(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 102867022 B
(45)授权公告日 2015.01.14

(21)申请号 201210285469.5
(22)申请日 2012.08.10
(73)专利权人 上海交通大学
地址 200240 上海市闵行区东川路800号
(72)发明人 朱其立 许信辉 贾枭 潘超
(74)专利代理机构 上海汉声知识产权代理有限公司 31236
代理人 郭国中
(51)Int.Cl.
G06F 17/30(2006.01)
(56)对比文件
CN 2010/127216 A2,2010.11.04, 全文.
CN 102542209 A,2012.07.04, 全文.
CN 101834872 A,2010.09.15, 全文.
CN 102156755 A,2011.08.17, 全文.
胡新平等. 基于敏感元组的隐私数据保护方法.《东南大学学报(自然科学版)》.2010, 第40卷(第5期),911－916.
王智慧等. 一种基于聚类的数据匿名方法.《软件学报》.2010, 第21卷(第4期),680－693.
许信辉等. 一种集合型数据匿名化的部分删除策略.《计算机工程》.2013, 第39卷(第11期),139－142.
韩建民等. 面向敏感值的个性化隐私保护.《电子学报》.2010, 第38卷(第7期),1723－1728.
审查员 杨春雨

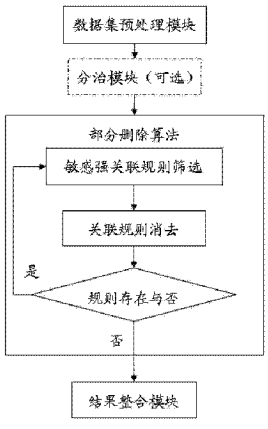
权利要求书2页 说明书4页 附图2页

(54)发明名称

通过部分删除某些项目达到对集合型数据匿名化的系统

(57)摘要

本发明提供通过部分删除某些项目达到对集合型数据匿名化的系统,其首先对集合型数据集进行预处理;其次利用多轮迭代方法对数据集中危险敏感的强关联规则进行消去并保证删除掉的项目尽量少。具体迭代的实施过程为:不断从数据集中筛选出敏感的强关联规则;从数据集中部分删除该规则中某些项目,以使得该危险敏感的强关联规则变为安全敏感的弱关联规则或不再存在于数据集中,直到最终数据集中不再存在危险敏感的强关联规则即可跳出该迭代过程。系统为了能让该匿名化处理过程以更快的速度进行,该系统结合了分而治之的思想,使得匿名化过程可以通过多个线程并发的执行,在保证不剧烈增加删除项目数目的前提下,匿名化处理过程效率大大提升。



CN 102867022 B

1. 一种通过部分删除某些项目达到对集合型数据匿名化的系统，其特征在于，包括数据集预处理模块、起到加速匿名化的分治模块、危险敏感的强关联规则筛选模块及通过部分删除方法实现的关联规则消去模块，还包括检测危险敏感的强关联规则存在与否模块和最终结果整合模块，其中：

- 数据集预处理模块，用于对原始集合型数据集进行前期处理，包括对数据集的信息统计，对项目的标识符进行正向哈希映射，对记录的排序及对记录的预删除处理；

- 危险敏感的强关联规则筛选模块，用于从数据集中筛选出危险敏感的强关联规则；

- 关联规则消去模块，用于对危险敏感的强关联规则筛选模块筛选出的敏感的强关联规则，利用部分删除策略使得危险敏感的强关联规则变为安全敏感的弱关联规则或不再存在于数据集中；

- 检测危险敏感的强关联规则存在与否模块，用于检查数据集中是否仍然存在危险敏感的强关联规则；

- 最终结果整合模块，用于将各个子数据集匿名化的结果进行整合，对项目的标识符进行反向哈希映射，并对整合后结果进行信息统计；

分治模块用于对数据集进行近似平均的划分，划分成大小近似的若干子数据集，并对各子数据集进行单独匿名化处理。

2. 根据权利要求 1 所述的通过部分删除某些项目达到对集合型数据匿名化的系统，其特征在于，所述数据集预处理模块对数据集进行信息统计，通过哈希映射对项目的标识符进行简化，再对记录进行排序和可配置的预删除处理，所得预处理结果传递给分治模块或危险敏感的强关联规则筛选模块进行下一步操作。

