



(12) 发明专利

(10) 授权公告号 CN 101635009 B

(45) 授权公告日 2015.06.17

(21) 申请号 200910042053.9

(22) 申请日 2009.08.21

(73) 专利权人 腾讯科技(深圳)有限公司

地址 518044 广东省深圳市福田区振兴路赛格科技园2栋东403室

(72) 发明人 林乐彬 陈川 凌国惠 孙阿利

(74) 专利代理机构 北京派特恩知识产权代理有限公司 11270

代理人 蒋雅洁 王黎延

(51) Int. Cl.

G06F 19/00(2011.01)

(56) 对比文件

CN 101061713 A, 2007.10.24,

CN 101251853 A, 2008.08.27,

CN 101360098 A, 2009.02.04,

审查员 徐春

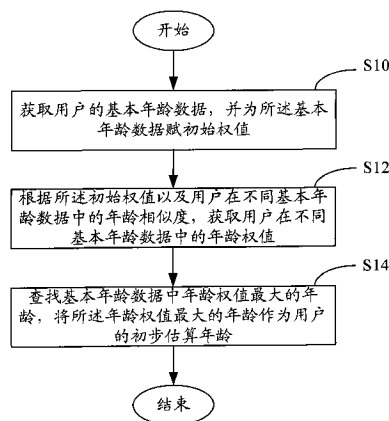
权利要求书2页 说明书6页 附图2页

(54) 发明名称

基于海量数据的用户年龄估算方法及系统

(57) 摘要

本发明提供了一种基于海量数据的用户年龄估算方法及系统。所述方法包括:获取用户的基本年龄数据,为所述基本年龄数据赋初始权值;根据所述初始权值以及用户在不同基本年龄数据中的年龄相似度,获取用户在不同基本年龄数据中的年龄权值;查找基本年龄数据中所述年龄权值最大的年龄,将年龄权值最大的年龄作为用户的初步估算年龄。采用本发明提供的基于海量数据的用户年龄估算方法及系统,能提高估算用户年龄的准确度。



1. 一种基于海量数据的用户年龄估算方法,其特征在于,所述方法包括:
获取用户的基本年龄数据,并为所述基本年龄数据赋初始权值;
将所述基本年龄数据进行两两对比,根据所述初始权值以及用户在不同基本年龄数据中的年龄相似度,设置用户的年龄权值加分,根据所述初始权值与年龄权值加分的和确定用户的年龄权值;
查找基本年龄数据中所述年龄权值最大的年龄,将年龄权值最大的年龄作为用户的初步估算年龄。
2. 根据权利要求1所述的基于海量数据的用户年龄估算方法,其特征在于,所述为所述基本年龄数据赋初始权值的步骤具体是:
获取用户的参考年龄数据;
将所述基本年龄数据与所述参考年龄数据进行对比,获取基本年龄数据的准确率;
根据所述准确率为所述基本年龄数据赋初始权值。
3. 根据权利要求1所述的基于海量数据的用户年龄估算方法,其特征在于,所述方法还包括:
获取同学关系链数据中的用户的初步估算年龄,并根据所述用户的初步估算年龄及其年龄权值调整所述同学关系链数据中的用户的初步估算年龄。
4. 根据权利要求1或3所述的基于海量数据的用户年龄估算方法,其特征在于,所述方法还包括:
比较所述用户的初步估算年龄的年龄权值与所述初始权值的大小,根据所述比较结果将所述用户的初步估算年龄的年龄权值划分为至少如下三个等级:权值为高、权值为中、权值为低。
5. 根据权利要求4所述的基于海量数据的用户年龄估算方法,其特征在于,所述方法还包括:
查找同学关系链数据中初步估算年龄的年龄权值为高且年龄相同的用户个数,判断所述用户个数是否满足预设条件,若是,则将所述同学关系链数据中初步估算年龄的年龄权值为中和年龄权值为低的用户的年龄调整为所述初步估算年龄的年龄权值为高且年龄相同的用户的年龄。
6. 一种基于海量数据的用户年龄估算系统,其特征在于,所述系统包括:
权值设置单元,用于获取用户的基本年龄数据,并为所述基本年龄数据赋初始权值;
权值处理单元,与所述权值设置单元相连,将所述基本年龄数据进行两两对比,根据所述初始权值以及用户在不同基本年龄数据中的年龄相似度,设置用户的年龄权值加分,根据所述初始权值与年龄权值加分的和确定用户的年龄权值;
年龄估算单元,与所述权值处理单元相连,查找基本年龄数据中年龄权值最大的年龄,将所述年龄权值最大的年龄作为用户的初步估算年龄。
7. 根据权利要求6所述的基于海量数据的用户年龄估算系统,其特征在于,所述权值设置单元还用于获取用户的参考年龄数据,将所述基本年龄数据与所述参考年龄数据进行对比,获取基本年龄数据的准确率,并根据所述准确率为所述基本年龄数据赋初始权值。
8. 根据权利要求6所述的基于海量数据的用户年龄估算系统,其特征在于,所述权值处理单元还用于比较所述用户的初步估算年龄与所述初始权值的大小,根据所述比较结果

