

[0034] 基于不同的实现，可以为每个具体新闻分配作为预定常数的热度值h；和/或基于具体新闻的发布媒体，为所述具体新闻分配不同的热度值h。

[0035] 在一个实施例中，本发明的新闻热度计算装置还可以包括：构造装置，后者可以用于为所述目标新闻事件构造按时间顺序排序的时间分片序列，其中，每个时间分片对应地包括所述累加热度值和所述最后更新时间值。

[0036] 由此，计算装置可以直接使用构造装置构造的时间分片序列生成所述目标新闻事件的热度-时间分布图。

[0037] 在其他实施例中，计算装置可以用于使用所述时间分片序列求取所述目标新闻事件的长期热度信息或短期爆发热度信息。

[0038] 具体地，计算装置求取所述目标新闻事件的长期热度信息可以包括：按照预定规则划分距当前时间或特定时间的多个热度计算时段，每个热度计算时段都包括在前更短的热度计算时段所包括的所有时间分片；求取每个热度计算时段的时段累加热度值；以及基于所述时段累加热度值加权求取所述长期热度信息。

[0039] 优选地，多个热度计算时段还可以包括距当前时间或特定时间最长热度计算时段之外的在前热度计算时间。

[0040] 优选地，基于时间分片的最后更新时间与所述当前时间或特定时间的时间差确定每个热度计算时段所包括的具体时间分片。

[0041] 计算装置求取所述目标新闻事件的短期爆发热度信息则可包括：选取距当前时间或特定时间的预定数量的连续时间分片；基于所述连续时间分片中取值小于预定阈值的时间分片将所述连续时间分片进行分组；以及基于每个分组的时间分片最后更新时间和累加热度值求取所述短期爆发热度信息。

[0042] 基于每个分组的时间分片最后更新时间和累加热度值求取所述短期爆发热度信息可以包括：基于每个时间分片分组，得到集合C＝{(t_i,h_i)|t_i为分组中所有时间分片最后更新时间的平均值，h_i为分组中所有时间分片的热度值的和}，并且基于下式获取短期爆发热度信息：

[0043] 短期爆发热度信息＝ $$\frac{\sum_{i=1}^{n} h_i * n}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |t_i - t_j|}$$

[0044] 其中n为分组个数。

[0045] 根据本发明的又一个方面，提出了一种在线新闻聚类服务器，包括：抓取装置，用于实时抓取新闻文章并进行聚合分类；如上任一项所述的新闻热度计算装置，用于获取所述抓取装置抓取的新闻文章的发布时间，并将其更新至对应目标新闻事件的时间分片中，并基于所述时间分片计算所述目标新闻事件的热度。

[0046] 根据本发明的一个方面，提出了一种计算设备，包括：处理器；以及存储器，其上存储有可执行代码，当所述可执行代码被所述处理器执行时，使所述处理器执行如上任一项所述的方法。

[0047] 根据本发明的另一个方面，提出了一种非暂时性机器可读存储介质，其上存储有可执行代码，当所述可执行代码被电子设备的处理器执行时，使所述处理器执行如上任一项所述的方法。

[0048] 本专利采用实时更新的时间分片方式，有效避免计算新闻事件热度时，需要存储新闻事件所对应的新闻文章发表时间序列的问题，节省存储空间并提高计算效率。同时长期热度信息与短期爆发热度信息可以描述一个新闻事件的发展状态，对新闻事件的重要性判定有重要的参考价值。

附图说明

[0049] 通过结合附图对本公开示例性实施方式进行更详细的描述，本公开的上述以及其它目的、特征和优势将变得更加明显，其中，在本公开示例性实施方式中，相同的参考标号通常代表相同部件。

[0050] 图1示出了根据本发明一个实施例的新闻热度计算方法的流程示意图。

[0051] 图2示出了根据本发明一个实施例的求取长期热度信息的流程示意图。

[0052] 图3示出了根据本发明一个实施例的求取短期爆发热度信息的流程示意图。

[0053] 图4示出了根据本发明一个实施例的新闻热度计算装置的结构示意图。

[0054] 图5示出了某一新闻事件在爆发后一段时间内的热度变化趋势。

[0055] 图6示出了用于实现上述新闻热度计算方法的计算设备的结构示意图。

具体实施方式

[0056] 下面将参照附图更详细地描述本公开的优选实施方式。虽然附图中显示了本公开的优选实施方式，然而应该理解，可以以各种形式实现本公开而不应被这里阐述的实施方式所限制。相反，提供这些实施方式是为了使本公开更加透彻和完整，并且能够将本公开的范围完整地传达给本领域的技术人员。

