

一种智能热点打散的方法、装置、存储介质及计算机设备

本申请要求在 2019 年 02 月 15 日提交中国专利局、申请号为 201910117615.5、发明名称为“一种智能热点打散的方法、装置、存储介质及计算机设备”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。

技术领域

本发明涉及计算机网络负载均衡领域，尤其涉及一种智能热点打散的方法、装置、存储介质及计算机设备。

背景技术

为了减小服务器端的处理时延，改善用户体验，通常都需要对 URL (Uniform Resource Location) (特别是热点 URL) 进行打散/均衡操作。

现有技术通常采用将集中的 URL 随机或均匀地分散到后端机器（例如，源站节点服务器集群、边缘节点服务器集群）的方式，来实现针对热点 URL 的负载均衡。

然而，现有技术方案通常只是在热点 URL 出现之后，才执行打散操作，从而将针对热点 URL 的请求分散地分配到多台机器，因此存在以下缺点：

1、没有在前期针对热点 URL 可能存在的趋势做出判断，做出预先处理。只是在发现热点之后直接打散，没有缓存的机器会直接回上层节点或者源造成大量流量浪费。

2、即使采用组内共享缓存方案，将所有回源流量重新导回统一端口，在热点文件较大（例如，文件大小超过 1G）时，在各个机器取完文件之前，热点堆积现象将无法解决，该统一端口的压力并没有减少。

为了解决上述问题，需要提出新的技术方案。

发明内容

根据本发明的智能热点打散的方法，包括：

基于人工智能学习模型学习 URL 的请求量曲线并预测 URL 的请求量；

将 URL 的预测请求量大于或等于该 URL 所对应的第一预定请求量阈值的 URL 确定为第一 URL，对 URL 执行第一热点打散操作；

将 URL 的请求量无法预测且实际请求量大于或等于该 URL 所对应的第二预定请求量阈值的 URL 确定为第二 URL，对 URL 执行第二热点打散操作。

根据本发明的智能热点打散的方法，其对 URL 执行第一热点打散操作的步骤包括：

将第一 URL 的请求随机或均匀地分配到多个缓存服务器和/或源站服务器的进程。

根据本发明的智能热点打散的方法，其对 URL 执行第二热点打散操作的步骤包括：

查找第二 URL 的落点进程；

将第一 URL 的新请求随机或均匀地分配到多个缓存服务器和/或源站服务器的进程中的、除了落点进程之外的进程；和/或

将第一 URL 的旧请求随机或均匀地重新分配到多个缓存服务器和/或源站服务器的进程中的、除了落点进程之外的进程。

根据本发明的智能热点打散的方法，其对 URL 执行第二热点打散操作的步骤还包括：

确定第二 URL 的请求文件的大小大于指定文件大小；和/或

在落点进程执行完针对第二 URL 的请求的、获取并发送请求文件或发送请求文件的响应之后，将第一 URL 的新请求随机或均匀地分配到多个缓存服务器和/或源站服务器的进程，和/或，将第一 URL 的旧请求随机或均匀地重新分配到多个缓存服务器和/或源站服务器的进程，和/或

基于人工智能学习模型学习 URL 的请求量曲线并预测 URL 的请求量的步骤包括：

采用聚类算法对 URL 进行聚类分析，将相同类别的 URL 在预设历史时段内的请求量或实时请求量自动绘制为对应于该类别的 URL 的请求量曲线，根据该类别的 URL 的请求量曲线来预测该类别的 URL 的请求量。

根据本发明的智能热点打散的装置，包括：

人工智能学习模型，用于学习 URL 的请求量曲线并预测 URL 的请求量；

第一热点打散模块，用于将 URL 的预测请求量大于或等于该 URL 所对应的第一预定请求量阈值的 URL 确定为第一 URL，对 URL 执行第一热点打散操作；

第二热点打散模块，用于将 URL 的请求量无法预测且实际请求量大于或等于该 URL 所对应的第二预定请求量阈值的 URL 确定为第二 URL，对 URL 执行第二热点打散操作。