3. 根据权利要求 1 所述的通过部分删除某些项目达到对集合型数据匿名化的系统，其特征在于，所述危险敏感的强关联规则筛选模块通过使用固定大小的缓冲区存储遍历数据集过程中产生出的固定数目的关联规则。

4. 根据权利要求 3 所述的通过部分删除某些项目达到对集合型数据匿名化的系统，其特征在于，所述危险敏感的强关联规则筛选模块包括以下装置：

第一处理模块，用于遍历数据集中每一个记录，根据当前记录产生存在于该记录中的关联规则，将该关联规则存储于所述固定大小的缓冲区中，更新记录对应关联规则强弱信息的数据结构；

第二处理模块，用于当缓冲区被填满时，不再往缓冲区增加新的关联规则而只是更新记录缓冲区中关联规则强弱信息的数据结构；

第三处理模块，用于当遍历到长度超过一个固定阈值的记录时，则不再通过第一处理模块和第二处理模块产生关联规则和更新记录关联规则的强弱信息的数据结构，而是通过对整个数据集中包含关联规则中具体项目的记录编号集合直接做交集运算以确定关联规则的强弱；

第四处理模块，用于从缓冲区中筛选出所有危险敏感的强关联规则供关联规则消去模块作下一步处理。

5. 根据权利要求 1 所述的通过部分删除某些项目达到对集合型数据匿名化的系统，其特征在于，所述关联规则消去模块包括以下装置：

第五处理模块，用于从缓冲区中挑选某一危险敏感的强关联规则进行消去，确定消去

该规则应从哪些记录中删除具体的哪个项目；

第六处理模块，用于在通过所述第五处理模块选定的记录中删除选定的项目后，更新记录缓冲区中关联规则强弱信息的数据结构；

第七处理模块，用于当缓冲区中不再存在危险敏感的强关联规则，则进入检测危险敏感的强关联规则存在与否模块；否则回到所述第五处理模块重复操作。

6. 根据权利要求 1 所述的通过部分删除某些项目达到对集合型数据匿名化的系统，其特征在于，所述检测危险敏感的强关联规则存在与否模块扫描数据集确定是否仍有危险敏感的强关联规则存在，若存在则回到敏感的强关联规则筛选模块重复新的迭代的过程；否则进入结果整合模块。

7. 根据权利要求 2 所述的通过部分删除某些项目达到对集合型数据匿名化的系统，其特征在于，分治模块的运行前提是保证不剧烈增加删除项目数目。

通过部分删除某些项目达到对集合型数据匿名化的系统

技术领域

[0001] 本发明涉及计算机技术领域的系统框架，具体是通过部分删除某些项目达到对集合型数据匿名化的系统。

背景技术

[0002] 随着计算机技术的飞快发展和迅速普及，海量的数字信息正在悄无声息地繁殖。无论是政府组织、社会机构，还是公司团体、个人都在不经意间制造并收集着丰富的数据信息。与此同时纷繁的数字信息也给数据分析师和相关科研人员带来了新的契机和挑战。科学家和工程师们通过利用数字信息进行各类统计分析、知识挖掘等活动，形成总结式的认识和规则，引导今后的相关活动和决定、并可做出相关预测，最终加速技术进步、提高人们的生活品质。然而数字信息的传递与流通过程中，必须要慎重考虑安全与隐私问题。集合型数据作为一类颇具价值的数据源，广泛存在于我们的日常生活中，例如超市 / 网购购物清单、提交搜索引擎搜索关键字等等。但原始集合型数据中存在数据属主的大量隐私信息，如何匿名化集合型数据以实现对数据属主隐私的保护，同时保证匿名化后数据的有效性，成为近些年来研究的热门课题。