将所述用户的初步估算年龄的年龄权值划分为至少如下三个等级：权值为高、权值为中、权值为低。

9. 根据权利要求 8 所述的基于海量数据的用户年龄估算系统,其特征在于,所述年龄估算单元还用于查找同学关系链数据中初步估算年龄的年龄权值为高且年龄相同的用户个数,判断所述用户个数是否满足预设条件,若是,则将所述同学关系链数据中初步估算年龄的年龄权值为中和年龄权值为低的用户的年龄调整为所述初步估算年龄的年龄权值为高且年龄相同的用户的年龄。

10. 根据权利要求 6 所述的基于海量数据的用户年龄估算系统,其特征在于,所述系统还包括:

年龄数据存储单元,与所述权值设置单元、权值处理单元及年龄估算单元相连,用于存储基本年龄数据和参考年龄数据;

同学关系链数据存储单元,与所述年龄估算单元相连,用于存储同学关系链数据。

基于海量数据的用户年龄估算方法及系统

技术领域

[0001] 本发明涉及海量数据处理技术领域,更具体地说,涉及一种基于海量数据的用户年龄估算方法及系统。

背景技术

[0002] 随着互联网的不断普及,网络已经成为人们生活中必不可少的一部分。通过互联网可以提供给用户各种各样的服务,例如网上购物、信息获取、游戏娱乐等。用户年龄是用户的基本属性,针对不同年龄的用户群体,可以为其提供个性化的互联网服务。然而通常情况下,由于网络的虚拟性,用户一般都不会填写真实准确的年龄,因此如何准确估算用户的真实年龄,已成为互联网业务急需解决的问题。

[0003] 目前,通常获取用户提供的年龄数据,通过简单的边界值过滤来估算用户年龄。具体地,是根据经验估计用户的年龄范围,将年龄范围之外的数值过滤掉,从而估算出用户年龄。然而,该方法过分依赖用户提供的年龄,因此准确度不高。

发明内容

[0004] 基于此,有必要提供一种能提高准确度的基于海量数据的用户年龄估算方法。

[0005] 此外,还有必要提供一种能提高准确度的基于海量数据的用户年龄估算系统。

[0006] 所述基于海量数据的用户年龄估算方法包括:获取用户的基本年龄数据,并为基本年龄数据赋初始权值;根据初始权值以及用户在不同基本年龄数据中的年龄相似度,获取用户在不同基本年龄数据中的年龄权值;查找基本年龄数据中年龄权值最大的年龄,将年龄权值最大的年龄作为用户的初步估算年龄。

[0007] 该设置基本年龄数据的初始权值的步骤具体是:获取用户的参考年龄数据;将基本年龄数据与所述参考年龄数据进行对比,获取基本年龄数据的准确率;根据准确率为基本年龄数据赋初始权值。

[0008] 该获取用户在不同基本年龄数据中的年龄权值的步骤具体可以是:将基本年龄数据进行两两对比;根据初始权值以及用户在不同基本年龄数据中的年龄相似度,设置用户的年龄权值加分;根据所述初始权值与年龄权值加分的和确定用户的年龄权值。

