



(12)发明专利

(10)授权公告号 CN 104021205 B

(45)授权公告日 2017.04.19

(21)申请号 201410272487.9

(51)Int.Cl.

(22)申请日 2014.06.18

G06F 17/30(2006.01)

(65)同一申请的已公布的文献号

审查员 姚晓斌

申请公布号 CN 104021205 A

(43)申请公布日 2014.09.03

(73)专利权人 中国人民解放军国防科学技术大学

地址 410073 湖南省长沙市开福区砚瓦池
正街47号

(72)发明人 杨树强 陈志坤 金松昌 尹洪
贾焰 韩伟红 周斌 李爱平

(74)专利代理机构 北京集佳知识产权代理有限公司 11227

代理人 王宝筠

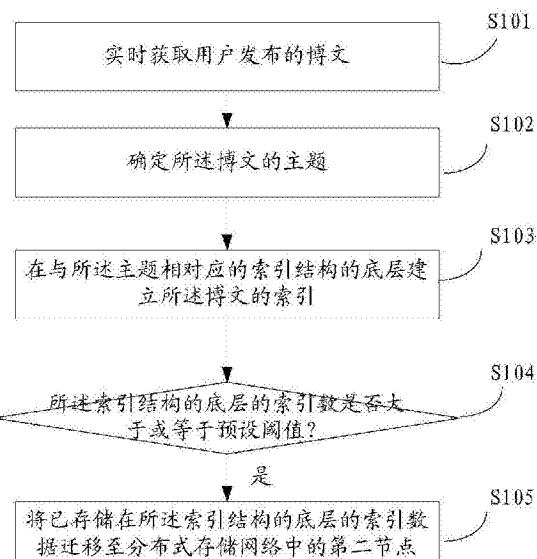
权利要求书2页 说明书8页 附图3页

(54)发明名称

一种建立微博索引的方法及装置

(57)摘要

本发明公开了一种建立微博索引的方法及装置，所述方法应用于分布式存储网络中的第一节点，包括：从微博系统中实时获取用户发布的博文；根据博文与其他博文的转发和/或回复关系并使用Twitter-LDA模型确定所述博文的主题；在与所述主题相对应的索引结构的底层建立所述博文的索引；判断所述索引结构的底层的索引数是否大于或等于预设阈值；如果是，将已存储在所述索引结构的底层的索引数据迁移至分布式存储网络中的第二节点。采用本发明的方法或装置，可以提高建立索引的效率，从而使最新博文在其发表后短时间内变为可搜索。



1. 一种建立微博索引的方法,其特征在于,所述方法应用于分布式存储网络中的第一节点,包括:

从微博系统中实时获取用户发布的博文;

当所述博文是对其他博文的转发和/或回复时,将所述博文的主题确定为所述博文所回复和/或所转发的原始博文的主题;

当所述博文与其他博文无转发和/或回复关系时,使用Twitter-LDA模型确定所述博文的主题;

在与所述主题相对应的索引结构的底层建立所述博文的索引;

判断所述索引结构的底层的索引数是否大于或等于预设阈值;

如果是,则将已存储在所述索引结构的底层的索引数据迁移至分布式存储网络中的第二节点。

2. 如权利要求1所述的方法,其特征在于,所述从微博系统中实时获取用户发布的博文之后,还包括:

确定发布所述博文的用户;

判断所述用户是否为恶意用户,如果否,才执行确定所述博文的主题的步骤。

3. 如权利要求1所述的方法,其特征在于,建立所述索引之后,还包括:

获取所述微博系统中的主节点下发的搜索任务;所述搜索任务是由所述主节点根据用户在搜索操作中给定的搜索关键字对应的主题所分配的;

在所述索引结构的底层搜索与所述关键字相匹配的索引,得到搜索结果;

将所述搜索结果发送至所述主节点,以便所述主节点综合所述第一节点的搜索结果和所述第二节点的搜索结果得到所述主题的搜索结果,综合所有主题的搜索结果,得到最终搜索结果。

4. 如权利要求3所述的方法,其特征在于,所述综合所有主题的搜索结果之前还包括:对所述所有主题的搜索结果进行排序。

5. 一种建立微博索引的装置,其特征在于,所述装置应用于分布式存储网络中的第一节点,包括:

博文获取模块:用于从微博系统中实时获取用户发布的博文;

主题确定模块:用于当所述博文是对其他博文的转发和/或回复时,将所述博文的主题确定为所述博文所回复和/或所转发的原始博文的主题,当所述博文与其他博文无转发和/或回复关系时,使用Twitter-LDA模型确定所述博文的主题;