[0003] 目前，过去的研究成果中对集合型数据的匿名化方法多集中使用全局删除和全局泛化方法。Y.Xu（参见 Y.Xu,K.Wang,A.W.-C.Fu,and P.S.Yu.Anonymizing transaction databases for publication KDD 2008）很早就研究了集合型数据匿名化问题，通过使用全局删除非隐私条目的方法保护数据属主的隐私；而 J.Cao（参见 J.Cao,P.Karras,C.Raissi,and K.-L.Tan. ρ-uncertainty:inference-proof transaction anonymization VLDB 2010）同时使用了全局删除和全局泛化方法消去数据集中所有危险敏感的强关联规则。但由于全局删除方法使用大量剧烈删除操作，造成信息失真严重；而全局泛化方法不但改变了数据本身的模样，并且使用了并不被数据使用者公认的泛化分类结构。

发明内容

[0004] 本发明针对现有技术中存在的上述不足，提供了一种通过部分删除某些项目达到对集合型数据进行匿名化的算法，及一整套基于该算法实现的集合型数据匿名化系统框架。在保证尽可能少地删除条目的前提下，确保消去集合型数据中所有的危险敏感的强关联规则。

[0005] 根据本发明的一个方面，提供一种通过部分删除某些项目达到对集合型数据进行匿名化的系统，包括数据集预处理模块、起到加速匿名化的分治模块、危险敏感的强关联规则筛选模块及通过部分删除方法实现的关联规则消去模块，还包括检测危险敏感的强关联规则存在与否模块和最终结果整合模块，其中：

[0006] - 数据集预处理模块，用于对原始集合型数据集进行前期处理，包括对数据集的信息统计，对项目的标识符进行正向哈希映射，对记录的排序及对记录的预删除处理；

[0007] - 危险敏感的强关联规则筛选模块，用于从数据集中筛选出危险敏感的强关联规

则；

[0008] －关联规则消去模块，用于对危险敏感的强关联规则筛选模块筛选出的敏感的强关联规则，利用部分删除策略使得危险敏感的强关联规则变为安全敏感的弱关联规则或不再存在于数据集中；

[0009] －检测危险敏感的强关联规则存在与否模块，用于检查数据集中是否仍然存在危险敏感的强关联规则；

[0010] －最终结果整合模块，用于将各个子数据集匿名化的结果进行整合，对项目的标识符进行反向哈希映射，并对整合后结果进行信息统计。

[0011] 优选地，所述数据集预处理模块对数据集进行信息统计，通过哈希映射对项目的标识符进行简化，再对记录进行排序和可配置的预删除处理，所得预处理结果传递给分治模块或危险敏感的强关联规则筛选模块进行下一步操作。

[0012] 优选地，还包括分治模块，其中，分治模块用于对数据集进行近似平均的划分，划分成大小近似的若干子数据集，并对各子数据集进行单独匿名化处理。

[0013] 优选地，所述危险敏感的强关联规则筛选模块通过使用固定大小的缓冲区存储遍历数据集过程中产生出的固定数目的关联规则。

[0014] 优选地，所述危险敏感的强关联规则筛选模块包括以下装置：

[0015] 第一处理模块，用于遍历数据集中每一个记录，根据当前记录产生存在于该记录中的关联规则，将该关联规则存储于所述固定大小的缓冲区中，更新记录对应关联规则强弱信息的数据结构；

[0016] 第二处理模块，用于当缓冲区被填满时，不再往缓冲区增加新的关联规则而只是更新记录缓冲区中关联规则强弱信息的数据结构；

[0017] 第三处理模块，用于当遍历到长度超过一个固定阈值的记录时，则不再通过第一处理模块和第二处理模块产生关联规则和更新记录关联规则的强弱信息的数据结构，而是通过对整个数据集中包含关联规则中具体项目的记录编号集合直接做交集运算以确定关联规则的强弱；

