[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
H04L 12/24 (2006.01)
H04L 29/06 (2006.01)

[12] 发明专利说明书

专利号 ZL 200610117927.9

[45] 授权公告日 2009 年 11 月 18 日

[11] 授权公告号 CN 100561938C

[22] 申请日 2006.11.3
[21] 申请号 200610117927.9
[73] 专利权人 盛大计算机(上海)有限公司
地址 201203 上海浦东新区郭守敬路 356 号
[72] 发明人 晏 飞
[56] 参考文献
US2004/0225553A1 2004.11.11
CN1494268A 2004.5.5
审查员 许 婵

[74] 专利代理机构 上海浦一知识产权代理有限公司
代理人 丁纪铁 李隽松

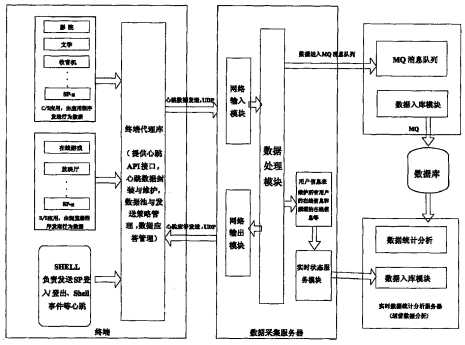
权利要求书 2 页 说明书 9 页 附图 3 页

[54] 发明名称

一种采集和统计分析数据的方法

[57] 摘要

本发明公开了一种采集和统计分析数据的方法，该方法包括终端数据采集方法和数据统计分析方法，终端数据采集方法是指终端与数据采集服务器使用约定的数据协议进行通讯，数据采集服务器以 UDP 方式提供服务，终端将状态和行为数据以 UDP 数据包的形式发送到数据采集服务器，数据统计分析方法，其统计分析的项目包括：各个栏目的实时在线人数、在一段时间内各个栏目的用户访问状况、每个终端用户的实时行为和历史行为。 本发明由于采取多层系统结构，且数据收集时采用 UDP 形式和约定数据协议，可准确收集终端用户的使用状态和行为数据，并对用户动态信息进行分析和统计。



1、一种采集和统计分析数据的方法，其特征在于，包括终端数据采集方法和数据统计分析方法；

其中所述终端数据采集方法是指终端与数据采集服务器使用约定的数据协议进行通讯，数据采集服务器以UDP方式提供服务，终端将状态和行为数据以UDP数据包的形式发送到数据采集服务器，包括如下步骤：

步骤一、首先终端在约定的时间向所述数据采集服务器发送数据，使数据采集服务器可监控并维持该终端的活跃状态；

步骤二、若在约定的时间内，终端状态因用户操作而发生变化时，终端则向数据采集服务器发送数据；若在约定的时间内，终端状态未发生改变时，终端则向数据采集服务器发送一个状态保持的数据；若终端退出，则发送带有退出标识的数据；上述数据中包含用户行为标识或用户当前所在服务栏目标识；

步骤三、所述数据采集服务器若在约定的时间内收到步骤二数据，将验证数据是否合法；若合法，则进入步骤四，若不合法则构造由终端接收的响应数据包并进入步骤七；若在约定的时间内，数据采集服务器未收到终端状态数据则将该终端状态标记为离线，进入步骤六；

步骤四、若数据带有退出标识，则数据采集服务器将该终端状态标记为离线，若数据无退出标识，将用户的最后在线时间设置为当前时间，并将用户所在栏目设置为终端提交的栏目标识；进入步骤五；

步骤五、将数据发送到MQ数据队列，进入步骤六，同时构造由终端接收的响应数据包，进入步骤七；

步骤六、将MQ数据队列读取数据保存到数据库，进行定时统计分析；

步骤七、将由终端接收的响应数据包返回给终端，终端计算丢包率，判断是否需重复进入步骤二；

所述数据采集服务器包括数据统计分析模块、实时状态服务模块、实时数据统计与分析服务器；

所述数据统计分析方法，其统计分析的项目包括：各个栏目的实时在线人数、在一段时间内各个栏目的用户访问状况、每个终端用户的实时行为和历史行为，实时数据统计与分析服务器以图表形式显示当前各栏目在线用户数据；包括如下步骤：