[0009] 该方法还可包括:获取同学关系链数据中的用户的初步估算年龄,并根据所述用户的初步估算年龄及其年龄权值调整所述同学关系链数据中的用户的初步估算年龄。

[0010] 该方法还可包括:比较用户的初步估算年龄的年龄权值与初始权值的大小,根据比较结果将用户的初步估算年龄的年龄权值划分为至少如下三个等级:权值为高、权值为中、权值为低。

[0011] 该方法还可包括:查找同学关系链数据中初步估算年龄的年龄权值为高且年龄相同的用户个数,判断用户个数是否满足预设条件,若是,则将同学关系链数据中初步估算年龄的年龄权值为中和年龄权值为低的用户的年龄调整为初步估算年龄的年龄权值为高且年龄相同的用户的年龄。

[0012] 所述基于海量数据的用户年龄估算系统包括：权值设置单元，用于获取用户的基本年龄数据，并为基本年龄数据赋初始权值；权值处理单元，与权值设置单元相连，根据初始权值以及用户在不同基本年龄数据中的年龄相似度，获取用户在不同基本年龄数据中的年龄权值；年龄估算单元，与权值处理单元相连，查找基本年龄数据中年龄权值最大的年龄，将年龄权值最大的年龄作为用户的初步估算年龄。

[0013] 该权值设置单元还可用于获取用户的参考年龄数据，将基本年龄数据与参考年龄数据进行对比，获取基本年龄数据的准确率，并根据准确率为基本年龄数据赋初始权值。

[0014] 该权值处理单元还可用于将基本年龄数据进行两两对比，根据初始权值以及用户在不同基本年龄数据中的年龄相似度，设置用户的年龄权值加分，根据所述初始权值与年龄权值加分的和确定用户的年龄权值。

[0015] 该权值处理单元还可用于比较用户的初步估算年龄与初始权值的大小，根据比较结果将用户的初步估算年龄的年龄权值划分为至少如下三个等级：权值为高、权值为中、权值为低。

[0016] 该年龄估算单元还可用于查找同学关系链数据中初步估算年龄的年龄权值为高且年龄相同的用户个数，判断用户个数是否满足预设条件，若是，则将同学关系链数据中初步估算年龄的年龄权值为中和年龄权值为低的用户的年龄调整为初步估算年龄的年龄权值为高且年龄相同的用户的年龄。

[0017] 另外，该系统还可包括：年龄数据存储单元，与权值设置单元、权值处理单元及年龄估算单元相连，用于存储基本年龄数据和参考年龄数据；同学关系链数据存储单元，与年龄估算单元相连，用于存储同学关系链数据。

[0018] 上述基于海量数据的用户年龄估算方法及系统，通过为基本年龄数据赋初始权值，并根据初始权值以及用户在不同基本年龄数据中的年龄相似度来获取用户在不同基本年龄数据中的年龄权值，以及取该年龄权值最高的年龄作为用户的初步估算年龄。由于对用户提供的多种基本年龄数据进行了综合评价，年龄权值最高的年龄更符合用户的真实年龄，因此能提高估算用户年龄的准确度。

附图说明

[0019] 图 1 是一个实施例中基于海量数据的用户年龄估算方法的流程图；

[0020] 图 2 是一个实施例中为基本年龄数据赋初始权值的方法流程图；

[0021] 图 3 是一个实施例中获取用户在不同基本年龄数据中的年龄权值的方法流程图；

[0022] 图 4 是一个实施例中利用同学关系链数据估算用户年龄的方法流程图；

[0023] 图 5 是一个实施例中基于海量数据的用户年龄估算系统的结构示意图；

[0024] 图 6 是另一个实施例中基于海量数据的用户年龄估算系统的结构示意图。

具体实施方式

[0025] 图 1 示出了一个实施例中基于海量数据的用户年龄估算方法流程，该方法流程具体包括以下步骤：

[0026] 在步骤 S10 中，获取用户的基本年龄数据，并为基本年龄数据赋初始权值。基本年龄数据是用户通过各种网络产品填写资料而提供的年龄数据，例如通过即时通讯工具或者

