

(12) 发明专利

(10) 授权公告号 CN 101986605 B

(45) 授权公告日 2013. 04. 24

(21) 申请号 201010537959. 0

CN 101170426 A, 2008. 04. 30,

(22) 申请日 2010. 11. 04

审查员 王玥

(73) 专利权人 北京迈朗世讯科技有限公司

地址 100080 北京市海淀区上地五街 7 号昊
海大厦一层 106C 室

(72) 发明人 王强

(74) 专利代理机构 中国国际贸易促进委员会专
利商标事务所 11038

代理人 孙宝海

(51) Int. Cl.

H04L 12/24 (2006. 01)

H04L 12/721 (2013. 01)

(56) 对比文件

CN 101409690 A, 2009. 04. 15,

CN 101790196 A, 2010. 07. 28,

US 6327266 B1, 2001. 12. 04,

CN 101431485 A, 2009. 05. 13,

权利要求书2页 说明书11页 附图9页

(54) 发明名称

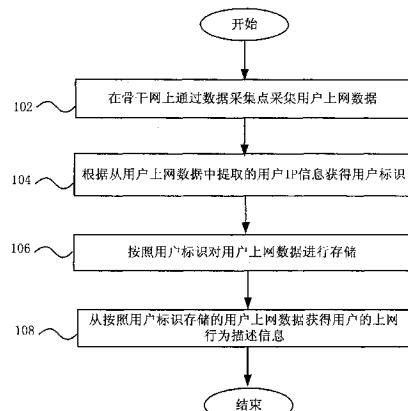
一种基于骨干网的用户上网数据处理方法和
系统

(57) 摘要

本发明公开一种基于骨干网的用户上网数据
处理方法和系统。该方法包括：在骨干网上通过
数据采集点采集用户上网数据；根据从用户上网
数据中提取的用户 IP 信息获得用户标识；按照用
户标识对用户上网数据进行存储；从按照用户标
识存储的用户上网数据获得用户的上网行为描
述信息。本发明的方法和系统实现了覆盖电信运
营商省级中心所有宽带用户，能够客观反映用户群
体上网行为的统计特征，客观反映单个用户上网
行为的统计特征。

B

CN 101986605



1. 一种基于骨干网的用户上网数据处理方法,其特征在于,包括:

根据网络拓扑层次结构和网络路由策略在骨干网上部署数据采集点,对于采用动态路由策略决定数据转发路径的路由器,根据 metric 信息选择所述路由器输出端口网络链路以部署所述数据采集点;和 / 或在传输距离短或链路状态好的路由器输出端口网络链路部署所述数据采集点;和 / 或对于路由器将数据按照负载均衡原则转发到多条网络链路上,从所述多条网络链路上选择任意一条链路部署所述数据采集点;和 / 或在拓扑结构上层的数据链路上部署所述数据采集点;

在骨干网上通过部署在路由器输入端或输出端的数据链路上的数据采集点采集用户上网数据;

根据从所述用户上网数据中提取的用户 IP 信息获得用户标识;

按照所述用户标识对所述用户上网数据进行存储;

从按照所述用户标识存储的所述用户上网数据获得所述用户的上网行为描述信息。

2. 根据权利要求 1 所述的用户上网数据处理方法,其特征在于,所述用户的上网行为描述信息包括访问时间、网站 IP 地址、网站 URL、页面文本标题、关键词、网站 cookie、和页面 Referrer 中的至少一个。

3. 根据权利要求 1 所述的用户上网数据处理方法,其特征在于,所述根据网络拓扑层次结构和网络路由策略在所述骨干网上部署所述数据采集点的步骤包括:

对于采用静态路由策略决定数据转发路径的路由器,在所述路由器的所有输出端口采集所述用户上网数据。

4. 根据权利要求 1 所述的用户上网数据处理方法,其特征在于,

所述数据采集点部署在数据流分散之前和 / 或数据流汇聚之后的网络路由设备的输入端口或输出端口上。

5. 一种基于骨干网的用户上网数据处理系统,其特征在于,包括:

多个数据采集设备,用于在骨干网上通过部署在路由器输入端或输出端的数据链路上的数据采集点采集用户上网数据;对于采用动态路由策略决定数据转发路径的路由器:所述数据采集设备部署在骨干网根据 metric 信息选择的所述路由器输出端口网络链路上;和 / 或所述数据采集设备部署在骨干网的传输距离短或链路状态好的路由器输出端口网络链路上;和 / 或对于路由器将数据按照负载均衡原则转发到多条网络链路上,所述数据采集设备部署在从所述多条网络链路上选择任意一条链路;和 / 或在拓扑结构上层的数据链路上部署所述数据采集点;

用户标识获取设备,用于根据从所述用户上网数据中提取的用户 IP 信息获得用户标识;

上网数据存储设备,用于按照所述用户标识对所述用户上网数据进行存储;

描述信息提取设备,用于从按照所述用户标识存储的所述用户上网数据获得所述用户的上网行为描述信息。

6. 根据权利要求 5 所述的用户上网数据处理系统,其特征在于,所述用户的上网行为描述信息包括访问时间、网站 IP 地址、网站 URL、页面文本标题、关键词、网站 cookie、和页面 Referrer 中的至少一个。

7. 根据权利要求 5 所述的用户上网数据处理系统,其特征在于,

所述数据采集点部署在数据流分散之前和 / 或数据流汇聚之后的网络路由设备的输入端口或输出端口上。

一种基于骨干网的用户上网数据处理方法和系统

技术领域

[0001] 本发明涉及网络数据处理技术,尤其涉及一种基于骨干网的用户上网数据处理方法和系统。

背景技术

[0002] 电信运营商通常拥有数百万的互联网宽带用户,相应骨干网的数据流总带宽在 TB 级别。多种应用需要在网络数据链路层面上采集宽带用户的上网数据,刻画用户的上网行为特征。

[0003] 为了实现在电信运营商的骨干网络上覆盖全范围内的宽带用户的上网行为,需要在电信运营商骨干网络上合理地部署数据采集点以尽量有效获得全体宽带用户的上网数据,并尽量获得用户上网行为的全面描述信息。

[0004] 目前业界还没有基于电信运营商骨干网络的用户上网行为数据采集解决方案。

发明内容

[0005] 本发明要解决的一个技术问题是提供一种用户上网数据处理方法,能够在骨干网上对个体用户上网行为进行描述。

[0006] 本发明提供一种基于骨干网的用户上网数据处理方法,包括:

[0007] 在骨干网上通过部署在路由器输入端或输出端的数据链路上的数据采集点采集用户上网数据;

[0008] 根据从用户上网数据中提取的用户 IP 信息获得用户标识;

[0009] 按照用户标识对用户上网数据进行存储;

[0010] 从按照用户标识存储的用户上网数据获得用户的上网行为描述信息。

[0011] 进一步,用户的上网行为描述信息包括访问时间、网站 IP 地址、网站 URL、页面文本标题、关键词、网站 cookie、和页面 Referrer 中的至少一个。

[0012] 进一步,根据网络拓扑层次结构和网络路由策略在骨干网上部署数据采集点。

[0013] 进一步,对于采用静态路由策略决定数据转发路径的路由器,在路由器的输出端口采集用户上网数据;

[0014] 和 / 或

[0015] 对于采用动态路由策略决定数据转发路径的路由器,根据 metric 信息选择路由器输出端口网络链路以部署数据采集点;

[0016] 和 / 或

[0017] 在传输距离短或链路状态好的路由器输出端口网络链路部署数据采集点;

[0018] 和 / 或

[0019] 对于路由器将数据按照负载均衡原则转发到多条网络链路上,每条链路以均等机会获得并传输数据的情况,从多条网络链路上选择任意一条链路部署数据采集点;

[0020] 和 / 或

- [0021] 在拓扑结构上层的数据链路上部署数据采集点。
- [0022] 进一步,该方法还包括:数据采集点部署在数据流分散之前和 / 或数据流汇聚之后的网络路由设备的输入端口或输出端口上。
- [0023] 本发明要解决的一个技术问题是提供一种用户上网数据处理系统,能够在骨干网上对个体用户上网行为进行描述。
- [0024] 本发明提供一种基于骨干网的用户上网数据处理系统,包括:
- [0025] 多个数据采集设备,用于在骨干网上通过部署在路由器输入端或输出端的数据链路上的数据采集点采集用户上网数据;
- [0026] 用户标识获取设备,用于根据从用户上网数据中提取的用户 IP 信息获得用户标识;
- [0027] 上网数据存储设备,用于按照用户标识对用户上网数据进行存储;
- [0028] 描述信息提取设备,用于从按照用户标识存储的用户上网数据获得用户的上网行为描述信息。
- [0029] 进一步,用户的上网行为描述信息包括访问时间、网站 IP 地址、网站 URL、页面文本标题、关键词、网站 cookie、和页面 Referrer 中的至少一个。
- [0030] 进一步,对于采用动态路由策略决定数据转发路径的路由器:
- [0031] 数据采集设备部署在骨干网根据 metric 信息选择的路由器输出端口网络链路上;
- [0032] 和 / 或
- [0033] 数据采集设备部署在骨干网的传输距离短或链路状态好的路由器输出端口网络链路上;
- [0034] 和 / 或
- [0035] 对于路由器将数据按照负载均衡原则转发到多条网络链路上,每条链路以均等机会获得并传输数据的情况,数据采集设备部署在从多条网络链路上选择任意一条链路。
- [0036] 进一步,数据采集点部署在数据流分散之前和 / 或数据流汇聚之后的网络路由设备的输入端口或输出端口上。
- [0037] 通过本发明实施例的用户上网数据处理方法和系统,在骨干网中采集用户上网数据,将用户的上网数据按照用户进行存储,并分析获得各个用户的上网行为描述信息,能够较好地获得个体用户上网行为描述。