步骤A、所述数据统计分析模块向所述实时状态服务模块采集实时在线用户数据；

步骤B、所述实时状态服务模块查询用户信息表，获取当前用户信息，得到当前用户数据；

步骤C、所述实时状态服务模块返还数据给所述实时数据统计与分析服务器；

步骤D、所述实时数据统计与分析服务器以图表形式显示当前各栏目在线用户数据。

一种采集和统计分析数据的方法

技术领域

本发明涉及一种服务器实时采集终端用户行为数据的方法，尤其涉及一种集成娱乐与服务系统中对所有终端用户的状态、行为进行监控、统计和分析的方法。

背景技术

目前互联网的发展使得基于网络的集成化娱乐应用和服务内容越来越丰富，人们可以通过这些应用及服务，享受到网络上的音乐、电影、文学、游戏等内容。作为提供这些应用服务的内容服务商，应充分了解用户使用这些产品的情况，掌握用户对产品所提供的各项应用服务的喜好程度。在此基础上进一步分析用户行为和习惯，以便提供更优质的服务，甚至提供针对不同的用户一对一的服务内容。如何准确地收集用户在集成了多个内容和服务的系统中的使用状态和行为数据，并对这些用户动态信息进行分析和统计，对内容服务商而言就是一个非常重要的问题。

发明内容

本发明要解决的技术问题是提供一种采集和统计分析数据的方法，能够准确地收集用户在集成了多个内容和服务的系统中的使用状态和行为数据，并对用户动态信息进行分析和统计。

为解决上述技术问题，本发明提供了一种采集和统计分析数据的方法，可用于上述的系统中，该方法包括终端数据采集方法和数据统计分析方法；终端数据采集方法是指终端与数据采集服务器使用约定的数据协议进行通

讯，数据采集服务器以UDP （User Datagram Protocol，用户数据报协议）方式提供服务，终端将状态和行为数据以UDP数据包的形式发送到数据采集服务器，包括：首先终端在约定的时间向数据采集服务器发送数据，使数据采集服务端可监控并维持该终端的活跃状态；然后若在约定的时间内，终端状态因用户操作而发生变化时，终端则向数据采集服务器发送数据，数据中包含用户行为标识或用户当前所在服务栏目标识，数据采集服务器收到此数据，将用户的最后在线时间设置为当前时间，并将用户所在栏目设置为终端提交的栏目标识；若在约定的时间内，终端状态未发生改变时，终端发送一个状态保持的到数据采集服务器，数据采集服务器收到此数据，将用户的最后在线时间设置为当前时间；若在约定的时间内，数据采集服务器未收到终端状态数据则将该终端状态标记为离线；当终端退出时，发送带有退出标识的数据，数据采集服务器将该终端状态标记为离线，并将终端的本次登入/登出记录保存到数据库；数据统计分析方法，其统计分析的项目包括：各个栏目的实时在线人数、在一段时间内各个栏目的用户访问状况、每个终端用户的实时行为和历史行为。

本发明由于在系统结构上采取多层结构，并且数据收集时采用 UDP 形式，并采用约定的数据协议，可准确地收集终端用户的使用状态和行为数据，并对用户动态信息进行分析和统计。

附图说明

图 1 是本发明的一个具体实施例的示意图；

图 2 是图 1 实施例中的终端行为数据采集流程；

图 3 是图 1 实施例中各栏目实时在线用户信息统计流程。

具体实施方式

下面结合附图和具体实施例对本发明作进一步详细的说明。

如图 1 所示，为本发明的一个具体实施例。

实施例：

本发明的目的是提供一种实时采集多终端用户状态和行为数据，并进行统计分析的系统和方法。

如图 1，是按本发明建立的多层结构的终端状态行为数据采集系统，包含以下几个单元：终端代理库、数据采集服务器、数据库、实时数据统计与分析服务器。按其结构及数据传输关系，各单元模块功能如下所述：终端代理库负责封装与服务端的通讯，并为终端提供接口，向服务端发送终端状态数据的函数库，该库提供给所有需要提交状态的终端；数据采集服务器可提供服务，获取终端代理库提交的数据，进行所有终端的实时状态维护，并将整理后的数据写入数据库的服务单元；数据库用于存储终端状态与行为数据的服务单元，存储的数据已经过数据采集服务器的处理；实时数据统计与分析服务器：从数据采集服务器和数据库获取实时数据，进行统计，并已数据表、图表和实时监控图等形态显示的服务单元。

