(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 104375826 A
(43)申请公布日 2015.02.25

(21)申请号 201410535111.2

(22)申请日 2014.10.11

(71)申请人 北京中搜网络技术股份有限公司
地址 100191 北京市海淀区学院路51号首亨科技大厦0902室

(72)发明人 王鹏

(74)专利代理机构 北京安博达知识产权代理有限公司 11271
代理人 徐国文

(51)Int.Cl.
G06F 9/44(2006.01)
G06F 17/30(2006.01)

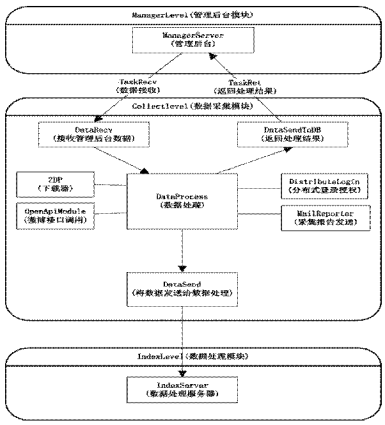
权利要求书2页 说明书5页 附图2页

(54)发明名称

一种高可用的微博采集平台及其方法

(57)摘要

本发明提供一种高可用的微博采集平台及其方法，所述平台包括采集系统、管理后台和数据处理系统；所述管理后台、采集系统和数据处理系统依次连接；所述采集系统包括命令交互模块、处理模块、数据发送模块、ZDP调用模块、OpenAPI调用模块。所述方法包括：(1)启动管理后台；(2)执行管理后台指令并进行数据采集；(3)马甲分布式登录和邮件发送。本发明认证模式更加高效，程序自动控制api的调用方法和调用次数，实现了大规模数据抽取；管理后台对抽取的数据进行人工干预，实现了高效的数据管理；用户只需要输入马甲、博主、应用，即可抽取数据，自动化程度极高。



CN 104375826 A

1. 一种高可用的微博采集平台，其特征在于，所述平台包括采集系统、管理后台和数据处理系统；

所述管理后台、采集系统和数据处理系统依次连接；

所述采集系统包括命令交互模块、处理模块、数据发送模块、ZDP 调用模块、OpenAPI 调用模块。

2. 如权利要求 1 所述的一种高可用的微博采集平台，其特征在于，所述平台包括分布式登陆模块，通过 Gearman 对马甲进行多机分布式验证。

3. 如权利要求 1 所述的一种高可用的微博采集平台，其特征在于，所述平台包括邮件发送模块，用于将日志统计信息发送到相关邮件组。

4. 一种高可用的微博采集方法，其特征在于，所述方法包括：

(1) 启动管理后台；

(2) 执行管理后台指令并进行数据采集；

(3) 马甲分布式登录和邮件发送。

5. 如权利要求 4 所述的一种高可用的微博采集方法，其特征在于，所述步骤 (1) 包括

(1.1) 对博主数据、应用数据、马甲数据分别进行增加、删除、修改和查询；

(1.2) 马甲和应用进行关联；

(1.3) 马甲登录验证；

(1.4) 管理后台将上述操作拼接成指令形式发至采集后台。

6. 如权利要求 4 所述的一种高可用的微博采集方法，其特征在于，所述步骤 (2) 包括

(2.1) 采集后台以指令的形式接到博主、马甲和应用的相关操作，并将博主、马甲和应用的操作结果写入本地数据文件；

(2.2) 采用马甲关注博主，并将数据更新到马甲文件中；

(2.3) 授权流程读取马甲本地文件，调用登录模块对马甲进行登录认证；

(2.4) 启动微博博文、博主信息和话题获取流程形成一个待下载的 URL；

(2.5) 将 URL 作为一个下载任务，提交至下载器，等待返回结果；

(2.6) 读取返回结果数据，并将数据按照类型存到对应的类对象中；

(2.7) 将解析出来的博主信息、博文和话题数据发送给数据处理。

7. 如权利要求6所述的一种高可用的微博采集方法，其特征在于，所述步骤(2.3)包括将认证参数一并写入马甲本地文件，用于授权后的应用调用微博开放 API。

8. 如权利要求 6 所述的一种高可用的微博采集方法，其特征在于，所述步骤 (2.4) 包括，获取一个用于博文、博主信息和话题下载的马甲，判断马甲的调度周期，按照微博开放 API 的要求，拼接请求参数，参数从马甲文件中读取，API 加请求参数形成一个待下载的 URL。

