改进的 Apriori 算法在智能温室大棚系统中的应用①

庄燕滨[1,2], 董 煜[1], 肖贤建[2], 潘 群[2]

[1](河海大学 计算机与信息学院, 南京 211100)
[2](常州工学院 计算机信息工程学院, 常州 213002)

摘 要: 针对智能温室大棚系统内局部传感器故障造成其不能及时有效向上推送准确数据的问题, 提出将关联规则中的 Apriori 算法应用于故障传感器数据的预测. 以温度传感器发生故障为例, 首先将关联规则中传统的 Apriori 算法进行优化, 然后将其运用到故障传感器参数的预测当中去. 实验仿真表明, 改进的 Apriori 算法能够快速的发现温室各参数之间的关联规则, 从而估计出故障传感器的参数的范围, 有一定的应用价值.

关键词: 温室大棚; 关联规则; Apriori; 传感器

Application of an Improved Apriori Algorithm in Intelligence Greenhouse System

ZHUANG Yan-Bin[1,2], DONG Yu[1], XIAO Xian-Jian[2], PAN Qun[2]

[1](College of Computer and Information Technology Engineering, Hohai University, Nangjing 211100, China)
[2](School of Computer Information and Engineering, Changzhou Institute of Technology, Changzhou 213002, China)

Abstract: To solve the problem that the accure data can't be pushed by the failure of local sensor in intelligence greenhouse system, it was presented that the Apriori algorithm which was based on association rule applied in the prediction of sensor fault data. Forcasting the greenhouse environment temperature is provided as an example in this paper, firstly, the classic Apriori algorithm is modified. Then it was used in the prediction of fault sensor data. The experimental results show that the improved Apriori algorithem could quickly find the association rule between the parameters in Greenhouse, thus estimated the range of parameters of the fault sensor and the method could be proved to be feasible.

Key words: greenhouse; association rule; Apriori; sensor

近年来, 随着物联网技术在农业中的广泛应用[1]以及精准农业[2]概念的提出, 以往只依靠人为经验进行农业耕作和管理的经验型农业已无法适应现代信息技术背景下农业生产和管理的需求. 具体到智能温室大棚系统作业中, 传感器通过实时采集温室内空气温湿度, CO2 浓度, 土壤温湿度以及光照, 叶面湿度, 露点温度等环境参数, 根据电磁阀和水泵、施肥系统等的目标值, 自动执行控制天窗、侧窗、内遮阳、外遮阳、风机、湿帘、外翻窗、加温、加湿、二氧化碳发生器和设备的开启/关闭时间等操作, 以保证温室内的农作物处于其最佳生长环境下. 由此可见, 传感器数据能否及时准确有效的向上推送, 直接关系到该大棚中的农作物是否始终处于自身的最佳生长环境中, 进而影响到大棚中的农作物的产量.

目前对传感器的故障数据的处理的方法比较多[3,4], 但大体上可以分为两个方向, 一是以现代控制理论为基础的分析冗余法[5], 该方法虽然能够依据系统的动态属性实现实时诊断, 但是难以精确建模, 仅适用于低维数据. 二是从计算机数据处理的角度, 现今多是采用基于知识的专家诊断系统[6], 其优点易于增加修改规则, 可信度较高, 但也存在领域知识提取困难的问题. 而在智能温室大棚控制系统中, 影响农作物生长的因素较多, 导致数据的维度较高, 采用分析冗余法建立该系统精确的数学模型将十分困难. 而如果采

① 基金项目:江苏省常州市武进区科技局科技支撑计划(农业)(WN201413)
收稿时间:2015-03-02;收到修改稿时间:2015-04-15

用基于知识的专家诊断系统, 训练样本的数据需要精心挑选, 这就需要大量的计算机科研人员和农业方面的专家共同协作完成, 极大的增加了工程预算.