附图说明

- [0038] 图 1 示出本发明的基于骨干网的用户上网数据处理方法的一个实施例的流程图;
- [0039] 图 2 示出路由器输入输出链路示意图;
- [0040] 图 3 示出本发明的基于骨干网的用户上网数据处理系统的一个实施例的结构图;
- [0041] 图 4 示出宽带用户群体对互联网网站的访问事件的集合在由时间、用户、和网站组成的三维空间中的示意图;
- [0042] 图 5 示出部署在网络链路上的采集点对用户访问网站事件在时间上进行均匀的随机采样所观察到的用户对网站的访问事件的集合;
- [0043] 图 6 示出部署在网络链路上的采集点对用户访问网站事件在时间上进行非均匀

的随机采样所观察到的用户对网站的访问事件的集合；

[0044] 图 7 示出特定用户的互联网访问数据由特定网络链路来传输的路由策略相对固定的情况下所观察到的用户对网站的访问事件的集合；

[0045] 图 8 示出在特定的电信运营商 IDC 机房部署采集点所观察到的用户对网站的访问事件的集合；

[0046] 图 9 示出三个宽带用户对多个网站的访问行为在由时间和网站组成的二维空间中的示意图；

[0047] 图 10 示出部署在固定网络链路上的采集点将对个体用户访问网站事件在时间上进行均匀的随机采样所观察到的个体用户对网站的访问事件的集合；

[0048] 图 11 示出部署在固定网络链路上的采集点将对个体用户访问网站事件在时间上进行非均匀的随机采样所观察到的个体用户对网站的访问事件的集合；

[0049] 图 12 示出部署在特定的电信运营商 IDC 机房的采集点所观察到的个体用户对网站的访问事件的集合；

[0050] 图 13 示出一个电信网络链路上数据采样点例子的示意图；以及

[0051] 图 14 示出一个电信运营商省级中心的骨干网络结构及其采集点部署示意图。

具体实施方式

[0052] 下面参照附图对本发明进行更全面的描述，其中说明本发明的示例性实施例。

[0053] 图 1 示出本发明的基于骨干网的用户上网数据处理方法的一个实施例的流程图。

[0054] 如图 1 所示，在步骤 102，在骨干网上通过数据采集点采集用户上网数据。例如，根据电信网络路由策略选择数据采集点的部署方式。

[0055] 在步骤 104，根据从用户上网数据中提取的用户 IP 信息获得用户标识。例如，从 AAA 服务器获取网络用户上下线信息，获得用户标识和 IP 地址的对应关系；根据用户上网数据中提取的 IP 地址以及用户标识和 IP 地址的对应关系，获得用户标识信息。

[0056] 在步骤 106，按照用户标识对用户上网数据进行存储。将采集的用户上网数据按照不同的用户标识分别存储，例如，存储在根据用户标识索引的各个用户目录中。

[0057] 在步骤 108，从按照用户标识存储的用户上网数据获得用户的上网行为描述信息。对不同用户的上网数据进行分析，获得各个用户的上网行为描述信息。用户的上网行为描述信息例如包括访问时间、网站 IP 地址、网站 URL、页面文本标题或用户提交的关键词、网站 Cookie、页面 Referrer 等信息。可以通过多个关键词来描述用户上网行为特征，作为用户上网行为描述信息，从用户访问页面的文本标题或者内容匹配各个关键词，从而体现用户上网行为特征。

[0058] 例如，采集代表宽带用户对网站页面的访问动作的 HTTP 请求数据以及相应网站页面的内容信息，宽带用户的上网行为描述信息可以通过用户对网站页面的访问事件来描述。每个访问事件记录了用户端信息和网站端信息，用户端信息包括用户 UserID、访问时间、用户 IP 地址，网站端信息包括网站 IP 地址、网站 URL、页面文本标题或用户提交的关键词。

[0059] 在上述实施例中，在骨干网中采集用户上网数据，将用户的上网数据按照用户进行存储，并分析获得各个用户的上网行为描述信息，能够较好地获得个体用户上网行为描

述，并根据个体用户上网信息描述获得用户群体上网行为描述。此外，由于区分用户进行数据存储和分析，可以在部分或者较少的链路上部署采集点，通过时间的积累获得个体用户上网行为描述，减少骨干网上数据采集点的部署，而同时仍能在统计意义上较准确地获得用户上网行为描述，从而减少了系统的成本，便于实施应用。

[0060] 网络拓扑结构上的关键节点由与之相关的链路连接和路由策略共同决定。本发明的一个实施例根据网络拓扑层次结构和网络路由策略在骨干网上部署数据采集点，从而实现在电信网络中部署适量数据采集点，并尽量获得该网络范围内所有个体用户在统计意义上的准确上网行为。一种实现方式是数据采集点部署在数据流分散之前和 / 或数据流汇聚之后的网络路由设备的输入端口或输出端口上，从而以尽量少的采集点部署获得尽可能多的用户数据。

[0061] 电信网络的骨干网由大量路由器彼此连接组成的，当一个数据包需要从网络链路的 A 节点传输到 B 节点时往往面临多条传输路径，这时电信网络需要采用路由策略来决定如何选择数据的传输路径。从大体上说路由策略分为动态路由策略和静态路由策略。

[0062] 动态路由策略是指电信网络能够根据当前网络状况确定最优的数据传输路径，常用的路由协议包括基于距离向量的 RIP 协议、基于链路状态的 OSPF 协议、和基于路径向量的 BGP 协议。RIP 协议采用 Bellman-Ford 算法确定跳跃计数 (Hop Count) 最少的网络链路为最优传输路径并写入路由表。OSPF 协议采用 Dijkstra's 算法确定带宽最大的网络链路为最优传输路径并写入路由表，该协议用于同一电信运营商运营的具有统一路由策略的自治系统网络内（参见 RFC1930）。BGP 协议采用经过修改的 Bellman-Ford 算法（参见 RFC1322）根据一系列与网络链路相关的参数来确定最优传输路径并写入路由表，该协议用于不同自治系统网络之间（参见 RFC1930）。如果多条可用网络链路具有相同的优先级顺序，那么路由器会根据负载均衡原则将数据均匀地转发到各条链路上。在采用动态路由策略的网络中路由器会根据不同网络状态选择不同网络链路传输数据。在这种情况下，如果网络结构状态稳定，则网络路由器的每个输出端口所连接的网络链路所传输的数据内容不变，如果网络状态结构发生变化，则路由器可使用网络路由协议探测到该变化，进而更新其路由表，这样路由器输出端口所连网络链路上传输的数据内容就会发生变化。这种数据内容的变化可能反映在时间、用户、和网站任何一个维度上。

[0063] 除了动态路由策略外，路由器还可以采用人工设置的特定数据转发规则来选择网络链路，即静态路由策略。比如电信运营商可以对具有不同源或目的 IP 地址段的数据选择不同的网络链路，或者根据不同的数据类型（比如 HTTP 数据、流媒体数据、P2P 数据）选择不同的网络链路。通常静态路由策略规则直接作用在数据转发过程中，优先级高于基于网络协议的动态路由策略。在这种情况下，网络路由器的不同输出端口所连接的网络链路传输的数据具有不同且固定的内容特征，并且不同转发规则决定了不同链路数据在时间、用户、和网站这三个维度上的差异。

[0064] 图 2 示出路由器输入输出链路示意图。如图 2 所示，路由器 21 通过输入端口有 m 条输入数据链路，通过输出端口有 m' 条输出数据链路，数据采集点部署在输出数据链路上。路由器 21 可以采用不同的路由策略。下面根据路由器 21 的路由策略介绍如何确定数据采集点的部署。

[0065] 为了既获得反映用户群体真实上网行为的数据，又能有效控制采集成本，可以采

用统计采样方法在电信骨干网络上选择合适的数据采集点部署采集设备，并且确保采集到的数据样本对整体的代表性。因此根据不同的电信网络路由策略采取不同方式在电信网络路由器的输出端口所连接的网络链路上部署数据采集设备。

[0066] 对以动态路由策略配置的路由器，数据被转发到各个输出端口所连网络链路上的转发规则是根据当时的网络状态动态确定，可以根据电信运营商提供的运营经验规则在可靠性高稳定性好的网络链路上部署数据采集点。不同的采集点部署方式对采集到的数据样本的影响有两种情况：

[0067] 在一般情况下为了获得对数据整体状况有代表性的数据样本，可以在尽可能多的路由器输出端口所连网络链路上部署采集点，使得数据的样本量大且受网络变化影响小。一个极端情况是在所有输出端口网络链路上部署采集点以获得全体数据，这样数据特征就不受网络变化影响。但是在实际操作中，所能部署的采集点个数受限于系统建设成本。

[0068] 如果路由器将数据按照负载均衡原则转发到多条网络链路上，每条链路都以均等机会获得并传输数据，这种情况下可在任何一条链路上部署采集点，而获得的数据样本在统计意义上都能代表数据整体状况。这样，就可以大大减少采集点的部署，减少系统建设成本。

[0069] 对以静态路由策略配置的路由器，相应路由转发规则已经事先确定，各条路由器输出端口所连网络链路上的数据内容特征也已确定，可以根据实际需要决定在哪条网络链路上采集什么样的用户上网行为数据。例如，如果路由器将从不同源 IP 地址发来的数据转发至不同输出端口所连的网络链路上，可以根据自身需要决定需要采集的数据样本范围：或者在所有链路上部署采集点以获得全部源 IP 地址对应的数据，或者选择性地在特定的链路上部署采集点以获得部分源 IP 地址段的数据。

