



(12) 发明专利申请

(10) 申请公布号 CN 105022808 A

(43) 申请公布日 2015. 11. 04

(21) 申请号 201510386278. 1

(22) 申请日 2015. 06. 29

(71) 申请人 程文举

地址 232000 安徽省淮南市安徽理工大学
(北校区)

(72) 发明人 程文举

(74) 专利代理机构 北京细软智谷知识产权代理
有限责任公司 11471

代理人 王金宝

(51) Int. Cl.

G06F 17/30(2006. 01)

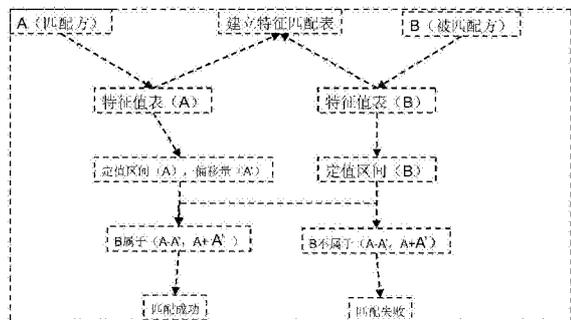
权利要求书1页 说明书7页 附图1页

(54) 发明名称

一种二进制定值区间匹配方法

(57) 摘要

本发明涉及一种二进制定值区间匹配方法，
(1) 建立特征值表 A 和 B；(2) 建立特征值匹配表；
(3) 提取特征值匹配表中特征值，进行格式化；
(4) 对特征值进行定值计算；(5) 计算偏移量；(6)
对计算结果进行匹配。本发明的有益效果为：本
发明减少服务器匹配数据或角色时的运算量；本
发明减少匹配时服务器的内存占用；本发明减少
匹配时繁琐的计算方法和步骤；本发明可控的数
据匹配度，对于不同的需要用户可以通过控制偏
移量来控制数据匹配度，达到用户需要的数据精
度。本发明高效的匹配效率。本发明具有匹配的
伸缩性，可以实时更新用户的状态和对应的特征
值，快速应对用户某些特性的改变。



1. 一种二进制定值区间匹配方法,其特征在于:

- (1) 建立特征值表 A 和 B;
- (2) 建立特征值匹配表;
- (3) 提取特征值匹配表中特征值,进行格式化;
- (4) 对特征值进行定值计算;
- (5) 计算偏移量;
- (6) 对计算结果进行匹配。

2. 根据权利要求 1 所述的一种二进制定值区间匹配方法,其特征在于:所述步骤(1)包括根据网站或搜索需求建立特征值表,并按照特征值的优先级降序排列。

3. 根据权利要求 1 所述的一种二进制定值区间匹配方法,其特征在于:所述步骤(2)包括对特征值表 A 和 B,建立用户对用户、用户对信息或信息对信息匹配,采用一一映射关系建立特征值匹配表。

4. 根据权利要求 1 所述的一种二进制定值区间匹配方法,其特征在于:所述步骤(3)包括根据用户行为或已有特征,提取特征值表中特征值,进行格式化生成二进制序列。

5. 根据权利要求 1 所述的一种二进制定值区间匹配方法,其特征在于:所述步骤(3)包括

(3.1) 选择二进制位宽,二进制位宽与特征值表中特征值数量相同;

(3.2) 根据二进制位宽生成一个二进制格式序列,二进制格式序列索引对应的默认值为 0;

(3.3) 已有特征值表中的特征值,则将其索引相同二进制格式序列中的位置 1 否则置 0,二进制格式序列位置默认值为 0。

6. 根据权利要求 1 所述的一种二进制定值区间匹配方法,其特征在于:所述步骤(4)包括

(4.1) 检索出二进制格式中除去开头全为 0 的剩余序列中每个 0 的位置下标,并储存下标序列,检索出的序列为空,则检索 1 的序列并去除 1 中最高位代替序列;

