(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108022006 B
(45)授权公告日 2020.07.24

(21)申请号 201711195575.3

(22)申请日 2017.11.24

(65)同一申请的已公布的文献号
申请公布号 CN 108022006 A

(43)申请公布日 2018.05.11

(73)专利权人 浙江大学
地址 310013 浙江省杭州市西湖区余杭塘路866号

(72)发明人 巫英才 翁荻 朱鹤鸣

(74)专利代理机构 杭州天勤知识产权代理有限公司 33224
代理人 徐敏

(51)Int.Cl.
G06Q 10/04(2012.01)
G06Q 50/30(2012.01)

(56)对比文件
CN 106570062 A,2017.04.19,
CN 103150326 A,2013.06.12,
CN 106407378 A,2017.02.15,
KR 20170016203 A,2017.02.13,
袁野、王国仁.面向不确定图的概率可达查询.《计算机学报》.2010,第33卷1-9.

审查员 钟福煌

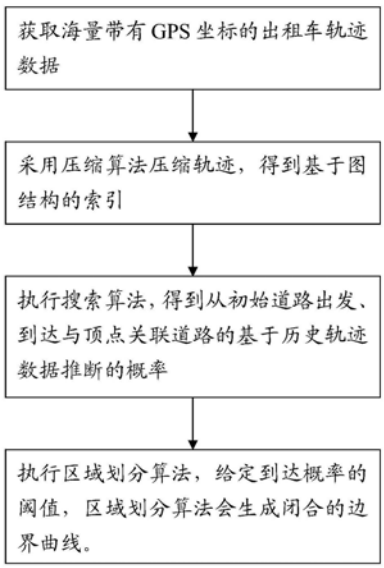
权利要求书2页 说明书4页 附图2页

(54)发明名称

一种数据驱动的可达性概率和区域生成方法

(57)摘要

本发明公开了一种数据驱动的可达性概率和区域生成方法,包括以下步骤:(1)处理出租车轨迹数据中的采样点,并将得到的采样点匹配到道路网络中,形成连续的轨迹数据;(2)扫描步骤(1)中处理得到的轨迹数据,生成基于图结构的轨迹索引;(3)通过步骤(2)建立的轨迹索引,对于用户发起的关于选定地点的可达性概率和区域的查询请求,采用带剪枝的算法搜索轨迹图中与查询请求相关的结点和边,以计算所选地点周围的可达性概率;(4)根据步骤(3)计算的可达性概率,结合预设的概率阈值,通过广度优先搜索算法搜索所述的可达性概率,在地图上划定在指定时间内从所选地点出发可达或可到达所述地点的区域。



CN 108022006 B

1.一种数据驱动的可达性概率和区域生成方法,其特征在于,包括以下步骤:

(1)处理出租车轨迹数据中的采样点,并将得到的采样点匹配到道路网络中,形成连续的轨迹数据;

(2)扫描步骤(1)中处理得到的轨迹数据,生成基于图结构的轨迹索引;

(3)通过步骤(2)建立的轨迹索引,对于用户发起的关于选定地点的可达性概率和区域的查询请求,采用带剪枝的算法搜索轨迹图中与查询请求相关的结点和边,以计算所选地点周围的可达性概率;

(4)根据步骤(3)计算的可达性概率,结合预设的概率阈值,通过广度优先搜索算法搜索所述的可达性概率,在地图上划定在指定时间内从所选地点出发可达或可到达所述地点的区域;

步骤(2)中,扫描步骤(1)中处理得到的轨迹数据,生成基于图结构的轨迹索引的具体步骤如下:

2-1、将一天分割成一个以v分钟为单位的时间片集合M＝{m_1,m_2,m_3...},其中每个时间片m_i的长度均为v分钟,i代表时间片的编号,给定道路网络的结构图G＝(V,E),定义轨迹图G_T＝(M×E,E_T),V代表交叉口,E代表道路,E_T代表轨迹图的边集;