索引建立模块:用于在与所述主题相对应的索引结构的底层建立所述博文的索引;

阈值判断模块:用于判断所述索引结构的底层的索引数是否大于或等于预设阈值,如果是,触发索引迁移模块;

索引迁移模块:用于将已存储在所述索引结构的底层的索引数据迁移至分布式存储网络中的第二节点。

6. 如权利要求5所述的装置,其特征在于,所述博文获取模块从微博系统中实时获取用户发布的博文之后还包括:

用户确定模块,用于确定发布所述博文的用户;

用户判断模块,用于判断所述用户是否为恶意用户,如果否,触发主题确定模块。

7. 如权利要求5所述的装置，其特征在于，所述索引建立模块建立所述索引之后还包括：

任务获取模块，用于获取所述微博系统中的主节点下发的搜索任务；所述搜索任务是由所述主节点根据用户在搜索操作中给定的搜索关键字对应的主题所分配的；

搜索执行模块，用于在所述索引结构的底层搜索与所述关键字相匹配的索引，得到搜索结果；

结果发送模块，用于将所述搜索结果发送至所述主节点，以便所述主节点综合所述第一节点和所述第二节点的搜索结果得到所述主题的搜索结果，综合所有主题的搜索结果，得到最终搜索结果。

8. 如权利要求7所述的装置，其特征在于，所述结果发送模块综合所述所有主题的搜索结果之前还包括：

排序模块，用于对所述所有主题的搜索结果进行排序。

一种建立微博索引的方法及装置

技术领域

[0001] 本发明涉及数据索引技术领域,更具体地说,涉及一种建立微博索引的方法及装置。

背景技术

[0002] 微博,是微型博客的简称,是一种基于用户关系分享、传播以及获取信息的平台。通过微博系统的实时搜索服务,用户可以快速得到新鲜的第一手草根信息,第一时间了解国内外事件。而实时搜索服务实现过程中,为了能够快速的获取实时微博的信息,需要对微博系统中的博文建立索引。

[0003] 目前,微博系统中建立索引的过程是这样的:只要有新博文进入微博系统,就为该博文建立一条索引,所有博文的索引以简单集合形式存在。

[0004] 发明人经研究发现,微博系统中实时产生的博文数量非常庞大,逐一为这些博文建立索引相当耗时,根本无法让最新的博文在其发表之后的几秒之内就变为可搜索;而且,由于微博系统中本身的博文数据量很大,故博文索引的数据量也不可小觑,如此庞大的数据对存储设备来说是极大的负荷,存储设备的读写速度会受影响,为新博文建立索引时速度会很慢,无法满足建立博文索引的实时性。

发明内容

[0005] 有鉴于此,本发明提供一种建立微博索引的方法及装置,能够快速的对最新博文建立索引,使最新博文在其发表后短时间内变为可搜索。

[0006] 为了实现上述目的,现提出的方案如下:

[0007] 一种建立微博索引的方法,所述方法应用于分布式存储网络中的第一节点,包括:

[0008] 从微博系统中实时获取用户发布的博文;

[0009] 当所述博文是对其他博文的转发和/或回复时,将所述博文的主题确定为所述博文所回复和/或所转发的原始博文的主题;

[0010] 当所述博文与其他博文无转发和/或回复关系时,使用Twitter-LDA模型确定所述博文的主题;

[0011] 在与所述主题相对应的索引结构的底层建立所述博文的索引;

[0012] 判断所述索引结构的底层的索引数是否大于或等于预设阈值;

[0013] 如果是,则将已存储在所述索引结构的底层的索引数据迁移至分布式存储网络中的第二节点。

[0014] 上述方法,所述从微博系统中实时获取用户发布的博文之后,还包括:

[0015] 确定发布所述博文的用户;

[0016] 判断所述用户是否为恶意用户,如果否,才执行确定所述博文的主题的步骤。

[0017] 上述方法,建立所述索引之后,还包括:

[0018] 获取所述微博系统中的主节点下发的搜索任务;所述搜索任务是由所述主节点根

据用户在搜索操作中给定的搜索关键字对应的主题所分配的；

[0019] 在所述索引结构的底层搜索与所述关键字相匹配的索引，得到搜索结果；

[0020] 将所述搜索结果发送至所述主节点，以便所述主节点综合所述第一节点的搜索结果和所述第二节点的搜索结果得到所述主题的搜索结果，综合所有主题的搜索结果，得到最终搜索结果。