[0057] 在线新闻聚类实时地将抓取到的新闻文章归并到一个具体的新闻事件当中。本发明提出的一种基于时间分片的新闻热度计算方案优选与上述新闻聚类抓取同时地实时进行。当一篇新闻文章归入一个具体的新闻事件中时，根据新闻文章的发表时间实时地更新该新闻文章所属的时间分片热度信息，然后遍历该新闻事件的所有时间分片热度信息，综合加权拟合出新闻事件的长期热度信息与短期爆发热度信息。

[0058] 图1示出了根据本发明一个实施例的新闻热度计算方法的流程示意图。基于本发明的新闻热度计算优选结合在线新闻聚类实时进行。例如，在线新闻聚类服务器在进行新

闻聚类时，可以为每一个目标新闻事件执行本发明的新闻热度计算方案。

[0059] 在步骤S110，记录与目标新闻事件相关的具体新闻的发布时间。例如，在线新闻聚类服务器可以在进行新闻文章抓取和归类时，至少同时获取该新闻文章的发布时间。优选地，记录多个媒体发布端下与目标新闻事件相关的具体新闻的发布时间

[0060] 在步骤S120，至少基于具体新闻的发布时间将该具体新闻的发布更新至所述发布时间所属时间分片的累加热度值。首先可以基于新闻文章的发布时间，确定其所属的时间分片。在一个实施例中，可以按小时整点分片。例如，一篇20:15:43发布的文章可被分入20-21点的时间分片内。在每篇文章权值相同且不考虑发布时间远近的影响时，每一篇文章的发布可以使得其对应时间分片的累加热度值直接加1。在更为复杂的实施例中，可以根据其他因素，例如下述的发布时间和该时间分片的最后更新时间的关系来确定其对累加热度值的影响。

[0061] 在步骤S130，使用时间分片及其相应热度值计算所述目标新闻事件的热度。例如，可以根据现有的多个时间分片及其内的热度值来求取该目标新闻事件本身的热度。

[0062] 如前所述，在线新闻聚类服务器可以在进行新闻文章抓取和归类时，至少记录该新闻文章的发布时间用于其所属目标新闻事件热度值的计算。在具体应用中，在线新闻聚类服务器可以采用各种策略进行文章抓取。相应地，步骤S110可以包括实时、轮询或以预定间隔获取每个媒体发布端下与目标新闻事件相关的具体新闻的发布时间。例如，针对每个媒体发布端，在线新闻聚类服务器可以每五分钟获取一次发布文章，将获取的文章按照不同的新闻事件加以归类，并更新该新闻事件下相应时间片的累加热度值(以及可选地最后更新时间)，并可实时更新该新闻事件的热度(例如，下述的长期或短期爆发热度值)。

[0063] 在不同的热度计算模型中，可以采取不同的热度值分配策略。在一个实施例中，可以为每个具体新闻分配作为预定常数的热度值h。例如，在线新闻聚类服务器可以向每一个抓取的具体新闻分配取值为常数1的热度值h。在其它实施例中，可以基于各种因素，来向具体新闻分配不同的热度值h。可以基于具体新闻的发布媒体，为具体新闻分配不同的热度值h。例如，大型权威网站发布的新闻文章，热度值为1，小型网站的热度值为0.8等。

[0064] 在一个实施例中，除了为每个时间分片保存累加热度值之外，还可以为其设置一个最后更新时间参数，以方便求取能够更为准确地反映目标新闻事件的整体热度的热度值。相应地，步骤S120可以包括使用所述具体新闻的发布时间和热度值更新所述发布时间所属时间分片的累加热度值和最后更新时间值。

[0065] 在一个实施例中，使用所述具体新闻的发布时间和热度值更新所述发布时间所属时间分片的累加热度值和最后更新时间值可以包括求取时间分片的当前最后更新时间t_0与具体新闻的发布时间t的时间差d，依据上述时间差d确定是否需要更新该时间分片的最后更新时间以及累加热度值的更新值大小。