SNS 社区服务等提供的年龄数据等。如图 2 所示,在一个实施方式中,为基本年龄数据赋初始权值的过程包括:

[0027] 在步骤 S100 中,获取用户的参考年龄数据。用户的参考年龄数据可以通过网络进行问卷调查而得到的用户年龄数据。由于问卷调查所设置的问题相对严谨,通过问卷调查得到的用户年龄会比用户直接填写的年龄更准确。

[0028] 在步骤 S102 中,将基本年龄数据与参考年龄数据进行对比,获取基本年龄数据的准确率。在各种基本年龄数据中查找与参考年龄相符的用户年龄个数,该用户年龄个数与用户总数的比值即为基本年龄数据的准确率。

[0029] 在步骤 S104 中,根据所述准确率为基本年龄数据赋初始权值。在一个实施方式中,将基本年龄数据的准确率分为三个等级:低、中、高。对应低、中、高的准确率分别设置基本年龄数据的初始权值分别为 P1、P2 和 P3,优选地,设置 $P1 = 1, P2 = 5, P3 = 9$ 。例如,通过即时通信工具获取 n 个用户的基本年龄数据 IM1、IM2、...、IMn;通过 SNS 服务获得的 n 个用户的基本年龄数据为 SNS1、SNS2、...、SNSn;通过网络进行问卷调查而得到的参考年龄数据为 R1、R2、...、Rn。通过对比 IM1、IM2、...、IMn 和 R1、R2、...、Rn,可以获得即时通信工具获得的基本年龄数据的准确率,假设该准确率的等级为低,则通过即时通信工具获得的基本年龄数据的初始权值为 P1。类似地,可以获得通过 SNS 服务获得的基本年龄数据的准确率,假设该准确率的等级为中,则通过 SNS 服务获得的基本年龄数据的初始权值为 P2。

[0030] 在另一个实施方式中,也可根据基本年龄数据的来源类型直接为基本年龄数据赋初始权值。例如,网络业务如校友录等的注册信息相对其它注册信息获取的用户年龄数据更准确,因此可设置这类基本年龄数据的初始权值比其它类型的基本年龄数据的初始权值高。

[0031] 在步骤 S12 中,根据初始权值以及用户在不同基本年龄数据中的年龄相似度,获取用户在不同基本年龄数据中的年龄权值。如图 3 所示,在一个实施方式中,获取用户在不同基本年龄数据中的年龄权值的具体过程如下:

[0032] 在步骤 S120 中,将基本年龄数据进行两两对比。对于通过各种方式得到的多种基本年龄数据,将该用户在不同基本年龄数据中的年龄进行两两对比。

[0033] 在步骤 S122 中,根据初始权值以及用户在不同基本年龄数据中的年龄相似度,设置用户的年龄权值加分。在一个实施方式中,用户在不同基本年龄数据中的年龄相似度可分为三类:年龄相同、年龄相近、年龄不同。其中,年龄相差在三岁以内表示年龄相近,年龄相差大于三岁表示年龄不同。比较不同基本年龄数据的初始权值,得到基本年龄数据之间的权重关系,该权重关系可分为三类:权重相同、权重相近和权重不同。其中,权重相同表示两种基本年龄数据的权重等级相同(即权重同为高、中或低);权重相近表示两种基本年龄数据的权重等级仅差一级(即两者的权重分别为高与中、或中与低);权重不同表示两种基本年龄数据的权重等级相差两级(即两者的权重分别为高与低)。在一个实施例中,设置用户的年龄权值加分如表 1 所示:

[0034] 表 1

[0035]	年龄相似度 权重关系	年龄相同	年龄相近	年龄不同
	权重相同	+A1	+A4	0
权重相近	+A2	+A5	0	
权重不同	+A3	+A6	0	

[0036] 优选地,设置 $A1 = 1, A2 = 2, A3 = 3, A4 = 4, A5 = 5, A6 = 6$ 。

[0037] 在步骤 S124 中,根据初始权值与年龄权值加分的和确定用户的年龄权值。在上述实施方式中,将各种基本年龄数据进行两两对比,对任意一种基本年龄数据,获取其与其它基本年龄数据之间的权重关系,以及在该权重关系下用户年龄之间的相似度,则在基本年龄数据中用户的年龄权值加分为该基本年龄数据与其它基本年龄数据进行对比后所得到的所有年龄权值加分的总和。