[0070] 根据本发明的一个实施例，在数据采集容量允许的情况下，在拓扑结构上层的数据链路上部署数据采集点，以覆盖更多的宽带用户。根据本发明的一个实施例，在由静态路由策略确定的所有数据传输链路上部署采集点，包括具有不同目的地址的数据由不同链路传输的情况（比如目的网站在省内设有站点），以实现完全覆盖上网行为在时间和网站维度上的分布。根据本发明的一个实施例，在多条负载均衡链路上任选一条部署采集点，即可获得在时间和网站维度上的具有准确统计意义的用户上网行为。根据本发明的一个实施例，在多条热备链路（即不同链路上数据相同）上均部署采集点，但在正常情况下只启用一条，若链路发生故障则启用其它链路上的采集点，以应对网络拓扑结构发生变化，实现对该链路上用户上网行为数据在时间和网站维度上的完全覆盖。对于冷备链路上可以不部署采集点，以节省成本。

[0071] 图 3 示出本发明的基于骨干网的用户上网数据处理系统的一个实施例的结构图。如图 3 所示，该系统包括多个数据采集设备 31、用户标识获取设备 32、上网数据存储设备 33 和描述信息提取设备 34。其中，数据采集设备 31 在骨干网上采集用户上网数据，将采集的用户上网数据发给用户标识获取设备 32。用户标识获取设备 32 根据从用户上网数据中提取的用户 IP 信息获得用户标识，将获得的用户标识发送给上网数据存储设备 33。上网数据存储设备 33 按照用户标识对用户上网数据进行存储。描述信息提取设备 34 从按照用户标识存储的用户上网数据获得用户的上网行为描述信息。例如，用户的上网行为描述信息包括访问时间、网站 IP 地址、网站 URL、页面文本标题、关键词、网站 cookie、和页面 Referrer

中的至少一个。

[0072] 根据本发明的一个实施例，对于采用动态路由策略决定数据转发路径的路由器：数据采集设备部署在骨干网根据 metric 信息选择的路由器输出端口网络链路上；和 / 或数据采集设备部署在骨干网的传输距离短或链路状态好的路由器输出端口网络链路上；和 / 或对于路由器将数据按照负载均衡原则转发到多条网络链路上，每条链路以均等机会获得并传输数据的情况，数据采集设备部署在从多条网络链路上选择任意一条链路。这种情况下可在任何一条链路上部署采集点，而获得的数据样本在统计意义上都能代表数据整体状况，就可以大大减少采集点的部署，减少系统建设成本。在路由策略与时间无关的情况下，数据采集设备部署在固定网络链路上对用户访问网站事件在时间上进行均匀的随机采样。这样随着采样时间的延长和样本数据的不断积累，采集得到的访问事件样本最终会在统计意义上趋于用户对网站的访问事件的全体集合。在这种情况下，可以减少采用时间，但仍然能够获得统计意义上用户对网站的访问事件的全体集合，减少了运营成本。

[0073] 下面从时间、用户、和网站三个维度描述宽带用户对互联网网站的访问行为。这种情况下，整个宽带用户群体对互联网网站的访问事件的集合可以表示在如图 4 所示的一个由时间、用户、和网站组成的三维空间中。在图 4 中，上述用户对网站的访问事件的三维图中只有时间坐标轴是连续有序排列的，用户和网站在相应坐标轴上的排列是离散且无序的，即不同的离散坐标值表示该维度属性上的不同个体，也就是说用户坐标轴上的每个离散坐标点表示一个用户、网站坐标轴上的每个离散坐标点表示一个网站。

[0074] 如果数据采集方法不能覆盖所有宽带用户对互联网网站的所有访问数据，那么其采集到的用户上网行为采集数据就是全部数据集合的子集。根据用户对网站的访问事件的三个描述维度，即时间、用户、和网站，宽带用户上网行为数据子集的采样效果可以用下列指标来衡量：

[0075] (1) 采集到的宽带用户群体访问事件样本的时间采样百分比 R_t ；

[0076] (2) 采集到的宽带用户群体访问事件样本的用户采样百分比 R_u ；

[0077] (3) 采集到的宽带用户群体访问事件样本的网站采样百分比 R_w 。

[0078] 【用户群体上网行为采样效果】

[0079] 下面分别从时间、用户、和网站三个维度来说明数据采样方案对所观察到的宽带用户群体对互联网网站的访问行为的影响。

[0080] 一. 时间维度采样

[0081] (1) 均匀采样

[0082] 无论是基于静态因素还是动态因素的路由策略，只要路由策略与时间没有关联，那么部署在固定网络链路上的采集点将对用户访问网站事件在时间上进行均匀的随机采样。此时所观察到的用户对网站的访问事件的集合将如图 5 所示。在这样的情况下所观察到的用户对网站的访问事件是全体集合的一个子集。但是随着采样时间的延长和样本数据的不断积累，采集得到的访问事件样本最终会在统计意义上趋于用户对网站的访问事件的全体集合。在这种情况下，可以减少采用时间，但仍然能够获得统计意义上用户对网站的访问事件的全体集合，减少了运营成本。

[0083] (2) 非均匀采样

[0084] 如果路由策略随时间而变化，那么部署在固定网络链路上的采集点将对用户访问

网站事件在时间上进行非均匀的随机采样。此时所观察到的用户对网站的访问事件的集合将如图 6 所示。在这样的情况下随着采样时间的延长和样本数据的不断积累,所观察到的用户对网站的访问事件子集在统计意义上反映了全体访问事件在时间维度上的简单或复杂的映射结果,而不会趋向于用户对网站的访问事件的全体集合。

[0085] 综合上述情况,在部署数据采集点时尽量选择优先级高、具备时间均匀特性路由策略的路由器输出数据链路作为数据采集路径,以确保获得充分的、且能代表整体用户上网行为特征的网络数据。这种情况下采集到的网络数据就是对其覆盖的宽带用户群体上网行为的一个估计,该估计的准确程度由采集到的用户群体对网站的访问时长占整个访问时间的百分比决定,即由采集到的宽带用户群体访问事件样本的时间采样百分比 R_t 决定。

[0086] 二, 用户维度采样

[0087] 根据电信网络的特点,特定用户的互联网访问数据是否由特定网络链路来传输的路由策略是相对固定的,即该用户的访问数据是否流经某条网络链路大多是由静态因素决定的,比如用户 IP 地址范围、用户所在区域的网络链路质量、用户所在区域的网络拓扑结构等。所以,在这样的情况下所观察到的用户群体采样也是固定的。如果观察到用户群体发生较大变化,那么很大程度上是因为静态的路由策略发生了变化所导致的;无论采样时间长短,所观察到的用户群体行为只描述了该用户群体采样的行为,而不能以此来估计未被观察到的用户群体的行为。

[0088] 此时所观察到的用户对网站的访问事件的集合将如图 7 所示。在部署数据采集点时尽量选择经过路由器汇聚的路由器输出端口所连接的数据链路作为数据采集路径,以确保覆盖该路由器输入端口所连接的数据链路所对应的所有宽带用户群体。这种情况下采集到的网络数据就是对其所应该覆盖的宽带用户群体上网行为的一个估计,该估计的准确程度由采集到的用户数目占整个用户群体的百分比决定,即由采集到的宽带用户群体访问事件样本的用户采样百分比 R_u 决定。

[0089] 三, 网站维度采样

[0090] 如果在特定的电信运营商 IDC 机房部署采集点,那么所获得的用户对网站的访问事件将只是全体访问事件集合的一个子集。由于特定 IDC 机房所包含的网站是相对固定的,于是与对用户采样的情况相似;

[0091] 在这样的情况下所观察到的用户访问的网站采样也是相对固定的。如果观察到网站发生较大变化,那么很大程度上是因为相应网络链路所连接的网站发生了变化所导致的;

[0092] 无论采样时间长短,所观察到的用户对网站的访问行为只描述了对相应网站集合的访问行为,而不能以此来估计用户在其所访问过但未被观察到的网站上的访问行为。

[0093] 此时所观察到的用户对网站的访问事件的集合将如图 8 所示。在这种情况下所观察到的宽带用户所访问的部分网站所对应的用户群体和访问时间仅是全体用户群体和全部访问时间的一个子集。因此无论采样时间的长短和样本数据的多少,采集到的子集数据只能反映这个子集所包含的用户群体在相应访问时间内的行为情况,而无法代表全部用户群体在任何时间访问全部网站的行为。这种情况下采集到的网络数据就是对访问这些网站的宽带用户群体的上网行为的一个估计,该估计的准确程度由采集到的网站占该用户群体所访问过的全部网站的百分比决定,即由采集到的宽带用户群体访问事件样本的网站采样

百分比 R_w 决定。

[0094] 因此,如果不能在用户数据汇聚的网络链路上部署数据采集点,那么会尽量选择连接着包含有大量大型互联网网站的电信运营商 IDC 机房的网络数据链路来部署数据采集点。

[0095] 由于在实际电信网络环境中对宽带用户群体上网行为的数据采样效果通常是在时间、用户、和网站三个维度上的组合形式,所以需要根据实际网络链路情况在数据采样效果和所需代价之间取得平衡。

[0096] 【用户个体上网行为采样效果】

[0097] 根据采集的宽带用户上网数据内容,个体用户的上网行为可以用其对网站页面的访问事件来描述。首先个体用户由其用户 UserID 标识,每个访问事件则记录了该用户访问某个网站时的时间信息和网站信息:

[0098] 时间信息:用户对单个网站的访问时间

[0099] 网站信息:网站 IP 地址、网站 URL、页面文本标题或用户提交的关键词、网站 Cookie、页面 Referrer

[0100] 因此将属于每个宽带用户的对互联网网站的访问事件归入到这个用户中,于是个体宽带用户的上网行为可以表示在一个由时间和网站组成的二维空间中。图 9 中显示了三个宽带用户对多个网站的访问行为。需要注意的是时间坐标轴是有序排列的,而网站坐标轴是无序排列的。

[0101] 如果数据采集方法不能覆盖所有宽带用户对互联网网站的所有访问数据,那么其采集到的用户上网行为采集数据就是全部数据集合的子集。根据个体用户对网站的访问事件的两个描述维度,即时间和网站,个体宽带用户上网行为数据子集的采样效果可以用下列指标来衡量:

[0102] (1) 采集到的个体宽带用户访问事件样本的时间采样百分比 R_t ;

[0103] (2) 采集到的个体宽带用户群体访问事件样本的网站采样百分比 R_w 。

[0104] 下面分别从时间和网站两个个维度来说明数据采样方案对所观察到的宽带用户个体对互联网网站的访问行为的影响。

[0105] 一,时间维度采样

[0106] (1) 均匀采样

[0107] 无论是基于网络因素的自适应路由策略还是基于人工设置的静态路由策略,只要路由策略与时间没有关联,那么部署在固定网络链路上的采集点将对个体用户访问网站事件在时间上进行均匀的随机采样。此时所观察到的个体用户对网站的访问事件的集合将如图 10 所示。图 10 中显示了三个宽带用户对多个网站的访问行为,因此在这样的情况下所观察到的个体用户对网站的访问事件是全体集合的一个子集,这意味着将不会知道任何在采样集合以外的访问事件。但是随着采样时间的延长和样本数据的不断积累,采集得到的访问事件样本最终会在统计意义上趋向于个体用户对网站的访问事件的全体集合。

[0108] (2) 非均匀采样

[0109] 如果路由策略随时间而变化,那么部署在固定网络链路上的采集点将对个体用户访问网站事件在时间上进行非均匀的随机采样。此时所观察到的个体用户对网站的访问事件的集合将如图 11 所示。图 11 中显示了三个宽带用户对多个网站的访问行为,因此在这

样的情况下随着采样时间的延长和样本数据的不断积累,所观察到的个体用户对网站的访问事件子集在统计意义上反映了全体访问事件在时间维度上的简单或复杂的映射结果,而不会趋向于用户对网站的访问事件的全体集合。

[0110] 综合上述情况,在部署数据采集点时尽量选择优先级高、具备时间均匀特性路由策略的路由器输出端口所连网络链路作为数据采集路径,以确保获得充分的、且能代表个体用户上网行为特征的网络数据。这种情况下采集到的网络数据就是对其覆盖的个体宽带用户上网行为的一个估计,该估计的准确程度由采集到的个体用户对网站的访问时长占整个访问时间的百分比决定,即由采集到的个体宽带用户访问事件样本的时间采样百分比 R_t 决定。

[0111] 二, 网站维度采样

[0112] 如果在特定的电信运营商 IDC 机房部署采集点,那么所获得的个体用户对网站的访问事件将只是全体访问事件集合的一个子集。由于特定 IDC 机房所包含的网站是相对固定的:

[0113] 在这样的情况下所观察到的个体用户访问的网站采样也是相对固定的。如果观察到网站发生较大变化,那么很大程度上是因为相应网络链路所连接的网站发生了变化所导致的;

[0114] 无论采样时间长短,所观察到的个体用户对网站的访问行为只描述了对相应网站集合的访问行为,而不能以此来估计个体用户在其所访问过但未被观察到的网站上的访问行为。

[0115] 此时所观察到的个体用户对网站的访问事件的集合将如图 12 所示。图 12 中显示了三个宽带用户对多个网站的访问行为,这些网站不含 [1, 2.5] 和 [4, 6] 两个区间内的网站。在这种情况下所观察到的个体宽带用户所访问的部分网站所对应的访问时间仅是该用户全部访问时间的一个子集。因此无论采样时间的长短和样本数据的多少,采集到的子集数据只能反映这个子集所包含的个体用户在相应访问时间内的行为情况,而无法代表该用户在任何时间访问全部网站的行为。这种情况下采集到的网络数据就是对该用户所访问网站的上网行为的一个采样样本,该样本的准确程度由采集到的网站占该用户所访问过的全部网站的百分比决定,即由采集到的个体宽带用户访问事件样本的网站采样百分比 R_w 决定。

[0116] 因此,可以在用户数据汇聚的网络链路上部署数据采集点,或者尽量选择连接着包含有大型互联网站的电信运营商 IDC 机房的网络数据链路来部署数据采集点。

[0117] 图 13 示出一个电信网络链路上数据采样点例子的示意图。例如在某个电信网络环境中,网络路由节点 H 的输入端口连接到三条网络链路路径 A → H、B → H、和 C → H,其输出端口连接到的另外三条网络链路路径 H → G、H → F、和 H → I。其中路径 H → F 具有比路径 H → G 和 H → I 更大的网络带宽,各条路径对应的网络链路成本值 (cost) 标注在图 3 的网络拓扑结构。

[0118] 对从节点 A、B、和 C 访问节点 E 的数据,网络路由节点 H 的路由策略如下:

[0119] ● 静态路由策略规定具有节点 A 的源 IP 地址段的数据由路径 H → G 传输;

[0120] ● 自适应路由策略规定节点 H 优先将数据转发到高带宽的网络链路即路径 H → F 上;

- [0121] ●因此这样的路由策略将形成如下从节点 A、B、和 C 访问节点 E 的数据转发情况：
- [0122] ●具有节点 A 的源 IP 地址段的数据由路径 H → G 传输；
- [0123] ●正常情况下其余数据由路径 H → F 传输，而路径 H → I 无数据传输成为备用链路；
- [0124] ●如果路径 H → F 中断，则：
- [0125] (1) 具有节点 A 的源 IP 地址段的数据仍由路径 H → G 传输；；
- [0126] (2) 其余数据则经路径 H → I 传输，因为路径 H → I → E 的成本值小于路径 H → G → F → E；
- [0127] ●如果路径 H → G 中断，则：
- [0128] (1) 具有节点 A 的源 IP 地址段的数据将丢失；
- [0129] (2) 其余数据仍经路径 H → F 传输，因为路径 H → F → E 的成本值小于路径 H → I → E；
- [0130] ●如果路径 H → F 和 H → G 同时中断，则：
- [0131] (1) 具有节点 A 的源 IP 地址段的数据将丢失；
- [0132] (2) 其余数据则经路径 H → I 传输；
- [0133] 在具有这样的路由策略的电信网络结构中，系统可采取如下方式部署数据采集点：
- [0134] ●根据静态路由策略的要求，系统必须在路径 H → G 上部署一个采集点以获得从节点 A 访问节点 E 的数据；
- [0135] ●根据自适应路由策略的要求，系统必须在路径 H → F 上部署一个采集点以获得在正常情况下从节点 B 和 C 访问节点 E 的数据；
- [0136] ●在网络结构因部分路径中断发生变化而使得自适应路由策略改变数据的传输路径的情况下，系统还需要在路径 H → I 上部署一个采集点，以确保采集到因网络结构变化而被重定向到这条路径上的数据；
- [0137] 在按照上述方式部署数据采集点时，在路由节点 H 输出端口采集到的宽带用户对互联网网站的访问行为在用户、时间、和网站三个维度上会呈现出如下效果特征：
- [0138] ●如果在网络路径 H → F 和 H → G 上都部署了采集点，那么在正常情况下系统采集到的数据将来自节点 A、B、和 C 的全部用户，并且覆盖全部用户的全部上网行为时间和全部访问过的网站。
- [0139] ●如果只在网络路径 H → F 而未在路径 H → G 上部署采集点，那么采集到的数据将不会包含来自节点 A 的用户群体、相应的访问时间、和访问过的网站，而对其余来自节点 B 和 C 的用户群体则覆盖他们全部上网行为时间和全部访问过的网站。
- [0140] ●如果在网络结构发生变化（比如路径 H → F 或 F → E 中断）使得数据被重定向到路径 H → I 上的情况下系统在该链路上部署了采集点：
- [0141] 那么系统仍将采集到所有被重定向的用户群体，以及覆盖他们全部上网行为时间和全部访问过的网站。否则这些数据将全部丢失，包含来相应的用户群体、访问时间、和访问过的网站。
- [0142] 同时如果是路径 F → E 中断，则由于经过路径 H → G 传输的数据将无法通过节点 F 到达 E，而且这部分数据的传输路径是由静态路由策略决定的，所以这部分数据将会丢失。

[0143] 下面举例说明电信骨干网络环境中的采集点部署。下文中, MTP(Media Technology Platform, 媒体技术平台)是本申请人的一个基于互联网宽带用户上网行为的为互联网网站提供根据用户偏好进行内容定制的智能化信息服务技术平台。