(4.2) 进行定值计算,计算出二进制格式对应的十进制数值。

7. 根据权利要求 1 所述的一种二进制定值区间匹配方法,其特征在于:所述步骤(5)包括根据匹配度需要通过检索序列来计算出偏移量的值,用于控制匹配成功率。

8. 根据权利要求 1 所述的一种二进制定值区间匹配方法,其特征在于:所述步骤(5)包括通过偏移量控制匹配精度,将计算结果对特征值表 A 和 B 进行匹配。

一种二进制定值区间匹配方法

技术领域

[0001] 本发明属于匹配技术,具体涉及一种二进制定值区间匹配方法。

背景技术

[0002] 随着网络生活的普及,信息量的繁多,人们想要查询一些成对或相似的互动信息越来越难。大部分这样信息的匹配是人工处理的,即:通过软件本身设置的供求区或跟贴区来实现自然的配对。如果是在海量的互联网信息中自动配对,则往往还是靠关键词匹配,匹配准确性差。这样的匹配格局就会给很多商家丢失一些潜在用户。

[0003] 现在还没有任何软件或方法来根据用户的特性向其推荐他所关心的产品配置,或者是基于信息本身来进行推送;当前软件大多数都是基于用户的浏览记录来做匹配,其中的信息的失真十分严重,匹配效率很低。本专利主要提供如:用户需要什么就推送什么;通过用户需求,判断潜在购买者是谁,如何根据购买者来定制包装方案,减少数据匹配中大量的数据抓取及运算。

发明内容

[0004] 为了解决现有技术存在的上述问题,本发明提供了一种二进制定值区间匹配方法。将需要提取或匹配的信息利用特征值表进行固化;利用二进制控制信息的特征值筛选;利用二进制和十进制的可转换性对信息进行定值计算;利用二进制的位置(计算出的偏移量)来确定信息的伸缩性;通过定值计算和偏移量进行匹配。

[0005] 本发明所采用的技术方案为:

[0006] 一种二进制定值区间匹配方法,其改进之处在于:

[0007] (1) 建立特征值表 A 和 B;

[0008] (2) 建立特征值匹配表;

[0009] (3) 提取特征值匹配表中特征值,进行格式化;

[0010] (4) 对特征值进行定值计算;

[0011] (5) 计算偏移量;

[0012] (6) 对计算结果进行匹配。

[0013] 优选的,所述步骤(1)包括根据网站或搜索需求建立特征值表,并按照特征值的优先级降序排列。

[0014] 优选的,所述步骤(2)包括对特征值表 A 和 B,建立用户对用户、用户对信息或信息对信息匹配,采用一一映射关系建立特征值匹配表。

[0015] 优选的,所述步骤(3)包括根据用户行为或已有特征,提取特征值表中特征值,进行格式化生成二进制序列。

[0016] 优选的,所述步骤(3)包括

[0017] (3.1) 选择二进制位宽,二进制位宽与特征值表中特征值数量相同;

[0018] (3.2) 根据二进制位宽生成一个二进制格式序列,二进制格式序列索引对应的默

认值为 0；

[0019] (3.3) 已有特征值表中的特征值,则将其索引相同二进制格式序列中的位置 1 否则置 0,二进制格式序列位置默认值为 0。

[0020] 优选的,所述步骤 (4) 包括

[0021] (4.1) 检索出二进制格式中除去开头全为 0 的剩余序列中每个 0 的位置下标,并储存下标序列,检索出的序列为空,则检索 1 的序列并去除 1 中最高位代替序列；

[0022] (4.2) 进行定值计算,计算出二进制格式对应的十进制数值。

[0023] 优选的,所述步骤 (5) 包括根据匹配度需要通过检索序列来计算出偏移量的值,用于控制匹配成功率。

[0024] 优选的,所述步骤 (5) 包括通过偏移量控制匹配精度,将计算结果对特征值表 A 和 B 进行匹配。

[0025] 本发明的有益效果为：

[0026] 本发明可以进行角色或信息的预处理。只要在用户或信息激发某种特征值时,对角色或信息的二进制定值区间进行一次计算即可用于以后的匹配。

[0027] 本发明减少服务器匹配数据或角色时的运算量。通过将信息或角色独立转化的方式先独立计算,然后只需要进行数匹配,避免大量的字符串匹配带来的服务压力,更加适用于小型的服务器。