我们知道温室中参数之间的数据紧密联系[7], 一个参数的波动会在另一个参数中体现出来. 例如, 空气温度的增加势必导致空气湿度以及土壤湿度的降低. 所以在智能温室大棚控制系统中各传感器采集的参数之间呈现出一定的关联性, 受数据挖掘中的关联规则分析的相关知识的启发, 挖掘出各传感器参数的关联规则, 当智能温室控制系统中某一传感器出现故障不能及时准确向上推送数据时, 此时我们可以根据相应的关联规则对该故障传感器数据范围进行合理估计, 从而判断作物是否处在适宜的生长环境之内. 本文结合实际的工程应用背景, 首先对传统的关联规则分析 Apriori 算法进行了改进, 然后再将其应用到温度传感器故障分析当中.

2 传统Apriori算法的优化改进

2.1 传统 Apriori 算法存在的问题

传统 Apriori 算法最典型的应用是购物篮分析, 其关联的规则的生成是基于事务数据库中对象相互之间的关系[8,9]. 例如, “在买了商品 A 和 B 的消费者中, 有80%的人会继续购买 C 和 D”该关联规则我们可以表述为 A∧B→C∧D. 然而, 在我们智能温室大棚系统中发生我们需要挖掘空气湿度, 光照度, 土壤温度, 土壤湿度, CO2 浓度和光照强度与空气温度之间的关联规则, 即发现形如A∧B∧C∧D→E的关联规则, 而使用传统的 Apriori 算法对同一属性参数不同的离散化区间进行连接将产生大量的冗余关联规则, 例如, A1, B1, C1 与 A1, B1, C2 连接产生 A1, B1, C1, C2, 而假设 C1 与 C2 是光照度这一参数不同的离散区间, 即 C1 与 C2 是互斥关系不能同时存在于同一关联规则中, 所以这种连接是毫无意义的, 并且将消耗大量的系统资源. 所以为了去除冗余的关联规则提高挖掘效率, 我们需要对传统的 Apriori 算法进行优化改进.

2.2 传统 Apriori 算法的改进

自从 Apriori 算法提出以来, 为了克服其自身的缺陷, 提高算法的性能. 许多学者进行大量研究, 提出了多种优化改进的算法. 大体可分为以下四个思路:

(1)基于散列(hash)的方法[10]: 该算法的主要思想是将频繁 k-1 项集产生的 k 项集通过散列函数映射到不同的桶中(映射地址), 并添加相应的桶计数. 当桶计数低于最小支持度阀值时候, 就把该桶中的 k 项集删除. 该方法显著的特点是压缩 2 项集效果明显, 能够提高挖掘的效率. 然而本次挖掘的数据维度是 6 维, 最终需要生成 6 项集, 每次项集的产生都需要通过散列函数产生映射地址跟桶计数, 所以与传统 Apriori 算法相比在本次挖掘中该算法的效率提升并不明显. 同时该算法并没有有效的去除冗余的机制, 如果冗余的关联规则所在桶的计数高于最小支持度阀值, 它将被保留产生下一频繁项集.

(2)基于采样的方法[11]: 其核心思想是在事务数据库中选取随机样本 S, 然后再 S 中产生频繁项集, 实质就是牺牲精度换取执行效率的提高, 特别适用于频繁密集的项集计算. 但是, 我们生成大棚温室内各环境参数之间关联规则的目的是预测故障传感器的参数范围, 需要尽可能多的训练样本才能比较准确的预测故障传感器参数的范围, 因此基于采样的这种精度换效率的优化方法并不符合我们这次实际工程应用的需求.

(3)减少交易的个数[12]: 根据不包含频繁 k 项集的事务必然不包含 k+1 项集这一结论, 把不包含候选项集的事务标记为删除, 从而能够减少扫描数据的数量. 该方法虽然压缩了事务数据集, 但是并没有减少扫描数据库的次数, 而在智能温室大棚控制系统中, 由于温室内的无线传感器的节点较多且采集的间隔时间较短(平均每隔一分钟向上推送一次数据), 这就导致数据不仅总量巨大而且实时性较强, 在生成关联规则的过程中需要不断的扫描服务器数据库更新的数据, 所以该优化算法对挖掘温室参数关联规则效率的提高并不明显.