[0038] 在一个具体的实施例中,获取到用户的三种基本年龄数据分别为 M、N、O。该实施例中,设置这三种基本年龄数据的初始权值分别为 P1、P2、P3。对其中的三个用户 a、b、c,假设 M 数据中各用户的年龄分别为 $M_a、M_b$ 和 M_c ,N 数据中各用户的年龄分别为 $N_a、N_b、N_c$,O 数据中各用户的年龄分别为 $O_a、O_b$ 和 O_c 。将 M、N、O 数据进行两两比较,由 M、N、O 的初始权值可知 M 与 N 的权重相近,与 O 的权重不同。对于用户 a,假设 $M_a = 25, N_a = 25, O_a = 23$,即 M_a 与 N_a 年龄相同, M_a 与 O_a 年龄相近, N_a 与 O_a 年龄相近。根据表 1 所设置的年龄权值加分可知, M_a 的年龄权值为 $P1+A2+A5$, N_a 的年龄权值为 $P2+A2+A5$, O_a 的年龄权值为 $P3+A5+A5$ 。同理,用户 b 和用户 c 的年龄权值也可按照上述方法原理计算得到。

[0039] 在步骤 S14 中,获取基本年龄数据中年龄权值最大的年龄,将所述年龄权值最大的年龄作为用户的初步估算年龄。上述实施例中,对于用户 a,则取 $M_a、N_a$ 和 O_a 的年龄权值最大的作为用户 a 的初步估算年龄。由于年龄权值最大的年龄最可能接近用户的真实年龄,因此所得到的初步估算年龄更准确。

[0040] 在一个实施方式中,得到用户的初步估算年龄后,比较用户的初步估算年龄的年龄权值与初始权值的大小,根据比较结果将用户的初步估算年龄的年龄权值划分为三个等级:权值为高、权值为中、权值为低。在一个实施例中,设置基本年龄数据的初始权值分别为 P1、P2 和 P3,当初步估算年龄的年龄权值小于等于 P2 时,权值为低;当初步估算年龄的年龄权值大于 P2 且小于等于 P3 时,权值为中;当初步估算年龄的年龄权值大于 P3 时,权值为高。

[0041] 图 4 示出了一个实施例中利用同学关系链数据估算用户年龄的方法流程,具体过程如下:

[0042] 在步骤 S20 中,查找同学关系链数据中初步估算年龄的年龄权值为高且年龄相同的用户个数。同学关系链数据是用户之间是同学关系的一个数据集合,具有同学关系的用户年龄通常相同或相近,可通过获取用户所在的同学群组成员及用户的好友分组来获取同学关系链数据。

[0043] 在步骤 S22 中,判断所述用户个数是否满足预设条件,若是,则进入步骤 S24,否则结束。在一个实施方式中,该预设条件为: $m > 3$ 且 $m/n \geq 1/4$,其中, m 为同学关系链数

据中初步估算年龄的年龄权值为高且年龄相同的用户个数, n 为同学关系链中的用户总数。

[0044] 在步骤 S24 中, 将同学关系链数据中初步估算年龄的年龄权值为中和为低的用户的年龄调整为年龄权值为高且年龄相同的用户的年龄。在一个实施例中, 当查找到同学关系链数据中初步估算年龄的年龄权值为高且年龄相同的用户个数满足上述预设条件时, 由于这些用户的初步估算年龄的年龄权值为高, 相对年龄权值为低和年龄取值为中的用户的初步估算年龄更准确, 而同学关系链数据中用户的年龄通常相同或相近, 因此利用年龄权值为高的用户的初步估算年龄去调整年龄权值为低及为中的用户年龄, 将初步估算年龄的年龄权值为中和为低的用户的年龄调整为年龄权值为高的用户年龄, 估算得到的用户年龄更准确。