根据本发明的智能热点打散的装置，其第一热点打散模块还用于：

将第一 URL 的请求随机或均匀地分配到多个缓存服务器和/或源站服务器的进程。

根据本发明的智能热点打散的装置，其第二热点打散模块还用于：

查找第二 URL 的落点进程；

将第一 URL 的新请求随机或均匀地分配到多个缓存服务器和/或源站服务器的进程中的、除了落点进程之外的进程；和/或

将第一 URL 的旧请求随机或均匀地重新分配到多个缓存服务器和/或源站服务器的进程中的、除了落点进程之外的进程，和/或

人工智能学习模型还用于：

采用聚类算法对 URL 进行聚类分析，将相同类别的 URL 在预设历史时段内的请求量或实时请求量自动绘制为对应于该类别的 URL 的请求量曲线，根据该类别的 URL 的请求量曲线来预测该类别的 URL 的请求量。

根据本发明的智能热点打散的装置，其第二热点打散模块还用于：

确定第二 URL 的请求文件的大小大于指定文件大小；和/或

在落点进程执行完针对第二 URL 的请求的、获取并发送请求文件或发送请求文件的响应之后，将第一 URL 的新请求随机或均匀地分配到多个缓存服务器和/或源站服务器的进程，和/或，将第一 URL 的旧请求随机或均匀地重新分配到多个缓存服务器和/或源站服务器的进程。

根据本发明的存储介质，该存储介质上存储有计算机程序，程序被处理器执行时实现上文所述方法的步骤。

根据本发明的计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，处理器执行程序时实现上文所述方法的步骤。

根据本发明的上述技术方案，能够自动预测非突发的热点 URL 请求、从而提前进行第一种打散操作，能够对不能预测的突发热点 URL 请求进行第二种打散操作，加快了热点业务的处理。

附图说明

此处所说明的附图用来提供对本发明实施例的进一步理解，构成本申请的一部分，本发明实施例的示意性实施例及其说明用于解释本发明实施例，并不构成对本发明实施例的不当限定。在附图中：

图 1 示例性地示出了根据本发明的智能热点打散的方法的示意流程图。

图 2 示例性地示出了根据本发明的智能热点打散的装置的示意框图。

图 3 示例性地示出了可以实现根据本发明的智能热点打散的方法的一个实施例的示意图。

图 4 示例性地示出了根据本发明的智能热点打散的方法可以使用的人工智能学习模型所匹配出来的模型图的示例。

具体实施方式

现结合附图和具体实施方式对本发明实施例进一步说明。

为使本发明实施例的目的、技术方案和优点更加清楚，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，所描述的实施例是本发明的一部分实施例，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做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。需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互任意组合。

为了解决背景技术部分所述的技术问题，本发明基于针对在（可预测的）热点还未突发之前提前预测（即，预测请求量大于等于指定 URL 的预定请求量时）并执行第一种处理操作、在（不可预测的）热点突发时（即，请求量无法预测且实际请求量大于或等于指定 URL 的预定请求量时）执行第二种处理（例

如，快速处理）操作的总体构思，提出了以下技术方案。

图 1 示例性地示出了根据本发明的智能热点打散的方法的示意图。

如图 1 所示，根据本发明的智能热点打散的方法，包括：

步骤 S102：基于人工智能学习模型学习 URL 的请求量曲线并预测 URL 的请求量；

步骤 S104：将 URL 的预测请求量大于或等于该 URL 所对应的第一预定请求量阈值的 URL 确定为第一 URL，对 URL 执行第一热点打散操作；

步骤 S106：将 URL 的请求量无法预测且实际请求量大于或等于该 URL 所对应的第二预定请求量阈值的 URL 确定为第二 URL，对 URL 执行第二热点打散操作。