[0066] 具体地，当d＝0，即时间分片的当前最后更新时间t_0与具体新闻的发布时间t相同，则可直接将所述具体新闻的热度值h累加至当前累加热度值h_0，以得到所述时间分片的更新的累加热度值。当d不等于0，即在时间分片的当前最后更新时间t_0与具体新闻的发布时间t之间存在时间差时，例如可以利用该时间差d的取值指数求取累加热度值。引入指数乘能够更为确切地反映时间邻近性对热度值的影响，符合新闻事件热度随时间衰减的趋势。

[0067] 更具体地，当d<0，即，具体新闻的发布时间t比时间分片的更新的最后更新时间t_0更晚，则可将t的值作为所述时间分片的更新的最后更新时间t_0，并且可以使用下式更新所述时间分片的累加热度值：

[0068] 更新的累加热度值＝h_0*exp(a*d)+h，

[0069] 其中exp(x)表示指数函数，a为预定的取值为正的系数。而当d>0，即，具体新闻的发布时间t比时间分片的最后更新时间t_0要早，则无需更新所述时间分片的更新的最后更新时间t_0，并且可以使用下式更新所述时间分片的累加热度值：

[0070] 更新的累加热度值＝h_0+h*exp(-a*d)。

[0071] 在这其中，a的取值可以配合d的单位而灵活变化。例如，当d以秒为单位且时间分片按小时分片时，a可以取较小的值，例如1/10000。而当d以分钟或小时为单位时，a的取值可以相应的增大60或3600倍。另外，可以根据经验值等合理规定a的取值，以获取能够更为准确的反映事件热度的指数曲线。

[0072] 在一个实施例中，本发明的新闻热度计算方法还可以包括为所述目标新闻事件构造按时间顺序排序的时间分片序列，其中，每个时间分片对应地包括所述累加热度值和所述最后更新时间值。例如，当在线新闻聚合服务器确定一个新的目标新闻事件时，便可为其划分一个用于存储时间分片序列的空间(通常很小)，并且随着时间流逝，逐个添加新的时间分片并更新当前时间分片内的具体取值。

[0073] 在一个实施例中，针对每一个新闻事件可以对应存储一个如表1所示的时间分片序列，每一个时间分片对应存储一个如表2所示的结构体。每个分片中存储的信息可以包括：最近一次更新时间，初始化为0；以及最近一次更新之后的热度值，初始化为0。当在线新闻聚类归并一篇新闻文章到某一个具体的新闻事件中时，首先根据该新闻文章的发表时间，选择所属的时间分片，更新对应的时间分片的结构体信息。

[0074]

0-1	1-2	2-3	……	21-22	22-23

[0075] 表1.时间按小时分片示例

[0076]

最近一次更新时间(更新时间)	2018-03-29 20:23:22
最近一次更新后的热度值(热度值)	1.5

[0077] 表2.分片信息示例

[0078] 例如，一篇发表于2018-03-29 10:30:42的新闻文章归并到某个新闻事件中，需要更新该新闻事件的时间分片序列信息，每一篇新闻文章对新闻事件的热度贡献为h。首先，根据新闻文章的发表时间，选择所属的时间分片，亦即时间分片10-11。该时间分片当前具有两种可能的状态，一是未曾更新，二是曾有过更新。

[0079] 当未曾更新时，初始化的更新时间为0，热度值为0。此时只需将该分片的更新时间更新为新闻文章发表时间，亦即2018-03-29 10:30:42；热度值更新为h即可。

[0080] 而在曾经有过更新的情况下，则可根据例如上文所述，基于时间差d的取值，给出针对累加热度值和/或最后更新时间的更新。

[0081] 基于如上构造并更新的时间分片序列，可以按需求取用于从不同层面反映目标新闻事件热度的各类热度信息。在一个实施例中，步骤S130可以包括直接使用所述时间分片

序列构造所述目标新闻事件的热度-时间分布图。在其他实施例中，步骤S130还可以包括使用所述时间分片序列求取所述目标新闻事件的长期热度信息或短期爆发热度信息。

[0082] 图2示出了根据本发明一个实施例的求取长期热度信息的流程示意图。该方法可以看作步骤S130的子步骤。

[0083] 在步骤S210，按照预定规则划分距当前时间或特定时间的多个热度计算时段，每个热度计算时段都包括在前更短的热度计算时段所包括的所有时间分片。

[0084] 在步骤S220，求取每个热度计算时段的时段累加热度值。

[0085] 在步骤S230，基于所述时段累加热度值加权求取所述长期热度信息。

[0086] 优选地，多个热度计算时段还包括距当前时间或特定时间最长热度计算时段之外的在前热度计算时间。

[0087] 优选地，每个热度计算时段所包括的具体时间分片可以基于时间分片的最后更新时间与所述当前时间或特定时间的时间差来确定。

[0088] 例如，在通过上述步骤得到某一新闻事件的实时更新的时间分片序列。在该实时更新的时间分片序列基础上，假设当前时间为t_c，第i个时间分片上存储的更新时间为t_i，热度值为h_i，则计算长期热度信息可以包括如下步骤：