[0028] 本发明减少匹配时服务器的内存占用,相对于现有的匹配方法,此方法匹配时需要存储的数值较少,同时可以实时的与数据库交互,减小服务器的内存压力。

[0029] 本发明减少匹配时繁琐的计算方法和步骤:直接将用户及信息固定成一个特定值,每次只需要对值进行匹配,而对于角色或数据本身不要做考虑。匹配只是关系的映射,与用户或信息本身的内容关系不大,所以只需要对其特征值进行处理。

[0030] 本发明可以利用闲时处理数据:只需要对用户或信息的特征值进行存储后,在服务繁忙是暂时挂起数据处理,在闲时启动数据处理,更加合理的利用服务空间。

[0031] 本发明可控的数据匹配度,对于不同的需要用户可以通过控制偏移量来控制数据匹配度,达到用户需要的数据精度。

[0032] 本发明高效的匹配效率。对于所有已经处理的数据或用户,服务器只需要执行简单的数值域的查询操作即可完成匹配任务。

[0033] 本发明具有匹配的伸缩性,可以实时更新用户的状态和对应的特征值,快速应对用户某些特性的改变。

[0034] 本发明流程控制更容易,只要拥有编程基础的开发人员都可以将此方法整合到相应的网站或者其他程序中。

附图说明

[0035] 图 1 是本发明的一种二进制定值区间匹配方法流程图；

图 2 是本发明的一种二进制定值区间匹配方法中算法转换流程图。

具体实施方式

[0036] 如图 1 所示,本发明提供了一种二进制定值区间匹配方法。

[0037] 注解及含义：

[0038] 特征值 (Eigenvalue, 以下简称为 E)、特征值有序表 (Ordered Eigenvalue, 以下简称为 OE)、特征值匹配表 (Eigenvalue Mapping, 以下简称为 EM)、特征值的索引、二进制位宽索引、二进制格式化 (Binary Format 一下简称 BF,)、检索序列 S, 十进制定值 D, 偏移量 flag 等名词解释及说明。

[0039] E:指的是能表示或者代表用户(数据)的描述。如:人、90后、学生、喜欢台球、关注科技等能标识的人的特征的词汇;又如:张三想要买一个手机,张三(人,继承人的部分特征)、买(行为特征)、手机(继承手机的特征)。E可以是缺省值(Default, 以下简称为 Def。缺省值代表匹配无关特征值,缺省值可以为多个,也可以与任何特征值建立映射关系, Def的优先级低于任何特征值)。

[0040] OE:用来列出某一类具体用户或信息(如学生)特征值的集合(非某一个特定的用户或信息)。用来建立某一类用户或信息的特征值匹配标准,同时对特征值进行优先级降序排序。如:针对某一类用户的OE为{90后,大学生,安徽,本科,男,女,缺省值1,缺省值2,}。

[0041] EM:特征值匹配表,建立用户对用户、用户对信息、信息对信息等匹配方式从前者到后者的一一映射关系(特征值条数以后者为主,前者数量不够将在前者OE中添加Def直到两者条数相等)。如一个卖游戏产品淘宝商家和其用户的特征值匹配表如下{游戏键鼠,游戏光盘,电脑配件,游戏充值,游戏配件,Def1};{喜欢游戏,注重游戏体验,关心游戏需要的配置,有充值意向,90后,大学生}。

[0042] 特征值的索引:特征值有序表中每个特征值对应的下标值,从优先级最小的的开始计数,起始索引为1;

[0043] 二进制位宽索引:二进制序列中为一位数值对应的下标值,从最低位开始计数,起始索引为1;

