



(12) 发明专利

(10) 授权公告号 CN 107229726 B

(45) 授权公告日 2021.02.19

(21) 申请号 201710417077.2

(22) 申请日 2017.06.06

(65) 同一申请的已公布的文献号
申请公布号 CN 107229726 A

(43) 申请公布日 2017.10.03

(73) 专利权人 成都索贝数码科技股份有限公司
地址 610041 四川省成都市高新区新园南
二路2号

(72) 发明人 吴春中 张浩阳

(74) 专利代理机构 成都金英专利代理事务所
(普通合伙) 51218

代理人 袁英

(51) Int.Cl.

G06F 16/22 (2019.01)

(56) 对比文件

US 2011282864 A1, 2011.11.17

CN 106599062 A, 2017.04.26

CN 1670729 A, 2005.09.21

EP 2608068 A1, 2013.06.26

陈燕. “浅谈如何提高基于WEB的软件系统的
查询速度”.《现代营销》.2011,第375页.

审查员 魏留强

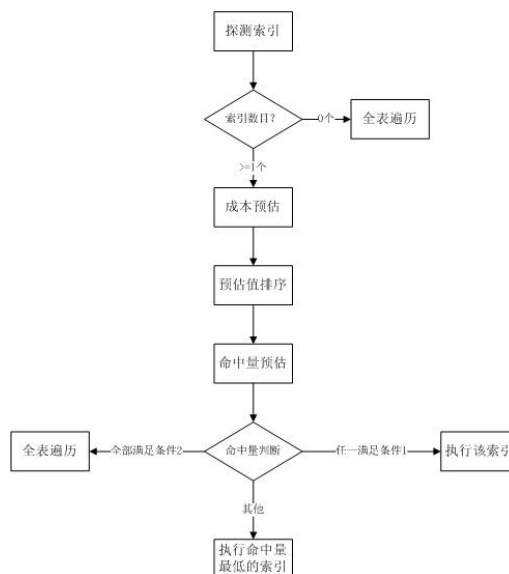
权利要求书2页 说明书6页 附图1页

(54) 发明名称

一种基于完全成本计算的数据索引方法及
系统

(57) 摘要

本发明公开了一种基于完全成本计算的数据索引方法及系统,涉及数据索引应用计算领域;一种基于完全成本计算的数据索引方法,包括:探测可能使用的索引、判断可使用的索引的数目、成本预估、预估值排序和命中量预估;一种基于完全成本计算的数据索引装置,其特征在于,它包括:索引探测模块、索引数目判断模块、成本预估模块、预估排序模块和命中量预估模块;一种终端系统,包括:处理器和存储器。本发明实现在每次执行查询操作时都对查询条件包含的索引进行成本的评估,选择操作成本更少的条件作为索引再进行下一步查询,能够有效提高检索的效率。



1. 一种基于完全成本计算的数据索引方法,其特征在于,它具体包括如下步骤:

S003成本预估:对每一个索引字段进行成本预估并得到预估值;针对于不同的索引字段类型的成本评估具体为:

整数型:条件扫描范围 $\div$ 索引有效数据范围=预估值;

浮点型:条件扫描范围 $\div$ 索引有效数据范围=预估值;

字符串、复合索引、二叉树索引:固定值 $\div$ 索引条件中的字符数=预估值;所述固定值为100;

S004预估值排序:将所得到的预估值按从小到大进行排序;

S005命中量预估:基于S004预估值排序的结果,依次使用对应的索引字段,直接定位索引中的开始和结束位置,预估命中数据量,并根据命中量结果符合以下条件之一进行判断执行相应的方案;

S501当任意一个索引满足预设条件1,即命中量小于1%且小于128条时,直接选择该索引执行;

S502当所有的索引都满足预设条件2,即命中量大于20%时,直接选择全表遍历;

S503其他情况下,执行命中量最低的索引。

2. 根据权利要求1所述的一种基于完全成本计算的数据索引方法,还包括:

S001探测可能使用的索引;

S002判断可使用的索引的数目:索引数目为0,直接全表遍历,反之进行下一个步骤。

3. 根据权利要求2所述的一种基于完全成本计算的数据索引方法,其特征在于:所述的步骤S001具体为:

S101若没有绑定变量,直接进行下一步骤,若有绑定变量,先解绑定再进行下一步骤;

S102生成BSON文件;