[0018] 第四处理模块，用于从缓冲区中筛选出所有危险敏感的强关联规则供关联规则消去模块作下一步处理。

[0019] 优选地，所述关联规则消去模块对危险敏感的强关联规则进行消去。

[0020] 优选地，所述关联规则消去模块包括以下装置：

[0021] 第五处理模块，用于从缓冲区中挑选某一危险敏感的强关联规则进行消去，确定消去该规则应从哪些记录中删除具体的哪个项目；

[0022] 第六处理模块，用于在通过所述第五处理模块选定的记录中删除选定的项目后，更新记录缓冲区中关联规则强弱信息的数据结构；

[0023] 第七处理模块，用于当缓冲区中不再存在危险敏感的强关联规则，则进入检测危险敏感的强关联规则存在与否模块；否则回到所述第五处理模块重复操作。

[0024] 优选地，所述检测危险敏感的强关联规则存在与否模块扫描数据集确定是否仍有危险敏感的强关联规则存在，若存在则回到敏感的强关联规则筛选模块重复新的迭代的过程；否则进入结果整合模块。

[0025] 优选地，所述最终结果整合模块将各个子数据集匿名化的结果进行整合。

[0026] 优选地，分治模块的运行前提是保证不剧烈增加删除项目数目。

[0027] 本发明工作时，先对原始数据集进行信息统计，通过哈希映射对项目的标识符进行简化，再对记录进行排序和预删除处理(可选)，所得预处理结果传递给分治模块或危险敏感的强关联规则筛选模块进行下一步操作。分治模块得到预处理的数据集后，进行近似平均的划分，划分成大小近似的若干子数据集，并对各子数据集单独进行随后的匿名化处理。危险敏感的强关联规则筛选模块通过使用固定大小的缓冲区存储遍历数据集过程中产生出的固定数目的关联规则，具体地，其通过子模块(第一处理模块、第二处理模块、第三处理模块、第四处理模块)实现的筛选功能如下：遍历数据集中每一个记录，根据当前记录产生存在于该记录中的关联规则，将该关联规则存储于如上描述的大小固定的缓冲区中，更新记录对应关联规则强弱等信息的数据结构；当缓冲区被填满时，不再往缓冲区增加新的关联规则而只是更新记录缓冲区中关联规则强弱等信息的数据结构；当遍历到长度超过一个固定阈值(算法的一个输入参数)的记录时，通过对整个数据集中包含关联规则中具体项目的记录编号集合直接做交集运算以确定关联规则的强弱；从缓冲区中筛选出所有危险敏感的强关联规则供关联规则消去模块作下一步处理。关联规则消去模块对危险敏感的强关联规则进行消去，具体地，其通过子模块(第五处理模块、第六处理模块、第七处理模块)实现的消去功能如下：从缓冲区中挑选某一危险敏感的强关联规则进行消去，确定消去该规则应从哪些记录中删除具体的哪个项目；在从选定的记录中删除选定的项目后，更新记录缓冲区中关联规则强弱等信息的数据结构；若缓冲区中不再存在危险敏感的强关联规则，则进入检测危险敏感的强关联规则存在与否模块；否则继续处理缓冲区中的危险敏感的强关联规则。检测危险敏感的强关联规则存在与否模块扫描数据集确定是否仍有危险敏感的强关联规则存在，若存在则回到敏感的强关联规则筛选模块重复新的迭代的过程；否则进入最终结果整合模块。而最终结果整合模块将各个子数据集匿名化的结果进行最终整合。

[0028] 与现有技术相比，本发明创造性的使用了部分删除条目的方法对集合型数据进行匿名化，弥补了现有全局删除和全局泛化的集合型数据匿名化方法的缺陷和不足。本发明引入一定大小的缓冲区来存储数目巨大的关联规则，使用迭代法不断消去危险关联规则，引入分而治之的思想通过多线程技术加速匿名化的处理过程，并将短记录与长记录区别处理，还提供了预删除处理的选项，使得该算法正确高效地完成了集合型数据的匿名化，并极大程度地保持了剩余数据的使用价值。