[0021] 上述方法，优选地，所述综合所有主题的搜索结果之前还包括：对所述所有主题的排序结果进行排序。

[0022] 一种建立微博索引的装置，所述装置应用于分布式存储网络中的第一节点，包括：

[0023] 博文获取模块：用于从微博系统中实时获取用户发布的博文；

[0024] 主题确定模块：用于当所述博文是对其他博文的转发和/或回复时，将所述博文的主题确定为所述博文所回复和/或所转发的原始博文的主题，当所述博文与其他博文无转发和/或回复关系时，使用Twitter-LDA模型确定所述博文的主题；

[0025] 索引建立模块：用于在与所述主题相对应的索引结构的底层建立所述博文的索引；

[0026] 阈值判断模块：用于判断所述索引结构的底层的索引数是否大于或等于预设阈值，如果是，触发索引迁移模块；

[0027] 索引迁移模块：用于将已存储在所述索引结构的底层的索引数据迁移至分布式存储网络中的第二节点。

[0028] 上述装置，优选地，所述博文获取模块从微博系统中实时获取用户发布的博文之后还包括：

[0029] 用户确定模块，用于确定发布所述博文的用户；

[0030] 用户判断模块，用于判断所述用户是否为恶意用户，如果否，触发主题确定模块。

[0031] 上述装置，优选地，所述索引建立模块建立所述索引之后还包括：

[0032] 任务获取模块，用于获取所述微博系统中的主节点下发的搜索任务；所述搜索任务是由所述主节点根据用户在搜索操作中给定的搜索关键字对应的主题所分配的；

[0033] 搜索执行模块，用于在所述索引结构的底层搜索与所述关键字相匹配的索引，得到搜索结果；

[0034] 结果发送模块，用于将所述搜索结果发送至所述主节点，以便所述主节点综合所述第一节点和所述第二节点的搜索结果得到所述主题的搜索结果，综合所有主题的搜索结果，得到最终搜索结果。

[0035] 上述装置，优选地，所述结果发送模块综合所述所有主题的搜索结果之前还包括：

[0036] 排序模块，用于对所述所有主题的搜索结果进行排序。

[0037] 本实施例公开的建立微博索引的方法，依据博文的主题，在与主题对应的索引结构中建立博文的索引，微博系统的内存中仅保存主题与索引结构的映射关系，映射关系的数据量相对博文索引量来说较小，各个主题的索引结构分布式存储于多个节点上，这样，属于不同主题的多个博文进入微博系统后可由多个节点同时处理，加快了索引建立速度；而且，所述博文的索引由所述索引结构的第一节点建立在所述索引结构的底层，当所述底层中索引数超过预设阈值时，将所述底层中索引数据移至所述索引结构的其他层，即交由所述索引结构的第二节点维护，也就是说，博文的索引在索引结构中分层存储，索引结构的底

层存储的都是为最新进入微博系统的博文建立的索引,这样就不会出现存储设备负荷过大的问题,保证了索引的快速建立;从而使最新博文在其发表后短时间内变为可搜索。

附图说明

[0038] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0039] 图1为本发明实施例公开的一种建立微博索引方法的流程图;

[0040] 图2为本发明实施例公开的一种索引结构工作过程示意图;

[0041] 图3为本发明实施例公开的一种基于索引结构搜索博文的流程图;

[0042] 图4为本发明实施例公开的一种建立微博索引装置的结构示意图;

[0043] 图5为本发明实施例公开的一种基于索引结构搜索博文装置的结构示意图。

具体实施方式

[0044] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0045] 本发明实施例一公开了一种建立微博索引的方法,参见图1所示,所述方法应用于分布式存储网络中的第一节点,包括步骤:

[0046] S101:从微博系统中实时获取用户发布的博文;

[0047] 任意注册用户只要发布博文,博文就会进入微博系统,也就是说博文系统中囊括了所有注册用户发布的所有博文。分布式存储网络中的第一节点,也就是为博文建立索引的节点,其首先要从微博系统中实时获取用户发布的博文,这意味着博文一进入到微博系统,就会被所述第一节点获取到,第一节点当前获取到的博文是最新的博文。

[0048] S102:确定所述博文的主题:

[0049] 当所述博文是对其他博文的转发和/或回复时,将所述博文的主题确定为所述博文所回复和/或所转发的原始博文的主题;

[0050] 当所述博文与其他博文无转发和/或回复关系时,使用Twitter-LDA模型确定所述博文的主题;