2-2、对于任意一辆出租车轨迹中的任意一条记录R_j＝(t_j,t_{j+1},r_j)∈T_i,j是出租车轨迹记录的编号,t_j是起始时间,t_{j+1}是终止时间,r_j是这辆出租车在这段时间内所在的道路编号;

确定t_j和t_{j+1}所对应的时间片m_j和m_{j+1},若$m_j \neq m_{j+1}$,则增加一条边<m_j,r_j,m_{j+1},r_j>至边集E'$_T$;

2-3、对于任意一辆出租车轨迹中的任意连续两条记录$\{R_j=(t_j,t_{j+1},r_j),R_{j+1}=(t_{j+1},t_{j+2},r_{j+1})\}\subseteq T_i$,确定$t_{j+1}$对应的时间片$m_{j+1}$,若$r_j \neq r_{j+1}$,则增加一条边<$m_{j+1}$,$r_j$,$m_{j+1}$,$r_{j+1}$>至边集E'$_T$;

2-4、将步骤2-2和2-3得到E'$_T$进行压缩后得到边集E_T,基于图结构的索引即为最后生成得到的轨迹图G_T＝(M×E,E_T)。

2.如权利要求1所述的数据驱动的可达性概率和区域生成方法,其特征在于,步骤2-4中,将步骤2-2和2-3得到E'$_T$进行压缩后得到边集E_T,基于图结构的索引即为最后生成得到的轨迹图G_T＝(M×E,E_T)的具体过程如下:

2-4-1、令E_T＝{<m_i,r_u;m_j,r_v,b>,...},其中b为二进制位组成的日期集合,长度为所有出租车轨迹覆盖的天数,m_i表示起始时间片,m_j表示终止时间片,r_u表示起始道路,r_v表示终止道路;

2-4-2、对于边集E'$_T$中的每条边e'＝<m_i,r_u;m_j,r_v>,e'对应的轨迹被记录的日期d,若存在一条边e与e'部分匹配,即e与e'前四个分量相同,则更新e使得b＝b∪d;否则,向E_T插入一条新的边e＝<e',{d}>。

3.如权利要求2所述的数据驱动的可达性概率和区域生成方法,其特征在于,如权利要求1所述的数据驱动的可达性概率和区域生成方法,其特征在于,步骤(3)中,基于步骤(2)生成的轨迹索引计算可达性概率的带剪枝图搜索算法具体步骤如下:

3-1、寻找一组连续的且与时间跨度对应的时间片M',计算在给定的开始时间t和持续时间L组成的时间跨度[t,t+L]中从所选地点r_0出发可达的或是到达所选地点r_0的概率;

3-2、给定步骤(2)生成的轨迹索引,从轨迹索引中检索一个顶点集合V＝{(m,r),...},

轨迹图中的顶点集合是包含多个时间片和道路组成的元组，满足m∈M'且r＝r_0;其中m表示时间片，r表示道路;轨迹索引中的每一个顶点通过时间片m与道路r构成的元组表示;

3-3、对每个顶点v∈V初始化广度优先搜索，其中，v表示轨迹图顶点集合中的顶点;

3-4、对每个顶点v∈V执行带剪枝的广度优先搜索;

3-5、遍历轨迹图中所有顶点，计算顶点关联道路的到达概率。

4.如权利要求3所述的数据驱动的可达性概率和区域生成方法，其特征在于，步骤(4)中，根据步骤(3)计算的可达性概率和用户设定的概率阈值，计算可达区域的划定具体过程如下：

4-1、从轨迹图中所选地点r_0出发开始广度优先搜索，终止条件为顶点关联的可达概率低于设定的概率阈值，得到终止顶点集合V;