[0089] (1)分别初始化1小时内、3小时内、7小时内、12小时内、1天内、3天内以及3天以外的热度值为t_{1h}，t_{3h}，t_{7h}，t_{12h}，t_{1d}，t_{3d}，t_{od}为0；

[0090] (2)根据当前时间t_c，确定当前时间所在的时间分片序列位置s_i，按时间向前循环递推，统计步骤(1)中定义的各个热度值。优选对分小时统计与分天统计采取不同的计算策略，其快速计算方法如下：

[0091] a)按小时统计时，根据位置s_i，按时间向前递推即可。此处以统计1小时内热度为例，说明具体统计方法，分为两种情况：

[0092] i)若s_i在时间分片0-1，则分别对比时间分片0-1、时间分片22-23的更新时间与当前时间t_c的差值，如时间差在1小时内，则对应时间分片的热度值累加到t_{1h}上；

[0093] ii)若s_i在除时间分片0-1的其他位置，则分别对比时间分片s_i、时间分片s_{i-1}的更新时间与当前时间t_c的时间差，如时间差在1小时内，则累加到t_{1h}上；

[0094] 其余t_{3h}，t_{7h}，t_{12h}可依照此方法类推，分别进行统计；

[0095] b)按天统计时，遍历整个时间分片，对比t_i与t_c的时间差d_c，如d_c在1天内则将h_i累加到t_{1d}，如d_c在3天内则将h_i累加到t_{3d}，如d_c在3天及3天外则将h_i累加到t_{od}；

[0096] (3)通过步骤(2)，得到了不同时间段内的新闻事件的热度信息值，随后可以基于如下公式，加权得到新闻事件的长期热度信息：

[0097] 长期热度信息＝$t_{1h}*a_{1h}+t_{3h}*a_{3h}+t_{7h}*a_{7h}+t_{12h}*a_{12h}$

[0098] $+t_{1d}*a_{1d}+t_{3d}*a_{3d}+t_{od}*a_{od}$

[0099] 其中，a_{1h}，a_{3h}，a_{7h}，a_{12h}，a_{1d}，a_{3d}，a_{od}分别为1小时内、3小时内、7小时内、12小时内、1天内、3天内及3天以外的热度信息对于长期热度信息的权重。应该理解的是，如上1小时内、3小时内、7小时内、12小时内、1天内、3天内以及3天以外的热度值的划分仅仅是一个例子，在具体应用中，可以根据经验值或具体应用场景灵活选择对具体时间段的划分，以及每个时间段的权值。由此，通过反复叠加接近时间片的累加值，可以对接近时间片的统计值叠加多层权重，从而能够更好地反映新闻事件的时间衰减性。

[0100] 图3示出了根据本发明一个实施例的求取短期爆发热度信息的流程示意图。该方法同样可以看作步骤S130的子步骤。

[0101] 与长期累积热度信息不同的是，短期爆发热度信息需要考虑文章发表的爆发集中程度，比如针对不同事件，1天内媒体发表了10篇文章与1小时内媒体发表了10篇文章，其爆发程度就不一样。

[0102] 由此，在步骤S310，选取距当前时间或特定时间的预定数量的连续时间分片。例如，可以选取距当前时间24小时之内的时间分片作为短期爆发热度信息的计算范围。

[0103] 在步骤S320，基于所述连续时间分片中取值小于预定阈值的时间分片将所述连续时间分片进行分组。在此，预定阈值可以为零或其他值，从而以爆发间歇作为分组的依据，在其他实施例中，还可以采取其他的分组依据。