[0051] 具体地,微博系统中的博文存在着回复以及转发等关系,例如,用户A在自己的博文中使用了“RT@B”或“via@B”,则说明用户A的这条博文是转发用户B的,而如果用户A发布的博文中包括“@B”,则说明用户A的这条博文是对用户B的博文的回复,可想而知,转发其他用户博文或回复其他用户博文的博文,其主题与其转发/回复的博文主题应该是一致的。具体实施过程中,按照微博系统中博文的回复以及转发关系,微博系统中以树的结构存储所有博文,确定主题时,属于同一棵微博树的所有博文的主题与该微博树的根节点博文的主题一致,也就是原始博文的主题。当然,其他实施例中还可以以其他便于博文主题确定的结构存储博文,例如无向图、队列等。

[0052] 对于那些与其他博文没有转发和/或回复关系的原始博文,使用Twitter-LDA模型确定博文主题,该模型下,一条博文有一个确定的主题。使用该模型确定博文主题时,主要包括两个阶段:离线训练阶段和在线推断阶段,离线训练阶段目的是得到一些全局统计参数,例如词-主题矩阵、背景词的主题向量以及背景词和主题词的分布,在线推断阶段根据离线训练阶段得到的参数对每个博文的主题进行推断,从而得出博文的主题。当然,其他实施例中还可以以其他基于统计学分类的模型确定博文主题。

[0053] S103:在与所述主题相对应的索引结构的底层建立所述博文的索引;

[0054] 这里,每个主题都有一个独立的索引结构,索引结构由一系列大小不同的倒排索引组成,分层存储,每层存储的索引量都有限制,即每层的索引量不能超过给定的阈值,而且高一层的阈值是第一层阈值的倍数,本实施例中设为2倍,属于同一主题的博文的索引均位于与该主题对应的索引结构中。另外,每个主题的索引结构以分布式方式存储于不同的节点上,微博系统的内存上会保存主题与索引结构的映射关系,在步骤S102确定了获取的博文的主题之后,微博系统的主节点会根据主题与索引结构的映射关系,将博文发送到与存储其主题对应的索引结构的节点上进行处理,而该节点又包括第一节点和第二节点,第一节点主要负责建立博文的索引,其在索引结构的底层建立所述博文的索引。

[0055] S104:判断所述索引结构的底层的索引数是否大于或等于预设阈值;如果是,则执行步骤S105;

[0056] S105:将已存储在所述索引结构的底层的索引数据迁移至分布式存储网络中的第二节点。

[0057] 这里,当所述索引结构的低一层索引的容量已经达到该层容量阈值时,如果仍然有新的索引加入,此时就需要将低一层的索引数据合并到高一层的索引数据中。第二节点承担了索引结构底层之外的其他层索引数据的维护工作,当所述索引结构的底层的索引数大于或等于预设阈值时,第一节点会将已存储在所述索引结构的底层的索引数据迁移至第二节点中,当除底层外低一层索引的索引数大于或等于该层的预设阈值时,第二节点会将该层的索引数据复制到该层的上一层,实现索引数据的合并,其它层索引的维护工作以此类推。

[0058] 假设有一个主题的索引结构,用L来表示索引结构的索引层次,索引结构底层的容量用m来表示,则在该索引结构中第i层的容量为 $2im$,每个主题索引的底层用0层来表示,所有新加入系统的博文的索引都建立在0层,而其他层($L \geq 1$)通过合并低一层索引而形成。可见,属于该主题的最新的博文的索引存储于索引结构的底层,该层的索引量不大,因此能够在极小的更新代价下为获取到的最新博文建立索引,从而使新博文能够及时可搜索。

[0059] 接下来将用一个简单的实例来对索引结构的具体工作过程进行简单的介绍,如图2所示。假设索引结构的索引层次 $L=3$,高一层索引容量阈值为下一层阈值的 $t=2$ 倍,图2中的圆边矩阵就是底层索引容量m的大小。首先,在第一阶段Stage1,第一节点在底层L0层中创建一个索引文件I01,而随着新博文的加入索引将会逐渐增大直到其大小达到阈值m(在第i阶段Stage i)。此时再有新的博文加入则分别在底层L0层以及底层的上一层L1层中创建新的索引文件I02、I11,并将I01的索引数据合并到I11中,然后再将I01删除。直到第m阶段Stage m,I02的索引文件也达到阈值了,则在第m+1阶段Stage m+1中,L0层需要创建一个新的索引文件I03,并且还需要将I02中的数据合并到L1层的I11索引文件中。到第n阶段Stage

n时,I03也已经达到了阈值,此时第n+1阶段Stage n+1中需要在L0层中创建新的索引文件I04用于吸收新博文的索引;同时需要将I03的索引合并到L1层中,而此时L1层的I11容量也已经达到了阈值,则需要创建更高一层的索引,即需要在L1层的上一层L2层中创建I21索引文件,并将I11的文件合并到I21中;然后需要在L1中创建新的索引文件I12,并将I03的索引文件合并到I12中;最后再将I03以及I11的索引文件删除。最后整个索引结构中保存的索引文件只有I04、I12以及I21三个。