(4)基于划分的方法[13]: 主要思想是根据相应的逻辑对事务数据库进行分块, 然后单独考虑每个分块生成频繁项集, 最后将每个分块产生的频繁项集合并. 本文就是基于划分的思想对 Apriori 算法进行了优化, 并做出了相应的修正. 首先我们以项集字典属性排序为依据划分事务数据库, 我们以某一 2 项候选集为例, 如图 1. 图 2 为对候选集中的项集进行字典排序后划分子集的子集.

然后我们是将子集 1 中的项集与子集 2 的项集做连接, 生成{A1, B1, C1}; {A1, B1, C2}; {A1, B1, C3}; {A1, B2, C1}; {A1, B2, C2}; {A1, B2, C3}; {A1, B3, C1}; {A1, B3, C1}和{A1, B3, C1}总共 9 个候选集. 而

如果使用传统的 Apriori 算法则会生成{A1, B1, C1}; {A1, B1, C2}; {A1, B1, C3}; {A1, B2, C1}; {A1, B2, C2}; {A1, B2, C3}; {A1, B3, C1}; {A1, B3, C1}; {A1, B3, C1}; {A1, C1, C2}; {A1, C1, C3}和{A1, C2, C3}总共 15 个候选集, 其中包含 6 个冗余规则.

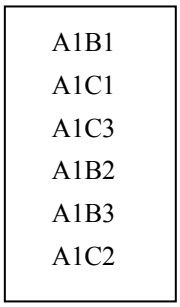

图 1　未排序

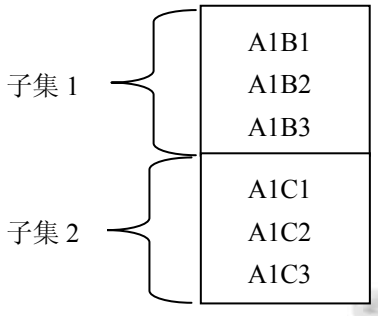

图 2　排序之后

由此可见与传统的 Apriori 算法相比该修正的 Apriori 优化算法能够很好的去除冗余规则, 从而提高了挖掘效率.

该修正算法的描述如下(标记为 IMApriori):

输入:事务数据库 D, 最小支持度 minsupport;

输出:D 中的频繁集 L;

(1)L1 =find_frequent_1-itemsets (D);

(2)for (k=2; Lk-1≠φ; k++) {

(3)　C_k=IMpriori_gen (L_{k-1},minsupport);

(4)　for each transaction t ∈ D {

(5)　C_t=subset(C_k,t);

(6)　　for each candidate c∈C_t

(7)　c.count++;

(8)　}

(9)L_k ={c∈C_k|c.count ≥ minsupport}

(10)}

(11)return L=∪$_k L_k$;

procedure IMApriori_gen (L_{k-1})

(1)I=Div_items(L_{k-1});　　//分割频繁集

(2)for each l_i∈I_i

(3)for each l_j∈I_j {

(4)　c=l_i ⋈ l_j　　// 连接步产生候选集

(5)　if has_infrequent_subset(c,L_{k-1})then

(6)　　delete c;　//剪枝步删除非频繁集的候选

(7)　else add c to C_k;

(8) }

(9)return C_k;

procedure Div_items (L_{k-1})

(1)i=1; n=1;

(2)while (i<N) {　　//N 为 L_{k-1} 的长度

(3)　add l_i∈L_{k-1} to I_n;

(4)　for each I_j∈L_{k-1} {　　//j>i

(5)　if ($l_{i[1]}$=$I_{j[1]}$]∧($l_{i[2]}$=$I_{j[2]}$)∧...∧$l_{i[k-2]}$=$I_{j[k-2]}$∧$l_{i[k-1]}$ ≠$I_{j[k-1]}$)

(6)　　if($l_{i[k-1]}$∈N∧$I_{j[k-1]}$∈N) //N 为同一属性集

(7)　　add I_j to I_n;

(8)　else break;

(9)　}

(10)i=j;

(11)n++;

(12)}

(13)return I=∪$_n I_n$;

Procedure has_infrequent_subset(c;L_{k-1})

(1)for each(k-1)-subset s of c

(2)　if s∉L_{k-1} then

(3)　　return TRUE;

(4)return FALSE;

2.2 算法效率分析

为了进一步验证本文所使用的算法的优越性, 我们将本文所提出的优化算法与传统的 Apriori 算法进行性能分析比较. 实验验证硬件环境为 Inter(R)core(MT) i3-2330M CPU 3.00GH, 内存为 2GB 的 PC 机. 软件环境为 win7 操作系统, 数据库系统为 oracle11, 开发语言为 R 语言.