可选地，在步骤 S104 中，对 URL 执行第一热点打散操作的步骤包括：

将第一 URL 的请求随机或均匀地分配到多个缓存服务器和/或源站服务器的进程。

可选地，在步骤 S106 中，对 URL 执行第二热点打散操作的步骤包括：

查找第二 URL 的落点进程；

将第一 URL 的新请求随机或均匀地分配到多个缓存服务器和/或源站服务器的进程中的、除了落点进程之外的进程；和/或

将第一 URL 的旧请求随机或均匀地重新分配到多个缓存服务器和/或源站服务器的进程中的、除了落点进程之外的进程。

可选地，在步骤 S106 中，对 URL 执行第二热点打散操作的步骤还包括：

确定第二 URL 的请求文件的大小大于指定文件大小；和/或

在落点进程执行完针对第二 URL 的请求的、获取并发送请求文件或发送请求文件的响应之后，将第一 URL 的新请求随机或均匀地分配到多个缓存服务器和/或源站服务器的进程，和/或，将第一 URL 的旧请求随机或均匀地重新分配到多个缓存服务器和/或源站服务器的进程，和/或

步骤 S102 包括：

采用聚类算法对 URL 进行聚类分析，将相同类别的 URL 在预设历史时段内的请求量或实时请求量自动绘制为对应于该类别的 URL 的请求量曲线，根据

该类别的 URL 的请求量曲线来预测该类别的 URL 的请求量。

可选地，请求量曲线采用三阶段模型曲线。

图 2 示例性地示出了根据本发明的智能热点打散的装置的示意框图。

如图 2 所示，根据本发明的智能热点打散的装置 200 包括：

人工智能学习模型 201，用于学习 URL 的请求量曲线并预测 URL 的请求量；

第一热点打散模块 203，用于将 URL 的预测请求量大于或等于该 URL 所对应的第一预定请求量阈值的 URL 确定为第一 URL，对 URL 执行第一热点打散操作；

第二热点打散模块 205，用于将 URL 的请求量无法预测且实际请求量大于或等于该 URL 所对应的第二预定请求量阈值的 URL 确定为第二 URL，对 URL 执行第二热点打散操作。

可选地，第一热点打散模块 203 还用于：

将第一 URL 的请求随机或均匀地分配到多个缓存服务器和/或源站服务器的进程。

可选地，第二热点打散模块 205 还用于：

查找第二 URL 的落点进程；

将第一 URL 的新请求随机或均匀地分配到多个缓存服务器和/或源站服务器的进程中的、除了落点进程之外的进程；和/或

将第一 URL 的旧请求随机或均匀地重新分配到多个缓存服务器和/或源站服务器的进程中的、除了落点进程之外的进程。

可选地，第二热点打散模块 205 还用于：

确定第二 URL 的请求文件的大小大于指定文件大小；和/或

在落点进程执行完针对第二 URL 的请求的、获取并发送请求文件或发送请求文件的响应之后，将第一 URL 的新请求随机或均匀地分配到多个缓存服务器和/或源站服务器的进程，和/或，将第一 URL 的旧请求随机或均匀地重新分配到多个缓存服务器和/或源站服务器的进程，和/或

人工智能学习模型 201 还用于：

采用聚类算法对 URL 进行聚类分析，将相同类别的 URL 在预设历史时段

内的请求量或实时请求量自动绘制为对应于该类别的 URL 的请求量曲线，根据该类别的 URL 的请求量曲线来预测该类别的 URL 的请求量。

结合根据本发明的上述方法和装置，还提出了一种存储介质，存储介质上存储有计算机程序，程序被处理器执行时实现上文所述方法的步骤。

结合根据本发明的上述方法和装置，还提出了一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，处理器执行程序时实现上文所述方法的步骤。

为了使本领域技术人员更清楚地理解根据本发明的上述技术方案，下文将结合具体实施例进行描述。

图 3 示例性地示出了可以实现根据本发明的智能热点打散的方法的一个实施例的示意图。

如图 3 所示，该实施例（“热点文件分析预取打散”，对应于结合图 1 所述的上述智能热点打散的方法）包括以下处理步骤：

1、（可预测的）“热点未突发时”（对应于上述步骤 S104）

1.1、人工智能学习模型预测可能的热点 URL（对应于上述步骤 S102）。

例如，热点 URL 预测主要核心可以为人工智能匹配出来的模型图（对应于上述请求量曲线），针对不同业务（例如，对应于不同的 URL）计算出参数（例如，上述 URL 所对应的第一预定请求量阈值）。