[0060] 另外,可选地,所述从微博系统中实时获取用户发布的博文之后,还包括对博文过滤过程,博文过滤的目的是把那些不希望处理的垃圾博文过滤掉,以便提高处理速度,具体地,首先确定发布所述博文的用户;然后判断所述用户是否为恶意用户,如果否,才执行确定所述博文的主题的步骤。

[0061] 上述博文过滤的原则是:将恶意用户发布的博文都定性为不希望处理的垃圾博文。因而,首先确定步骤S101中获取到的博文是哪个用户发布的,然后判断发布所述博文的用户是否为恶意用户,如果是,则所述博文为垃圾博文,不再对所述博文进行后续处理,如果否,对博文执行后续索引建立操作。实际应用中,可以将发送过违规言论的用户或者重复发送无意义信息的用户判定为恶意用户。

[0062] 本实施例公开的建立微博索引的方法,依据博文的主题,在与主题对应的索引结构中建立博文的索引,微博系统的内存中仅保存主题与索引结构的映射关系,映射关系的数据量相对博文索引量来说较小,各个主题的索引结构分布式存储于多个节点上,这样,属于不同主题的多个博文进入微博系统后可由多个节点同时处理,加快了索引建立速度;而且,所述博文的索引由所述索引结构的第一节点建立在所述索引结构的底层,当所述底层中索引数超过预设阈值时,将所述底层中索引数据移至所述索引结构的其他层,即交由所述索引结构的第二节点维护,也就是说,博文的索引在索引结构中分层存储,索引结构的底层存储的都是为最新进入微博系统的博文建立的索引,这样就不会出现第一节点的存储负荷过大的问题,保证了索引的快速建立。

[0063] 建立博文索引,形成这种与博文主题对应的分布式多层索引结构之后,利用该索引结构搜索博文的具体步骤参见图3,其示出了本发明实施例二公开的一种基于索引结构搜索博文的流程,该流程,具体包括:

[0064] S301:获取所述微博系统中的主节点下发的搜索任务;所述搜索任务是由所述主节点根据用户在搜索操作中给定的搜索关键字对应的主题所分配的;

[0065] 其中,微博系统的主节点在接收到用户的搜索操作后,推断用户搜索操作中给定的搜索关键字的主题,用户提供的搜索关键字一般都是很短的,因此如果只将其分类到一个特定的主题的话是不准确的,本实施例中使用传统的LDA模型作为关键字的主题分类模型,LDA分类模型会返回一个主题概率的向量,通过该主题概率向量就可以知道该关键字可能涉及到的主题,这样,一个搜索关键字至少与一个主题相对应。确认了关键字的主题之后,主节点会下发搜索任务到存储与主题相对应的索引结构的节点中,所述主节点下发的搜索任务由所述索引结构的第一节点获取,同时所述索引结构的第二节点也会获取到所述主节点下发的搜索任务。可见,搜索操作以分布式操作的形式来完成,由多个节点共同完成搜索请求。

[0066] S302:在所述索引结构的底层搜索与所述关键字相匹配的索引,得到搜索结果;

[0067] 其中,所述索引结构的第一节点在所述索引结构的底层进行搜索,所述索引结构的第二节点在所述索引结构的其它层进行搜索,这样,第二节点分担了第一节点的工作,底层索引数据量相对其他层来说较小,且存储的索引是最新博文的索引,第一节点只负责在底层进行搜索,搜索速度快,不会影响索引创建以及更新的效率。另外值得一提的是,本发明实施例中的索引结构是具有时间顺序的,存储在高层的索引比低层的索引的建立时间要早,并且索引结构的每层上记录有该层中索引建立的起始时间戳,这样,更有利于针对特定时间范围的查找。

[0068] S303:将所述搜索结果发送至所述主节点,以便所述主节点综合所述第一节点的搜索结果和所述第二节点的搜索结果得到所述主题的搜索结果,综合所有主题的搜索结果,得到最终搜索结果。