附图说明

[0029] 通过阅读参照以下附图对非限制性实施例所作的详细描述，本发明的其它特征、目的和优点将会变得更明显：

[0030] 图 1 示出本发明的系统框架模块框图；

[0031] 图 2 示出本发明的数据集预处理模块和分治模块的实施细节；

[0032] 图 3 示出本发明的危险敏感的强关联规则筛选模块中关联规则的生成细节；

[0033] 图 4 示出本发明的关联规则消去模块消去敏感规则的实施细节。

具体实施方式

[0034] 下面结合附图对本发明的实施例作详细说明，本实施例在以发明技术方案为前提

下进行实施，给出了详细的实施方式和具体的操作过程，但本发明的保护范围不限于下述的实施例。

[0035] 本实施例的任务是对一简化集合型数据集进行匿名化，该数据集为记录一 (a)、记录二 (a，b)、记录三 (a，d，c)、记录四 (b，c)、记录五 (d)，其中项目 a、c、d 为隐私条目，仅项目 b 为非隐私条目，且要求对该数据集匿名化后的结果中所有敏感关联规则的置信度 (confidence) 不高于 0.5。

[0036] 如图 1 所示，本实施例包括 6 个模块：数据集预处理模块、起到加速匿名化的分治模块、危险敏感的强关联规则筛选模块及通过部分删除方法实现的关联规则消去模块，还包括检测危险敏感的强关联规则存在与否的模块和最终结果整合模块。所述数据集预处理模块，用于对原始集合型数据集进行前期处理，包括对数据集的信息统计，对项目的标识符进行正向哈希映射，对记录的排序及对记录的预删除处理等。所述分治模块，用于对数据集进行近似平均的划分，划分成大小近似的若干子数据集，并对各子数据集进行单独匿名化处理。所述危险敏感的强关联规则筛选模块，用于从数据集中筛选出危险敏感的强关联规则。所述关联规则消去模块，利用部分删除策略使得危险敏感的强关联规则变为“安全”敏感的弱关联规则或不再存在于数据集中。所述检测危险敏感的强关联规则存在与否的模块，检查数据集中是否仍然存在危险敏感的强关联规则。所述最终结果整合模块，用于将各个子数据集匿名化的结果进行整合，对项目的标识符进行反向哈希映射，并对整合后结果进行信息统计等。

[0037] 在本实施例的一个优选例中，所述分治模块可以被省略。

[0038] 如图 2 所示，数据集预处理模块首先将原始项目编号经哈希映射后映射到简易的整数编号，再将数据集中的记录按照记录长度递增序进行排序。随后分治模块将数据集划分成大小近似的两个子数据集，等待紧接着的其它模块分别对两个子数据集进行处理。图 2 省略了预处理模块中对数据集进行信息统计的结果展示。

[0039] 如图 3 所示，危险敏感的强关联规则筛选模块分别对两个子数据集进行处理，筛选出各子数据集中的敏感的关联规则。具体敏感关联规则的生成和对应置信度的计算参照“发明内容”部分中敏感的强关联规则筛选模块的描述。

[0040] 如图 4 所示，关联规则消去模块不断从缓冲区中挑选某一危险敏感的强关联规则进行消去，确定消去该规则应从哪些记录中删除具体的哪个项目；在从选定的记录中删除选定的项目后，更新记录缓冲区中关联规则强弱等信息的数据结构；若缓冲区中不再存在危险敏感的强关联规则，则进入检测危险敏感的强关联规则存在与否模块。

[0041] 对各子数据集分别进行匿名化过程后，利用结果整合模块，将各个子数据集匿名化的结果进行整合，对项目的标识符进行反向哈希映射，并对整合后结果进行信息统计等。这样就完成了对原始数据集的匿名化任务。

[0042] 以上对本发明的具体实施例进行了描述。需要理解的是，本发明并不局限于上述特定实施方式，本领域技术人员可以在权利要求的范围内做出各种变形或修改，这并不影响本发明的实质内容。