4-2、对步骤4-1得到的终止顶点集合V计算凹包，得到可达区域。

5.如权利要求4所述的数据驱动的可达性概率和区域生成方法，其特征在于，步骤4-2中，采用滚球法或Delaunay三角化法计算凹包。

6.如权利要求3所述的数据驱动的可达性概率和区域生成方法，其特征在于，步骤3-3中，对每个顶点v∈V初始化广度优先搜索具体如下：

在起源于v的搜索过程开始时，将长度为D的位集合b与顶点v关联，且b中所有位被初始化为1，表示所有时间均可到达顶点v代表的道路。

7.如权利要求6所述的数据驱动的可达性概率和区域生成方法，其特征在于，步骤3-4中，对每个顶点v∈V执行带剪枝的广度优先搜索具体如下：

3-4-1、令当前节点为v，当前位集合为与v关联的b，检索v所有的邻居顶点，令b与连接邻居顶点的边上的位集合相交，将v所有的邻居节点的位集合与当前位集合求交，作为新的位集合;

3-4-2、若位集合为非空，将邻居节点加入搜索队列，否则作为剪枝条件之一，拒绝此顶点入队;

3-4-3、若预计的行驶时间超出时间片长度，算法将停止在此时间片内的搜索;

3-4-4、重复步骤3-4-1～3-4-3，直至遍历所有入队顶点。

8.如权利要求7所述的数据驱动的可达性概率和区域生成方法，其特征在于，3-5、遍历轨迹图中所有顶点，计算顶点关联道路的到达概率具体如下：

令总天数为D，当前遍历的顶点为v，将v关联的位集合b中的元素个数除以D，即得当前顶点对应道路的到达概率。

9.如权利要求1所述的数据驱动的可达性概率和区域生成方法，其特征在于，步骤(1)中，处理得到的轨迹数据以一系列在时间上连续的四元组格式存储于本地文件中，所述四元组包括车辆编号、到达时间、离开时间和道路编号的信息。

一种数据驱动的可达性概率和区域生成方法

技术领域

[0001] 本发明涉及数据库及数据挖掘领域，尤其涉及一种数据驱动的可达性概率和区域生成方法。

背景技术

[0002] 可达性的概率和区域计算允许人们预测在一定的时间范围内从城市中的一个地点到其他区域的概率，这使得它在城市空间中的选址问题、预测车辆行驶时间、业务覆盖率分析等方面有着广泛的应用。然而，现有的可达性估计算法主要基于物理距离计算。例如，许多酒店或住房推荐系统允许用户根据离地标建筑的远近过滤候选的地点；许多导航软件仅根据规划路线的距离和沿路的拥堵情况粗略估计时间。

[0003] 在大多数情况下，由于存在多种可能的行驶路线、变化的交通流量、恶劣的天气情况等原因，算法无法准确地通过物理距离估计可达性。随着传感器和数据采集技术的发展，通过城市大数据，监控城市的运行秩序、解决出行或城市地理规划相关的问题、地理上的发展满足人类的潜在需求等一系列设想已经逐渐成为现实。先前的研究表明，长时间、大规模的出租车轨迹数据能够有效地揭示隐藏在城市道路流量中的交通模式。因此，利用这种在许多城市可以公开获取的数据，本专利所描述的一种数据驱动的可达性概率和区域生成方法能够帮助人们精确地估计城市多变的环境下可达概率和区域。

发明内容

[0004] 本发明提供了一种数据驱动的可达性概率和区域生成方法，提供了可靠、高效地可达性计算方法，在城市空间中的选址问题、预测车辆行驶时间、业务覆盖率分析等方面有着广泛的应用，并可扩展解决其他与地理距离相关的问题。

[0005] 一种数据驱动的可达性概率和区域生成方法，包括以下步骤：

[0006] (1)处理出租车轨迹数据中的采样点，并将得到的采样点匹配到道路网络中，形成连续的轨迹数据；

[0007] (2)扫描步骤(1)中处理得到的轨迹数据，生成基于图结构的轨迹索引；

[0008] (3)通过步骤(2)建立的轨迹索引，对于用户发起的关于选定地点的可达性概率和区域的查询请求，采用带剪枝的算法搜索轨迹图中与查询请求相关的结点和边，以计算所选地点周围的可达性概率，同时还计算得到概率密度信息；概率密度信息一般在计算可达性概率时同时得到。

