



(12)发明专利

(10)授权公告号 CN 103685198 B

(45)授权公告日 2017.03.01

(21)申请号 201210356956.6

(22)申请日 2012.09.21

(65)同一申请的已公布的文献号

申请公布号 CN 103685198 A

(43)申请公布日 2014.03.26

(73)专利权人 阿里巴巴集团控股有限公司

地址 英属开曼群岛大开曼岛资本大厦一座
四层847号邮箱

(72)发明人 赵雪波

(74)专利代理机构 北京同达信恒知识产权代理
有限公司 11291

代理人 郭润湘

(51)Int.Cl.

H04L 29/06(2006.01)

(56)对比文件

US 7761558 B1, 2010.07.20,

US 6141778 A, 2000.10.31,

CN 102143091 A, 2011.08.03,

审查员 谭美玲

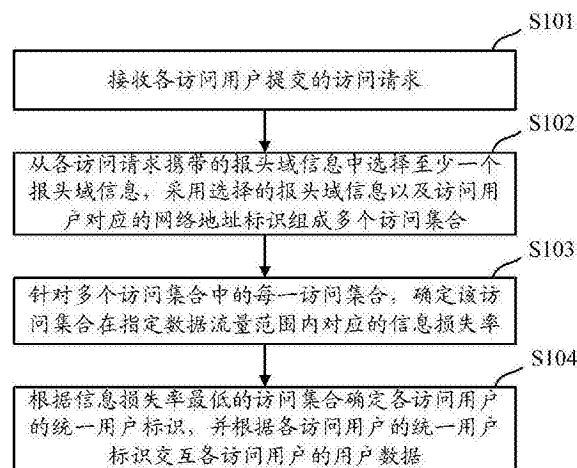
权利要求书2页 说明书9页 附图2页

(54)发明名称

一种交互用户数据的方法和装置

(57)摘要

本申请公开了一种交互用户数据的方法和装置,用于提高网站访问的安全性。该方法从各访问用户提交的访问请求中选择至少一个报头域信息和网络地址标识,组成多个访问集合,并确定每个访问集合对应的信息损失率,根据信息损失率最低的访问集合确定各访问用户的统一用户标识,并据此交互各访问用户的用户数据。由于上述过程中,网站无需向第三方服务器申请各访问用户对应的统一用户标识,因此减少了用户信息泄密机会,增加了网站访问的安全性。



1. 一种交互用户数据的方法,其特征在于,包括:

接收各访问用户提交的访问请求,访问请求中携带访问用户的初始用户标识、至少一个报头域信息和网络地址标识;

从各访问请求携带的报头域信息中选择至少一个报头域信息,采用选择的报头域信息以及访问用户对应的网络地址标识组成多个访问集合;

针对多个访问集合中的每一访问集合,确定该访问集合在指定数据流量范围内对应的信息损失率,其中,信息损失率表征了采用该访问集合确定各访问用户的统一用户标识的准确性;

根据信息损失率最低的访问集合确定各访问用户的所述统一用户标识,并根据各访问用户的所述统一用户标识交互各访问用户的用户数据。

2. 如权利要求1所述的方法,其特征在于,针对多个访问集合中的每一访问集合,确定该访问集合在指定数据流量范围内对应的信息损失率,具体包括:

针对每一访问集合,根据该访问集合包含的报头域信息的内容,将该访问集合重新组合为多个访问子集合,每一访问子集合的元素为报头域信息的内容和网络地址标识;

在所述指定数据流量范围内,统计每一访问子集合在各自对应的所有初始用户标识上承载的数据流量;

根据统计的每一访问子集合在各自对应的所有初始用户标识上承载的数据流量,确定该访问集合在所述指定数据流量范围内对应的信息损失率。

3. 如权利要求2所述的方法,其特征在于,所述确定该访问集合在指定数据流量范围内对应的信息损失率,根据以下公式公式确定:

$$1 - \frac{R}{W},$$

其中,

R表示该访问集合包含的每一访问子集合在各自对应的所有初始用户标识上承载的数据流量之和;

W表示指定数据流量范围内包含的所有数据流量之和。

4. 如权利要求3所述的方法,其特征在于,统计一个访问子集合在其对应的所有初始用户标识上承载的数据流量,具体包括:

在所述指定数据流量范围内,统计该访问子集合承载的数据流量、该访问子集合对应的初始用户标识数量,以及该访问子集合对应的每一初始用户标识对应的访问子集合的数量;

针对该访问子集合对应的每一初始用户标识,确定该访问子集合在该初始用户标识承载的数据流量;

将该访问子集合在其对应的每一初始用户标识上承载的数据流量之和,确定为该访问子集合在其对应的所有初始用户标识上承载的数据流量。

5. 如权利要求4所述的方法,其特征在于,按照以下公式确定该访问子集合在该初始用户标识承载的数据流量:

$$\frac{N}{i * j},$$

其中：

N为该访问子集合承载的数据流量；

i该访问子集合对应的初始用户标识数量；

j为该初始用户标识对应的访问子集合的数量。

6.如权利要求1所述的方法，其特征在于，根据信息损失率最低的访问集合确定各访问用户的统一用户标识之前，还包括：

确定信息损失率最低的访问集合对应的信息损失率不超过预设阈值。

7.如权利要求1所述的方法，其特征在于，当访问请求为超文本传输协议http访问请求时，该访问请求携带的报头域信息包括以下信息中的至少一种：用户接受的语言Accept-Language信息、用户接受的编码方式Accept-Encoding信息、用户代理User-Agent信息以及用户客户端信息。