图 4 示例性地示出了根据本发明的智能热点打散的方法可以使用的人工智能学习模型所匹配出来的模型图的示例。

例如，实现了使用该模型图的上述智能热点打散的方法的存储介质，可以存储 apk 等形式的软件安装包，以便于产品的发布。

如图 4 所示的示例模型图所示，其包含 A、B、C 三个阶段（即，上述三个阶段模型曲线）。在阶段 A，业务还未突发，刚开始上量时请求较小。在阶段 B，请求数升高，达到较高的业务值。在阶段 C，业务突发，请求量在瞬间达到高点。

例如，模型会在 A->B 过程中预测出可能会有 C 突发的趋势。例如，上述 URL 所对应的第一预定请求量阈值可以被设置为阶段 B 和阶段 C 的分界点 T 所对应的纵坐标的数值。

可选地,可以将决策出的可能 URL(即,上述预测请求量大于或等于该 URL 所对应的第一预定请求量阈值的 URL)提交给组(例如,多个缓存服务器组成的集群)内的预取中心(即,服务器集群中指定的、作为统一端口、执行预取操作的缓存服务器)。

1.2、提交 URL 到组内预取中心、将 URL 预取到组内指定进程。

例如,组内预取中心可以执行以下操作:

1.2.1、将获取的 URL 通过计算获取到 URL 的上层节点,父或者同节点省份的其他机器,此时尝试对父和源发起探测,计算两边的下载速度哪个更快,作为挑选出来的上层节点。

1.2.2、挑选整组机器中空闲的进程,作为要预取到的那个进程 a。让进程 a 执行 http 的请求将文件取下,作为缓存基础文件 afile(本地文件),待文件取完,执行下一步。

1.2.3、计算出打散后该 URL 即将落在机器的进程号 pid,预取程序告知这些进程都到之前的进程 a 上拿文件 afile,从而保证只有一次回上层节点(即,父节点)或源节点的请求。

1.3、将 URL 打散。

例如,可以按照打散规则(例如,随机或均匀地等),将该 URL 在调度服务器上做打散操作,在热点还未突发时将其快速解决。

然而,有些热点是前期没有任何预兆(即,其请求量是不可预测的),此时需要采用下文所述的方法。

2、(不可预测的)“热点突发”时(对应于上述步骤 S106)

2.1、热点打散与组内共享缓存。

例如,热点突发时,服务器上的热点检测装置(例如,可以对应于上述第二热点打散模块 205)会识别出突发的热点将其打散。例如,在没有缓存的服务器上,还可以将请求导向原先 hash 的组内端口上,从而实现组内共享缓存。

然而,当文件超过 1G(对应于上述指定文件大小)时,请求量大的请求下,此时组内其他服务器无法快速取完 URL,会造成请求堆积,没有达到打散的效果,不能快速解决问题。此时,可以执行以下操作,以减小处理时延,改善用

户体验。

2.2、查找热点 URL。

2.3、计算 URL 落点进程，对其他 URL 做进程屏蔽。

例如，可以计算该热点 URL 的落点进程 b，将其他业务的落点为该进程的请求重新做 hash，将业务导向其他空闲进程，保护其他业务不受影响，此时热点 URL 还是正常运作在该繁忙进程中。

可选地，针对下载类型业务或其他支持断点续传的业务，可以强行 kill 掉保留在屏蔽进程 b 中的链接，让链接重新接入服务器时可以享受预取，且在另行打散后自动实现业务的流畅运行。

即，此时能够通过进程隔离，在单个业务突发且无法瞬间恢复时，将其他业务先迁移出执行突发业务的进程，保证其他业务的正常运行。

2.4、将 URL 提交组内预取中心，将 URL 预取到组内指定进程。

该步骤包含步骤 1.2 中的相同操作。

2.5、恢复屏蔽的进程状态。

在进程 b 的预取操作完成之后，恢复屏蔽的进程 b 的（可被用于分配和执行其他业务的）状态，从而恢复正常业务。

根据本发明的上述技术方案，具有以下优点：

1、能够自动预测非突发的热点 URL 请求，从而提前进行第一种打散操作（包括预取到缓存），在热点还没造成实质的业务影响（例如，大的时延）时解决问题，能够对不能预测的突发热点 URL 请求进行第二种打散操作，加快了热点业务的处理。