[0009] (4)根据步骤(3)计算的可达性概率以及概率密度信息，结合预设的概率阈值，通过广度优先搜索算法搜索所述的可达性概率，在地图上划定在指定时间内从所选地点出发可达或可到达所述地点的区域。

[0010] 本发明可处理规模极大的、包含上千万个有效GPS坐标点的出租车轨迹数据，且处理的数据规模可在保证查询效率的情况下，根据运行平台的硬件配置弹性变化。优选的，步骤(1)中，处理得到的轨迹数据以一系列在时间上连续的四元组(车辆编号、到达时间、离开

时间、道路编号)格式存储于本地文件中。四元组即为一种轨迹数据。

[0011] 由于轨迹数据量极其庞大,传统的线性扫描算法无法在短时间内产生可达性结果。为了高效地计算可达概率,必须先对轨迹数据建立索引,优选的,优选的,步骤(2)中,扫描步骤(1)中处理得到的轨迹数据,生成基于图结构的轨迹索引的具体步骤如下:

[0012] 2-1、将一天分割成一个以v分钟为单位的时间片集合M=$\{m_1,m_2,m_3...\}$,其中每个时间片m_i的长度均为v分钟,给定道路网络的结构图G=(V,E),定义轨迹图$G_T=(M\times E,E_T)$,V代表交叉口,E代表道路,E_T代表轨迹图的边集;

[0013] 计算轨迹图的边集ET之前,算法需要先扫描所有步骤(1)中匹配得到的轨迹T={T1,T2,T3,...}以构造一个未被压缩的边集E'T=$\{<mi,ru;mj,rv>,...\}$,压缩E'T得到ET,具体包括步骤2-2～2-4。

[0014] 2-2、对于任意一辆出租车轨迹中的任意一条记录$R_j=(t_j,t_{j+1},r_j)\in T_i$,$t_j$是起始时间,$t_{j+1}$是终止时间,$r_j$是这辆出租车在这段时间内所在的道路编号;

[0015] 确定t_j和t_{j+1}所对应的时间片m_j和m_{j+1},若$m_j\neq m_{j+1}$,则增加一条边$<m_j,r_j,m_{j+1},r_j>$至边集E'_T;

[0016] 2-3、对于任意一辆出租车轨迹中的任意连续两条记录$\{R_j=(t_j,t_{j+1},r_j),R_{j+1}=(t_{j+1},t_{j+2},r_{j+1})\}\subseteq T_i$,确定$t_{j+1}$对应的时间片$m_{j+1}$,若$r_j\neq r_{j+1}$,则增加一条边$<m_{j+1},r_j,m_{j+1},r_{j+1}>$至边集$E'_T$,其中i表示起始时间篇与道路的编号,j表示终止时间片与道路的编号;

[0017] 2-4、将步骤2-2和2-3得到E'_T进行压缩后得到边集E_T,基于图结构的索引即为最后生成得到的轨迹图$G_T=(M\times E,E_T)$。

[0018] 为了提高压缩的效率,同时保证数据的完整性,优选的,步骤2-4中,将步骤2-2和2-3得到E'_T进行压缩后得到边集E_T,基于图结构的索引即为最后生成得到的轨迹图$G_T=(M\times E,E_T)$的具体过程如下:

[0019] 2-4-1、令$E_T=\{<m_i,r_u;m_j,r_v,b>,...\}$,其中b为二进制位组成的日期集合,长度为所有出租车轨迹覆盖的天数;

[0020] 2-4-2、对于边集E'_T中的每条边$e'=<m_i,r_u;m_j,r_v>$,e'对应的轨迹被记录的日期d,若存在一条边e与e'部分匹配,即e与e'前四个分量相同,则更新e使得$b=b\cup d$;否则,向E_T插入一条新的边$e=<e',\{d\}>$;u、v为道路编号,m_i表示起始时间片,m_j表示终止时间片,r_u表示起始道路,r_v表示终止道路。

[0021] 通过这种压缩数据中冗余轨迹的方法,可以获取极小的、可以直接存放在内存中、支持高速查询的索引,生成的索引可使用Boost标准库提供的序列化库存放在磁盘上。