[0104] 在步骤S330，基于每个分组的时间分片最后更新时间和累加热度值求取所述短期爆发热度信息。

[0105] 例如，假设当前时间为t_c，第i个时间分片上存储的更新时间为t_i，热度值为h_i，则计算短期爆发热度信息的步骤如下：

[0106] (1)遍历时间分片序列，对比t_i与t_c的时间差，如时间差在1天之内，挑选出来组成新的序列SEQ＝{(t_k,h_k)|t_k与t_c的时间差在1天内}；

[0107] (2)将SEQ根据时间序列进行分组，在时间序列中同一组内的时间分片在原始时间分片序列中的位置在时间上是相邻的；例如，时间分片0-1与时间分片1-2是相邻的，时间分片0-1与时间分片22-23在时间上也是相邻的，于是可以直接基于时间分片的取值是否超过阈值进行分组，例如，可以将热度值小于阈值α的时间分片作为分组的临界分片，或者直接将无更新时间段之间的分片聚合为一个组；

[0108] (3)基于每个时间分片分组，得到集合C＝{(t_i,h_i)|t_i为分组中所有时间分片最后更新时间的平均值，h_i为分组中所有时间分片的热度值的和}，并且基于下式获取短期爆发热度信息：

[0109] 短期爆发热度信息= $$\frac{\sum_{i=1}^{n} h_i * n}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}|t_i - t_j|}$$

[0110] 其中n为分组个数。

[0111] 由此，使得求取的短期爆发热度信息能够较为准确的反映时间衰减和集中爆发程度。

[0112] 如上已结合图1-3描述了根据本发明的新闻热度计算方法。如下将结合图4描述根据本发明的新闻热度计算装置。

[0113] 图4示出了根据本发明一个实施例的新闻热度计算装置的结构示意图。如图4所示，新闻热度计算装置400可以包括：记录装置410、更新装置420和计算装置430。

[0114] 记录装置410可以用于记录与目标新闻事件相关的具体新闻的发布时间。更新装置420可以用于至少基于具体新闻的发布时间将该具体新闻的发布更新至所述发布时间所属时间分片的累加热度值。计算装置430则可使用时间分片及其相应热度值计算所述目标新闻事件的热度。

[0115] 在一个实施例中，记录装置410可以进一步用于实时、轮询或以预定间隔获取每个

媒体发布端下与目标新闻事件相关的具体新闻的发布时间。

[0116] 在一个实施例中，更新装置410可以进一步用于使用所述具体新闻的发布时间和热度值更新所述发布时间所属时间分片的累加热度值和最后更新时间值。

[0117] 在一个实施例中，更新装置410可以进一步用于：求取所述当前最后更新时间t_0与所述具体新闻的发布时间t的时间差d；当d＝0，直接将所述具体新闻的热度值h累加至当前累加热度值h_0，以得到所述时间分片的更新的累加热度值；当d<0，将t的值作为所述时间分片的更新的最后更新时间t_0，并且使用下式更新所述时间分片的累加热度值：

[0118] 更新的累加热度值＝h_0*exp(a*d)+h，

[0119] 其中exp(x)表示指数函数，a为预定的取值为正的系数；

[0120] 当d>0，使用下式更新所述时间分片的累加热度值：

[0121] 更新的累加热度值＝h_0+h*exp(-a*d)。

[0122] 基于不同的实现，可以为每个具体新闻分配作为预定常数的热度值h；和/或基于具体新闻的发布媒体，为所述具体新闻分配不同的热度值h。

[0123] 在一个实施例中，本发明的新闻热度计算装置400还可以包括：构造装置440，后者可以用于为所述目标新闻事件构造按时间顺序排序的时间分片序列，其中，每个时间分片对应地包括所述累加热度值和所述最后更新时间值。

[0124] 由此，计算装置430可以直接使用构造装置440构造的时间分片序列生成所述目标新闻事件的热度-时间分布图。

[0125] 在其他实施例中，计算装置430可以用于使用所述时间分片序列求取所述目标新闻事件的长期热度信息或短期爆发热度信息。

[0126] 具体地，计算装置430求取所述目标新闻事件的长期热度信息可以包括：按照预定规则划分距当前时间或特定时间的多个热度计算时段，每个热度计算时段都包括在前更短的热度计算时段所包括的所有时间分片；求取每个热度计算时段的时段累加热度值；以及基于所述时段累加热度值加权求取所述长期热度信息。