2、例如，可以结合将所有回源流量重新导回统一端口的操作，减少多份回上层（即，上层的源站节点）取资源的流量，节省成本，提高边缘节点与上层节点的承载能力。还可以结合挑选较为空闲的链路和较为空闲的进程执行一次快速回源预取、再将文件快速复制到该统一端口对应的、通过内网连接的其他机器上的操作，节省多份回上层节点的回源流量。

3、能够在热点请求处理完成之后，为原本用于热点请求处理的进程分配其他请求，尽快修复热点产生的影响。

4、能够通过进程隔离，在单个业务突发且无法瞬间恢复时，将其他业务先迁移出执行突发业务的进程，保证其他业务的正常运行。

上面描述的内容可以单独地或者以各种方式组合起来实施，而这些变型方式都在本发明的保护范围之内。

本领域普通技术人员可以理解，上文中所公开方法中的全部或某些步骤、系统、装置中的功能模块/单元可以被实施为软件、固件、硬件及其适当的组合。在硬件实施方式中，在以上描述中提及的功能模块/单元之间的划分不一定对应于物理组件的划分；例如，一个物理组件可以具有多个功能，或者一个功能或步骤可以由若干物理组件合作执行。某些组件或所有组件可以被实施为由处理器，如数字信号处理器或微处理器执行的软件，或者被实施为硬件，或者被实施为集成电路，如专用集成电路。这样的软件可以分布在计算机可读介质上，计算机可读介质可以包括计算机存储介质（或非暂时性介质）和通信介质（或暂时性介质）。如本领域普通技术人员公知的，术语计算机存储介质包括在用于存储信息（诸如计算机可读指令、数据结构、程序模块或其他数据）的任何方法或技术中实施的易失性和非易失性、可移除和不可移除介质。计算机存储介质包括但不限于 RAM、ROM、EEPROM、闪存或其他存储器技术、CD-ROM、数字多功能盘（DVD）或其他光盘存储、磁盒、磁带、磁盘存储或其他磁存储装置、或者可以用于存储期望的信息并且可以被计算机访问的任何其他的介质。此外，本领域普通技术人员公知的是，通信介质通常包含计算机可读指令、数据结构、程序模块或者诸如载波或其他传输机制之类的调制数据信号中的其他数据，并且可包括任何信息递送介质。

最后应说明的是：以上实施例仅用以说明本发明的技术方案，而非对其限制。尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例的技术方案的精神和范围。

工业实用性

本发明的实施例提供了一种智能热点打散的方法、装置。可以基于人工智能学习模型学习 URL 的请求量曲线并预测 URL 的请求量；将 URL 的预测请求

量大于或等于该 URL 所对应的第一预定请求量阈值的 URL 确定为第一 URL，对 URL 执行第一热点打散操作；将 URL 的请求量无法预测且实际请求量大于或等于该 URL 所对应的第二预定请求量阈值的 URL 确定为第二 URL，对 URL 执行第二热点打散操作。能够自动预测非突发的热点 URL 请求、从而提前进行第一种打散操作，能够对不能预测的突发热点 URL 请求进行第二种打散操作，加快了热点业务的处理。