[0022] 为了从生成的轨迹索引中准确计算可达性概率,优选的,步骤(3)中,基于步骤(2)生成的轨迹索引计算可达性概率的带剪枝图搜索算法具体步骤如下:

[0023] 3-1、寻找一组连续的且与时间跨度对应的时间片M',计算在给定的开始时间t和持续时间L组成的时间跨度[t,t+L)中从所选地点r_0出发可达的或是到达所选地点r_0的概率;

[0024] 3-2、给定步骤(2)生成的轨迹索引,从轨迹索引中检索一个顶点集合V={(m,r),...},满足$m\in M'$且$r=r_0$;其中m表示时间片,r表示道路;轨迹索引中的每一个顶点通过时间片m与道路r构成的元组表示;

[0025] 3-3、对每个顶点$v\in V$初始化广度优先搜索,其中,v表示轨迹图顶点集合中的顶

点；

[0026] 3-4、对每个顶点v∈V执行带剪枝的广度优先搜索；

[0027] 3-5、遍历轨迹图中所有顶点，计算顶点关联道路的到达概率。

[0028] 通过这个搜索算法可以得到起始道路相对于所有道路的到达概率。对应的，该算法可简易扩展用于计算所有道路相对于起始道路的到达概率。作为直观的可达性可视化表示，优选的，步骤(4)中，根据步骤(3)计算的可达性概率和用户设定的概率阈值，计算可达区域的划定具体过程如下：

[0029] 4-1、从轨迹图中所选地点r_0出发开始广度优先搜索，终止条件为顶点关联的可达概率低于设定的概率阈值，得到终止顶点集合V；

[0030] 4-2、对步骤4-1得到的终止顶点集合V计算凹包，得到可达区域。

[0031] 为了提高计算效率和准确性，优选的，步骤4-2中，采用滚球法或Delaunay三角化法计算凹包。

[0032] 为了提高计算效率和准确性，优选的，步骤3-3中，对每个顶点v∈V初始化广度优先搜索具体如下：

[0033] 在起源于v的搜索过程开始时，将长度为D的位集合b与顶点v关联，且b中所有位被初始化为1，表示所有时间均可到达顶点v代表的道路。

[0034] 为了提高计算效率和准确性，优选的，步骤3-4中，对每个顶点v∈V执行带剪枝的广度优先搜索具体如下：

[0035] 3-4-1、令当前节点为v，当前位集合为与v关联的b，检索v所有的邻居顶点，令b与连接邻居顶点的边上的位集合相交，将v所有的邻居节点的位集合与当前位集合求交，作为新的位集合；

[0036] 3-4-2、若位集合为非空，将邻居节点加入搜索队列，否则作为剪枝条件之一，拒绝此顶点入队；

[0037] 3-4-3、若预计的行驶时间超出时间片长度，算法将停止在此时间片内的搜索；

[0038] 3-4-4、重复步骤3-4-1～3-4-3，直至遍历所有入队顶点。

[0039] 为了提高计算效率和准确性，优选的，3-5、遍历轨迹图中所有顶点，计算顶点关联道路的到达概率具体如下：

[0040] 令总天数为D，当前遍历的顶点为v，将v关联的位集合b中的元素个数除以D，即得当前顶点对应道路的到达概率。

[0041] 本发明从大规模出租车轨迹数据中，计算在指定时间段内从城市中某个地点出发到达其他地点、或从城市中其他地点到达某个地点形成的可达性概率以及概率密度与分布区域，提高了计算的准确性、可靠性以及计算效率。

[0042] 本发明的有益效果：

[0043] 本发明的数据驱动的可达性概率和区域生成方法提供了一种可靠、高效地可达性计算方法，在城市空间中的选址问题、预测车辆行驶时间、业务覆盖率分析等方面有着广泛的应用，并可扩展解决其他与地理距离相关的问题。