[0045] 图 5 示出了一个实施例中基于海量数据的用户年龄估算系统, 该系统包括权值设置单元 10、权值处理单元 20、年龄估算单元 30。其中:

[0046] 权值设置单元 10 用于获取用户的基本年龄数据, 并设置所述基本年龄数据的初始权值。

[0047] 权值处理单元 20 与权值设置单元 10 相连, 根据初始权值以及用户在不同基本年龄数据中的年龄相似度, 获取用户在不同基本年龄数据中的年龄权值。

[0048] 年龄估算单元 30 与权值处理单元 20 相连, 用于查找基本年龄数据中年龄权值最大的年龄, 将所述年龄权值最大的年龄作为用户的初步估算年龄。

[0049] 图 6 示出了另一个实施例中基于海量数据的用户年龄估算系统, 该系统除了包括上述权值设置单元 10、权值处理单元 20 和年龄估算单元 30 外, 还包括年龄数据存储单元 40 和同学关系链数据存储单元 50。其中:

[0050] 年龄数据存储单元 40 与权值设置单元 10、权值处理单元 20 及年龄估算单元 40 相连, 用于存储基本年龄数据和参考年龄数据。基本年龄数据是用户通过各种网络产品填写资料而提供的年龄数据; 参考年龄数据可以通过网络进行问卷调查得到的用户年龄数据。由于问卷调查所设置的问题相对严谨, 所得到的参考年龄数据比基本年龄数据更准确。

[0051] 同学关系链数据存储单元 50 与年龄估算单元 30 相连, 用于存储同学关系链数据。具有同学关系的用户年龄通常相同或相近, 可通过获取用户所在的同学群组成员及用户的好友分组来获取同学关系链数据。

[0052] 在一个实施方式中, 权值设置单元 10 还用于获取用户的参考年龄数据, 将基本年龄数据与参考年龄数据进行对比, 获取基本年龄数据的准确率, 并根据该准确率设置基本年龄的初始权值。可通过在各种基本年龄数据中查找与参考年龄相符的年龄个数, 基本年龄数据的准确率则为该年龄个数与用户总数的比值。权值设置单元 10 可将准确率划分为三个等级: 低、中、高, 并对应不同等级的准确率设置基本年龄数据的初始权值。

[0053] 在一个实施方式中, 权值处理单元 20 还用于将基本年龄数据进行两两对比, 根据初始权值以及用户在不同基本年龄数据中的年龄相似度, 设置用户的年龄权值加分, 则用户的年龄权值为初始权值与年龄权值加分的和。权值处理单元 20 将各种基本年龄数据进行两两对比, 对任意一种基本年龄数据, 获取其与其它基本年龄数据之间的权重关系, 以及在该权重关系下用户年龄之间的相似度, 则在基本年龄数据中用户的年龄权值加分为该基本年龄数据与其它基本年龄数据进行对比后所得到的所有年龄权值加分的总和。权值处理单元 20 计算得到用户的年龄权值后, 年龄估算单元 30 则查找年龄权值最大的年龄, 并将该

年龄权值最大的年龄作为用户的初步估算年龄。

[0054] 在一个实施方式中,年龄估算单元 30 得到用户的初步估算年龄后,权值处理单元 20 还比较用户的初步估算年龄与初始权值的大小,并根据比较结果将用户的初步估算年龄的年龄权值划分为至少如下三个等级:权值为高、权值为中、权值为低。

[0055] 在一个实施方式中,年龄估算单元 30 还用于查找同学关系链数据中初步估算年龄的年龄权值为高且年龄相同的用户个数,并判断该用户个数是否满足预设条件,若是,则将同学关系链数据中初步估算年龄的年龄权值为中和年龄权值为低的用户的年龄调整为所述初步估算年龄的年龄权值为高且年龄相同的用户的年龄。在一个实施例中,所述预设条件为: $m > 3$ 且 $m/n \geq 1/4$,其中, m 为同学关系链数据中初步估算年龄的年龄权值为高且年龄相同的用户个数, n 为同学关系链数据中的用户总数。由于同学关系链数据中的用户年龄相同或相近,利用年龄权值为高的用户的初步估算年龄去调整年龄权值为低及为中的用户年龄,将初步估算年龄的年龄权值为中和为低的用户的年龄调整为年龄权值为高的用户年龄,估算得到的用户年龄更准确。