权利要求

1. 一种智能热点打散的方法，包括：

基于人工智能学习模型学习 URL 的请求量曲线并预测 URL 的请求量；

将 URL 的预测请求量大于或等于该 URL 所对应的第一预定请求量阈值的 URL 确定为第一 URL，对 URL 执行第一热点打散操作；

将 URL 的请求量无法预测且实际请求量大于或等于该 URL 所对应的第二预定请求量阈值的 URL 确定为第二 URL，对 URL 执行第二热点打散操作。

2. 如权利要求 1 所述的智能热点打散的方法，其中，所述对 URL 执行第一热点打散操作的步骤包括：

将所述第一 URL 的请求随机或均匀地分配到多个缓存服务器和/或源站服务器的进程。

3. 如权利要求 1 所述的智能热点打散的方法，其中，所述对 URL 执行第二热点打散操作的步骤包括：

查找所述第二 URL 的落点进程；

将所述第一 URL 的新请求随机或均匀地分配到多个缓存服务器和/或源站服务器的进程中的、除了所述落点进程之外的进程；和/或

将所述第一 URL 的旧请求随机或均匀地重新分配到多个缓存服务器和/或源站服务器的进程中的、除了所述落点进程之外的进程。

4. 如权利要求 3 所述的智能热点打散的方法，其中，所述对 URL 执行第二热点打散操作的步骤还包括：

确定所述第二 URL 的请求文件的大小大于指定文件大小；和/或

在所述落点进程执行完针对所述第二 URL 的请求的、获取并发送所述请求文件或发送所述请求文件的响应之后，将所述第一 URL 的新请求随机或均匀地分配到多个缓存服务器和/或源站服务器的进程，和/或，将所述第一 URL 的旧请求随机或均匀地重新分配到多个缓存服务器和/或源站服务器的进程，和/或

所述基于人工智能学习模型学习 URL 的请求量曲线并预测 URL 的请求量

的步骤包括：

采用聚类算法对 URL 进行聚类分析，将相同类别的 URL 在预设历史时段内的请求量或实时请求量自动绘制为对应于该类别的 URL 的请求量曲线，根据该类别的 URL 的请求量曲线来预测该类别的 URL 的请求量。

5. 一种智能热点打散的装置，包括：

人工智能学习模型，用于学习 URL 的请求量曲线并预测 URL 的请求量；

第一热点打散模块，用于将 URL 的预测请求量大于或等于该 URL 所对应的第一预定请求量阈值的 URL 确定为第一 URL，对 URL 执行第一热点打散操作；

第二热点打散模块，用于将 URL 的请求量无法预测且实际请求量大于或等于该 URL 所对应的第二预定请求量阈值的 URL 确定为第二 URL，对 URL 执行第二热点打散操作。

6. 如权利要求 5 所述的智能热点打散的装置，其中，所述第一热点打散模块还用于：

将所述第一 URL 的请求随机或均匀地分配到多个缓存服务器和/或源站服务器的进程。

7. 如权利要求 5 所述的智能热点打散的装置，其中，所述第二热点打散模块还用于：

查找所述第二 URL 的落点进程；

将所述第一 URL 的新请求随机或均匀地分配到多个缓存服务器和/或源站服务器的进程中的、除了所述落点进程之外的进程；和/或

将所述第一 URL 的旧请求随机或均匀地重新分配到多个缓存服务器和/或源站服务器的进程中的、除了所述落点进程之外的进程，和/或

所述人工智能学习模型还用于：

采用聚类算法对 URL 进行聚类分析，将相同类别的 URL 在预设历史时段内的请求量或实时请求量自动绘制为对应于该类别的 URL 的请求量曲线，根据该类别的 URL 的请求量曲线来预测该类别的 URL 的请求量。

8. 如权利要求 7 所述的智能热点打散的装置，其中，所述第二热点打散模

块还用于：

确定所述第二 URL 的请求文件的大小大于指定文件大小；和/或

在所述落点进程执行完针对所述第二 URL 的请求的、获取并发送所述请求文件或发送所述请求文件的响应之后，将所述第一 URL 的新请求随机或均匀地分配到多个缓存服务器和/或源站服务器的进程，和/或，将所述第一 URL 的旧请求随机或均匀地重新分配到多个缓存服务器和/或源站服务器的进程。

9. 一种存储介质，所述存储介质上存储有计算机程序，所述程序被处理器执行时实现权利要求 1 至 4 中任意一项所述方法的步骤。

10. 一种计算机设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述程序时实现权利要求 1 至 4 中任意一项所述方法的步骤。