附图说明

[0044] 图1是本发明的数据驱动的可达性概率和区域生成方法的流程线框图。

[0045] 图2是本发明的数据驱动的可达性概率和区域生成方法步骤(1)处理后的轨迹数据示意图。

[0046] 图3是本发明的数据驱动的可达性概率和区域生成方法步骤(2)基于图结构的轨迹索引示意图。

具体实施方式

[0047] 下面通过大规模出租车轨迹数据集的案例,结合附图详细描述本发明,本发明的目的和效果将变得更加明显。

[0048] 如图1所示,本实施例的数据驱动的可达性概率和区域生成方法包括以下步骤:

[0049] (1)获取海量带有GPS坐标的出租车轨迹数据,通过轨迹匹配算法(MapMatching)和轨迹修复技术得到四元组数据,并存储在本地磁盘上,处理后的数据如图2所示,图2中仅包含作为举例的2条轨迹T_1和T_2,四元组($T_{currrent}$,m_i,m_j,r_1)表示从时间片m_i到时间片m_j,$T_{current}$对应的出租车在道路r_1上;为了方便表示,图上略去了轨迹标号;其中current表示当前轨迹编号,i表示起始时间片编号,j表示终止时间片编号,l表示当前所在的道路编号。

[0050] (2)采用压缩算法压缩轨迹,得到基于图结构的索引,图3表示由图2中两条样例轨迹压缩得到的索引,可以看到(T_1,m_1,m_2,r_3)和(T_2,m_1,m_2,r_3)、(T_1,m_2,m_2,r_4)和(T_2,m_2,m_2,r_4)在图中被压缩为同一条边,通过这种压缩数据中冗余轨迹的方法,可以获取极小的、可以直接存放在内存中、支持高速查询的索引,生成的索引可使用Boost标准库提供的序列化库存放在磁盘上。

[0051] (3)执行搜索算法,得到从初始道路出发、到达与顶点关联道路的基于历史轨迹数据推断的概率,具体步骤如下:

[0052] 3-1、寻找与时间跨度对应的时间片集合;

[0053] 3-2、初始化广度优先搜索;

[0054] 3-3、对步骤(2)生成的图索引执行广度优先搜索,具体如下:

[0055] 3-3-1、执行剪枝操作,减少搜索空间大小;

[0056] 3-3-2、访问邻居顶点;

[0057] 3-3-3、更新邻居顶点的位集合,将需要访问的顶点加入搜索队列;

[0058] 3-4、处理搜索过程记录的信息,为所有顶点计算到达概率。

[0059] 生成的概率分布可以用热力图的形式呈现,颜色深的地方表示更容易从初始地点到达;对应的,颜色浅的地方表示概率更低。

[0060] (4)执行区域划分算法。给定到达概率的阈值,区域划分算法会生成闭合的边界曲线。闭合的曲线有助于用户直观地理解区域大小,以便进一步分析。

[0061] 本实施例方法阐述了将本发明应用于实际出租车轨迹数据的流程,该流程提供了一种把数据转化为可供用户直接分析的形式的途径。本发明应用过程简单,适应场景广泛。通过本发明,相关用户可以更好的了解城市动态变化的结构和脉络,为解决城市发展带来的问题打下坚实的基础。