[0044] BF:用户或信息的二进制格式化的序列和规则。选择合适的二进制位宽(如选择64位或者8位等,根据特征值大小确定,后文中为描述方便统一选择8bit);将OE中优先级最小的对应BF最低位,优先级第二小的对应BF次低位。以此类推;如果指定角色或信息中具有OE中的性质,则将对应BF中的对应位置1(如果某一项性质可能的概率大于50%即认为具有该性质),否则置0,Def的值为0。

[0045] 检索序列S:检索出BF中每个非打头0(即第一个1后面出现的0)的位置下标,并储存下标序列S,如果检索出的S为空,则检索1的序列并去除1中最高位代替S,S按照升序排列,同时建立索引。

[0046] 十进制定值D:定值计算,根据二进制与十进制之间的转换计算出BF对应的十进制数值。

[0047] 偏移量flag:用来控制匹配的效率和结果的准确性;偏移量(flag)计算,偏移量的计算需根据检索序列S来计算,根据匹配度的需要,在S中选取合适的下标值,计算出偏移量的值(对于选取的下标值为i; $flag = 2^{(i-1)}$),用来控制在后面的匹配算法中的匹配成功率;

[0048] 其中,匹配中偏移量flag值对匹配效果影响如下:

[0049] 对于一个二进制序列 $B = \{10001001 \cdots \cdots 11101110\}$;

[0050] 对应的检索序列为 $S = \{s_1, s_2, s_3, \dots, s_n\}$, 计算方法如上文说明,

[0051] 选择 S 中的一个值用于计算下偏移量, 记此值为 i , 此值在 S 中对应的索引为 t , S 中的下标值总数为 n ;

[0052] 则匹配率 $= (n-t+1)/n$;

[0053] 其中, 对于 i 的选择给出做如下说明:

[0054] 设: 用户 u 对应的 $BF(u)$ (BF 中第一次出现 1 的索引为 K);

[0055] $S(u) = \{s_1, s_2, s_3, \dots, s_n\}$ ($n < K$) $flag = 2^{(i-1)}$ $i \in S(u)$;

[0056] ① $2^{s_n} > (2^{(s_1-1)} + 2^{(s_2-1)} + 2^{(s_3-1)} + \dots + 2^{(s_n-1)})$;

[0057] ② $D(u) > \sum 2^{(S_i-1)}$;

[0058] ③ $2^{(K-1)} > \sum 2^{(S_i-1)}$;

[0059] ④对于任何 $i \in S(u)$, 若十进制数 $D' \in (D(u)-flag, D(u)+flag)$ 则 $D(u)$ 和 D' 的二进制表示中 i (不包括) 到 K 位上的二进制序列片段是相同的;

[0060] ⑤对于 $i \in S(u)$, i 的值越小, 则 D' 和 $D(u)$ 上相同的二进制序列片段越长 (即匹配率越高), 反之, 越大。当 $i = 1$ 时则 D' 和 $D(u)$ 完全相同。

[0061] 本发明一种二进制定值区间匹配方法, 具体方法如下:

[0062] (1) 建立特征值表 A 和 B ;

[0063] 确定用户特征值, 建立 OE , 根据网站或者搜索需求建立有效的 OE 。

[0064] OE 中 E 的排列按照降序排列, 关于排列的优点说明如下:

[0065] 设一个用户或信息二进制 (8bit) 格式化 (格式化规则在后面会具体描述) 后为 00110010; 按照优先级降序排列转化为十进制为 $2^5 + 2^4 + 2$ 可以看出位数最高位对整个十进制数据的值影响最大。如果按照升序 (01001100) 即 $2^6 + 2^3 + 2^2$ 可以看出优先级较大的对整个数据的变化影响不大, 同时在建立 EM 双方特征值数量不匹配时, 优先级降序对于 Def 的添加控制十分简单。

[0066] (2) 建立特征值匹配表;

[0067] 建立匹配双方的 EM 。

[0068] (3) 提取特征值匹配表中特征值, 进行格式化;

[0069] 根据角色或信息的 OE 进行数据的二进制格式化规则 (Binary Format, 以下简称 BF);