[0127] 优选地，多个热度计算时段还可以包括距当前时间或特定时间最长热度计算时段之外的在前热度计算时间。

[0128] 优选地，基于时间分片的最后更新时间与所述当前时间或特定时间的时间差确定每个热度计算时段所包括的具体时间分片。

[0129] 计算装置430求取所述目标新闻事件的短期爆发热度信息则可包括：选取距当前时间或特定时间的预定数量的连续时间分片；基于所述连续时间分片中取值小于预定阈值的时间分片将所述连续时间分片进行分组；以及基于每个分组的时间分片最后更新时间和累加热度值求取所述短期爆发热度信息。

[0130] 基于每个分组的时间分片最后更新时间和累加热度值求取所述短期爆发热度信息可以包括：基于每个时间分片分组，得到集合C＝{(t_i,h_i)|t_i为分组中所有时间分片最后更新时间的平均值，h_i为分组中所有时间分片的热度值的和}，并且基于下式获取短期爆发热度信息：

[0131] 短期爆发热度信息= $$\frac{\sum_{i=1}^{n} h_i * n}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}|t_i - t_j|}$$

[0132] 其中n为分组个数。

[0133] 本发明的技术方案还可以实现为一种在线新闻聚类服务器，包括：抓取装置，用于实时抓取新闻文章并进行聚合分类；以及如上所述的新闻热度计算装置，用于获取所述抓取装置抓取的新闻文章的发布时间，并将其更新至对应目标新闻事件的时间分片中，并基于所述时间分片计算所述目标新闻事件的热度。

[0134] 由此，本发明采用时间分片方式来有效避免计算新闻事件热度时，需要存储一个新闻事件所对应的新闻文章发表时间序列的问题，节省存储空间并提高计算效率。同时长期热度信息与短期爆发热度信息可以描述一个新闻事件的发展状态，对新闻事件的重要性判定有重要的参考价值。

[0135] [应用例]

[0136] 以事件“中美贸易战”为例，图5示出了某一新闻事件在爆发后一段时间内的热度变化趋势。图5所示例如可以基于本发明的时间分片序列直接得出。

[0137] 进一步地，表3示出了关键词“中美贸易战”进行检索的相关事件排序。

[0138]

序号	事件Id	一级类目	二级类目	新闻事件(点击查看聚类详情)	事件开始时间	热点聚类时间	热度	score
1	2f726b4fbc9633b9a1dfec6216d71150	未分类 财经 国内	时政	商务部: 希望美方不要一意孤行	1522332480->2018-03-29 22:08:00	1522332886->2018-03-29 22:14:46	47.1724	3.71578908731875
2	22954bb7c07af69c9623dc8655e163d6	国内	时政	如何解决中美贸易争端? 商务部: 解铃还须系铃人	1522293180->2018-03-29 11:13:00	1522294373->2018-03-29 11:32:53	27.5682	3.47164720439993
3	47b2dd56dbb693748f1845b160ea5c9a	社会 未分类 国际	国际时政	[老外街访评]在中美贸易争端这件事上, 老外们这样力挺中国	1522284126->2018-03-29 08:42:06	1522284596->2018-03-29 08:49:56	6.1703 1	2.98314019820672
4	a51e678ff02a456d976c2435e67af579	财经	宏观经济 经济民生	WTO总干事: 如果美中贸易战全面打响, 世界经济增长可能急剧萎缩	1522267200->2018-03-29 04:00:00	1522274634->2018-03-29 06:03:54	8.8921 2	2.84929518485231
5	aff937481a712a03550ac830a8f6733a	财经	股票 经济民生	全球股市波动加大 投资者该如何应对	1522275600->2018-03-29 06:20:00	1522277209->2018-03-29 06:46:49	6.583	2.79730957356798
6	ac95bf72e1a2019f36e9402f11c62e05	财经 未分类		打还是谈? 中美贸易重要关头, 你不能糊涂	1522313880->2018-03-29 16:58:00	1522314130->2018-03-29 17:02:10	5.2748 9	2.64852434711224

[0139] 表3.新闻事件检索热度信息应用例

[0140] 由此，利用本发明的新闻热度计算方案，能够以极小的存储和计算代价实现对新闻热度的计算，尤其是实时计算。

[0141] 进一步地，图6示出了用于实现上述新闻热度计算方法的计算设备的结构示意图。

[0142] 参见图6，计算设备600包括存储器610和处理器620。

[0143] 处理器620可以是一个多核的处理器，也可以包含多个处理器。在一些实施例中，处理器620可以包含一个通用的主处理器以及一个或多个特殊的协处理器，例如图形处理器(GPU)、数字信号处理器(DSP)等等。在一些实施例中，处理器620可以使用定制的电路实现，例如特定用途集成电路(ASIC，Application Specific Integrated Circuit)或者现场可编程逻辑门阵列(FPGA，Field Programmable Gate Arrays)。