我们从以下三个方面对两种算法进行了比较分析①在相同支持度下两种算法在数量不同的数据库事务集下运行所消耗的时间. ②在支持度相同数据库中事务集数量也相等的情况下, 产生的频繁集数量对比. ③在数据库中事务集数量相等支持度不同的情况下, 两种算法运行所消耗的时间.

图 3 为分别用这两种算法对含有 10000 条, 20000 条, 30000 条, 40000 条, 50000 条数据的传感器采集的数据集进行关联规则挖掘所消耗的时间, 其中红色线代表本文改进的 IMApriori 算法, 蓝色线代表传统

Apriori 的算法, 最小支持度为 10%. 从图 3 我们可以看出, 随着数据集规模的扩大, 改进的 IMApriori 算法的运算时间明显少于传统的 Apriori 算法. 这是因为由于数据集中数据的增加, 产生的冗余关联规则越来越多, 改进的 IMApriori 算法能够有效的过滤掉这部分关联规则从而运算速度加快.

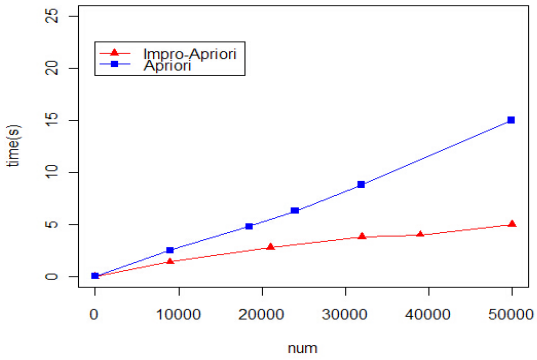

图 3 两种算法运行在不同数据集合消耗时间对比

表 1 两种算法所产生的频繁集数量

频繁集	IMApriori	Apriori
1-item	29	29
2-item	301	406
3-item	2504	3654
4-item	3407	8701
5-item	4407	10219

表 1 为 IMApriori 算法和 Apriori 算法在数据库事务集为 10000 条时所生成的频繁集数量对比, 由表 1 我们可以看到频繁集属性个数为 3 项以下时候, IMApriori 算法与 Aprioi 算法生成的频繁集数量有所减少但不够明显, 但是当属性个数超过 3 项时候, IMApriori 算法所产生的频繁集数量规模远远小于 Aprioi 算法, 这能够大大减少去除冗余关联所消耗的时间, 使得算法的效率更高.

图 4 为这两种算法在最小支持度分别为 5%, 10%, 15%, 20%, 25%下挖掘关联规则所消耗的时间, 其中数据库事务集数量为 10000. 由图 4 我们可以看到随着最小支持度的降低改进 IMApriori 算法的运算时间明显少于传统 Apriori 算法的运算时间.

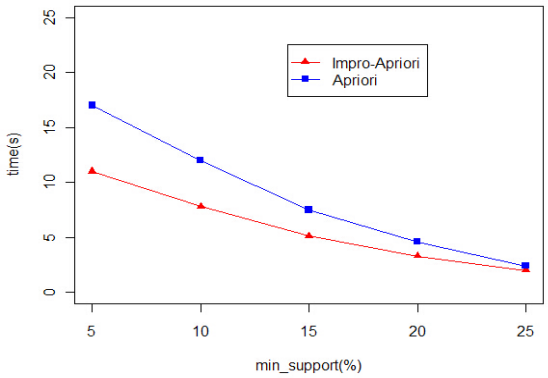

图 4 两种算法运行在不同支持度下消耗时间对比

综上所述, 我们可以认为, 在生成与温室温度预测有关的关联规则这一背景下, 改进的 IMApriori 算法性能优于传统的 Aprior 算法性能.