9. 如权利要求6所述的一种高可用的微博采集方法，其特征在于，所述步骤(2.6)包括得到的返回结果是 json 格式的，将 json 数据加载到 json 容器中，从 json 容器中按字段读取数据，并将数据按照类型存到对应的类对象中。

10. 如权利要求 4 所述的一种高可用的微博采集方法，其特征在于，所述步骤 (3) 马甲分布式登录包括采用多机登陆，登录任务分配采用 Gearman。

11. 如权利要求 4 所述的一种高可用的微博采集方法，其特征在于，所述步骤 (3) 邮件

发送包括

(3.1) 对采集系统采集数据的日志进行统计；

(3.2) 对四大微博媒体网站的数据下载次数，下载失败次数，解析成功次数进行计算；

(3.3) 生成采集系统的数据采集量报告；

(3.4) 启动邮件发送程序，将报告发送给负责人。

一种高可用的微博采集平台及其方法

技术领域

[0001] 本发明属于一种微博采集系统，具体讲涉及一种高可用的微博采集平台及其方法。

背景技术

[0002] 微博作为新生网络应用形式，在最近几年得到了迅猛的发展，随着微博用户群体的增长，微博数据的获取在微博搜索领域扮演着至关重要的角色。

[0003] 目前微博网页抽取方式多样，主要分成两类：基于微博页面解析的数据获取方法和基于微博 api 数据获取方法。

[0004] 基于微博页面解析的数据获取方法：这种方法主要是通过网络爬虫实现的，程序按照模板要求将网页内容以文本文件的形式保存在本地存储系统中，直到爬行完毕或者满足既定条件后终止。

[0005] 基于微博 api 数据获取方法：这种方法主要是通过微博开放平台提供的接口，然后对得到的数据按照格式要求进行解析。

[0006] 传统基于微博页面解析的数据获取方法，需要人工编写模板，如果模板有变化，维护成本比较高，且抽取得到的数据多种类型掺杂在一起，数据不够简洁，需要再写程序加以区分，效率比较低。

[0007] 基于微博 api 数据获取方法，首先要解决的是用户认证的问题，并且四大微博媒体网站认证方法各不相同，这些不利于大规模数据抽取。

发明内容

[0008] 针对现有技术的不足，本发明提出一种高可用的微博采集平台及其方法，对微博用户自动授权的机制，并对四大微博媒体网站的认证方法进行了规整，针对基于微博页面解析获取数据方法的缺陷，采用基于微博 api 数据获取方法，程序逻辑控制 api 调用方法和频率，获取 json 对象并解析实现数据高效获取。