8.如权利要求2所述的方法，其特征在于，根据信息损失率最低的访问集合确定各访问用户的统一用户标识，具体包括：

分别以信息损失率最低的访问集合中包含的不同访问子集合作为不同访问用户的统一用户标识。

9.一种交互用户数据的装置，其特征在于，包括：

接收单元，用于接收各访问用户提交的访问请求，访问请求中携带访问用户的初始用户标识、至少一个报头域信息和网络地址标识；

选择单元，用于从各访问请求携带的报头域信息中选择至少一个报头域信息以及访问用户对应的网络地址标识组成多个访问集合；

确定单元，用于针对每一访问集合，确定该访问集合在指定数据流量范围内对应的信息损失率，其中，信息损失率表征了采用该访问集合确定各访问用户的统一用户标识的准确性；

数据交互单元，用于根据信息损失率最低的访问集合确定各访问用户的所述统一用户标识，并根据各访问用户的所述统一用户标识交互各访问用户的用户数据。

10.如权利要求9所述的装置，其特征在于，所述确定单元，包括：

组合子单元，用于针对每一访问集合，根据该访问集合包含的报头域信息的内容，将该访问集合重新组合为多个访问子集合，每一访问子集合的元素为报头域信息的内容和网络地址标识；

统计子单元，用于在所述指定数据流量范围内，统计每一访问子集合在各自对应的所有初始用户标识上承载的数据流量；

确定子单元，用于根据统计的每一访问子集合在各自对应的所有初始用户标识上承载的数据流量，确定该访问集合在所述指定数据流量范围内对应的信息损失率。

一种交互用户数据的方法和装置

技术领域

[0001] 本申请涉及互联网技术领域,尤其涉及一种交互用户数据的方法和装置。

背景技术

[0002] 随着互联网技术的发展,各互联网网站之间通过在一定程度上交互访问用户的用户数据,从而能够更全面地了解用户的访问信息,使得各互互联网站能够根据用户的访问信息,为用户提供更加符合用户需求的产品。

[0003] 为了实现不同的互联网网站之间交互同一用户的用户数据的目的,各互联网网站需要确定一个唯一标识用于识别用户,并在交互用户数据时,根据该唯一标识交互该用户的用户数据。具体的,各互联网网站在用户访问时,可向统一服务器发送请求,以申请该用户的唯一标识,统一服务器根据网站发送的请求中携带的用户客户端特征生成该用户对应的唯一标识,并返回给各网站,各网站将该唯一标识写入用户客户端Cookie中,各互联网网站再根据该唯一标识交互该用户的用户数据。

[0004] 上述过程中,由于各网站均需要向作为第三方的统一服务器申请访问用户对应的唯一标识,并将申请到的唯一标识存储于用户客户端Cookie中,这存在用户信息泄密风险,降低了网站访问的安全性。

发明内容

[0005] 本申请提供一种交互用户数据的方法和装置,用于提高网站访问的安全性。

[0006] 本申请实施例提供一种交互用户数据的方法,包括:

[0007] 接收各访问用户提交的访问请求,访问请求中携带访问用户的初始用户标识、至少一个报头域信息和网络地址标识;

[0008] 从各访问请求携带的报头域信息中选择至少一个报头域信息,采用选择的报头域信息以及访问用户对应的网络地址标识组成多个访问集合;

[0009] 针对多个访问集合中的每一访问集合,确定该访问集合在指定数据流量范围内对应的信息损失率;

[0010] 根据信息损失率最低的访问集合确定各访问用户的统一用户标识,并根据各访问用户的统一用户标识交互各访问用户的用户数据。

[0011] 本申请实施例提供一种交互用户数据的装置,包括:

[0012] 接收单元,用于接收各访问用户提交的访问请求,访问请求中携带访问用户的初始用户标识、至少一个报头域信息和网络地址标识;

[0013] 选择单元,用于从各访问请求携带的报头域信息中选择至少一个报头域信息以及访问用户对应的网络地址标识组成多个访问集合;

[0014] 确定单元,用于针对每一访问集合,确定该访问集合在指定数据流量范围内对应的信息损失率;

[0015] 数据交互单元,用于根据信息损失率最低的访问集合确定各访问用户的统一用户

标识,并根据各访问用户的统一用户标识交互各访问用户的用户数据。

[0016] 本申请实施例提供一种交互用户数据的方法及装置,该方法从各访问用户提交的访问请求中选择至少一个报头域信息和网络地址标识,组成多个访问集合,并确定每个访问集合对应的信息损失率,根据信息损失率最低的访问集合确定各访问用户的统一用户标识,并据此交互各访问用户的用户数据。由于上述过程中,网站无需向第三方服务器申请各访问用户对应的统一用户标识,因此减少了用户信息泄密机会,增加了网站访问的安全性。

附图说明

[0017] 图1为本申请实施例提供的交互用户数据的过程;

[0018] 图2为本申请实施例提供的确定一个访问集合在指定流量范围内对应的信息损失率的过程;

[0019] 图3为本申请实施例提供的访问子集合与初始用户标识之间的对应关系示意图;