[0144] 图 14 示出一个电信运营商省级中心的骨干网络结构及其采集点部署示意图。在图 14 所示的电信运营商的省级中心网络中, 省内宽带用户先通过分别在各地市的汇聚层路由器汇总后接入省级中心, 并与 2 个省网路由器相连。一方面省网路由器与 2 个省内 IDC 机房的路由器相连, 使得省内用户可以直接访问 IDC 机房中的网站, 另一方面省网路由器也与电信运营商的 2 个集团路由器相连, 并通过该集团路由器与电信运营商在其它省份的省级中心相连, 这样省内宽带用户就可以通过省网路由器访问其它省份的网络资源。通常情况下为了增强网络结构的可靠性, 各同级路由器之间也直接相连, 比如省网路由器和 IDC 机房路由器。而且集团路由器直接与 IDC 机房相连, 使得外省用户可以不必经过省网路由器。同时各个机房路由器与网站之间(黄色连线)也采用直连方式, 以缩短传输路径, 提高传输效率和可靠性。

[0145] 对于路由选择, 通常情况下, 网络路由器采用自适应路由策略来选择数据转发路径, 如果多条转发路径的优先级相同, 则路由器采用负载均衡的方式随机且均匀地选择转发路径。在图 14 所示的省级中心网络中, 省网路由器采用多条 10G 的 POS 链路与 IDC 机房和集团路由器相连, 并采用自适应路由策略转发数据, 各条链路均匀地负担数据传输任务。

[0146] 对于采集点部署, 为了获得省内宽带用户的 HTTP 请求数据, MTP 在两个地方部署数据采集点:

[0147] (1) 在省网路由器与省内 IDC 机房路由器之间的连接链路上部署采集点, 以获得省内宽带用户对 IDC 机房内的网站访问所产生的 HTTP 请求数据;

[0148] (2) 在省网路由器与集团路由器之间的连接链路上部署采集点, 以获得省内宽带用户对省外网站访问所产生的 HTTP 请求数据。

[0149] 由于各条网络链路均匀地承载网络数据流量, 所以 MTP 在两个采集点上均只从多条网络链路中的任意一条上采用户访问网站时向网站发出的 HTTP 请求数据。

[0150] 对于数据采样效果, MTP 系统从上述两个采集点获得的用户对网站访问产生的 HTTP 请求数据样本具有如下特征:

[0151] (1) MTP 采集到的用户对网站的访问行为覆盖全部省内宽带用户, 无论用户访问的是省内还是省外的网站;

[0152] (2) MTP 采集到的用户对网站的访问事件样本在统计意义上逐渐趋近于用户对网站的访问事件的全体集合, 趋近速度与采样时间成正比;

[0153] (3) MTP 采集到的用户对网站的访问行为覆盖所有位于省内 IDC 机房的网站以及省外网站。

[0154] 本发明实施例的方法和系统, 实现了覆盖电信运营商省级中心所有宽带用户, 能够客观反映用户群体上网行为的统计特征, 客观反映单个用户上网行为的统计特征。

[0155] 本发明的描述是为了示例和描述起见而给出的, 而并不是无遗漏的或者将本发明限于所公开的形式。很多修改和变化对于本领域的普通技术人员而言是显然的。选择和描述实施例是为了更好说明本发明的原理和实际应用, 并且使本领域的普通技术人员能够理解本发明从而设计适于特定用途的带有各种修改的各种实施例。