[0009] 本发明的目的是采用下述技术方案实现的：

[0010] 一种高可用的微博采集平台，其改进之处在于，所述平台包括采集系统、管理后台和数据处理系统；

[0011] 所述管理后台、采集系统和数据处理系统依次连接；

[0012] 所述采集系统包括命令交互模块、处理模块、数据发送模块、ZDP 调用模块、OpenAPI 调用模块。

[0013] 优选的，所述平台包括分布式登陆模块，通过 Gearman 对马甲进行多机分布式验证。

[0014] 优选的，所述平台包括邮件发送模块，用于将日志统计信息发送到相关邮件组。

[0015] 本发明基于另一目的提供的一种高可用的微博采集方法，其改进之处在于，所述方法包括：

[0016] (1) 启动管理后台；

[0017] (2) 执行管理后台指令并进行数据采集；

[0018] (3) 马甲分布式登录和邮件发送。

[0019] 优选的，所述步骤 (1) 包括

[0020] (1.1) 对博主数据、应用数据、马甲数据分别进行增加、删除、修改和查询；

[0021] (1.2) 马甲和应用进行关联；

[0022] (1.3) 马甲登录验证；

[0023] (1.4) 管理后台将上述操作拼接成指令形式发至采集后台。

[0024] 优选的，所述步骤 (2) 包括

[0025] (2.1) 采集后台以指令的形式接到博主、马甲和应用的相关操作，并将博主、马甲和应用的操作结果写入本地数据文件；

[0026] (2.2) 采用马甲关注博主，并将数据更新到马甲文件中；

[0027] (2.3) 授权流程读取马甲本地文件，调用登录模块对马甲进行登录认证；

[0028] (2.4) 启动微博博文、博主信息和话题获取流程形成一个待下载的 URL；

[0029] (2.5) 将 URL 作为一个下载任务，提交至下载器，等待返回结果；

[0030] (2.6) 读取返回结果数据，并将数据按照类型存到对应的类对象中；

[0031] (2.7) 将解析出来的博主信息、博文和话题数据发送给数据处理。

[0032] 进一步地，所述步骤 (2.3) 包括将认证参数一并写入马甲本地文件，用于授权后的应用调用微博开放 API。

[0033] 进一步地，所述步骤 (2.4) 包括，获取一个用于博文、博主信息和话题下载的马甲，判断马甲的调度周期，按照微博开放 API 的要求，拼接请求参数，参数从马甲文件中读取，API 加请求参数形成一个待下载的 URL。

[0034] 进一步地，所述步骤 (2.6) 包括得到的返回结果是 json 格式的，将 json 数据加载到 json 容器中，从 json 容器中按字段读取数据，并将数据按照类型存到对应的类对象中。

[0035] 优选的，所述步骤 (3) 马甲分布式登录包括采用多机登陆，登录任务分配采用 Gearman。

[0036] 优选的，所述步骤 (3) 邮件发送包括

[0037] (3.1) 对采集系统采集数据的日志进行统计；

[0038] (3.2) 对四大微博媒体网站的数据下载次数，下载失败次数，解析成功次数进行计算；

[0039] (3.3) 生成采集系统的数据采集量报告；

[0040] (3.4) 启动邮件发送程序，将报告发送给负责人。

[0041] 与现有技术比，本发明的有益效果为：

[0042] (1) 四大微博媒体认证虽然各不相同，此程序对此进行了规整，认证模式更加高效。

[0043] (2) 程序自动控制 api 的调用方法和调用次数，实现了大规模数据抽取。

[0044] (3) 管理后台对抽取的数据进行人工干预，实现了高效的数据管理。

[0045] (4) 用户只需要输入马甲、博主、应用，即可抽取数据，自动化程度极高。

附图说明

[0046] 图 1 为本发明提供的一种高可用的微博采集平台结构图。

[0047] 图 2 为本发明提供的用户（马甲）分布式登录图。

具体实施方式

[0048] 下面结合附图对本发明的具体实施方式作进一步的详细说明。

[0049] 如图 1 所示，本发明一种高可用的微博采集平台包括：采集系统和与之交互的管理后台、数据处理系统。按交互流程采集系统可分为命令交互模块、处理模块、数据发送模块、ZDP 调用模块、OpenAPI 调用模块。

[0050] 1. 三个系统具体如下：

[0051] a.ManagerServer（管理后台）：管理后台服务系统，采集系统需要处理管理后台的命令。

[0052] b.IndexServer（数据处理系统）：数据处理系统，采集的数据需要发送的数据处理系统进行下一步处理。

[0053] c.MicCollector（采集系统）：采集系统。

[0054] 2. 采集系统包括：

[0055] a.DataRecv（数据接收模块）&DataSendToDB（数据发送模块）：命令交互模块，与管理后台进行信息交互。

[0056] b.DataProcess（数据处理模块）：处理模块，进行数据采集及后台命令执行。

[0057] c.DataSend（数据发送模块）：数据发送模块，向数据处理系统发送数据。

[0058] d.ZDP（下载器）：ZDP 调用模块。

[0059] e.OpenAPIModule（开放 API 调用模块）：OpenAPI 调用模块。

[0060] f.Gearman（任务分配系统）：用来把工作委派给其他机器、分布式的调用更适合做某项工作的机器、并发的做某项工作在多个调用间做负载均衡的函数的系统。

[0061] 3. 其他附属模块

[0062] a.DistributeLogin（分布式登录模块）：分布式登陆模块，利用 pearman 对马甲进行多机分布式验证。

[0063] b.MailReporter（邮件发送模块）：邮件发送模块，将日志统计信息发送到相关邮件组。

[0064] 其中，本发明一种高可用的微博采集平台采集系统如下模块，具体为：

[0065] 1）命令交互模块，处理与管理后台的交互信息，管理后台主要是对采集系统中的数据进行人工管理，例如博主的关注，删除，修改，拉黑等操作，DataRecv（管理后台数据接收）类用于接收管理后台发送过来的指令，并将指令按照指定格式存于本地队列 m_pJobQueue（任务存放队列）中，DataSendToDB（处理结果发送）类将采集系统对管理后台发送过来的指令的处理结果回送给管理后台。