[0020] 图4为本申请实施例提供的交互用户数据的装置结构示意图。

具体实施方式

[0021] 由于现有技术中各网站之间交互用户数据时,对于每一个访问用户,需要利用该访问用户的一些相关信息向第三方服务器申请该访问用户对应的统一用户标识,因此,在此过程中存在一定的泄密风险,降低了网站访问的安全性。有鉴于此,本申请实施例利用各访问用户提交的访问请求中携带的报头域信息确定统一用户标识,而无需向第三方服务器申请统一用户标识,以达到降低用户信息泄漏风险、提高网站访问的安全性的目的。

[0022] 以下结合说明书附图对本申请的优选实施例进行说明,应当理解,此处所描述的优选实施例仅用于说明和解释本申请,并不用于限定本申请,并且在不冲突的情况下,本申请中的实施例及实施例中的特征可以相互组合。

[0023] 图1为本申请实施例提供的交互用户数据的过程,包括以下步骤:

[0024] S101:接收各访问用户提交的访问请求。

[0025] 其中,一个访问用户提交的访问请求中携带该访问用户的初始用户标识、至少一个报头域信息和网络地址标识。网络地址标识可以为访问用户对应的网络互联协议(Internet Protocol,IP)地址;用户的初始用户标识可以为用户cookie标识;当访问请求是超文本传输协议(hypertext transport protocol,http)访问请求时,该访问请求携带的报头域信息包括以下信息中的至少一种:Accept-Language(用户接受的语言)信息、Accept-Encoding(用户接受的编码方式)信息、User-Agent(用户代理)信息以及用户客户端信息,用户客户端信息包括:UA-Pixels(用户客户端像素)信息、UA-Color(用户客户端色彩)信息、UA-OS(用户客户端操作系统)信息、UA-CPU(用户客户端中央处理单元)信息等。另外,访问请求也可以是其他协议的访问请求,本申请中不具体限定。

[0026] S102:从各访问请求携带的报头域信息中选择至少一个报头域信息,采用选择的报头域信息以及访问用户对应的网络地址标识组成多个访问集合。

[0027] 为了便于描述,分别以A,B,C,D表示Accept-Language,Accept-Encoding,User-Agent和用户客户端信息,例如,对于网络地址相同的访问用户来说,假设访问用户对应的网络地址标识均为IP1,则根据该四种报头域信息之间的排列组合,可以组成15(

$C_4^1 + C_4^2 + C_4^3 + C_4^4$)种访问集合,分别为:[A,IP1],[B,IP1],[C,IP1],[D,IP1],[A,B,IP 1],[A,C,IP 1],[A,D,IP1],[B,C,IP1],[B,D,IP1],[C,D,IP1],[A,B,C,IP1],[A,B,D,IP1],[A,C,D,IP1],[B,C,D,IP1],[A,B,C,D,IP1]。

[0028] S103:针对多个访问集合中的每一访问集合,确定该访问集合在指定数据流量范围内对应的信息损失率。

[0029] 在本申请实施例中,首先可设置该指定数据流量范围,具体可以根据实际需要设置该数据流量范围。例如,该指定数据流量范围可以设置为指定时间段内(如1天内)的数据流量。又如,对于电子商务网站来说,其数据流量来源主要包括:登录页面与用户交互的数据流量、商品浏览页面与用户交互的数据流量、商品交易页面与用户交互的数据流量,因此,也可以将该指定数据流量范围设置为某一种或几种类型的页面与用户交互的数据流量。

[0030] 在设置了该指定数据流量范围之后,则可以分别确定通过上述步骤S102组成的每个访问集合在该指定数据流量范围内的信息损失率。

[0031] 对于一个访问集合来说,该访问集合在该指定数据流量范围内的信息损失率可以表征:采用该访问集合确定各访问用户的统一用户标识的准确性,或者说可信度。也即,该访问集合在该指定数据流量范围内的信息损失率越低,采用该访问集合确定各访问用户的统一用户标识的准确性越高,或者说可信度越大,反之,该访问集合在该指定数据流量范围内的信息损失率越高,采用该访问集合确定各访问用户的统一用户标识的准确性越低,或者说可信度越小。

[0032] S104:根据信息损失率最低的访问集合确定各访问用户的统一用户标识,并根据各访问用户的统一用户标识交互各访问用户的用户数据。

[0033] 由于一个访问集合在指定数据流量范围内的信息损失率越低,采用该访问集合确定各访问用户的统一用户标识的准确性越高,因此,在通过步骤S102组成的各访问集合中,选择出在指定数据流量范围内的信息损失率最低的访问集合,并根据选择出的该访问集合确定各访问用户的统一用户标识,可以尽量的提高确定的统一用户标识的准确性。后续则可以根据确定的各访问用户的统一用户标识交互各访问用户的用户数据。

[0034] 具体的,在上述图1所示的步骤S103中,针对每一访问集合,确定该访问集合在指定数据流量范围内对应的信息损失率的方法如图2所示。

[0035] 图2为本申请实施例提供的确定一个访问集合在指定流量范围内对应的信息损失率的过程,具体包括以下步骤:

[0036] S1031:根据该访问集合包含的报头域信息的内容,将该访问集合重新组合为多个访问子集合。

[0037] 其中,每一访问子集合的元素为报头域信息的内容和网络地址标识。

[0038] 例如,假设有3个用户提交了访问请求,这3个用户的初始用户标识分别为B₁、B₂、B₃,这3个访问用户的网络地址标识均为IP1,指定数据流量范围内的数据流量的数量为50个,现在要确定包含的元素为Accept-Language信息、Accept-Coding信息以及IP1的访问集合[Accept-Language信息,Accept-Coding信息,IP1]在指定流量范围内的信息损失率。