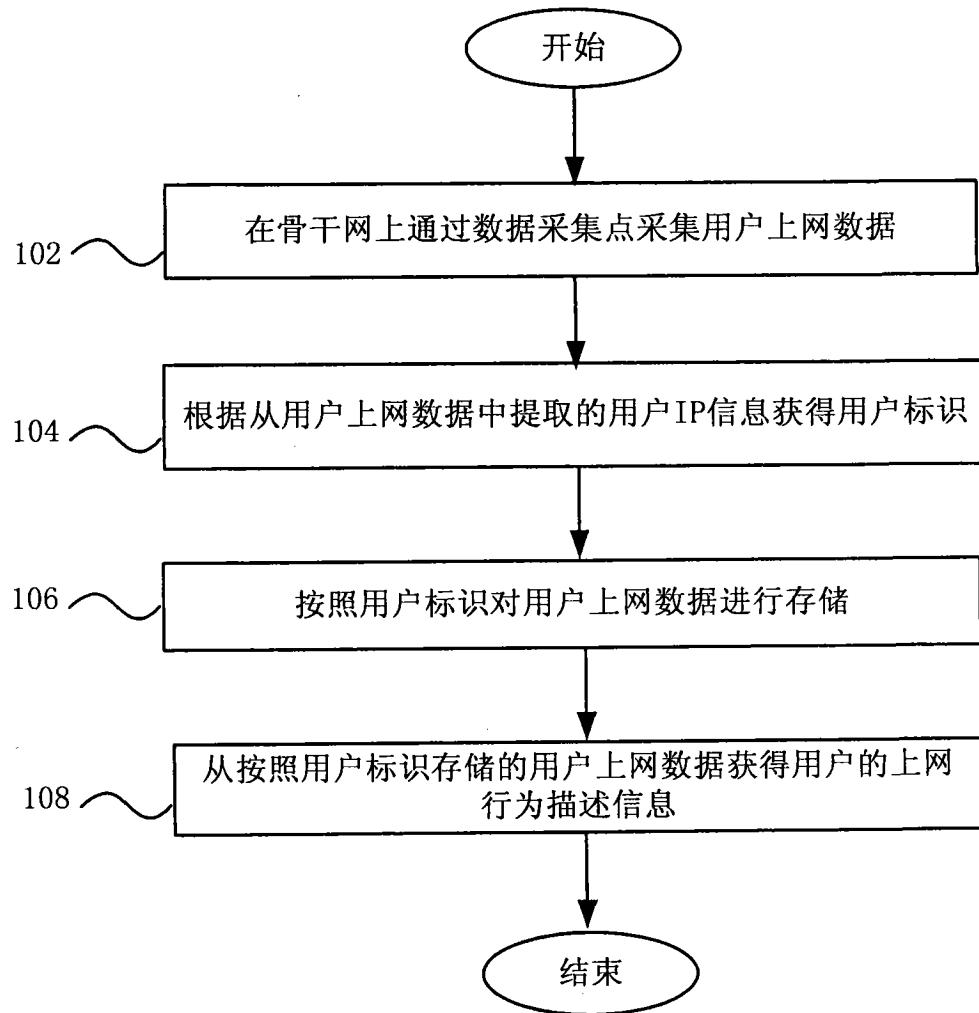


图 1

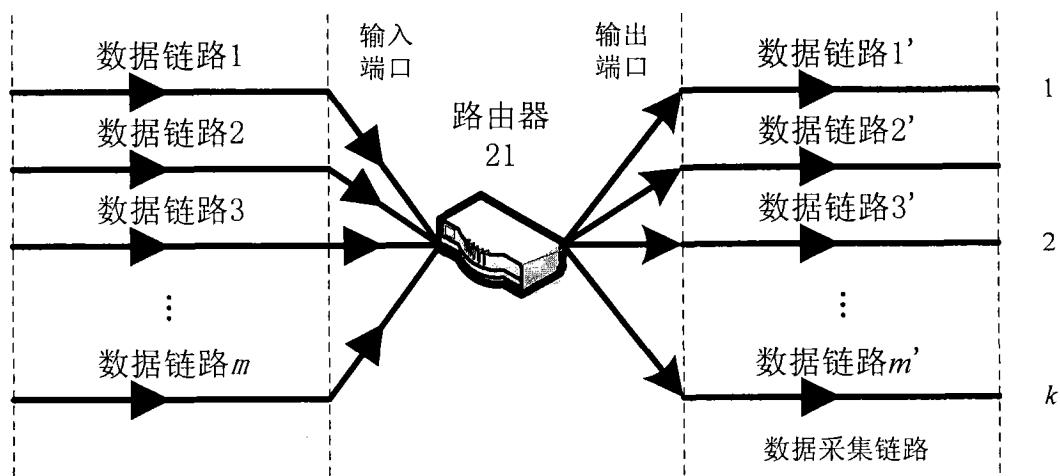


图 2

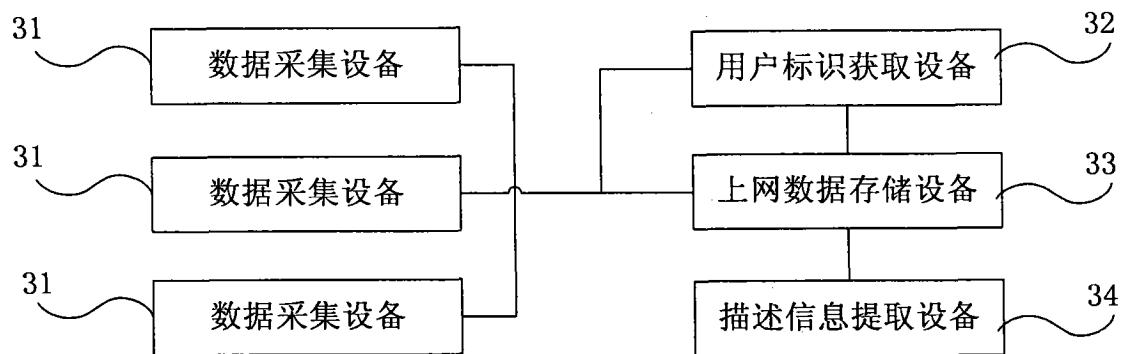


图 3

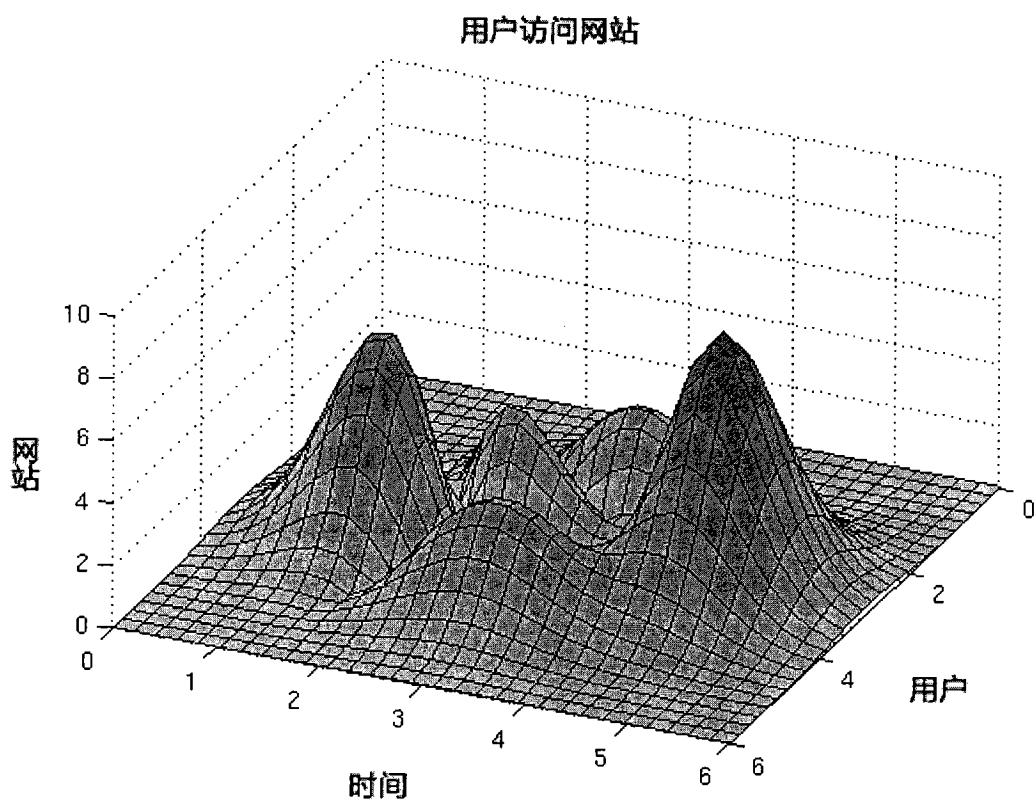


图 4

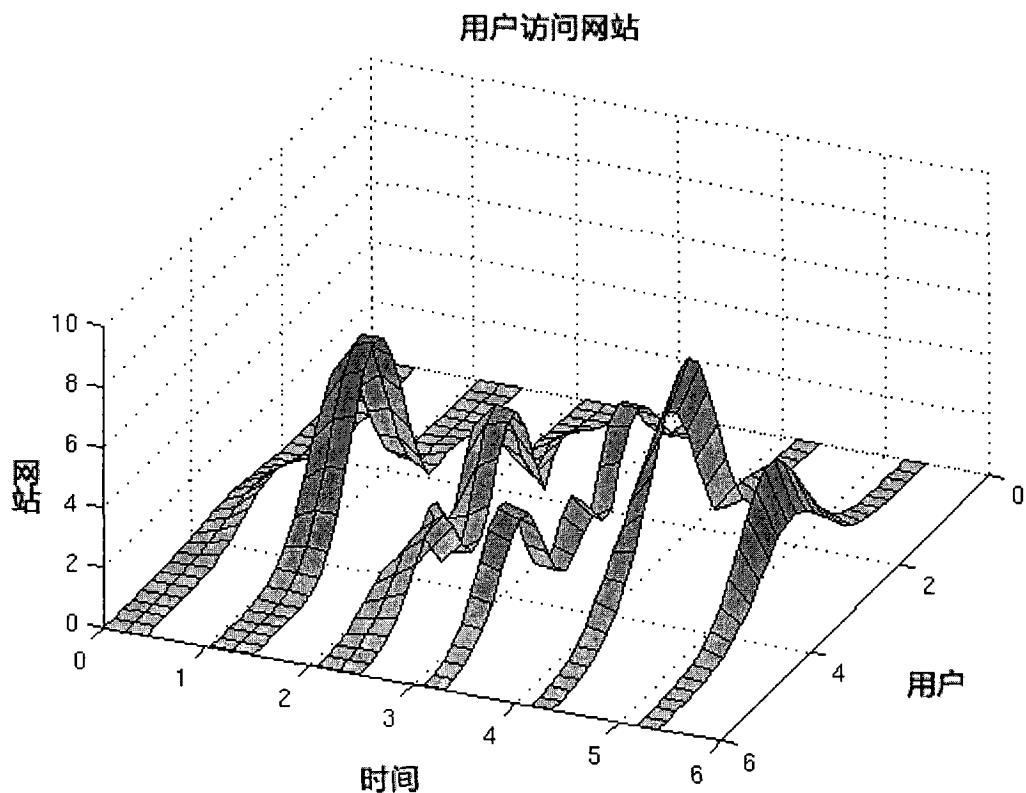