[0066] 2）处理模块，DataProcess（数据处理）类主要包括两部分，数据采集和管理后台指令的执行，主要包括以下流程：

[0067] a. 自动授权，由于微博用户的认证有使用期限，新浪用户需要定期登录获取认证；搜狐和网易的认证获取一次即可，不需要定期刷新；拿到认证之后，将认证信息与用户绑定

到一起

[0068] b. 刷新认证，腾讯用户的认证比较特殊，只需要登录一次，拿到认证之后，定期调接口刷新认证即可；

[0069] c. 任务处理工作流程，这个线程的主要工作是从任务队列获得任务，进行协议解析，对任务包中每一个任务解析并处理，当所有 Job 处理完成后，将处理结果返回给后台处理程序（采用接收任务的 Socket 链接）。

[0070] d. 博主关注与取消关注，博主信息下载，博文下载，这些数据的下载，需要按照微博开放平台对应接口的要求，拼接参数列表，然后将拼接好的 url（数据下载链接）按照指定格式提交给 zdp（下载器）下载器。

[0071] e. 下载结果处理流程，提交给 zdp（下载器）任务后，zdp（下载器）会返回下载结果，得到的结果是 json（数据存储格式）格式的，将 json（数据存储格式）数据加载到 json（数据存储格式）容器中，从 json（数据存储格式）容器中按字段取数据，并将数据按照类型（博主信息，博文）存到对应的类对象中。

[0072] f.DataSend（数据发送给数据处理）数据发送流程，将下载好的数据（博文、博主信息、话题）按照协议要求的格式拼接好后，发送给数据处理。

[0073] g.ZDP（下载器）下载，整个采集系统中，所有数据下载都采用 zdp 下载器。

[0074] h.OpenAPIModule（微博开放接口调用）模块管理所有微博 api（开放接口）的调用。

[0075] 3) 分布式登录模块和邮件发送模块：

[0076] a. 新浪用户需要每天登录重新认证，根据现在的登陆间隔限制，每个 ip 在不小于 5min 登陆，ip 不会被限制，用户登陆周期是一天，多天运行，比较稳定。鉴于以上限制，用户数目比较多时，单机不能完成在认证可用周期内全部登陆一次，所以采用多机登陆。登录任务分配采用 gearman（任务分配系统），如图 2 所示；腾讯用户只需要在用户录入系统时，登录一次，以后调接口刷新认证即可；搜狐和网易登录后得到的认证永不过期，不需要刷新。

[0077] b. 邮件发送模块，对每天采集系统采集到的数据进行统计，分别对四大微博媒体网站的数据下载次数，下载失败次数，解析成功次数进行计算，客观反应采集系统的运行状态，得到一份报告，将报告发送给负责人。

[0078] 实施例

[0079] 1) 启动管理后台：

[0080] 管理后台主要管理三种类型的数据，博主、应用、马甲。

[0081] (1) 对博主数据、应用数据、马甲数据分别进行增加、删除、修改、查询；

[0082] (2) 马甲和应用进行关联；

[0083] (3) 马甲登录验证，保证马甲可用；

[0084] (4) 管理后台将以上操作拼接成指令的形式发给采集后台。

[0085] 2) 执行管理后台指令并进行数据采集：

[0086] (1) 采集后台以指令的形式接到博主、马甲、应用的相关操作，并将博主、马甲、应用的操作结果写入对应的本地数据文件；

[0087] (2) 启动关注流程，用马甲关注一批博主，并将数据更新到马甲文件中；

[0088] (3) 授权流程读取马甲本地文件，调用登录模块对马甲进行登录认证，并将认证参

数一并写入马甲本地文件，授权后的应用才能调用微博开放 API ；

[0089] (4) 授权完成后，启动微博博文、博主信息、话题获取流程，获取一个用于博文、博主信息、话题下载的马甲，要判断马甲的调度周期，按照微博开放 API 的要求，拼接请求参数，参数从马甲文件中读取，API 加请求参数形成一个待下载的 url ；