[0070] 格式化的规则按照 OE 的排序制定:

[0071] 1、选择合适的二进制位宽 (如选择 64 位或者 8 位等, 根据特征值大小确定, 后文中为描述方便统一选择 8bit);

[0072] 2、将 OE 中优先级最小的对应 BF 最低位, 优先级第二小的对应 BF 次低位, 以此类推;

[0073] 3、如果指定角色或信息中具有 OE 中的性质, 则将对应 BF 中的对应位置 1 (如果某一项可能的概率大于 50% 即认为具有该性质), 否则置 0, Def 的值为 0;

[0074] 例如: {90 后, 大学生, 安徽, 本科, 男, 女, 缺省值 1, 缺省值 2,} 对应的一种可能的 BF 为 11010100。

[0075] (4) 对特征值进行定值计算和偏移量计算;

[0076] 1、检索出二进制格式中除去开头全为 0 的剩余序列中 (第一次出现 1 以后的序

列) 每个 0 的位置下标,并储存下标序列,检索出的序列为空,则检索 1 的序列并去除 1 中最高位代替序列;

[0077] 如:01010101 $S = \{2, 4, 6\}$

[0078] 00011111 $S = \{1, 2, 3, 4\}$

[0079] 2、定值计算,计算出 BF 对应的十进制数值(简写为 D, $D(u)$ 表示为对于信息或用户 u 的十进制数值);

[0080] 3、偏移量(flag)计算,偏移量的计算需根据检索序列 S 来计算,根据匹配度的需要计算出偏移量的值,用来控制在后面的匹配算法中的匹配成功率。

[0081] (5) 通过偏移量控制匹配精度,对计算结果进行匹配;

[0082] 通过计算出匹配方的定值区间 A 和偏移量 A', 被匹配方的 B, 只要满足 B 属于 $(A-A', A+A')$, 则算匹配成功。

[0083] 本发明提供不同用户匹配的方案。如:给一个指定的淘宝商家匹配 n 个潜在用户;

[0084] 本发明提供数据匹配方案,根据不同用户的需求提供合适的数据。如:用户需要一条关于 iPhone6 的数据,根据用户的习惯向其推荐他所关心的产品配置,而不是全部信息推送;

[0085] 本发明提供新的搜索匹配方案。根据算法的定值区间匹配推送用户需要的信息,而不是根据信息本身的相关度推送(如现有的字符串模式匹配,或是树图匹配等);

[0086] 本发明提供精准定制,如:用户需要什么就推送什么;

[0087] 本发明提供电子商务中产品与潜在用户的配对。如:我有一个 iPhone, 谁将是潜在的购买者;

[0088] 本发明提供生产中的流程定制。如:包装一批苹果,潜在购买者是谁,如何根据购买者来定制包装方案;

[0089] 本发明提供新的市场调查方案。如:新产品是针对某一特定用户群,如何避免大量的非用户群人员参与调查,或者说让所有参与的调查者都是用户群人员;

[0090] 本发明提供数据挖掘中用户有效数据掘取控制方案。如:用户想购买新手机,如何准确推断用户需要的手机类型,价格,而这些应该基于用户本身特征和需求而不是手机性能和价格。

[0091] 本发明提供快速匹配方案,减少数据匹配中大量的数据抓取及运算。

[0092] 本发明提供信息解析方法。

[0093] 对海量数据的匹配方式,可以通过多张 OE 进行二次定值区间,二次定值区间 OE 中每一条是否满足需根据一次定值中是 OE 否匹配确定,如果一次定值中匹配成功则将对应位置的值设为 1, 否则为 0。后续匹配按照上述原则。

[0094] 实施例

[0095] 针对一个销售游戏相关产品的店家和其用户建立 EM;