图 5

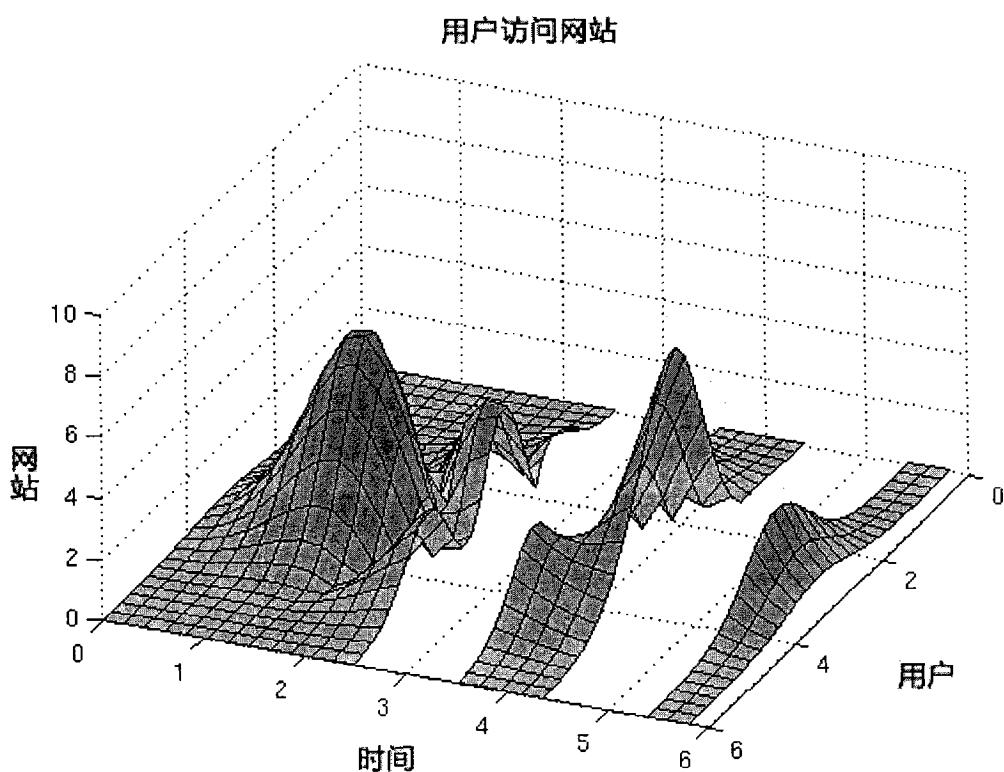


图 6

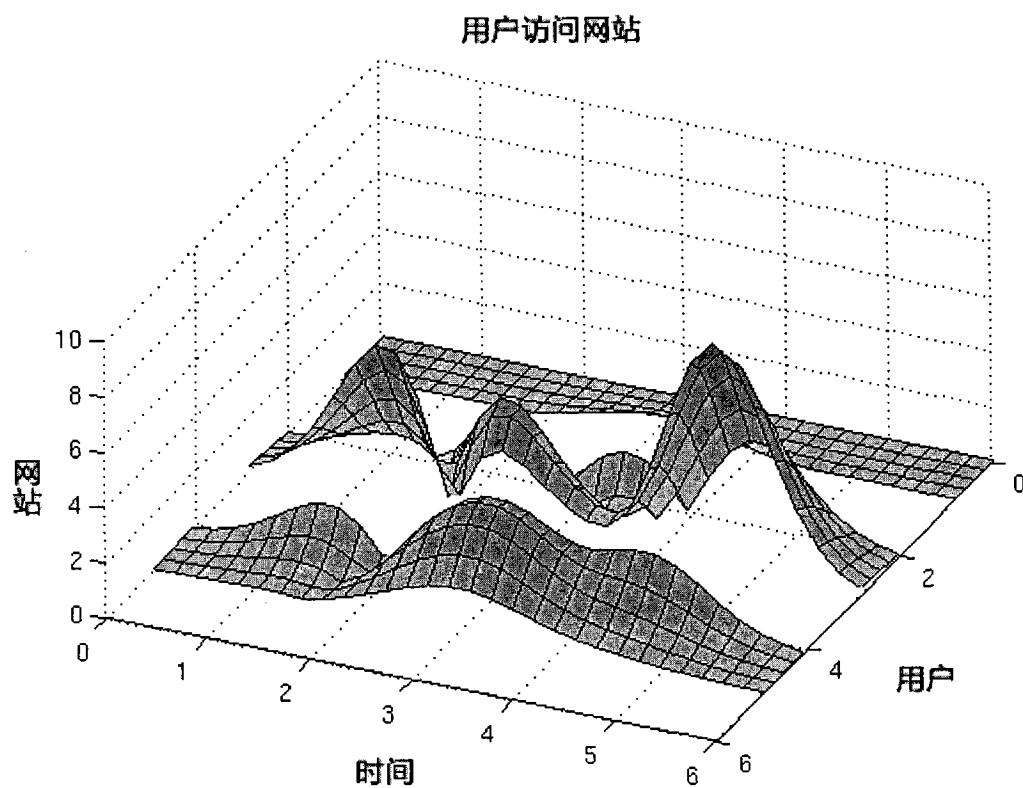


图 7

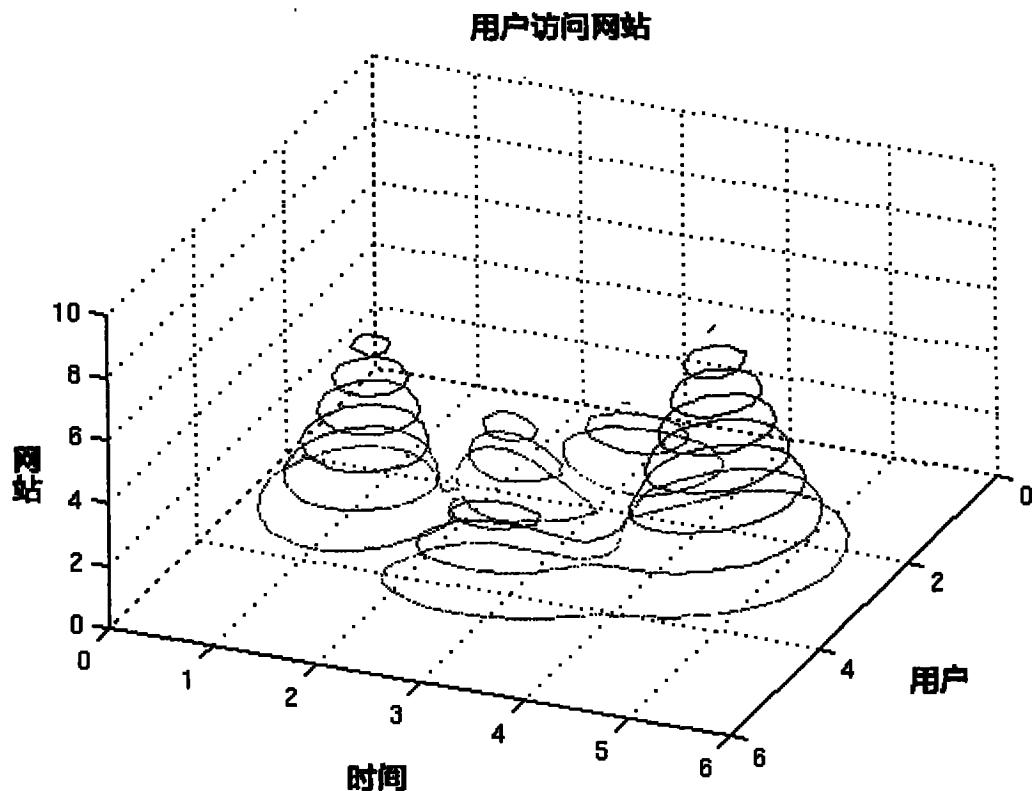


图 8

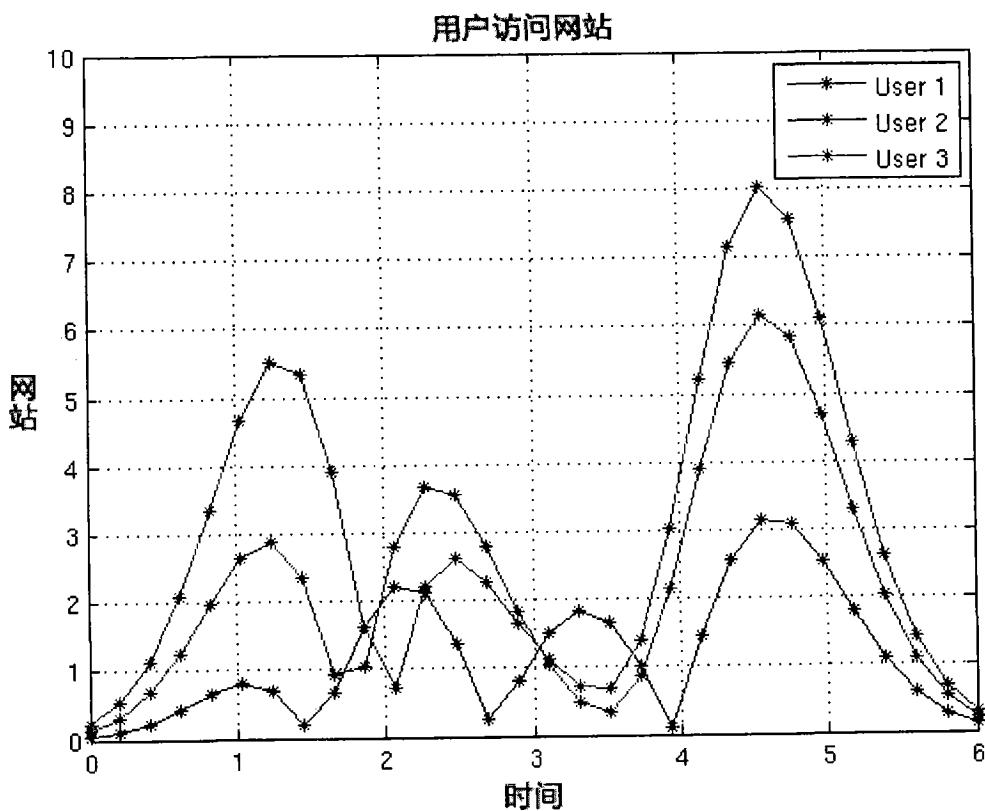


图 9