在上述系统中的数据采集采用如下的终端数据采集策略：

终端与数据采集服务器使用约定的数据协议进行通讯，数据格式协议见表1；数据采集服务器以UDP 网络服务的方式提供服务，所有终端将状态和行为数据以UDP数据包的形势发送到心跳服务器；终端在约定的时机向数据采集服务器发送心跳数据，使数据采集服务端可监控并维持该终端的活跃状态；终端状态在因用户操作改变而发生变化时，向数据采集服务器发

送心跳数据，数据中包含用户行为标识[或用户当前所在服务栏目标识]，数据采集服务器收到此数据，将用户的最后在线时间设置为当前时间，并将用户所在栏目设置为终端提交的栏目标识；终端状态在约定的时间内（默认为1分钟）未发生改变（用户在1分钟内未做任何操作）时，发送一个状态保持的心跳数据到数据采集服务器，数据采集服务器收到此数据，将用户的最后在线时间设置为当前时间；终端用户退出终端系统时，发送带有退出标识的心跳数据，数据采集服务器将该终端状态标记为离线，并将用户的本次登入/登出记录保存到数据库；终端用户异常断开，未发送退出状态数据时：数据采集服务器定时检测所有终端最后在线时间，在某终端发送心跳数据超时（在指定的时间，通常为2分钟内，某终端未向数据采集服务器发送心跳数据），数据采集服务器将该终端状态标记为离线；采用UDP通讯方式，一台心跳服务器可同时为5000个以上的终端提供服务。

图2是本实施例中的终端行为数据采集流程。包括：一、首先终端在约定的时间向数据采集服务器发送数据，使数据采集服务端可监控并维持该终端的活跃状态；二、若在约定的时间内，终端状态因用户操作而发生变化时，终端则向数据采集服务器发送数据；若在约定的时间内，终端状态未发生改变时，终端则向数据采集服务器发送一个状态保持的数据；若终端退出，则发送带有退出标识的数据；上述数据中包含用户行为标识或用户当前所在服务栏目标识；三、数据采集服务器若在约定的时间内收到二的数据，将验证数据是否合法；若合法，则进入四，若不合法则构造终端响应数据包并进入七；若在约定的时间内，数据采集服务器未收到终端状态数据则将该终端状态标记为离线，进入六；四、若数据带有退出标识，

则数据采集服务器将该终端状态标记为离线，若数据无退出标识将用户的最后在线时间设置为当前时间，并将用户所在栏目设置为终端提交的栏目标识；进入五；五、将数据发送到MQ数据队列，进入六，同时构造终端响应数据包，进入七；六、将MQ数据队列读取数据保存到数据库，进行定时统计分析；七、将终端响应数据包返回给终端，终端计算丢包率，判断是否需重复进入二；

图3是本实施例中各栏目实时在线用户信息统计流程，包括如下步骤：A、数据统计分析模块向实时状态服务模块调用实时在线用户数据；B、实时状态服务模块查询用户信息表，获取当前用户信息，得到当前用户数据；C、实时状态服务模块返还数据给实时数据统计与分析服务器；D、实时数据统计与分析服务器以图表形式显示当前各栏目在线用户数据。

下面以一个具体的例子来讲述如何使用本发明的系统和方法，该例如下：

用户A登陆到集成化在线娱乐终端（集成了游戏、VOD视频点播、新闻、教育等内容）。此时，收集该用户的行为数据的模块被启动，并向数据采集服务发送一个用户登陆的UDP消息。

消息格式：表1中将Type字段设为1001。

数据采集服务器收到用户登陆的消息，立即在用户状态维护表（内存中）将用户状态标志为在线，并将用户当前所在栏目标志为主界面。然后将本条数据写入MSMQ队列。（另有程序定时从该队列里提取数据写入数据库）

用户开始在集成了众多娱乐内容的娱乐终端里浏览寻找他所感兴趣的

栏目。他选择了休闲游戏，并按下‘确定’键进入。终端通过代理库向服务器发送栏目跳转的UDP消息。

消息格式：表3中将20、21位的用户当前所在栏目数设为1，22、23位栏目标识设为37（休闲游戏栏目的标识号）。

数据采集服务器收到栏目跳转的消息，修改用户状态维护表（内存中），将用户状态标志为在线（解决因异常导致未收到用户登陆消息的情况），并将用户当前所在栏目设为终端提交的栏目。然后将本条数据写入MSMQ队列。