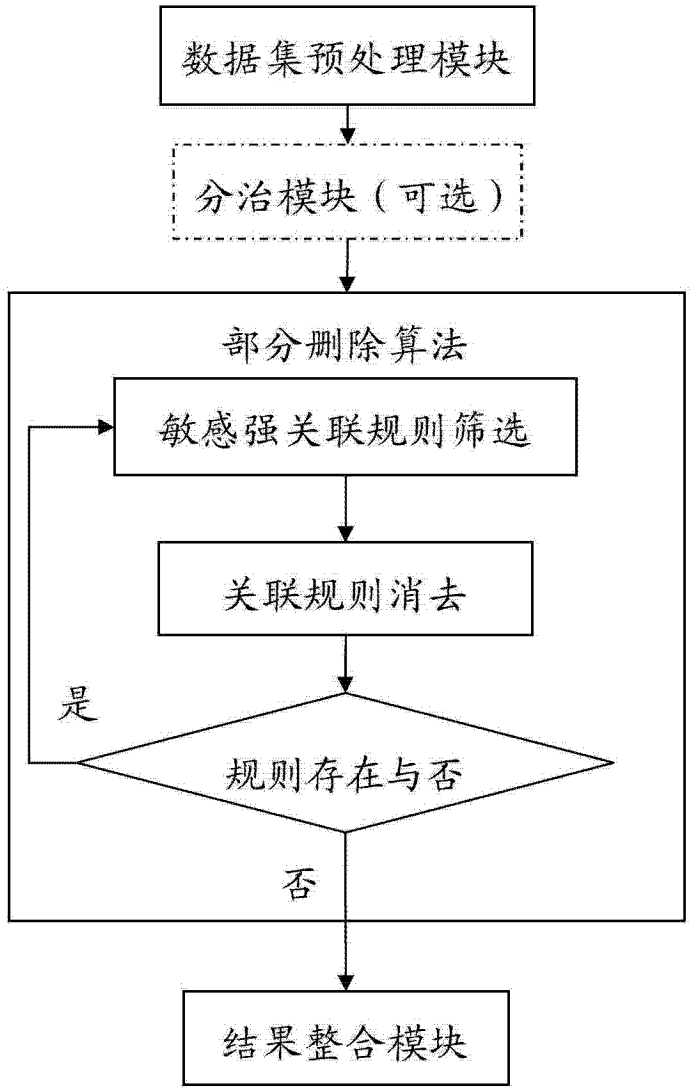

数据集预处理模块
分治模块（可选）
部分删除算法
敏感强关联规则筛选
关联规则消去
是
规则存在与否
否
结果整合模块

图 1

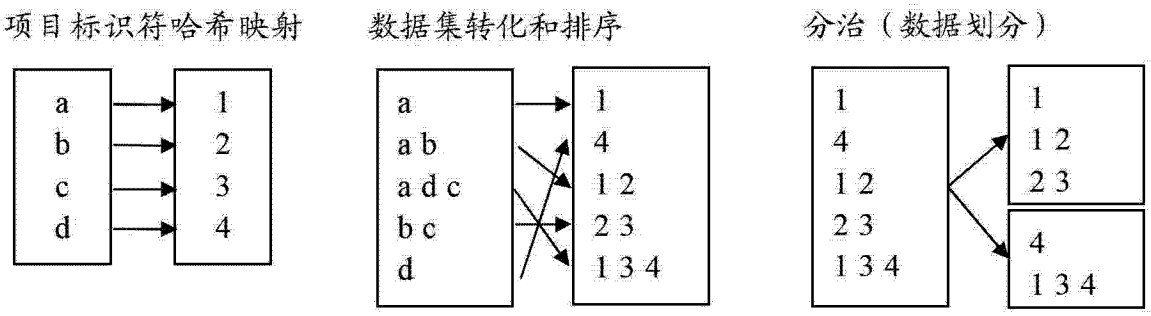

项目标识符哈希映射
a
b
c
d
1
2
3
4
数据集转化和排序
a
a b
a d c
b c
d
1
4
1 2
2 3
1 3 4
分治（数据划分）
1
4
1 2
2 3
1 3 4
1
1 2
2 3
4
1 3 4