[0039] 如果这3个用户提交的访问请求中,Accept-Language信息的具体内容均为中文,而Accept-Coding信息的具体内容为编码方式1和编码方式2共两种,那么,对于网络地址标

识均为IP1的这3个访问用户,针对访问集合[Accept-Language信息,Accept-Coding信息,IP1]重组的访问子集合即为:[中文,编码方式1,IP1]和[中文,编码方式2,IP1]。

[0040] 为了便于描述,下面以A₁表示中文和英文,以C₁和C₂分别表示编码方式1和编码方式2进行说明。

[0041] 也即重组的访问子集合为:[A₁,C₁,IP1]和[A₁,C₂,IP1]。

[0042] S1032:在该指定数据流量范围内,统计每一访问子集合在各自对应的所有初始用户标识上承载的数据流量。

[0043] 由于访问请求中携带的报头域信息全部来自于标准协议,因此,对于不同访问用户来说,其提交的访问请求中携带的请求报头域信息的内容可能相同,也可能不同,也即,对于一个访问子集合来说,其对应的访问用户可能不只一个。另外,对于同一访问用户来说,其提交的不同的访问请求中所携带的报头域信息可能相同,也可能不同。

[0044] 继续沿用上例,假设:

[0045] 在初始用户标识B₁对应的访问用户提交的访问请求中,存在采用中文、编码方式1、IP1提交的访问请求,也存在采用中文、编码方式2、IP1提交的访问请求;

[0046] 在初始用户标识B₂对应的访问用户提交的访问请求中,存在采用中文、编码方式1、IP1提交的访问请求,也存在采用中文、编码方式2、IP1提交的访问请求;

[0047] 在初始用户标识B₃对应的访问用户提交的访问请求中,只存在采用中文、编码方式1、IP1提交的访问请求。

[0048] 则:访问子集合[A₁,C₁,IP1]和[A₁,C₂,IP1],与用户初始标识B₁、B₂、B₃的对应关系如图3所示。图3为本申请实施例提供的访问子集合与初始用户标识之间的对应关系示意图,在图3中,访问子集合[A₁,C₁,IP1]与初始用户标识B₁、B₂、B₃对应,访问子集合[A₁,C₂,IP1]与初始用户标识B₁、B₂对应。

[0049] 因此,在指定数据流量范围内,也即在50个数据流量中,统计[A₁,C₁,IP1]在对应的B₁、B₂、B₃上承载的数据流量,统计[A₁,C₂,IP1]在对应的B₁、B₂上承载的数据流量。

[0050] 具体的,统计一个访问子集合在其对应的所有初始用户标识上承载的数据流量的方法为:在指定数据流量范围内,统计该访问子集合承载的数据流量、该访问子集合对应的初始用户标识数量,以及该访问子集合对应的每一初始用户标识对应的访问子集合的数量;针对该访问子集合对应的每一初始用户标识,确定该访问子集合在该初始用户标识承载的数据流量;将该访问子集合在其对应的每一初始用户标识上承载的数据流量之和,确定为该访问子集合在其对应的所有初始用户标识上承载的数据流量。

[0051] 其中,可以采用公式 $\frac{N}{i*j}$ 确定一个访问子集合在其对应的一个初始用户标识承载的数据流量,N为该访问子集合承载的数据流量,i为该访问子集合对应的初始用户标识数量,j为该初始用户标识对应的访问子集合的数量。

[0052] 继续沿用上例,假设在这指定的50个数据流量中,有30个承载在[A₁,C₁,IP1]上,有20个承载在[A₁,C₂,IP1]上,则:

[0053] 针对[A₁,C₁,IP1],N=30,其对应的初始用户标识分别为B₁、B₂、B₃,i=3,初始用户标识B₁对应的访问子集合分别为[A₁,C₁,IP1]和[A₁,C₂,IP1],j=2,因此,[A₁,C₁,IP1]承载在B₁上的数据流量为((30/3)/2)=5个;

[0054] 类似的, $[A_1, C_1, IP1]$ 承载在 B_2 上的数据流量为 $((30/3)/2)=5$ 个, $[A_1, C_1, IP1]$ 承载在 B_3 上的数据流量为 $(30/3)=10$ 个, 从而, 访问子集合 $[A_1, C_1, IP1]$ 在其在对应的所有初始用户标识上承载的数据流量为 $(5+5+10)=20$ 个。

[0055] 相应的, $[A_1, C_2, IP1]$ 承载在 B_1 上的数据流量为 $((20/2)/2)=5$ 个, $[A_1, C_2, IP1]$ 承载在 B_2 上的数据流量为 $((20/2)/2)=5$ 个, 因此访问子集合 $[A_1, C_2, IP1]$ 在其对应的所有初始用户标识上承载的数据流量为 $(5+5)=10$ 个。

[0056] S1033: 根据统计的每一访问子集合在各自对应的所有初始用户标识上承载的数据流量, 确定该访问集合在指定数据流量范围内对应的信息损失率。

[0057] 具体的, 可以采用公式 $1 - \frac{R}{W}$ 确定该访问集合在指定数据流量范围内对应的信息损失率, R 表示该访问集合包含的每一访问子集合在各自对应的所有初始用户标识上承载的数据流量之和, W 表示指定数据流量范围内包含的所有数据流量之和。