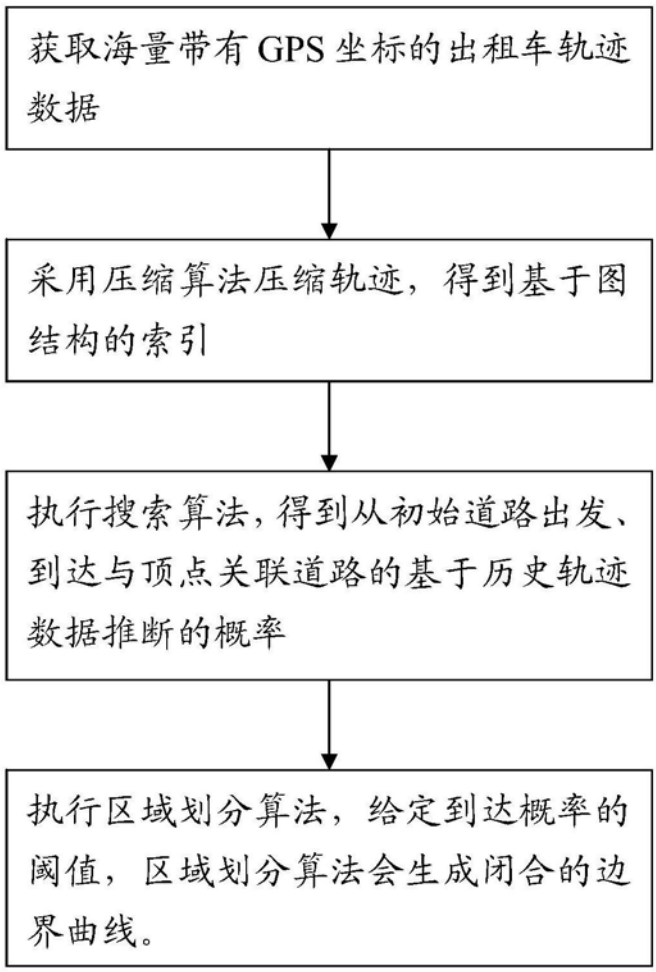

获取海量带有GPS坐标的出租车轨迹数据
采用压缩算法压缩轨迹，得到基于图结构的索引
执行搜索算法，得到从初始道路出发、到达与顶点关联道路的基于历史轨迹数据推断的概率
执行区域划分算法，给定到达概率的阈值，区域划分算法会生成闭合的边界曲线。

图1

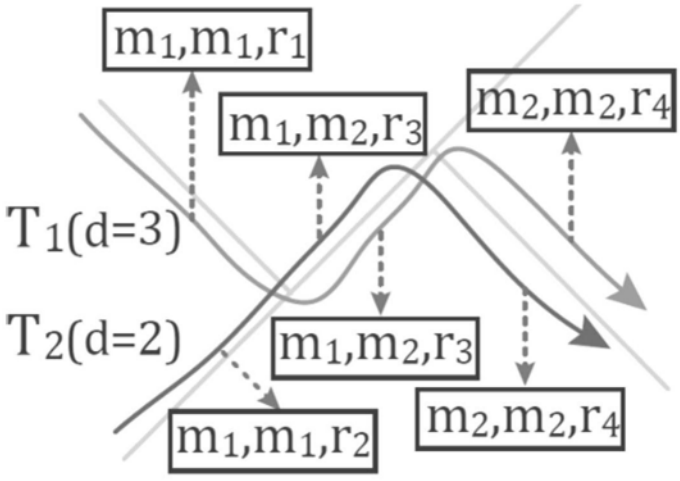

m1,m1,r1
m1,m2,r3
m2,m2,r4
T1(d=3)
T2(d=2)
m1,m2,r3
m2,m2,r4
m1,m1,r2

图2

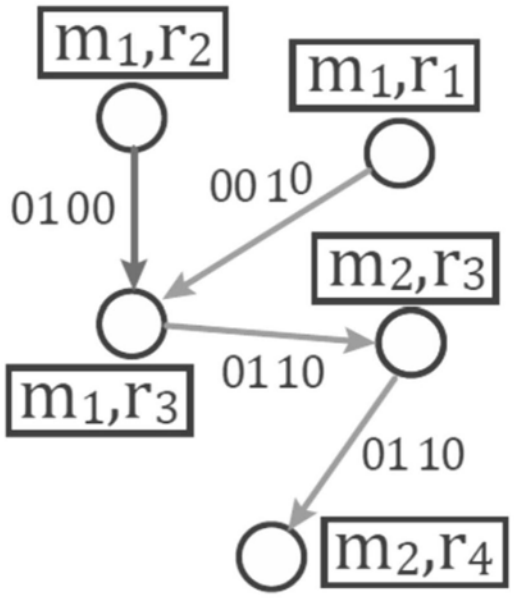


图3