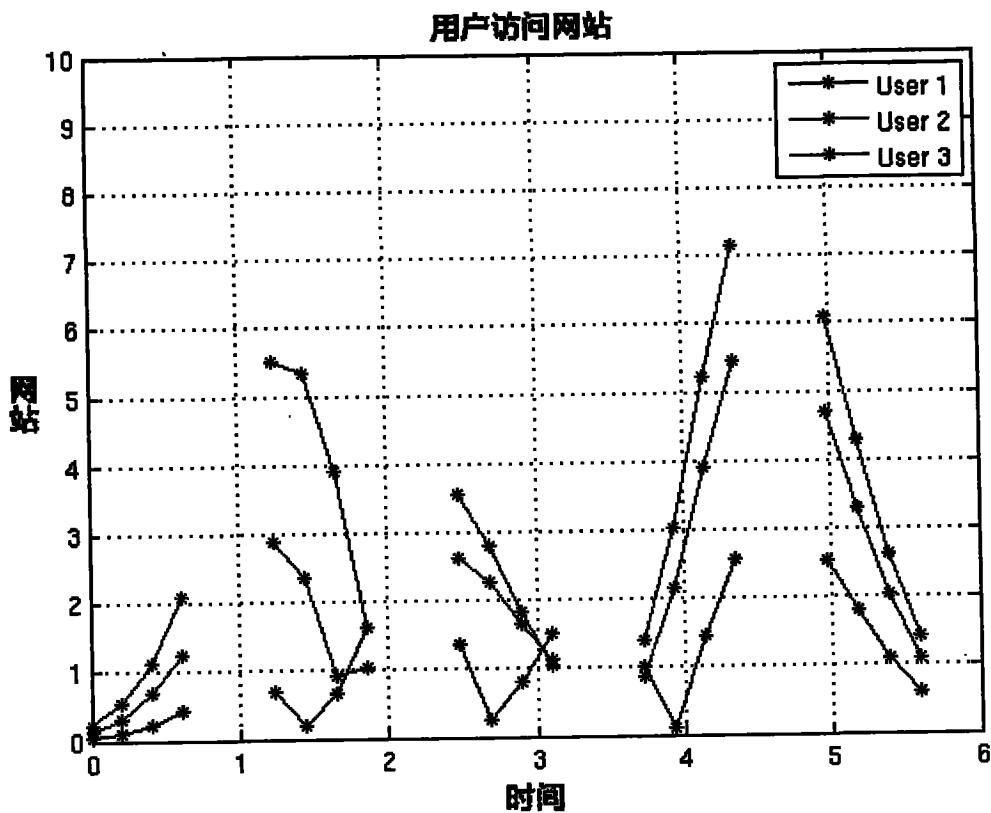


图 10

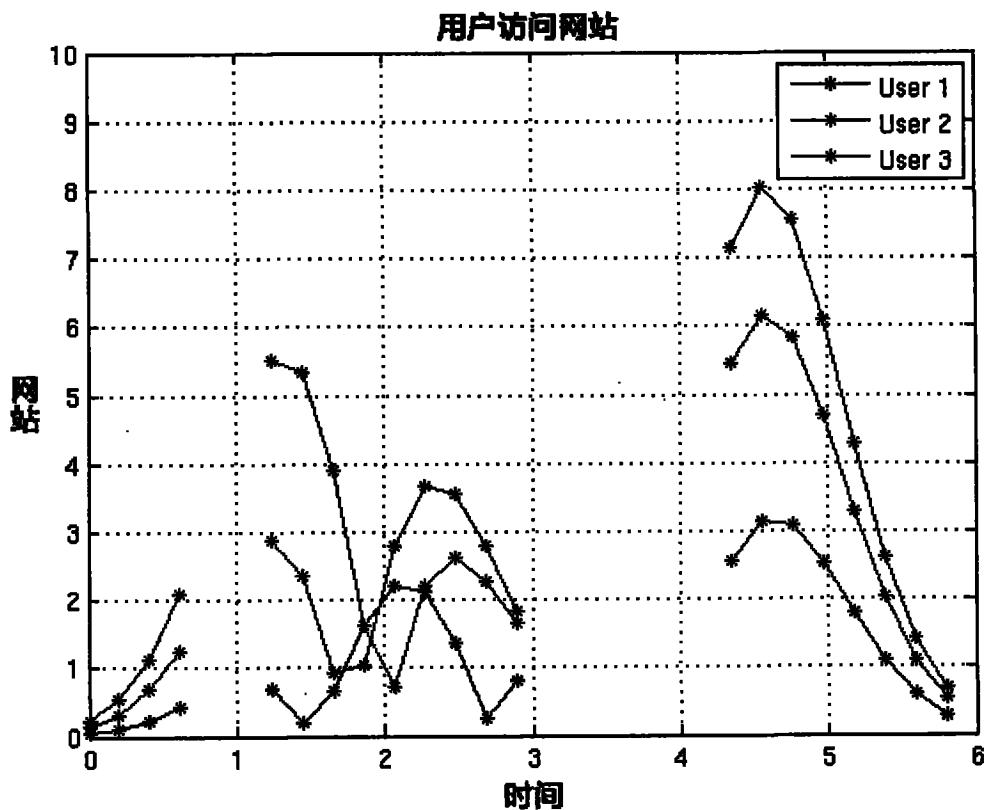


图 11

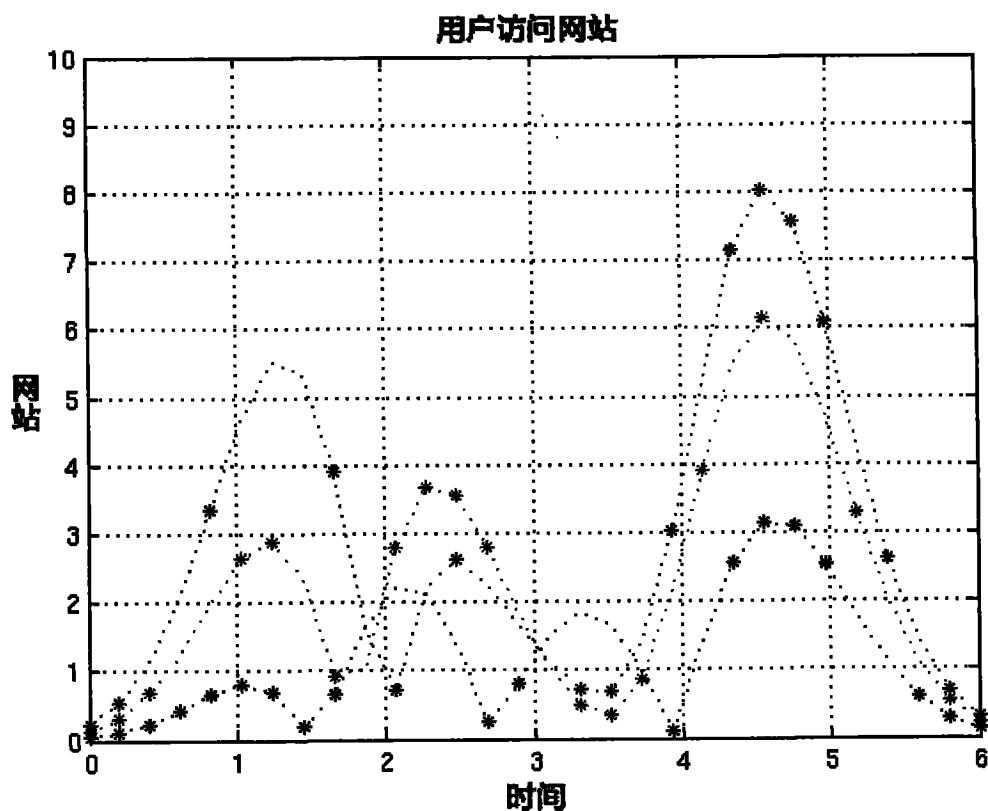


图 12

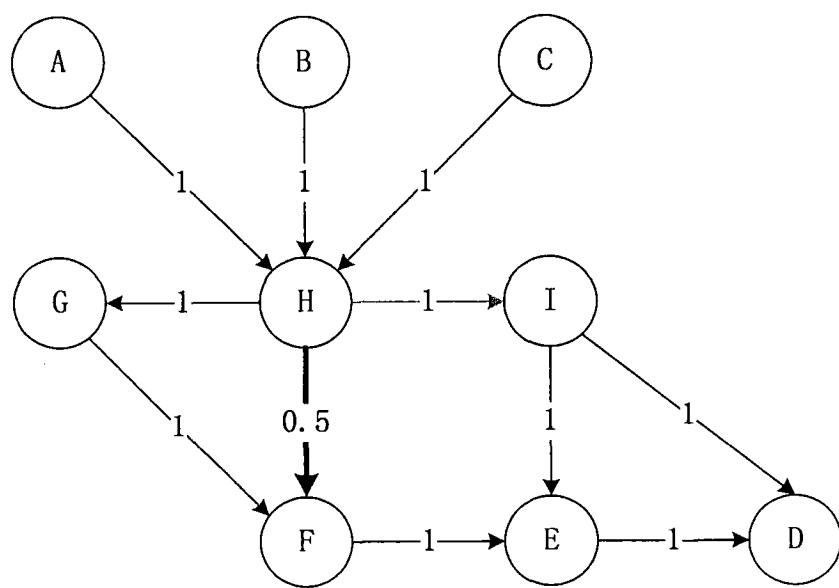


图 13

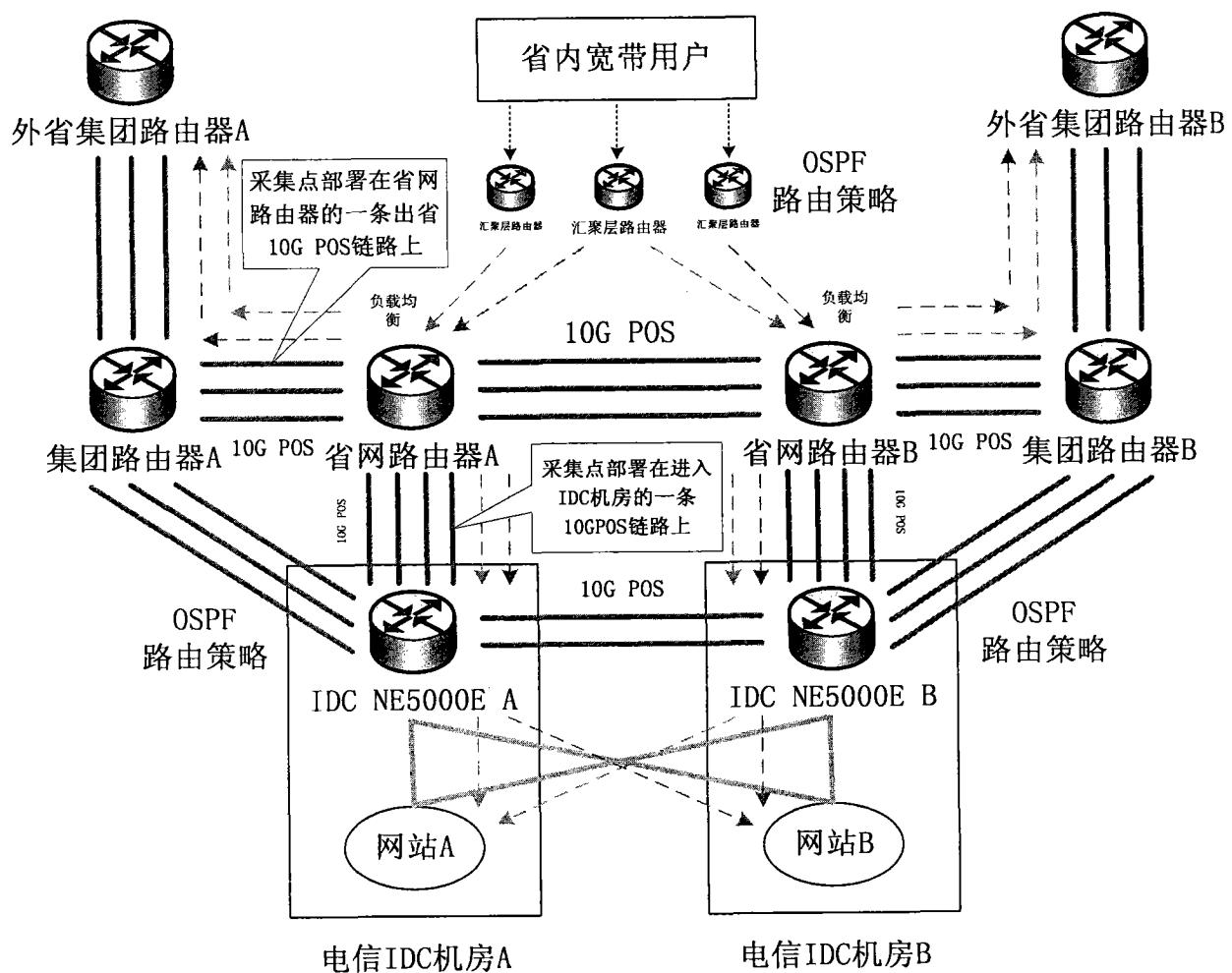


图 14