[0058] 继续沿用上例, 由于 $[A_1, C_1, IP1]$ 在其在对应的所有初始用户标识上承载的数据流量为 20 个, 访问子集合 $[A_1, C_2, IP1]$ 在其对应的所有初始用户标识上承载的数据流量为 10 个, 因此 $R=20+10=30$, $W=50$, 从而, 访问集合 $[Accept-Language$ 信息, $Accept-Coding$ 信息, $IP1]$ 在这 50 个数据流量内对应的信息损失率为: $1-30/50=0.4$ 。

[0059] 以上是本申请实施例提供的确定一个访问集合在指定数据流量范围内的信息损失率的方法。在后续过程中, 则可以选择出信息损失率最低的访问集合确定各访问用户的统一用户标识, 并根据各访问用户的统一用户标识交互各访问用户的用户数据。

[0060] 其中, 根据信息损失率最低的访问集合确定各访问用户的统一用户标识的方法具体为: 分别以信息损失率最低的访问集合中包含的不同访问子集合作为不同访问用户的统一用户标识。

[0061] 继续沿用上例, 如果访问集合 $[Accept-Language$ 信息, $Accept-Coding$ 信息, $IP1]$ 的信息损失率最低, 则以该访问集合 $[Accept-Language$ 信息, $Accept-Coding$ 信息, $IP1]$ 中包含的访问子集合 $[A_1, C_1, IP1]$ 作为一个访问用户的同一用户标识, 以访问子集合 $[A_1, C_2, IP1]$ 作为另一个访问用户的同一用户标识。

[0062] 需要说明的是, 在上例中, 初始用户标识有 3 个, 也即有 3 个访问用户, 而最终确定出的统一用户标识只有 2 个, 用这 2 个统一用户标识只能标识出 2 个访问用户, 所以, 采用一个访问集合中包含的每个访问子集合标识访问用户时, 其标识的准确性会下降。而其准确性下降的程度, 就是通过上述计算的信息损失率来表征的, 因此, 本申请实施例中要根据信息损失率最小的访问集合来确定统一用户标识。

[0063] 进一步的, 理想的情况是: 找到一种访问集合, 这种访问集合中包含的每个访问子集合与每个初始用户标识是完全一一对应的关系。这样, 用该访问集合中包含的不同访问子集合来标识不同访问用户的准确率是 100%。但是, 实际应用中几乎找不到这样的访问集合, 也即, 对于一个访问集合来说, 如果其包含的每个访问子集合与每个初始用户标识并不是一一对应的关系, 则用该访问集合中的访问子集合标识访问用户时, 就会造成准确率的下降。造成准确率下降的情况有以下两种:

[0064] 情况一、不同的访问用户采用同一个访问子集合承载数据流量时, 该同一个访问子集合并未标识出不同的用户。例如, 上例中的初始用户标识为 B_1 、 B_2 、 B_3 的 3 个访问用户, 这

3个访问用户都采用访问子集合 $[A_1, C_1, IP1]$ 承载了数据流量,但是访问子集合 $[A_1, C_1, IP1]$ 只能标识一个访问用户,也即,同样采用 $[A_1, C_1, IP1]$ 承载数据流量的3个访问用户并未被区分出来。

[0065] 情况二、同一个访问用户采用不同的访问子集合承载数据流量时,该同一个用户承载的数据流量会被标识为不同的用户承载的数据流量。例如,上例中的初始用户标识为 B_1 的访问用户,该访问用户采用访问子集合 $[A_1, C_1, IP1]$ 和 $[A_1, C_2, IP1]$ 承载了数据流量,但是, $[A_1, C_1, IP1]$ 和 $[A_1, C_2, IP1]$ 分别标识了不同的用户,也即,同样是初始用户标识为 B_1 的访问用户承载的数据流量却被标识成了两个不同的用户承载的数据流量。

[0066] 因此,上述计算的一个访问集合在指定数据流量访问内的信息损失量,就是在计算采用这个访问集合中的每个访问子集合标识每个访问用户时,由于上述情况一和情况二所带来的准确率的下降。

[0067] 从而,针对上例中的计算方法也可以这样理解:

[0068] 如图3所示,针对访问子集合 $[A_1, C_1, IP1]$,由于3个访问用户都采用了 $[A_1, C_1, IP1]$ 承载数据流量,因此基于情况一,无论采用 $[A_1, C_1, IP1]$ 标识哪个访问用户,其准确率也只有 $\frac{1}{3}$ 。进一步的,在基于情况一的基础上,当采用 $[A_1, C_1, IP1]$ 标识 B_1 对应的访问用户时,基于情况二, B_1 对应的访问用户采用 $[A_1, C_2, IP1]$ 承载的数据流量也会被标识为其他用户承载的数据流量,因此采用 $[A_1, C_1, IP1]$ 标识 B_1 对应的访问用户的准确率为 $\frac{1}{3} \times \frac{1}{2}$ 。同样的,采用 $[A_1,$