用户选择了休闲游戏频道中的赛车游戏，按下‘确定’时。终端通过代理库向服务器发送栏目/应用跳转消息。

消息格式：表3中将20、21位的用户当前所在栏目数设为1，22、23位栏目标识设为371（赛车应用的标识号）。

数据采集服务器收到栏目/应用跳转消息，修改用户状态为在线，并将用户当前所在栏目设为赛车。然后将本条消息写入MSMQ队列。（注：因为服务端是保存了终端显示的所有栏目和应用的树状结构图的，所以，只要终端提交用户的当前栏目/应用标识，服务端便可查出用户现在的栏目具体路径。）

用户退出栏目或应用时（退出赛车游戏，回到休闲游戏频道），终端向服务器发送栏目/应用跳转消息。

消息格式：表3中将20、21位的用户当前所在栏目数设为1，22、23位栏目标识设为37（休闲游戏栏目的标识号）。

数据采集服务器收到栏目/应用跳转消息，此消息意义除了包含用户进入了某一频道，也包含了用户退出了前一个栏目。

用户长时间在赛车游戏应用中娱乐（或静止在其他栏目中），在这种情况下，为使服务器不会错误的认为终端发生异常意外退出了，则需要定时（例如1分钟）向服务器发送行为保持消息，消息格式可完全和上一次发送到服务器的消息一致。

消息格式：表3中将20、21位的用户当前所在栏目数设为1，22、23位栏目标识设为37（休闲游戏栏目的标识号）。

数据采集服务器收到终端的行为保持消息，将用户的最后发送消息的时间更改为当前时间，并将此条消息写入MSMQ数据库。

用户退出娱乐终端程序时，终端向服务器发出系统退出的消息。

消息格式：表1中将Type字段设为1002。

数据采集服务器收到系统退出消息，将用户的状态设置为离线。并将本次用户登陆[时间]－退出[时间]的记录写入数据库，然后将本条退出消息写入MSMQ队列。

如果终端发生网络异常端口，或程序异常中止的情况，且未能通知服务器终端已退出，则服务器根据预定策略自动检查终端退出状态。

服务器将有一个线程定时（2分钟，可配置）检查所有终端上一次发送行为数据的时间，如发现某终端在2分钟内未向服务器发送行为消息，则将此终端状态标志为离线。并将终端本次上线－离线记录写入数据库。

下表所示的数据格式协议包括数据头（HEAD，16固定长度）、数据体（BODY，长度和内容在HEAD中指定）、数据扩展段（SPID & EXTEND），

0~15	16~x	x+1~y
HEAD	BODY	SPID & EXTEND
数据头，16 固定长度	数据体，长度和内容在 HEAD 中指定	数据扩展段

表1

而其中HEAD数据格式定义见下表：

0	1	2	3	4	5	6	7	8	9	10	11
'E'	'Z'	'A'	'P'	Version		Size		Type			
12	13	14	15								
Timestamp											

表2

其中0~3字节为数据通讯协议标志，用于标志该UDP数据包是属于本系统数据采集协议包，如定为“PRAP”；4~5为两字节的版本号；6~7为协议包BODY部分的字节长度；8~11为协议包BODY的类型标识，四字节的编码，该编码是全局统一的；12~15为timestamp,即时间戳。

BODY数据格式定义：（仅举例说明）

BODY数据体格式可变，数据的意义由HEAD中的Type位来标识。

当Type为1时，如下表3：

0~15		16	17	18	19	20	21	22	23	24	25
HEAD		终端用户标识（ID）				用户当前所在栏目数量		栏目1的标识		栏目2的标识	
26	27	28	29	30	31	32	33	34	35	36~n	
扩展数据长度		事件ID				状态ID				Extend	

表3

其中16~19为USERID，即终端用户标识；20~21为用户当前所在的应用栏目数量（用户可以同时在使用多个应用），后面紧接每个栏目的标识，每个栏目标识固定占2个字节；22~23为栏目1的标识；24~25为栏目2的标识；26~27为Extend扩展位的长度；28~31为事件ID（标识此数据包的含义，如登陆、栏目变更、登出等）；32~35为状态ID（标识本次事件发生的状态，如成功、失败及原因等）；36~n为扩展数据，其长度由26~27扩展数据长度指出，以xml格式描述。扩展数据中可包含用户所在各个栏目当前的运行状态，已经在栏目中的哪个子模块中。