[0144] 存储器610可以包括各种类型的存储单元，例如系统内存、只读存储器(ROM)，和永久存储装置。其中，ROM可以存储处理器620或者计算机的其他模块需要的静态数据或者指令。永久存储装置可以是可读写的存储装置。永久存储装置可以是即使计算机断电后也不会失去存储的指令和数据的非易失性存储设备。在一些实施方式中，永久性存储装置采用大容量存储装置(例如磁或光盘、闪存)作为永久存储装置。另外一些实施方式中，永久性存

储装置可以是可移除的存储设备(例如软盘、光驱)。系统内存可以是可读写存储设备或者易失性可读写存储设备,例如动态随机访问内存。系统内存可以存储一些或者所有处理器在运行时需要的指令和数据。此外,存储器610可以包括任意计算机可读存储媒介的组合,包括各种类型的半导体存储芯片(DRAM,SRAM,SDRAM,闪存,可编程只读存储器),磁盘和/或光盘也可以采用。在一些实施方式中,存储器610可以包括可读和/或写的可移除的存储设备,例如激光唱片(CD)、只读数字多功能光盘(例如DVD-ROM,双层DVD-ROM)、只读蓝光光盘、超密度光盘、闪存卡(例如SD卡、min SD卡、Micro-SD卡等等)、磁性软盘等等。计算机可读存储媒介不包含载波和通过无线或有线传输的瞬间电子信号。

[0145] 存储器610上存储有可执行代码,当可执行代码被处理器620处理时,可以使处理器620执行上文述及的新闻热度计算方法。

[0146] 上文中已经参考附图详细描述了根据本发明的新闻热度计算方案。

[0147] 此外,根据本发明的方法还可以实现为一种计算机程序或计算机程序产品,该计算机程序或计算机程序产品包括用于执行本发明的上述方法中限定的上述各步骤的计算机程序代码指令。

[0148] 或者,本发明还可以实施为一种非暂时性机器可读存储介质(或计算机可读存储介质、或机器可读存储介质),其上存储有可执行代码(或计算机程序、或计算机指令代码),当所述可执行代码(或计算机程序、或计算机指令代码)被电子设备(或计算设备、服务器等)的处理器执行时,使所述处理器执行根据本发明的上述方法的各个步骤。

[0149] 本领域技术人员还将明白的是,结合这里的公开所描述的各种示例性逻辑块、模块、电路和算法步骤可以被实现为电子硬件、计算机软件或两者的组合。

[0150] 附图中的流程图和框图显示了根据本发明的多个实施例的系统和方法的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段或代码的一部分,所述模块、程序段或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意,在有些作为替换的实现中,方框中所标记的功能也可以以不同于附图中所标记的顺序发生。例如,两个连续的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意的是,框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合,可以用执行规定的功能或操作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

[0151] 以上已经描述了本发明的各实施例,上述说明是示例性的,并非穷尽性的,并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下,对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择,旨在最好地解释各实施例的原理、实际应用或对市场中的技术的改进,或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。

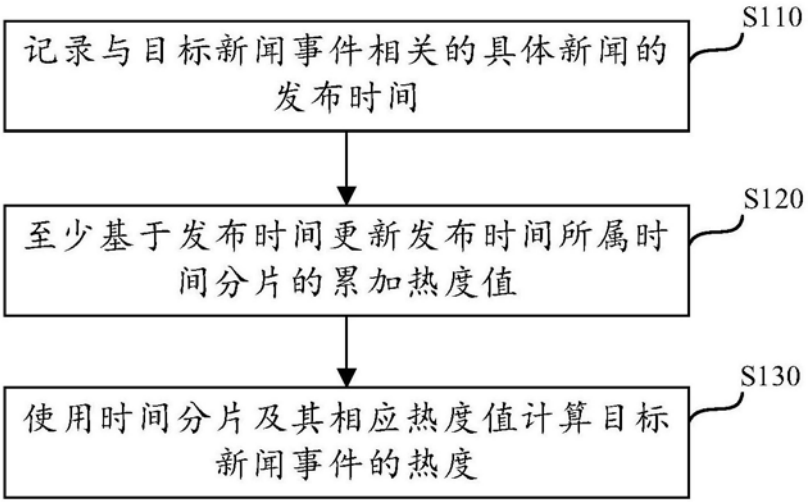

记录与目标新闻事件相关的具体新闻的发布时间
S110
至少基于发布时间更新发布时间所属时间分片的累加热度值
S120
使用时间分片及其相应热度值计算目标新闻事件的热度
S130