[0069] 这里,所述索引结构的第一节点和第二节点得到针对所述索引结构的搜索结构后,会根据排名原则对与该主题对应的搜索结果进行排序,然后将搜索结果发送至所述微博系统的主节点。所述微博系统的主节点首先会综合与所述主题对应的第一节点的搜索结果和第二节点的搜索结果,进而类似地,综合与用户搜索关键字对应的所有主题的搜索结果,得出最终完整的搜索结果。微博系统的主节点综合了所有主题的搜索结果之后,以层次结构的形式将搜索结果(也就是搜索到的博文)展现出来,尤其是对那些原本就属于同一微博树的博文,这样能够更加清晰的展现一些重要事件或者突发事件的演化和发展过程。

[0070] 可选地,所述综合所有主题的搜索结果之前还包括对所述所有主题的排序结果进行排序的步骤。

[0071] 本实施例中采用的排序算法考虑了博文的时间、用户的权威性和主题的受欢迎度,其排序表达式为:

$$[0072] \text{Rank}(d, q) = \omega_1 \cdot \text{sig}(d.\text{user}) + \omega_2 \cdot \text{sim}(d, q) + \omega_3 \cdot \text{fresh}(\text{ts}_d, \text{ts}_q)$$

[0073] 其中:

$$[0074] \omega_1 + \omega_2 + \omega_3 = 1 \text{ 且 } \omega_1, \omega_2, \omega_3 > 0;$$

[0075] $\text{sig}(d.\text{user})$ 表示发布博文d的用户的权威度;

[0076] $\text{sim}(d, q)$ 表示博文d与查询处理q的相似度;

[0077] $\text{fresh}(\text{ts}_d, \text{ts}_q)$ 表示基于博文d以及查询q的时间戳来判断d在查询q中的新鲜度。

[0078] 上述实施例中表明,本发明公开的建立微博索引的方法能够实时的对微博系统中的博文建立索引,通过推断用户给定的搜索关键字的主题,根据所建立的索引结构的特点,分布式的在多个节点上执行搜索任务,确保在搜索处理过程中快速、准确的将用户需要的数据返回。

[0079] 本发明实施例三公开了一种建立微博索引的装置,参见图4所示,所述装置应用于分布式存储网络中的第一节点,包括:

[0080] 博文获取模块401:用于从微博系统中实时获取用户发布的博文;

[0081] 主题确定模块402:用于当所述博文是对其他博文的转发和/或回复时,将所述博文的主题确定为所述博文所回复和/或所转发的原始博文的主题,当所述博文与其他博文无转发和/或回复关系时,使用Twitter-LDA模型确定所述博文的主题;

[0082] 索引建立模块403:用于在与所述主题相对应的索引结构的底层建立所述博文的索引;

[0083] 阈值判断模块404：用于判断所述索引结构的底层的索引数是否大于或等于预设阈值，如果是，触发索引迁移模块405；

[0084] 索引迁移模块405：用于将已存储在所述索引结构的底层的索引数据迁移至分布式存储网络中的第二节点。

[0085] 其中，可选地，所述博文获取模块从微博系统中实时获取用户发布的博文之后还包括博文过滤模块406：

[0086] 用户确定模块461，用于确定发布所述博文的用户；

[0087] 用户判断模块461，用于判断所述用户是否为恶意用户，如果否，触发主题确定模块402。

[0088] 本实施例公开的建立微博索引的装置，依据博文的主题，在与主题对应的索引结构中建立博文的索引，属于不同主题的多个博文进入微博系统后可由多个节点同时处理，加快了索引建立速度；而且，博文的索引在索引结构中分层存储，索引结构的底层存储的都是为最新进入微博系统的博文建立的索引，这样就不会出现存储设备负荷过大的问题，保证了索引的快速建立。

[0089] 本发明实施例四公开了一种建立微博索引的装置，参见图5所示，其示出了本发明实施例公开的一种基于索引结构搜索博文的装置结构示意图，具体地，建立微博索引的装置还包括：

[0090] 任务获取模块501，用于获取所述微博系统中的主节点下发的搜索任务；所述搜索任务是由所述主节点根据用户在搜索操作中给定的搜索关键字对应的主题所分配的；

[0091] 搜索执行模块502，用于在所述索引结构的底层搜索与所述关键字相匹配的索引，得到搜索结果；

[0092] 结果发送模块503，用于将所述搜索结果发送至所述主节点，以便所述主节点综合所述第一节点和所述第二节点的搜索结果得到所述主题的搜索结果，综合所有主题的搜索结果，得到最终搜索结果。