[0096] $A: \{ \text{游戏键鼠, 游戏光盘, 电脑配件, 游戏充值, 其他配件, Def1} \};$

[0097] $B: \{ \text{喜欢游戏, 注重游戏体验, 关心游戏需要的配置, 有充值意向, 90 后, 大学生} \};$

[0098] 分析 OE:对于店家来说,以游戏产品为辐射的销售体系,根据自己产品的主次建立适合的 OE。

[0099] 对于用户来说,不同的用户对产品的需求不同,但是用户潜在的一些特性决定了用户的潜在购买力。

[0100] 分析 EM:对于上面的 EM,两个 OE 间的 0 看起来并没有明显的映射关系,但是通过对用户的一些特性分析,比如对于一个不玩游戏的人正常来说是不会购买游戏相关产品,所以用户是否喜欢玩游戏具有最高优先级,同样对于店家来说,随着游戏注重 pc 端的特性,键鼠等具有最高优先级,不注重游戏体验的人购买游戏相关产品的可能性也是比较小的等等。当然对于 EM 之间的映射并不是简单的数学函数关系,这个必须满足对用户特性的分析和建立合适的特征值,同样良好的行为映射的建立是匹配成功的前提条件。

[0101] 设有一个商家 M,其产品满足:游戏键鼠,游戏光盘,游戏充值,其他,这四个特征;

[0102] 则 $BF(M):00110110$;

[0103] $S(M) = \{1, 4\}$;

[0104] $D(M) = 2^5 + 2^4 + 2^2 + 2$;

[0105] 对于偏移量的选择:对于有多个偏移量可供选择时,选择 S 中元素最大,则匹配率越低,但是选择最小元素虽然匹配率高,但是又会无法与相似用户匹配,所以配对原则可以根据需要自定义设定,如果对匹配率要求不高的话建议可以选择最高位或者次高位。反之如果严格要求匹配率则可以选择最低位或次低位。本例中选择最高位。

[0106] $flag(M) = 2^3$;

[0107] 有多个用户 a, b, c, d 等;

[0108] a:偶尔游戏,不关心游戏体验,关心游戏需要的配置,有充值意向,90 后,大学生;

[0109] 则 $BF(a):00001111$;

[0110] $S(a) = \{1, 2, 3\}$;

[0111] $D(a) = 2^3 + 2^2 + 2 + 1$ $flag(a) = 4$;

[0112] b:喜欢游戏,不关心游戏体验,关心游戏需要的配置,有充值意向,90 后,大学生;

[0113] 则 $BF(b):00101111$;

[0114] $S(b) = \{5\}$;

[0115] $D(b) = 2^5 + 2^3 + 2^2 + 2 + 1$ $flag(b) = 2^4$;

[0116] c:喜欢游戏,不关心游戏体验,不关心游戏的配置,没有充值意向,90 后,大学生;

[0117] 则 $BF(c):00100011$;

[0118] $S(c) = \{3, 4, 5\}$;

[0119] $D(c) = 2^5 + 2^1 + 1$ $flag(c) = 2^4$;

[0120] d:喜欢游戏,关心游戏体验,关心游戏需要的配置,有充值意向,非 90 后,非大学生;

[0121] 则 $BF(d):00111100$;

[0122] $S(a) = \{1, 2\}$;

[0123] $D(d) = 2^5 + 2^4 + 2^3 + 2^2$ $flag(d) = 2$;

[0124] 匹配: $M\{a, b, c, d\}$;

[0125] $k \in \{a, b, c, d\}$;

[0126] $if(D(k) \in (D(M) - flag(M), D(M) + flag(M))) \{ // D(k) \in (46, 62) \}$;

[0127] 得到 $\{e | e \in k\}$ // 匹配成功的用户集合；

[0128] }

[0129] 则：得到 $k = \{b, d\}$ 。

[0130] 本发明不局限于上述最佳实施方式，任何人在本发明的启示下都可得出其他各种形式的产品，但不论在其形状或结构上作任何变化，凡是具有与本申请相同或相近似的技术方案，均落在本发明的保护范围之内。