[0090] (5) 将 URL 作为一个下载任务，提交给下载器，等待返回结果 ；

[0091] (6) 得到的返回结果是 json 格式的，将 json 数据加载到 json 容器中，从 json 容器中按字段读取取数据，并将数据按照类型（博主信息，博文，话题）存到对应的类对象中 ；

[0092] (7) 将解析出来的博主信息，博文，话题数据发送给数据处理。

[0093] 3) 马甲分布式登录和邮件发送

[0094] a. 四大微博的应用用户的帐号，访问其微博帐户，进行内容读写，都需要在使用时，得到用户（马甲）本人授权。这就涉及到用户（马甲）登录认证的问题，根据现在的登陆间隔限制，每个 ip 在不小于 5min 登陆，ip 不会被限制，所有用户（马甲）在固定周期内要全部登录一次。鉴于以上限制，用户数目比较多时，单机不能完成在认证可用周期内全部登陆一次，所以采用多机登陆。登录任务分配采用 gearman（任务分配系统），如图 2 所示。

[0095] b. 邮件发送模块

[0096] (1) 对采集系统采集数据的日志进行统计 ；

[0097] (2) 对四大微博媒体网站的数据下载次数，下载失败次数，解析成功次数进行计算 ；

[0098] (3) 得到一份采集系统的数据采集量报告 ；

[0099] (4) 启动邮件发送程序，将报告发送给负责人。

[0100] 最后应当说明的是 ：以上实施例仅用以说明本发明的技术方案而非对其限制，所属领域的普通技术人员参照上述实施例依然可以对本发明的具体实施方式进行修改或者等同替换，这些未脱离本发明精神和范围的任何修改或者等同替换，均在申请待批的本发明的权利要求保护范围之内。