[0093] 其中，可选地，所述结果发送模块503综合所述所有主题的搜索结果之前还包括：

[0094] 排序模块504，用于对所述所有主题的搜索结果进行排序。

[0095] 上述实施例中表明，本发明公开的建立微博索引的装置能够实时的对微博系统中的博文建立索引，通过推断用户给定的搜索关键字的主题，根据所建立的索引结构的特点，分布式的在多个节点上执行搜索任务，确保在搜索处理过程中快速、准确的将用户需要的数据返回。

[0096] 最后，还需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，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、方法、物品或者设备不仅包括那些要素，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、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0097] 本说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似部分互相参见即可。

[0098] 对所公开的实施例的上述说明，使本领域专业技术人员能够实现或使用本发明。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下，在其它实施例中实现。因此，本发明将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

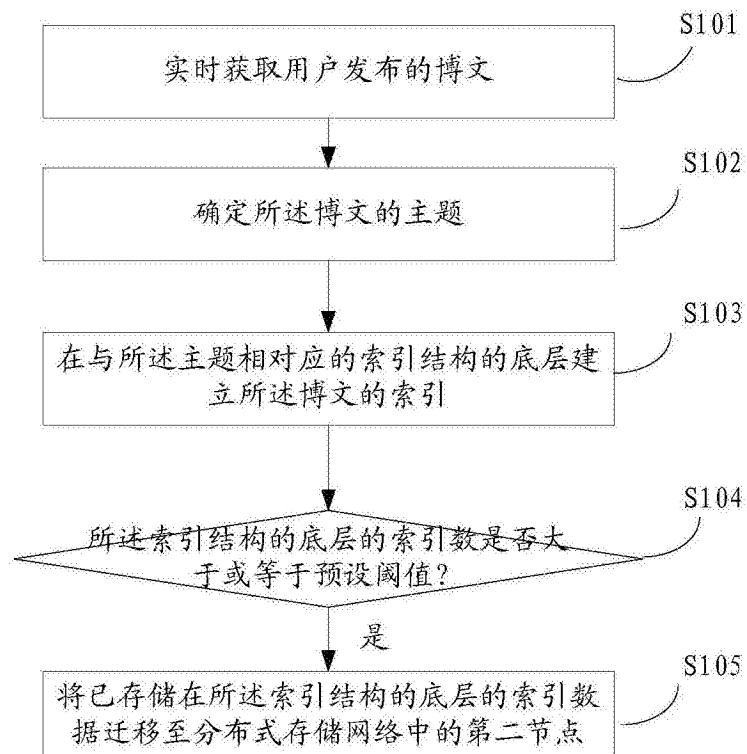


图1

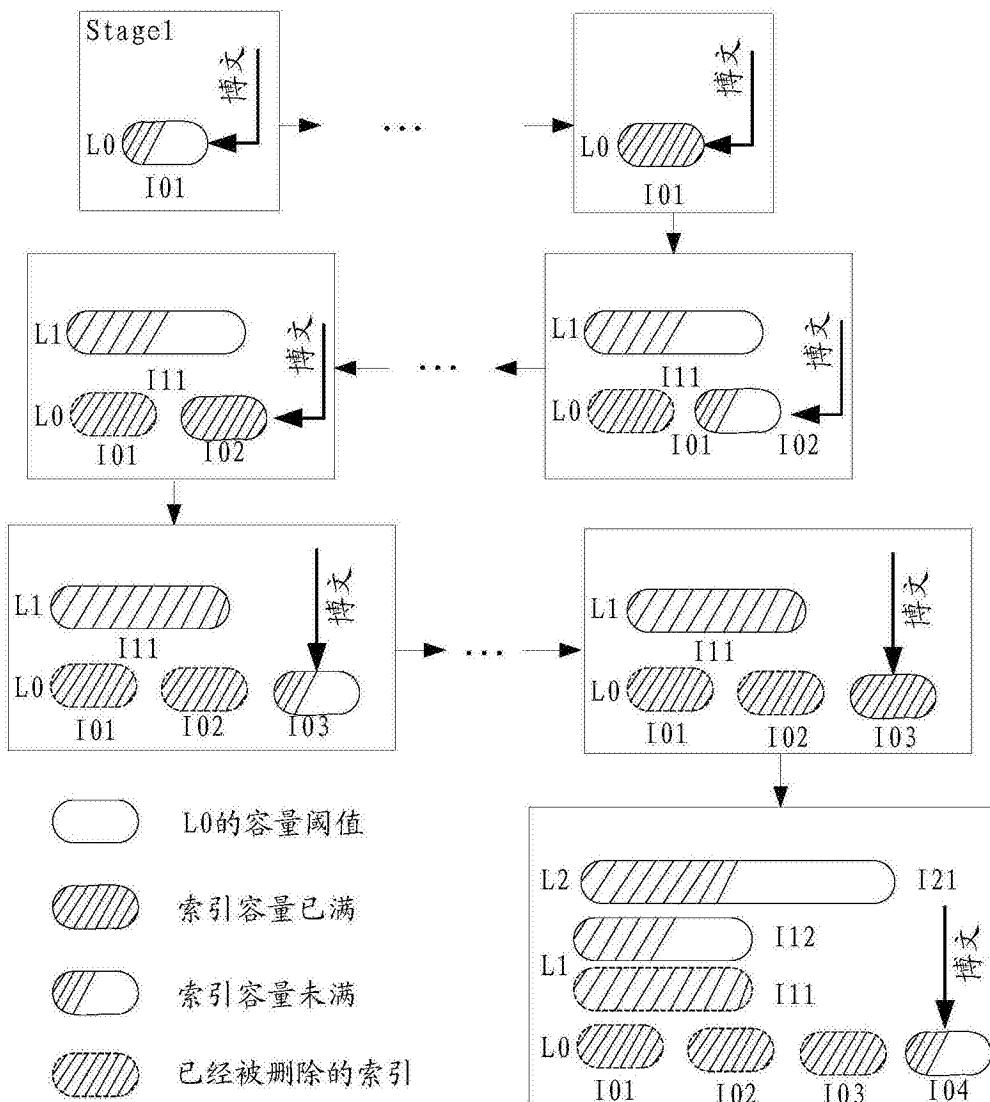


图2

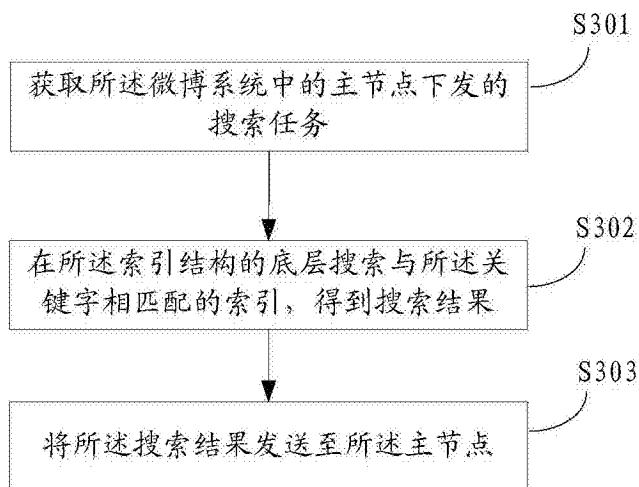


图3

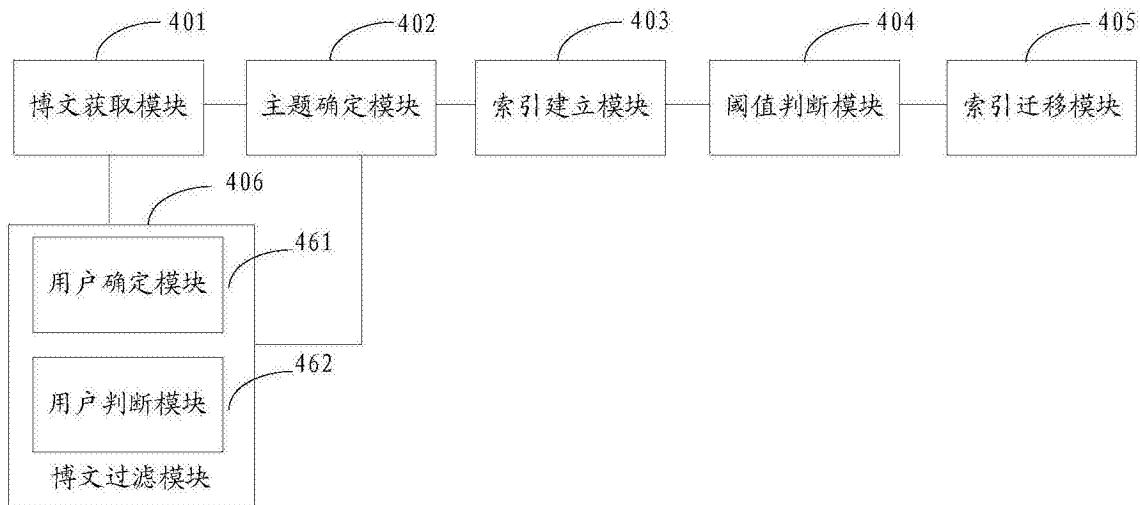


图4



图5