3 改进的IMApriori算法在温室温度预测中的应用

3.1 温室传感器采集数据的预处理

本文采用常州市邹区果蔬生产基地智能温室大棚 2014 年 4 月 5 号一天的传感器节点采集的数据作为原始数据集, 表 2 所示为部分原始数据. 其中 METER_ID 属性栏中, SARH 代表环境湿度, SSTP 代表土壤温度, SSRH 代表土壤湿度, SACD 代表 CO2 浓度, SSUN 代表光照强度, SATD 代表环境温度.

表 2 部分原始数据

DATA_ID	COLLECTOR_ID	METER_ID	PARAM_ID	VALUE	COLLECT_TIME
37811	320402789	SACD	1	400.1	2014-4-5 0:48:55
37812	320402789	SSTP	1	15.1	2014-4-5 0:48:55
37813	320402789	SSRH	1	18.4	2014-4-5 0:48:55
37814	320402789	SARH	1	17.7	2014-4-5 0:48:55
37815	320402789	SSUN	1	83.6	2014-4-5 0:48:55
37816	320402789	SATD	1	16.4	2014-4-5 0:48:55
37817	320402789	SSRH	1	17.6	2014-4-5 0:48:55
37818	320402789	SSUN	1	82	2014-4-5 0:48:55
37819	320402789	SATD	1	0	2014-4-5 0:49:55

由表 2 我们可以看到在 0:49:55 这一时刻 SATD 的 VALUE 值为 0, 也就是说在这一时刻温度传感器出现故障未能向上推送准确的数据. 而温度作为影响农作物生长的重要条件, 农作物如果在适合自己生长的温度范围内, 能够较好的生长发育, 当高于或者低于该范围后农作物的生长活性就会降低, 其生长发育将受到影响最终会导致该作物产量或者质量下降. 如果此时而温度传感器发生故障(向上推送温度值为 0), 温室大棚内的自动执行机构会执行错误的操作, 使得温室温度不再处于作物的最佳生长环境中. 这个时候我们就需要根据关联规则对温度传感器数值范围作出合理估计, 为我们人工操作大棚执行机构提供科学依据.

首先要对数据进行预处理, 我们根据传感器自身的测量范围对传感器采集的数据进行离散化等级划分, 表 3 为温室内各传感器性能参数.

环境湿度(单位%): A1 代表 10~15, A2 代表 15~20, A3 代表 20~25, A4 代表 25~30, A5 代表 30~35, A6 代表 35~40, A7 代表其它(表示不常出现的数值下同).

土壤温度(单位℃): B1 代表 5~10, B2 代表 10~15, B3 代表 15~20, B4 代表 20~25, B5 代表其它.

土壤湿度(单位%): C1 代表 18~19, C2 代表 19~20, C3 代表 20~21, C4 代表 21~22, C5 代表 22~23, C6 代表 23~24, C7 代表其它.

CO2 浓度(单位 mg/L): D1 代表 350~450, D2 代表 450~550, D3 代表 550~650, D4 代表 650~750, D5 代表 750~850, D6 代表 850~950, D7 代表其它.

光照强度(单位 KLUX): E1 代表光照等级弱(0~100), E2 代表光照等级中(100~300), E3 代表光照等级强(300 以上).

环境温度(单位℃): F1 代表 10~15, F2 代表 15~20, F3 代表 20~25, F4 代表 25~30, F5 代表 30~35, F6 代表 35~40, F7 代表其它.

表 3　传感器性能参数

传感器类别	安装位置	测量范围	功耗
环境湿度	大棚中部	0~100%	0~15mA
土壤温度	浸于土壤 15cm	-40~120℃	80μW
土壤湿度	浸于土壤 15cm	0~100%	80μW
CO2 浓度	大棚底部	0~2000mg/L	4~20mA
光照强度	大棚顶部	0~20 万 lux	2.5vA
环境温度	大棚中部	0~50℃	0~15mA

预处理后的部分数据如表 4 所示.