S103根据SQL语句中的条件结合表索引情况初步筛选可使用的索引。

4. 根据权利要求3所述的一种基于完全成本计算的数据索引方法,其特征在于:所述的SQL语句中的条件,包括:检索、更新、删除语句的一种或多种。

5. 根据权利要求1所述的一种基于完全成本计算的数据索引方法,其特征在于,所述的预估值和命中量还包括:

在一个索引字段中同时出现多个AND关系的条件选择成本小的条件使用索引,计算预估值和命中量;

在一个索引字段中同时出现多个OR关系的条件分别使用索引后合并,整体计算预估值和命中量。

6. 一种基于完全成本计算的数据索引装置,其特征在于,它包括:索引探测模块、索引数目判断模块、成本预估模块、预估排序模块和命中量预估模块;

所述的索引探测模块探测可能使用的索引;

所述的索引数目判断模块判断可使用的索引的数目;

所述的成本预估模块对每一个索引字段进行成本预估;所述成本预估具体为:

整数型:条件扫描范围 $\div$ 索引有效数据范围=预估值;

浮点型:条件扫描范围 $\div$ 索引有效数据范围=预估值;

字符串、复合索引、二叉树索引:固定值 $\div$ 索引条件中的字符数=预估值;所述固定值为

100;

所述的预估排序模块将所得到的预估值按从小到大进行排序;

所述的命中量预估模排序依次使用索引直接定位开始和结束位置,预估命中数据量,并根据命中量结果进行判断;所述根据命中量结果进行判断过程具体为:当任意一个索引满足预设条件1,即命中量小于1%且小于128条时,直接选择该索引执行;当所有的索引都满足预设条件2,即命中量大于20%时,直接选择全表遍历;其他情况下,执行命中量最低的索引。

7.一种终端系统,其特征在于,所述终端系统包括:

处理器,用于执行程序;

存储器,用于存储由处理器执行的程序,其中所述程序在执行时包括以下步骤:

S001探测可能使用的索引;

S002判断可使用的索引的数目:索引数目为0,直接全表遍历,反之进行下一个步骤;

S003成本预估:对每一个索引字段进行成本预估并得到预估值;针对于不同的索引字段类型的成本评估具体为:

整数型:条件扫描范围 $\div$ 索引有效数据范围=预估值;

浮点型:条件扫描范围 $\div$ 索引有效数据范围=预估值;

字符串、复合索引、二叉树索引:固定值 $\div$ 索引条件中的字符数=预估值;所述固定值为100;

S004预估值排序:将所得到的预估值按从小到大进行排序;

S005命中量预估:基于S004预估值排序的结果,依次使用对应的索引字段,直接定位索引中的开始和结束位置,预估命中数据量,并根据命中量结果符合以下条件之一进行判断执行相应的方案;

S501当任意一个索引满足预设条件1,即命中量小于1%且小于128条时,直接选择该索引执行;

S502当所有的索引都满足预设条件2,即命中量大于20%时,直接选择全表遍历;

S503其他情况下,执行命中量最低的索引。

一种基于完全成本计算的数据索引方法及系统

技术领域

[0001] 本发明涉及数据索引应用计算领域,尤其是一种基于完全成本计算的数据索引方法及系统。

背景技术

[0002] 索引的性能是数据库性能提升的关键,传统SQL数据库架构设计处于内存昂贵的时代,为了减少对内存的需求,SQL语句的执行过程如下:

[0003] 检查语法是否有拼写错误;

[0004] 检查访问对象是否有相应权限;

[0005] 利用内部的hash算法来取得该SQL的hash值,然后和缓存中的hash值对比;

[0006] 若存在相同的hash值则直接执行缓存中的执行计划,称为软解析;

[0007] 若不存在,则对该SQL进行解析,生成解析树及执行计划,称为硬解析;

[0008] 返回执行结果。

[0009] 硬解析会进行基于成本的索引选择,但对内存的需求高;软解析确实降低了对内存的依赖,传统的数据库建议尽量使用绑定变量进行软解析,但是软解析最大的问题在于直接执行缓存中的执行计划,而不会进行基于成本的索引选择。

[0010] 一方面,现在内存已经不是奢侈品,另一方面,现在已经步入大数据时代,数据量大而且更新很快,仍然采用软解析很容易发生这样的情况:一条查询语句包含索引A和索引B,索引A指向1000条数据,索引B指向100万条数据,第一次执行该查询时是硬解析,根据一定的规则选定索引A执行,形成了执行计划;一段时间后,索引A指向1000万条数据,索引B指向2万条数据,这时执行该查询语句时是软解析,直接读取缓存中的执行计划执行索引A。显然这不是最优的执行方案。