$C_1, IP1]$ 标识 B_2 对应的访问用户的准确率为 $\frac{1}{3} \times \frac{1}{2}$,采用 $[A_1, C_1, IP1]$ 标识 B_3 对应的访问用户的准确率为 $\frac{1}{3} \times 1$ 。因此,采用访问子集合 $[A_1, C_1, IP1]$ 标识一个访问用户的准确率为

$$\frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times 1。$$

[0069] 类似的,针对访问子集合 $[A_1, C_2, IP1]$,由于两个访问用户(B_1 对应的访问用户和 B_2 对应的访问用户)采用了 $[A_1, C_2, IP1]$ 承载数据流量,因此基于情况一,无论采用 $[A_1, C_2, IP1]$ 标识这两个中的哪个访问用户,其准确率也只有 $\frac{1}{2}$ 。进一步的,在基于情况一的基础上,

当采用 $[A_1, C_2, IP1]$ 标识 B_1 对应的访问用户时,基于情况二, B_1 对应的访问用户采用 $[A_1, C_1, IP1]$ 承载的数据流量也会被标识为其他用户承载的数据流量,因此采用 $[A_1, C_2, IP1]$ 标识 B_1 对应的访问用户的准确率为 $\frac{1}{2} \times \frac{1}{2}$ 。同样的,采用 $[A_1, C_2, IP1]$ 标识 B_2 对应的访问用户的

准确率为 $\frac{1}{2} \times \frac{1}{2}$ 。因此,采用访问子集合 $[A_1, C_2, IP1]$ 标识一个访问用户的准确率为

$$\frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2}。$$

[0070] 再考虑到 $[A_1, C_1, IP1]$ 承载的数据流量在总流量中所占的比例 $\frac{30}{50}$ 和 $[A_1, C_2, IP1]$ 承

载的数据流量在总流量中所占的比例 $\frac{20}{50}$,则在这50个数据流量中,采用访问集合[Accept-Language信息,Accept-Coding信息,IP1]中包含的两个访问子集合分别标识不同的访问用户的准确率即为: $\frac{30}{50}\left(\frac{1}{3}\times\frac{1}{2}+\frac{1}{3}\times\frac{1}{2}+\frac{1}{3}\times 1\right)+\frac{20}{50}\left(\frac{1}{2}\times\frac{1}{2}+\frac{1}{2}\times\frac{1}{2}\right)=0.6$ 。因此,其准确率下降了 $1-0.6=0.4$,也即,信息损失率是0.4。

[0071] 从而,本申请实施例实际上是寻找一种包含的每个访问子集合尽可能与每个初始用户标识接近于一一对应的访问集合,也即,采用信息损失率最低的访问集合确定访问用户的统一用户标识。

[0072] 较佳地,为了保证采用访问集合对访问用户进行标识的可用性,本申请实施例中,可以预先设置信息损失率阈值,当信息损失率最低的访问集合对应的信息损失率不超过预设阈值时,则信息损失率最低的访问集合中包含的访问子集合可以用于标识访问用户,当信息损失率最低的访问集合对应的信息损失率超过预设阈值时,说明采用该访问集合中包含的访问子集合标识访问用户的准确率过低,因此该访问集合中包含的访问子集合无法用于标识访问用户。

[0073] 在本申请实施例中,由于同一用户在不同网站有不同的用户数据,例如在某搜索网站有对关键词搜索的数据,在门户网站有对各频道页面的访问数据,在电子商务网站有商品浏览、收藏以及交易等行为数据,因此在每个网站可以分别保存同一用户的不同数据。在确定出统一用户标识后,就可以根据该用户的统一用户标识对该用户在不同网站的不同数据进行拼接,使得用户数据更加完整,后续各网站根据自身需要可以用于优化搜索、个性化页面推送、广告推送优化等应用中。

[0074] 基于同一申请构思,本申请实施例中还提供了一种交互用户数据的装置,由于该装置解决问题的原理与网站访问用户标识确定方法相似,因此该装置的实施可以参见方法的实施,重复之处不再赘述。

[0075] 图4为本申请实施例提供的交互用户数据的装置结构示意图,具体包括:

[0076] 接收单元401,用于接收各访问用户提交的访问请求,访问请求中携带访问用户的初始用户标识、至少一个报头域信息和网络地址标识;

[0077] 选择单元402,用于从各访问请求携带的报头域信息中选择至少一个报头域信息以及访问用户对应的网络地址标识组成多个访问集合;

[0078] 确定单元403,用于针对每一访问集合,确定该访问集合在指定数据流量范围内对应的信息损失率;

[0079] 数据交互单元404,用于根据信息损失率最低的访问集合确定各访问用户的统一用户标识,并根据各访问用户的统一用户标识交互各访问用户的用户数据。

[0080] 所述确定单元403,包括:

[0081] 组合子单元4031,用于针对每一访问集合,根据该访问集合包含的报头域信息的内容,将该访问集合重新组合为多个访问子集合,每一访问子集合的元素为报头域信息的内容和网络地址标识;

[0082] 统计子单元4032,用于在所述指定数据流量范围内,统计每一访问子集合在各自对应的所有初始用户标识上承载的数据流量;

[0083] 确定子单元4033,用于根据统计的每一访问子集合在各自对应的所有初始用户标识上承载的数据流量,确定该访问集合在所述指定数据流量范围内对应的信息损失率。

[0084] 所述确定子单元4033具体用于,按照以下公式确定该访问集合在所述指定数据流量范围内对应的信息损失率: $1 - \frac{R}{W}$,其中:R表示该访问集合包含的每一访问子集合在各自对应的所有初始用户标识上承载的数据流量之和;W表示指定数据流量范围内包含的所有数据流量之和。

[0085] 所述统计子单元4032,包括:

[0086] 统计模块40321,用于针对每一访问子集合,在所述指定数据流量范围内,统计该访问子集合承载的数据流量、该访问子集合对应的初始用户标识数量,以及该访问子集合对应的每一初始用户标识对应的访问子集合的数量;

[0087] 第一确定模块40322,用于针对该访问子集合对应的每一初始用户标识,确定该访问子集合在该初始用户标识承载的数据流量;

[0088] 第二确定模块40323,用于将该访问子集合在其对应的每一初始用户标识上承载的数据流量之和,确定为该访问子集合在对应的所有初始用户标识上承载的数据流量。

[0089] 所述第一确定模块40322具体用于,按照以下公式确定该访问子集合在该初始用户标识承载的数据流量: $\frac{N}{i*j}$,其中:N为该访问子集合承载的数据流量;i为该访问子集合对应的初始用户标识数量;j为该初始用户标识对应的访问子集合的数量。

[0090] 所述数据交互单元404还用于,在根据信息损失率最低的访问集合确定各访问用户的统一用户标识之前,确定信息损失率最低的访问集合对应的信息损失率不超过预设阈值。

[0091] 所述数据交互单元404具体用于,分别以信息损失率最低的访问集合中包含的不同访问子集合作为不同访问用户的统一用户标识。

[0092] 具体的上述交互用户数据的装置可以位于网站的服务器中。

[0093] 本申请实施例提供一种交互用户数据的方法及装置,该方法从各访问用户提交的访问请求中选择至少一个报头域信息和网络地址标识,组成多个访问集合,并确定每个访问集合对应的信息损失率,根据信息损失率最低的访问集合确定各访问用户的统一用户标识,并据此交互各访问用户的用户数据。由于上述过程中,网站无需向第三方服务器申请各访问用户对应的统一用户标识,因此减少了用户信息泄密机会,增加了网站访问的安全性。

[0094] 本领域内的技术人员应明白,本申请的实施例可提供为方法、系统、或计算机程序产品。因此,本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且,本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0095] 本申请是参照根据本申请实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产

生一个机器,使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0096] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中,使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品,该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0097] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上,使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理,从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0098] 尽管已描述了本申请的优选实施例,但本领域内的技术人员一旦得知了基本创造性概念,则可对这些实施例做出另外的变更和修改。所以,所附权利要求意欲解释为包括优选实施例以及落入本申请范围的所有变更和修改。

[0099] 显然,本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的精神和范围。这样,倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内,则本申请也意图包含这些改动和变型在内。