图 2

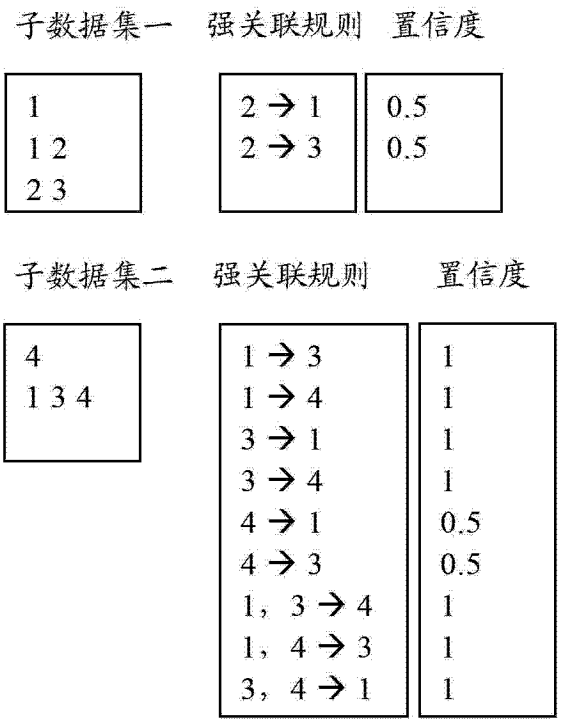

子数据集一
强关联规则
置信度
1
1 2
2 3
2 → 1
2 → 3
0.5
0.5
子数据集二
强关联规则
置信度
4
1 3 4
1 → 3
1 → 4
3 → 1
3 → 4
4 → 1
4 → 3
1，3 → 4
1，4 → 3
3，4 → 1
1
1
1
1
0.5
0.5
1
1
1

图 3

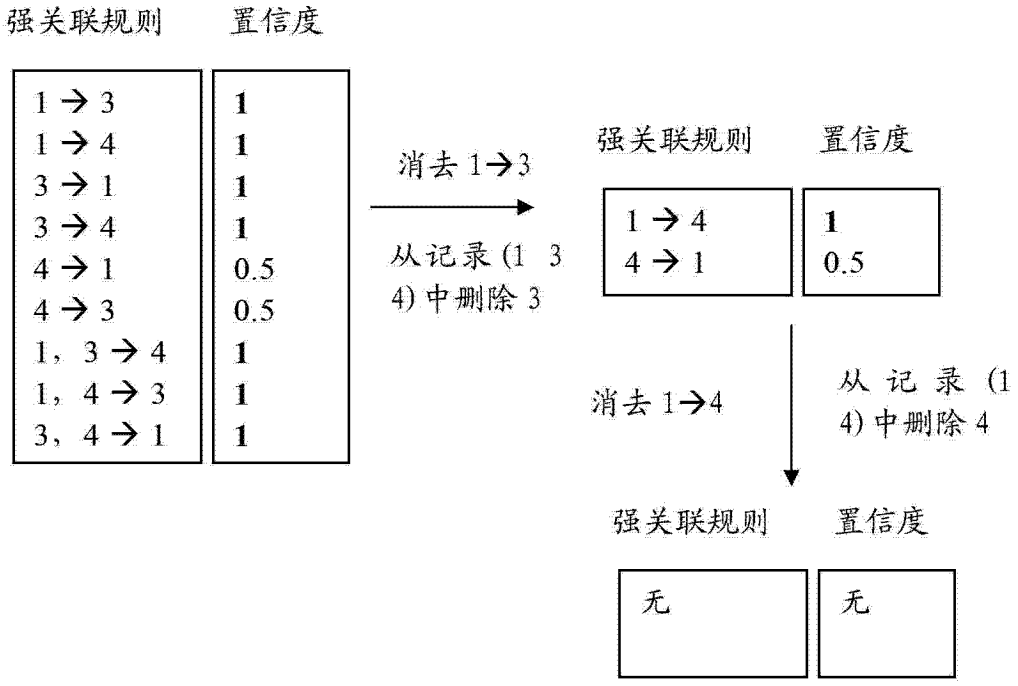

强关联规则
置信度
1 → 3
1 → 4
3 → 1
3 → 4
4 → 1
4 → 3
1，3 → 4
1，4 → 3
3，4 → 1
1
1
1
1
0.5
0.5
1
1
1
消去 1→3
从记录 (1 3 4) 中删除 3
强关联规则
置信度
1 → 4
4 → 1
1
0.5
消去 1→4
从记录 (1 4) 中删除 4
强关联规则
置信度
无
无

图 4