[0011] 另外,现有的成本计算方法中并没有考虑到一个索引中含多个条件的情况。

[0012] (1)“AND”连接,现有的方法中多是任选其一进行成本评估:同一个索引,包含条件1和条件2,条件1指向100万条数据,条件2指向200条数据,选择条件1还是条件2严重影响了这条索引的成本评估;

[0013] (2)“OR”连接,现有的方法中多为直接全表遍历,检索效率很低。

发明内容

[0014] 本发明的目的在于克服现有技术的不足,提供一种基于完全成本计算的数据索引方法及系统,实现在每次执行查询操作时都对查询条件包含的索引进行成本的评估,选择操作成本更少的条件作为索引再进行下一步查询,能够有效提高检索的效率。

[0015] 本发明的目的是通过以下技术方案来实现的:一种基于完全成本计算的数据索引方法,其特征在于,它具体包括如下步骤:

[0016] S003成本预估:对每一个索引字段进行成本预估并得到预估值;

[0017] S004预估值排序:将所得到的预估值按从小到大进行排序;

- [0018] S005命中量预估:排序依次使用索引直接定位开始和结束位置,预估命中数据量,并根据命中量结果进行判断。
- [0019] 一种基于完全成本计算的数据索引方法,还包括:
- [0020] S001探测可能使用的索引;
- [0021] S002判断可使用的索引的数目:索引数目为0,直接全表遍历,反之进行下一个步骤;
- [0022] 进一步限定,所述的步骤S001具体为:
- [0023] S101若没有绑定变量,直接进行下一步骤,若有绑定变量,先解绑定再进行下一步骤;
- [0024] S102生成BSON文件;
- [0025] S103根据SQL语句中的条件结合表索引情况初步筛选可使用的索引。
- [0026] 进一步限定,所述的SQL语句中的条件,包括:检索、更新、删除语句的一种或多种。
- [0027] 进一步限定,所述的步骤S003针对于不同的索引字段类型的成本评估具体为:
- [0028] 整数型:条件扫描范围 $\div$ 索引有效数据范围=预估值;
- [0029] 浮点型:条件扫描范围 $\div$ 索引有效数据范围=预估值;
- [0030] 字符串、复合索引、二叉树索引:固定值 $\div$ 条件字符数=预估值;
- [0031] 进一步限定,所述的步骤S005命中量预估,具体为:
- [0032] S501任意一个索引满足预设条件1,直接选择该索引执行;
- [0033] S502所有的索引都满足预设条件2,直接选择全表遍历;
- [0034] S503其他情况下,执行命中量最低的索引。
- [0035] 所述的预估值和命中量还包括:
- [0036] 在一个索引字段中同时出现多个AND关系的条件选择成本小的条件使用索引,计算预估值和命中量;
- [0037] 在一个索引字段中同时出现多个OR关系的条件分别使用索引后合并,整体计算预估值和命中量。
- [0038] 一种基于完全成本计算的数据索引装置,它包括:索引探测模块、索引数目判断模块、成本预估模块、预估排序模块和命中量预估模块;
- [0039] 所述的索引探测模块探测可能使用的索引;
- [0040] 所述的索引数目判断模块判断可使用的索引的数目;
- [0041] 所述的成本预估模块对每一个索引字段进行成本预估;
- [0042] 所述的预估排序模块将所得到的预估值按从小到大进行排序;
- [0043] 所述的命中量预估模块排序依次使用索引直接定位开始和结束位置,预估命中数据量,并根据命中量结果进行判断。
- [0044] 一种终端系统,其特征在于,所述终端系统包括:
- [0045] 处理器,用于执行程序;
- [0046] 存储器,用于存储由处理器执行的程序,其中所述程序在执行时包括以下步骤:
- [0047] S001探测可能使用的索引;
- [0048] S002判断可使用的索引的数目:索引数目为0,直接全表遍历,反之进行下一个步骤;