本实施例中，对收集数据的进行统计分析，主要统计分析的项目包括：各个栏目的实时在线人数、各个栏目的用户访问状况（在一段时间内）、每个终端用户的实时行为和历史行为等。

数据统计分析的具体数据源为：

各个栏目的实时在线人数通过数据采集服务器提供的服务。数据采集服务器实时维护监控所有终端用户的当前状态和所在栏目，并提供Socket接口，监控程序定时调用此接口，得到当前在线人数并以曲线图的方式显示在监控图表上；

各个栏目的总在线时长、平均在线时长等数据由数据库数据计算得出，数据来源于数据采集服务器；

所有用户的行为历史记录保存在数据库，此数据由数据采集服务器在用户的一次登入登出过程中，全程记录到数据库。此用户用于整理每个用户的行为规律，分析出用户的习惯和爱好，便于对用户提供一对一的服务。

综上所述，本发明提出的一种进行数据采集和统计分析的系统和方法，能够准确地收集用户在集成了多个内容和服务的系统中的使用状态和行为数据。并对这些用户动态信息进行分析和统计，该系统尤其适用于宽带数字家庭娱乐系统中采集和分析用户行为，并为不同用户提供完美的个性化服务系统中。

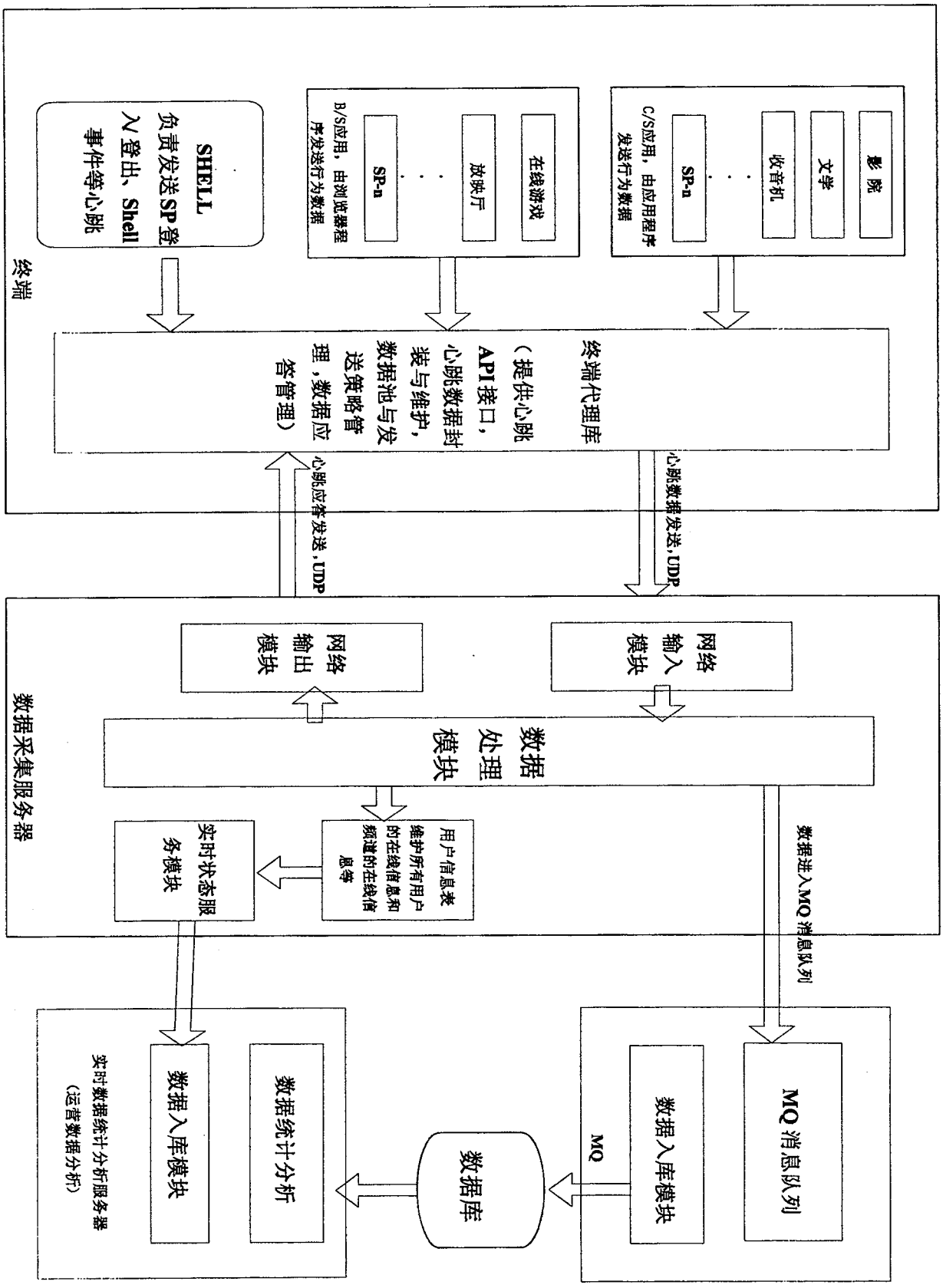

影 院
文学
收音机
SP-n
C/S应用，由应用程序发送行为数据
在线游戏
放映厅
SP-n
B/S应用，由浏览器程序发送行为数据
SHELL
负责发送SP登入/登出、Shell事件等心跳
终端
终端代理库
（提供心跳API接口，心跳数据封装与维护，数据池与发送策略管理，数据应答管理）
心跳数据发送，UDP
心跳应答发送，UDP
网络输入模块
网络输出模块
数据处理模块
用户信息表 维护所有用户的在线信息和频道的在线信息等
实时状态服务模块
数据进入MQ消息队列
数据采集服务器
MQ消息队列
数据入库模块
MQ
数据库
数据统计分析
数据入库模块
实时数据统计分析服务器（运营数据分析）