[0056] 以上所述仅为本发明的较佳实施例而已,并不用以限制本发明,凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等,均应包含在本发明的保护范围之内。

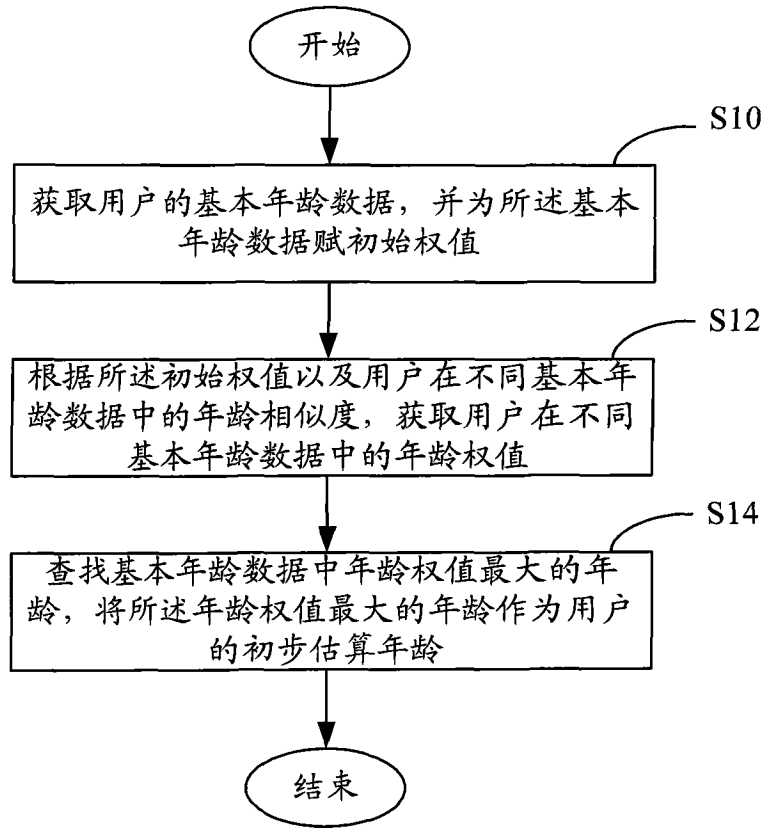