- [0049] S003成本预估:对每一个索引字段进行成本预估并得到预估值;
- [0050] S004预估值排序:将所得到的预估值按从小到大进行排序;
- [0051] S005命中量预估:排序依次使用索引直接定位开始和结束位置,预估命中数据量,并根据命中量结果进行判断。
- [0052] 进一步限定,一种存储器,其中程序在执行时包括以下步骤:
- [0053] S001探测可能使用的索引;
- [0054] S002判断可使用的索引的数目:索引数目为0,直接全表遍历,反之进行下一个步骤;
- [0055] S003成本预估:对每一个索引字段进行成本预估并得到预估值;
- [0056] S004预估值排序:将所得到的预估值按从小到大进行排序;
- [0057] S005命中量预估:排序依次使用索引直接定位开始和结束位置,预估命中数据量,并根据命中量结果进行判断。
- [0058] 本发明的有益效果是:
- [0059] (1)本发明提出的方法中对所有的SQL都采用硬解析的方式,对绑定变量的查询先还原成完整的查询语句再进行硬解析,规避了大数据时代下传统数据库索引方法软解析和绑定变量带来的对成本评估产生巨大误差的可能性,能够提高检索效率;
- [0060] (2)本发明提出的方法改善了现有方法中对于一个索引字段中同时出现多个AND关系连接的条件的索引成本的评估的空缺:“同一个索引,包含条件1和条件2,条件1指向10万条数据,条件2指向100条数据”,该方法先按照上述方法分别判断条件1和条件2的执行成本,选择成本低的条件2的成本作为这条索引的成本,再和整条SQL语句包含的其他索引进行成本的判断;
- [0061] (3)现有的方法在处理这种情况时几乎采用的是直接全表遍历,检索效率很低,本发明提供的方法中,将同一个索引中多个OR连接的不同条件分别进行成本评估后再合并进行成本评估,比起全表遍历大大缩短了范围,能够提升检索效率。

附图说明

- [0062] 图1为一种基于完全成本计算的数据索引方法流程图;
- [0063] 图2为一种基于完全成本计算的数据索引装置框架图。

具体实施方式

- [0064] 下面将详细描述本发明的具体实施例,应当注意,这里描述的实施例只用于举例说明,并不用于限制本发明。在以下描述中,为了提供对本发明的透彻理解,阐述了大量特定细节。然而,对于本领域普通技术人员显而易见的是:不必采用这些特定细节来实行本发明。在其他实例中,为了避免混淆本发明,未具体描述公知的电路,软件或方法。
- [0065] 在整个说明书中,对“一个实施例”、“实施例”、“一个示例”或“示例”的提及意味着:结合该实施例或示例描述的特定特征、结构或特性被包含在本发明至少一个实施例中。因此,在整个说明书的各个地方出现的短语“在一个实施例中”、“在实施例中”、“一个示例”或“示例”不一定都指同一实施例或示例。此外,可以以任何适当的组合和/或子组合将特定的特征、结构或特性组合在一个或多个实施例或示例中。此外,本领域普通技术人员应当理

解,在此提供的示图都是为了说明的目的,并且示图不一定是按比例绘制的。

[0066] 如图1所示,一种基于完全成本计算的数据索引方法,它具体包括如下步骤:

[0067] S001探测可能使用的索引

[0068] S101判断是否含有绑定变量,如果含有绑定变量则需要先将绑定变量还原成完整的SQL语句。

[0069] S102生成BSON文件。

[0070] S103根据SQL语句中的条件(含检索、更新、删除语句)结合表索引情况初步筛选可使用的索引。

[0071] S002判断可使用的索引的数目

[0072] S201索引数目为0,直接全表遍历

[0073] S202索引数目不为0,进行下一个步骤

[0074] S003成本预估

[0075] 对每一个索引字段进行成本预估。

[0076] 成本预估的方法如下:

[0077] (1)整数型:条件扫描范围 $\div$ 索引有效数据范围=预估值;

[0078] (2)浮点型:条件扫描范围 $\div$ 索引有效数据范围=预估值;

[0079] (3)字符串、复合索引、二叉树索引:固定值(以100为例) $\div$ 条件字符数=预估值

[0080] S004预估值排序;

[0081] 将步骤3算出的预估值按从小到大进行排序,预估值越小,这条索引越优。

[0082] S005命中量预估

[0083] 按步骤4的排序依次使用索引直接定位开始和结束位置,预估命中数据量,并根据命中量结果进行判断:

[0084] S501任意一个索引满足预设条件1,直接选择该索引执行;

[0085] 在一个实施例中,预设条件1为小于1%且小于128条,直接选择该索引执行;

[0086] S502所有的索引都满足预设条件2,直接选择全表遍历;