图1

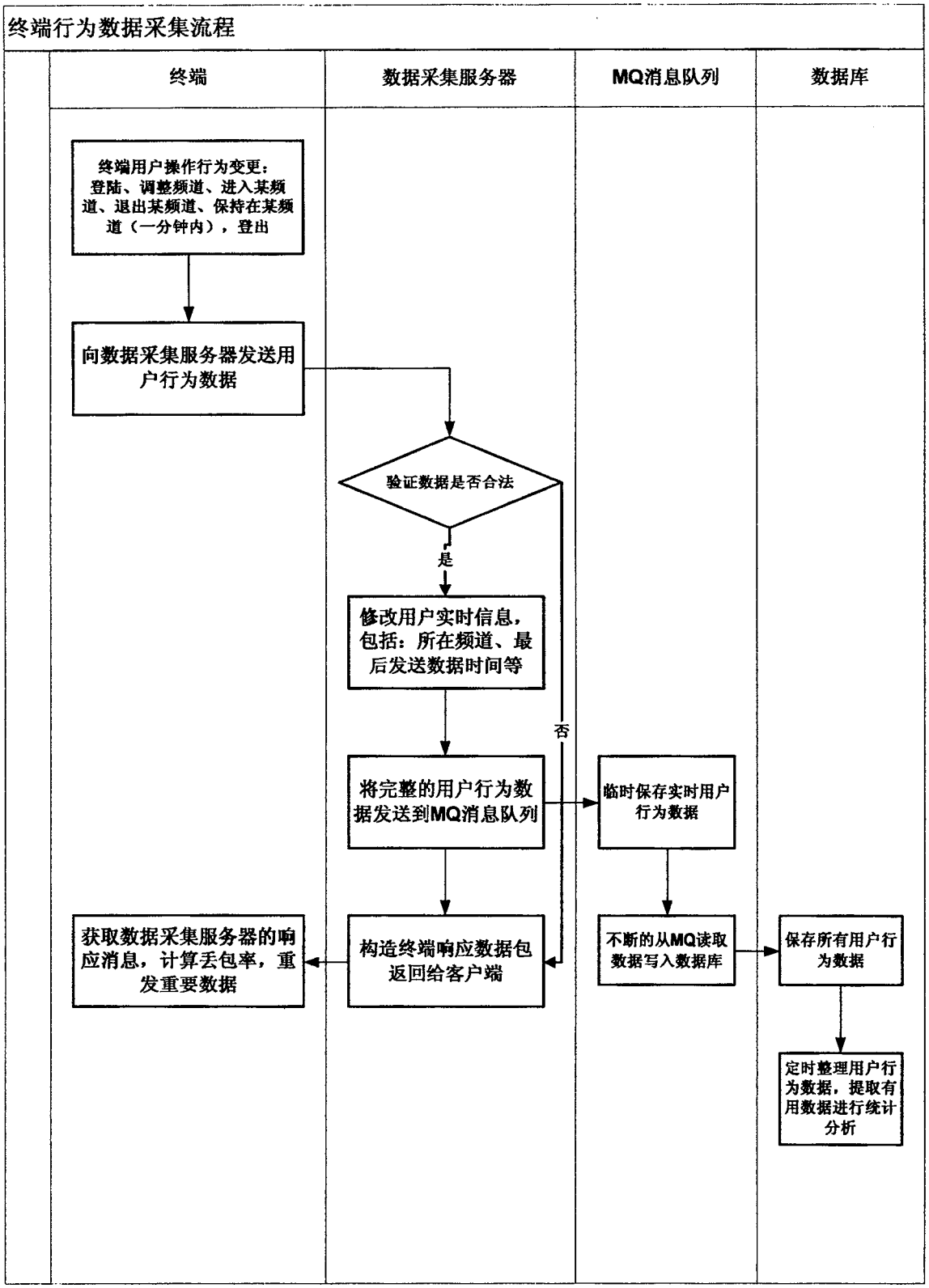

终端行为数据采集流程
终端
数据采集服务器
MQ消息队列
数据库
终端用户操作行为变更：登陆、调整频道、进入某频道、退出某频道、保持在某频道（一分钟内），登出
向数据采集服务器发送用户行为数据
验证数据是否合法
是
否
修改用户实时信息，包括：所在频道、最后发送数据时间等
将完整的用户行为数据发送到MQ消息队列
临时保存实时用户行为数据
获取数据采集服务器的响应消息，计算丢包率，重发重要数据