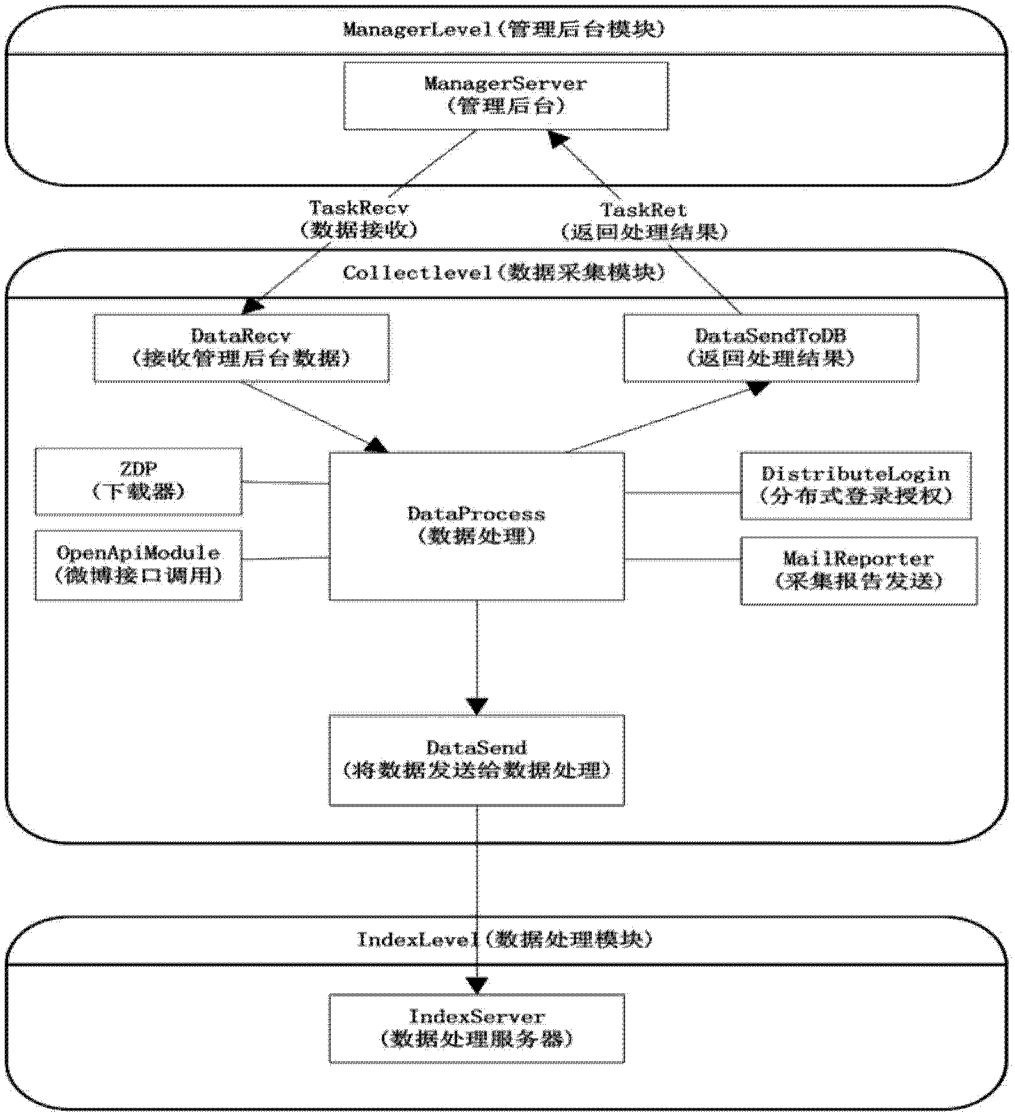

ManagerLevel(管理后台模块)
ManagerServer
(管理后台)
TaskRecv
(数据接收)
TaskRet
(返回处理结果)
Collectlevel(数据采集模块)
DataRecv
(接收管理后台数据)
DataSendToDB
(返回处理结果)
ZDP
(下载器)
OpenApiModule
(微博接口调用)
DataProcess
(数据处理)
DistributeLogin
(分布式登录授权)
MailReporter
(采集报告发送)
DataSend
(将数据发送给数据处理)
IndexLevel(数据处理模块)
IndexServer
(数据处理服务器)

图 1

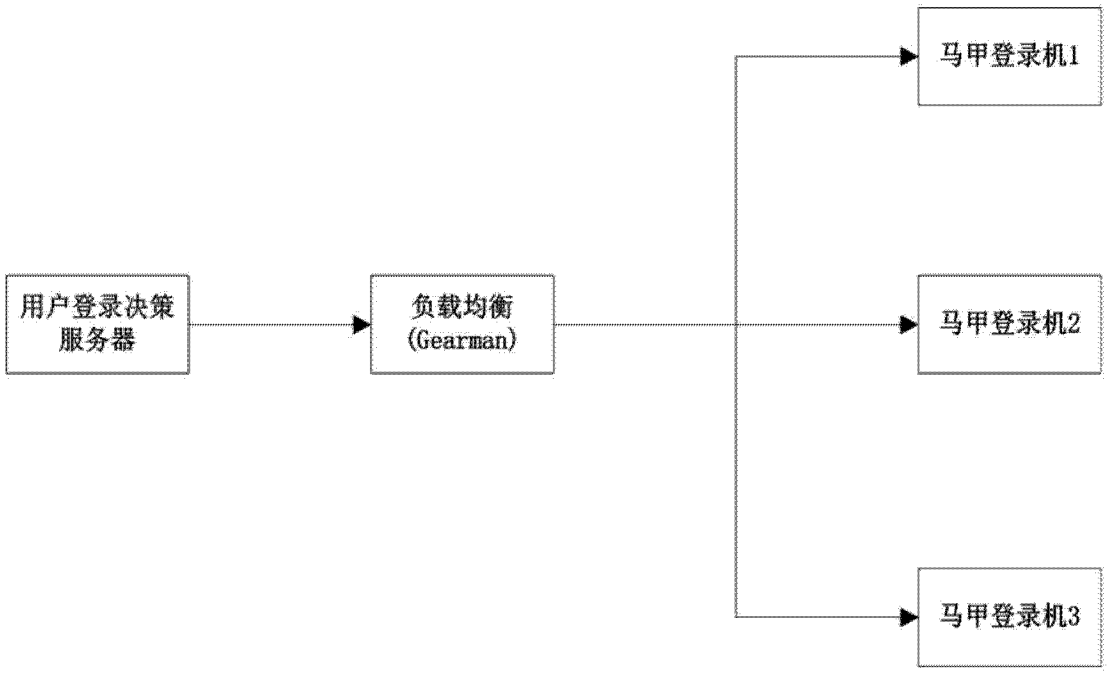


图 2