图 1

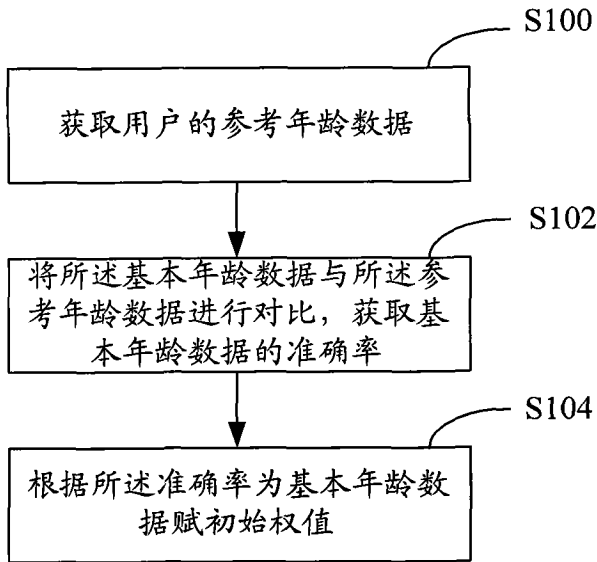


图 2

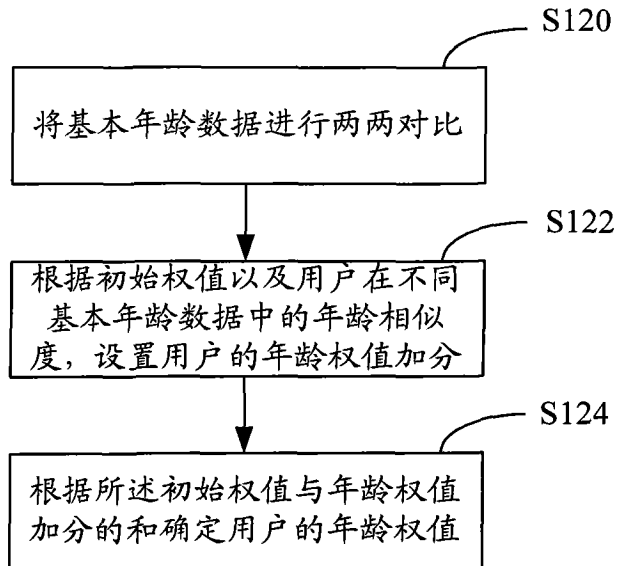


图 3

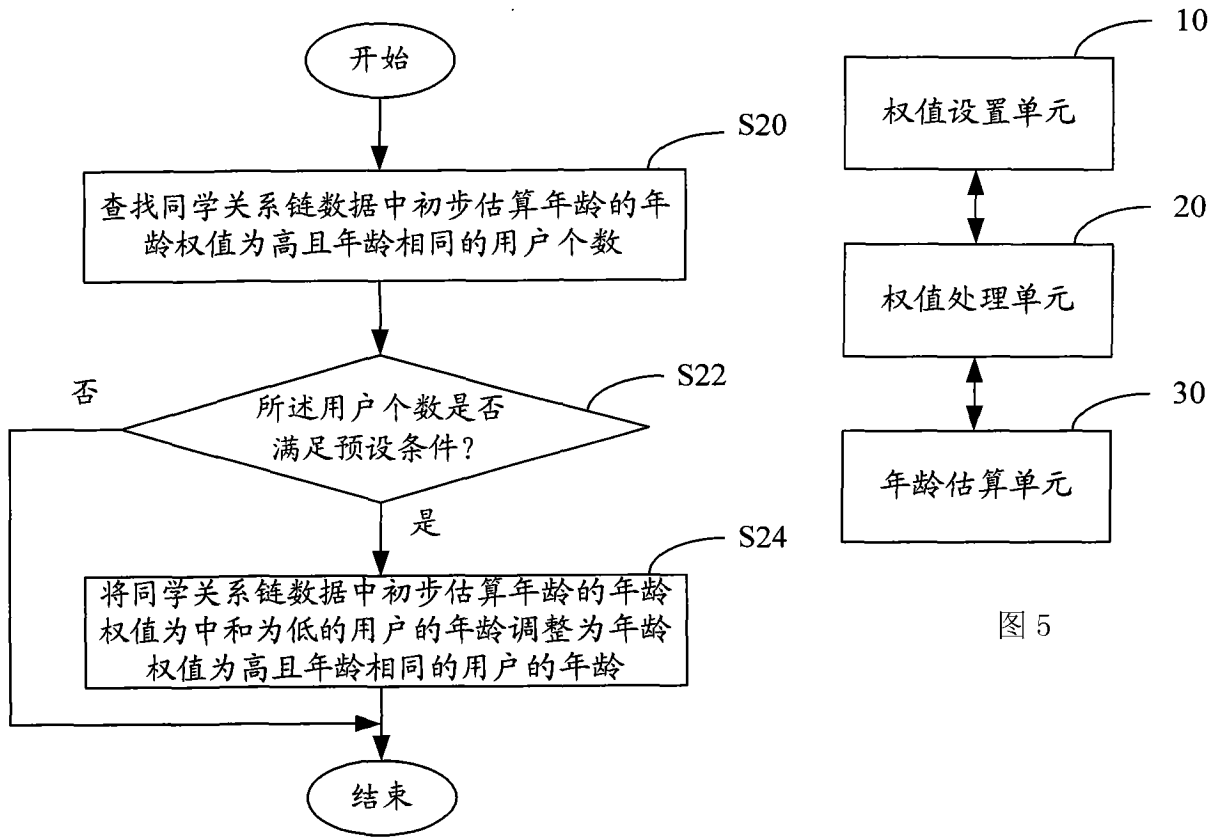


图 4