[0087] 在另外一个实施例中,预设条件2为以大于20%,直接选择全表遍历;

[0088] S503其他情况下,执行命中量最低的索引。

[0089] 所述的预估值和命中量:

[0090] 优选的,在一个索引字段中同时出现多个AND关系的条件选择成本小的条件使用索引,计算预估值和命中量;

[0091] 优选的,在一个索引字段中同时出现多个OR关系的条件分别使用索引后合并,整体计算预估值和命中量。

[0092] 如图2所示,一种基于完全成本计算的数据索引装置,它包括:索引探测模块、索引数目判断模块、成本预估模块、预估排序模块和命中量预估模块;

[0093] 所述的索引探测模块探测可能使用的索引;

[0094] 所述的索引数目判断模块判断可使用的索引的数目;

[0095] 所述的成本预估模块对每一个索引字段进行成本预估;

[0096] 所述的预估排序模块将所得到的预估值按从小到大进行排序;

[0097] 所述的命中量预估模排序依次使用索引直接定位开始和结束位置,预估命中数据

量,并根据命中量结果进行判断。

[0098] 一种终端系统,其特征在于,所述终端系统包括:

[0099] 处理器,用于执行程序;

[0100] 存储器,用于存储由处理器执行的程序,其中所述程序在执行时包括以下步骤:

[0101] S001探测可能使用的索引;

[0102] S002判断可使用的索引的数目:索引数目为0,直接全表遍历,反之进行下一个步骤;

[0103] S003成本预估:对每一个索引字段进行成本预估并得到预估值;

[0104] S004预估值排序:将所得到的预估值按从小到大进行排序;

[0105] S005命中量预估:排序依次使用索引直接定位开始和结束位置,预估命中数据量,并根据命中量结果进行判断。

[0106] 一种存储器,其中程序在执行时包括以下步骤:

[0107] S001探测可能使用的索引;

[0108] S002判断可使用的索引的数目:索引数目为0,直接全表遍历,反之进行下一个步骤;

[0109] S003成本预估:对每一个索引字段进行成本预估并得到预估值;

[0110] S004预估值排序:将所得到的预估值按从小到大进行排序;

[0111] S005命中量预估:排序依次使用索引直接定位开始和结束位置,预估命中数据量,并根据命中量结果进行判断。

[0112] 本发明对软解析和绑定变量查询的完全成本评估:

[0113] 本发明提出的方法中对所有的SQL都采用硬解析的方式,对绑定变量的查询先还原成完整的查询语句再进行硬解析,规避了大数据时代下传统数据库索引方法软解析和绑定变量带来的对成本评估产生巨大误差的可能性,能够提高检索效率。

[0114] 本发明对同一个索引AND连接的不同条件的成本评估:

[0115] 本发明提出的方法改善了现有方法中对于一个索引字段中同时出现多个AND关系连接的条件的索引成本的评估的空缺:“同一个索引,包含条件1和条件2,条件1指向10万条数据,条件2指向100条数据”,该方法先按照上述方法分别判断条件1和条件2的执行成本,选择成本低的条件2的成本作为这条索引的成本,再和整条SQL语句包含的其他索引进行成本的判断。

[0116] 本发明对对同一个索引OR连接的不同条件的成本评估

[0117] 现有的方法在处理这种情况时几乎采用的是直接全表遍历,检索效率很低。本发明提供的方法中,将同一个索引中多个OR连接的不同条件分别进行成本评估后再合并进行成本评估,比起全表遍历大大缩短了范围,能够提升检索效率。

[0118] 所属领域的技术人员可以清楚地了解到,为了描述的方便和简洁,上述描述的方法、系统和模块的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0119] 本发明所揭露的方法、系统和模块,可以通过其它的方式实现。例如,以上所描述的实施例仅是示意性的,例如,所述模块的划分,可以仅仅是一种逻辑功能划分,实际实现时可以有另外的划分方式,例如多个模块或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另一点,所显示或讨论的相互之间的耦合或直接耦合或通信

连接可以说通过一些接口,系统或模块的间接耦合或通信连接,可以是电性,机械或其它的形式。

[0120] 所述分立部件说明的模块可以是或者也可以不是物理上分开的,作为模块显示的部件可以是或者可以不是物理模块,即可以位于一个地方,或者也可以分布到多个网络模块上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例的方案目的。