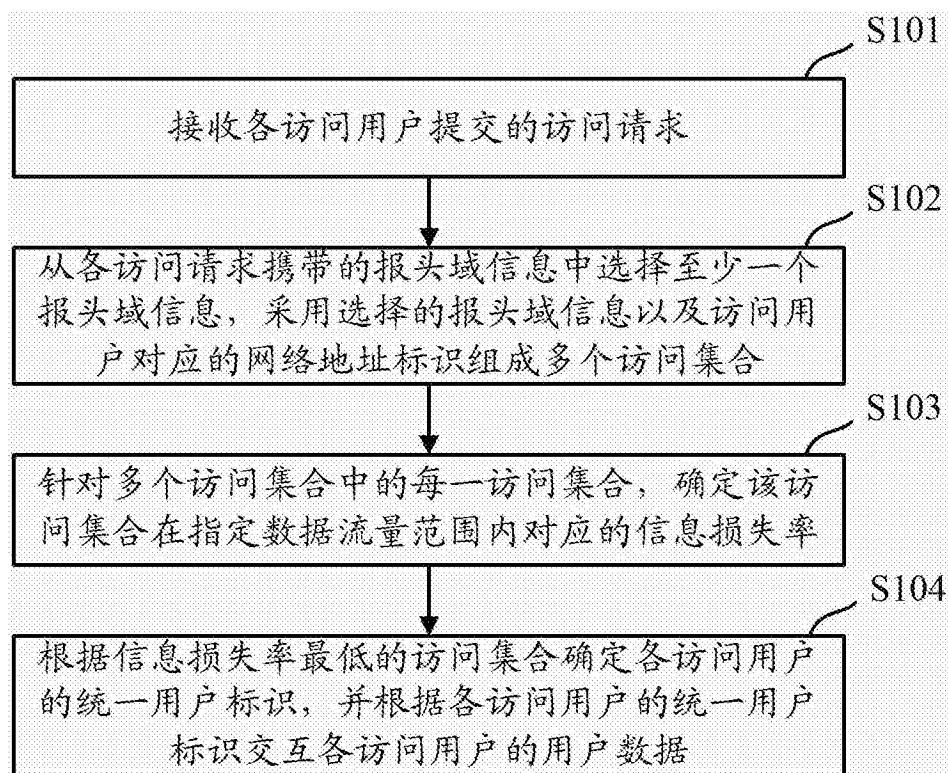


图1

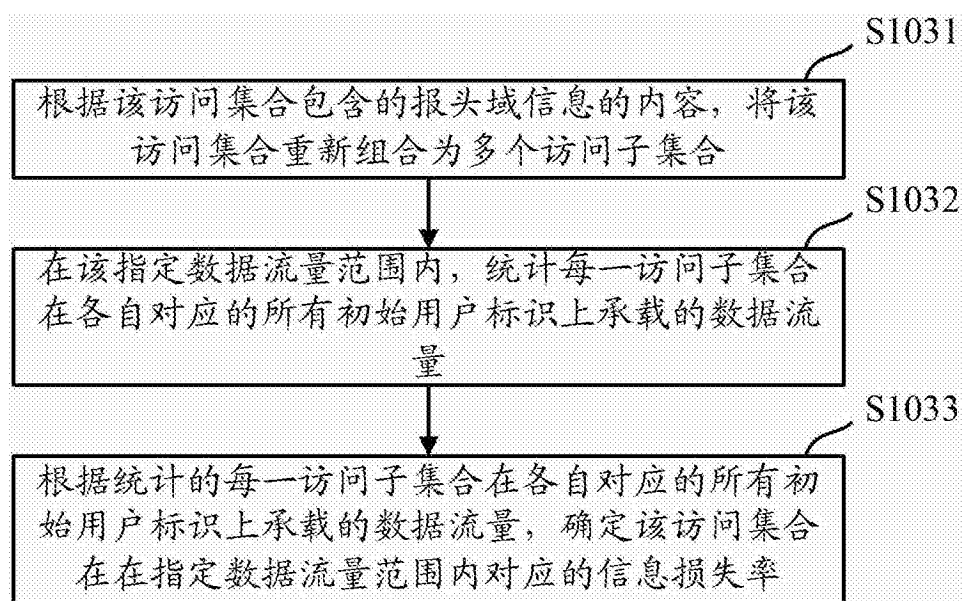


图2

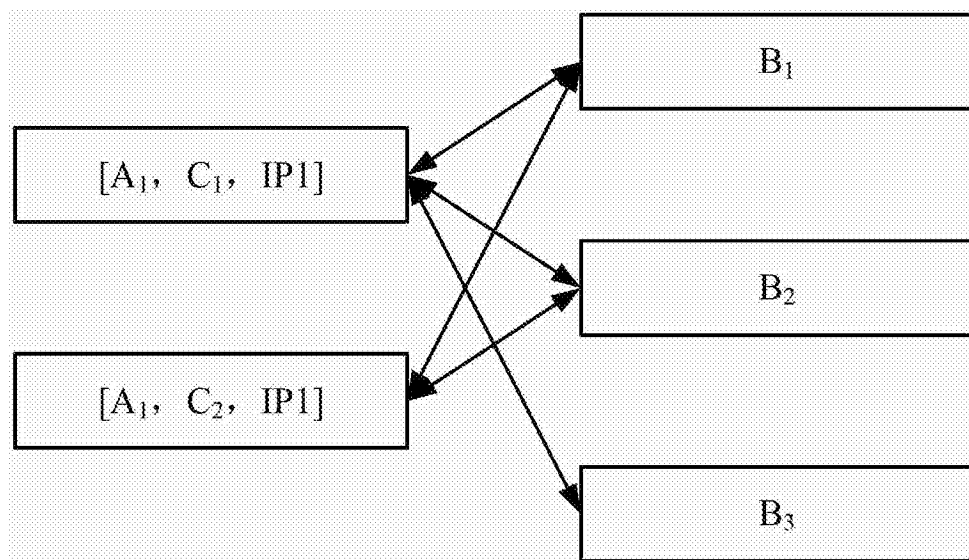


图3

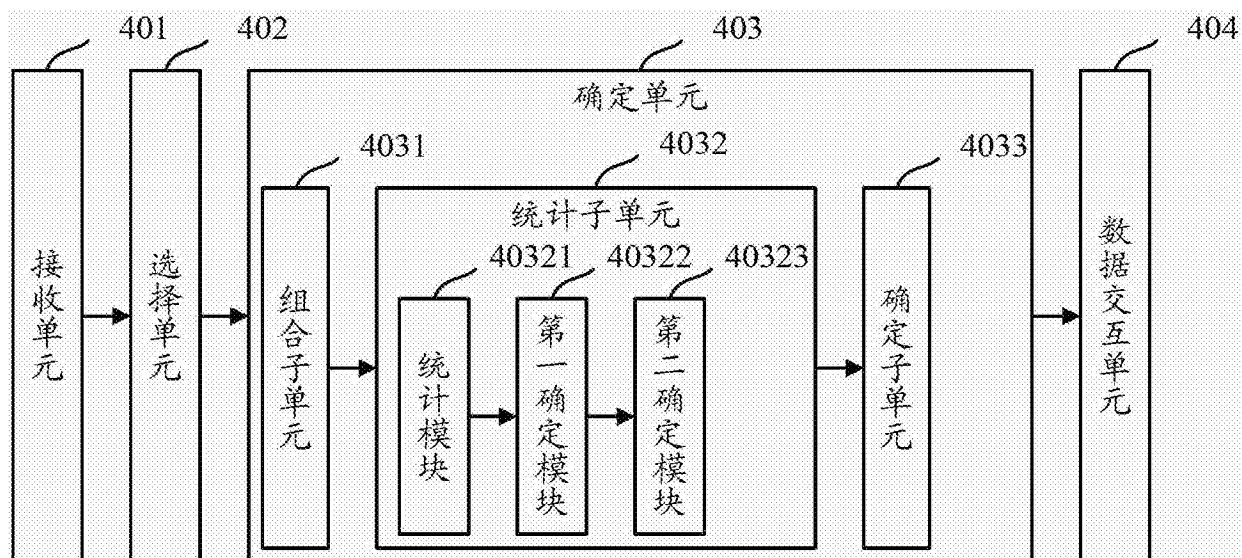


图4