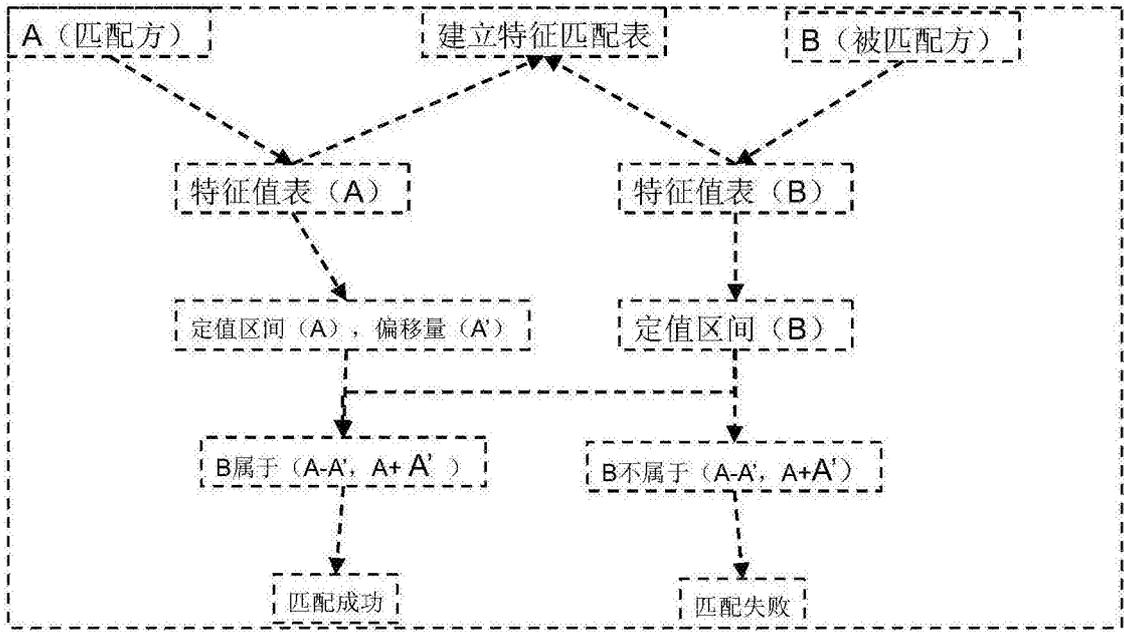


图 1

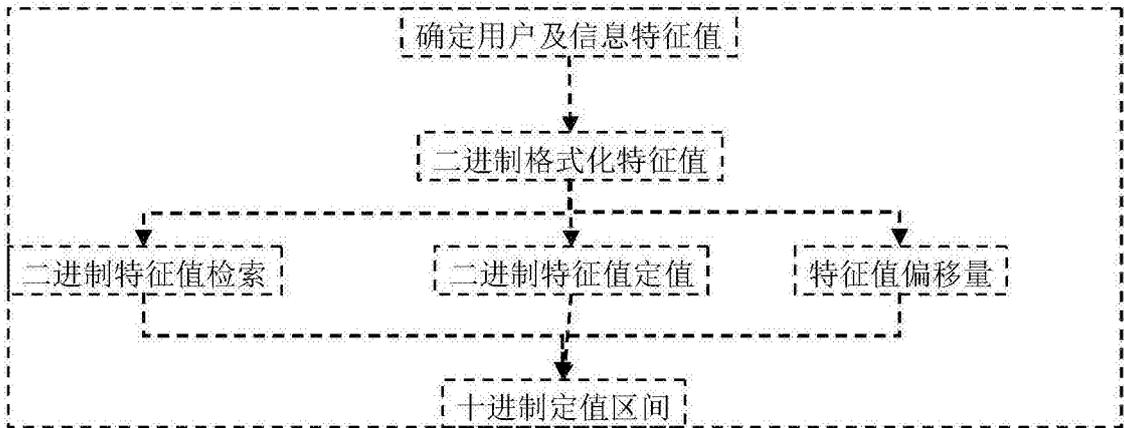


图 2