[0121] 另外,在本发明各个实施例中的各功能模块可以集成在一个处理模块中,也可以是各个模块单独物理存在,也可以两个或两个以上模块集成在一个模块中。

[0122] 所述功能如果以软件功能模块的形式实现并作为独立的产品销售或使用,可以存储在一个计算机可读存储介质中。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(Read-Only Memory,ROM)、随机存取存储器(Random Access Memory,RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0123] 以上所述仅是本发明的优选实施方式,应当理解本发明并非局限于本文所披露的形式,不应看作是对其他实施例的排除,而可用于各种其他组合、修改和环境,并能够在本文所述构想范围内,通过上述教导或相关领域的技术或知识进行改动。而本领域人员所进行的改动和变化不脱离本发明的精神和范围,则都应在本发明所附权利要求的保护范围内。

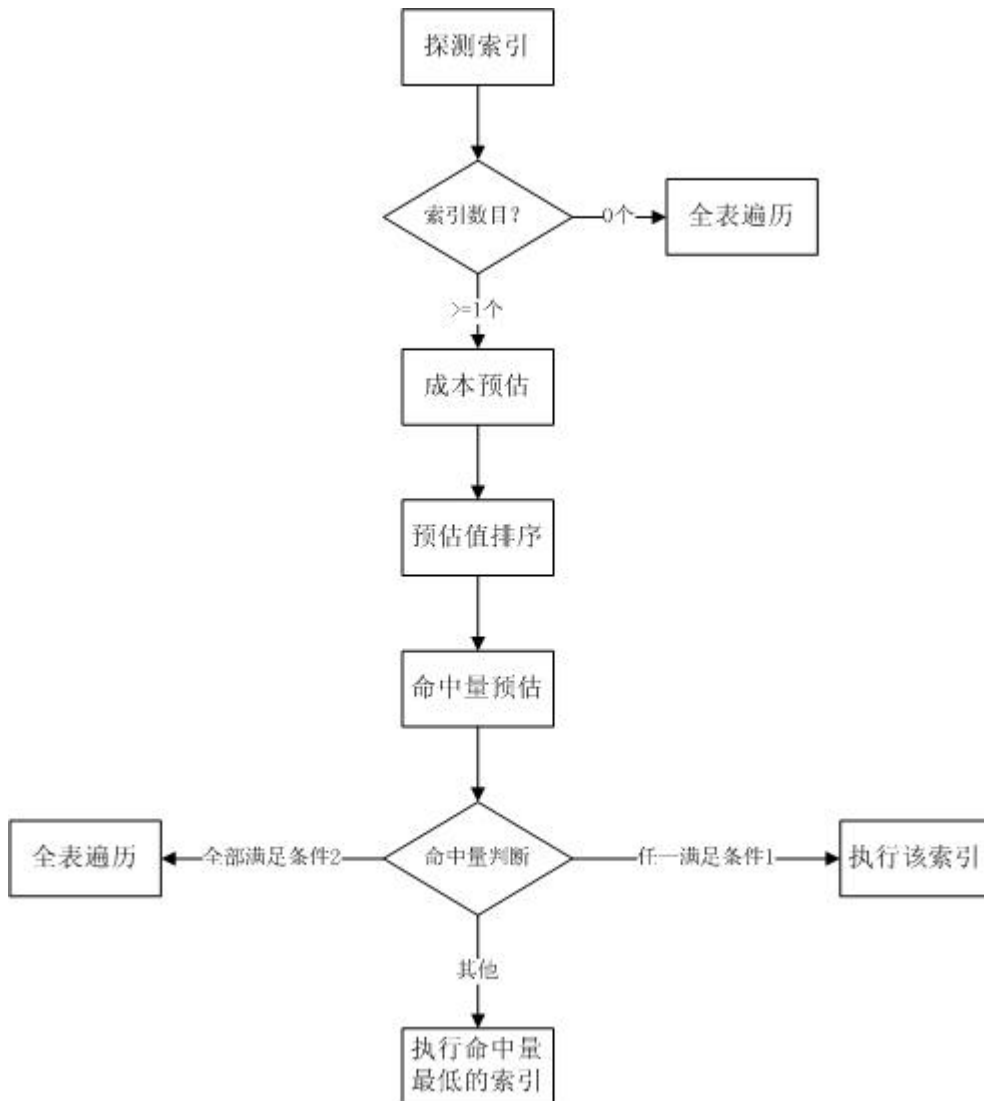


图1



图2