构造终端响应数据包返回给客户端
不断的从MQ读取数据写入数据库
保存所有用户行为数据
定时整理用户行为数据，提取有用数据进行统计分析

图2

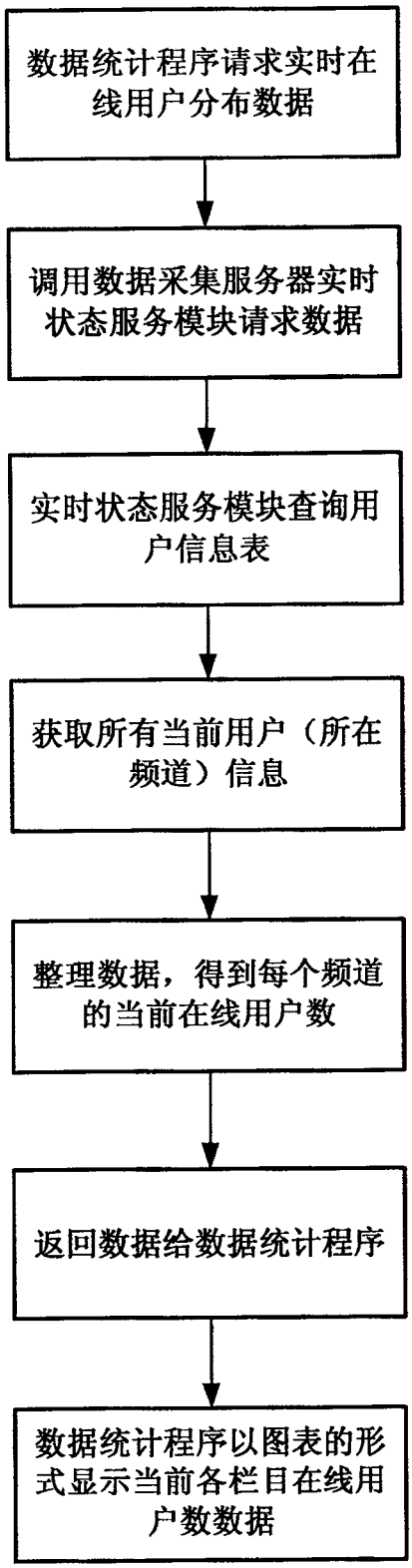

数据统计程序请求实时在线用户分布数据
调用数据采集服务器实时状态服务模块请求数据
实时状态服务模块查询用户信息表
获取所有当前用户（所在频道）信息
整理数据，得到每个频道的当前在线用户数
返回数据给数据统计程序
数据统计程序以图表的形式显示当前各栏目在线用户数数据

图3