图1

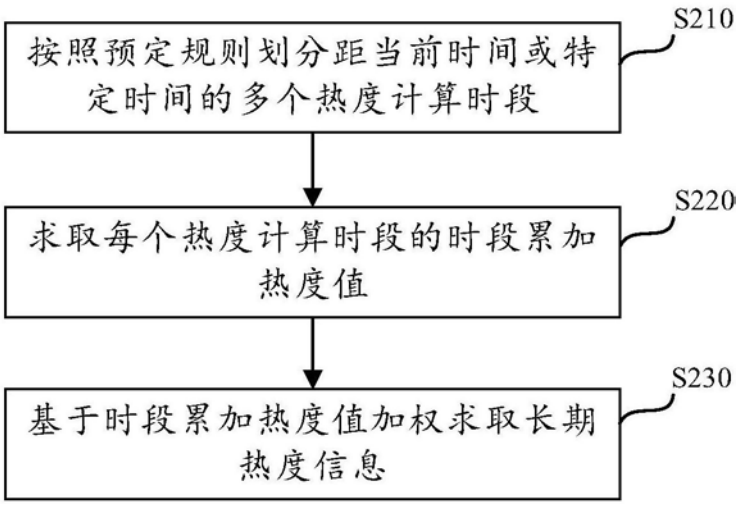

按照预定规则划分距当前时间或特定时间的多个热度计算时段
S210
求取每个热度计算时段的时段累加热度值
S220
基于时段累加热度值加权求取长期热度信息
S230

图2

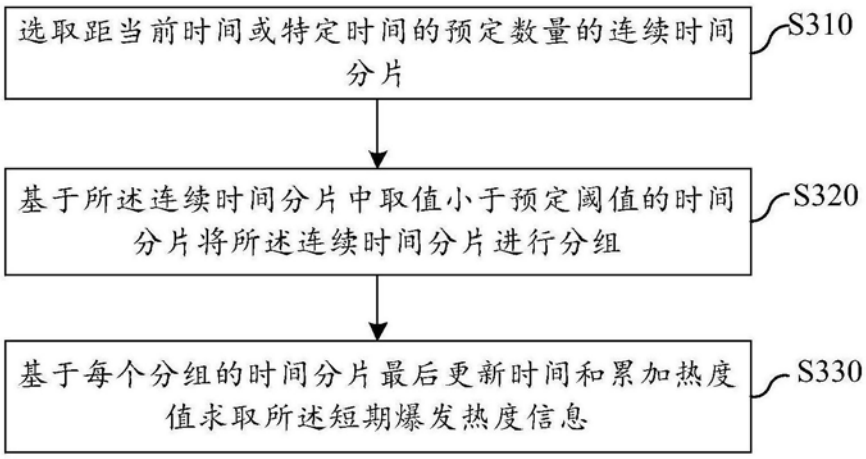

选取距当前时间或特定时间的预定数量的连续时间分片
S310
基于所述连续时间分片中取值小于预定阈值的时间分片将所述连续时间分片进行分组
S320
基于每个分组的时间分片最后更新时间和累加热度值求取所述短期爆发热度信息
S330

图3


新闻热度计算装置400
记录装置410
更新装置420
计算装置430
构造装置440

图4


97.6424
11.2874
43.4638
24.4218
15.7257
22.4824
15.9558
14.8002
24.9
0
125
100
75
50
25
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30

图5

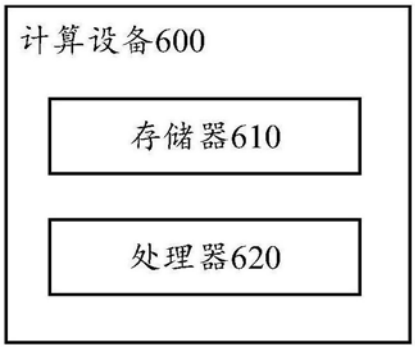


图6