表 4　预处理的部分数据

序号	空气相对湿度	土壤温度	土壤湿度	CO2 浓度	光照强度	空气温度	时间
1	A1	B2	C1	D1	E1	F1	2014-4-5 0:48:54
2	A1	B3	C1	D1	E1	F2	2014-4-5 0:48:55
3	A2	B4	C2	D1	E1	F2	2014-4-5 8:32:33
4	A2	B4	C2	D1	E1	F2	2014-4-5 8:32:36
5	A5	B4	C3	D2	E1	F3	2014-4-5 12:23:36
6	A5	B4	C4	D2	E2	F3	2014-4-5 12:23:50
7	A5	B3	C1	D2	E3	F3	2014-4-5 13:00:21
8	.....	.....	.....	.....			

3.2 关联规则的生成及结果分析

将改进 IMApriori 算法应用到预处理后的数据集中进行关联规则的挖掘, 设最小支持度为 5%, 最小置信度为 50%, 得到满足最小支持度和最小置信度的关联规则. 部分关联规则如表 5 所示.

(1)由关联规则 1 和 2 可知, 光照强度对温室温度影响较为敏感, 在空气相对湿度, 土壤温度, CO2 浓度这四个参数不变的情况下, 光照强度等级由弱变为中, 温室空气温度从区间 20℃~25℃变为 25℃~30℃. 如果温室中空气传感器发生故障, 在连续数个时间无法接收到温度传感器的温度值, 此时假定我们已经根据关联规则 1 得出温室温度大概在 20℃~25℃范围内, 而 35℃左右才是我们需要的作物最佳生长温度, 这时候我们就可以采取相应的补光措施将光照强度由弱变为中.

(2)由关联规则 3 和 4 可知, 当空气相对湿度在 20%~25%, 土壤温度在 15%~20%, 土壤湿度在 20%~21%这一范围内时候, 此时 CO2 浓度如果由

20%~30%升高到 40%~50%, 但是温室内温度依然维持在 20℃~25℃之间, 由此我们可以断定 CO2 浓度这一因素对温室温度的影响不太明显. 当上次传感器推送过来的数据与当前传感器采集的数据比较, 如果仅有 CO2 浓度的变化, 我们可以做出当前温度值仍在上次温度范围内的预测.

表 5 生成的部分关联规则

序号	关联规则	支持度(%)	置信度(%)
1	A3∧B2∧C3∧D4∧E1→F3	14.4	50
2	A3∧B2∧C3∧D4∧E2→F4	14.4	50
3	A3∧B3∧C3∧D2∧E1→F3	27.8	100
4	A3∧B3∧C3∧D4∧E1→F3	27.8	80
5	A3∧B3∧C3∧D4∧E3→F5	34.6	80
6	A3∧B3∧C3∧D2∧E3→F5	20.0	50
7	.....	.....	.....

4 结语

本文在温室环境温度预测这一实际工程背景下改进了传统 Apriori 算法, 实验结果证明改进的 IMApriori 算法能够较好的应用到智能温室大棚系统中去, 生成的关联规则能能够帮助温室大棚管理人员科学的分析温室环境各参数之间的关系指导其进行合理的人工操作. 下一步的工作是进一步进行关联规则的挖掘和分析, 寻找到如果多个传感器同时出现故障时的解决办法.

参考文献

1 张海江,赵建民,朱信英,徐慧英.基于云计算的物联网数据挖掘.微型电脑应用,2012,28(6),10–13.

2 陈桂芬,曹丽英,马丽.数据挖掘在精准农业中的应用现状及发展趋势.吉林农业大学学报,2008,30(4),621–626.

3 张娅玲,陈伟民,章鹏,胡顺仁,黄晓微,郑伟.传感器故障诊断技术概述.传感器与微系统,2009,28(1),4–6,12.

4 杨建平.传感器故障诊断的研究与应用[学位论文].北京:华北电力大学(北京),2004.

5 房方,孙万云,牛玉广.控制系统传感器故障的检测与诊断研究.华北电力大学学报,2001,28(4),47–52.

6 张荣标,蔡兰,王贵成.基于时间序列和专家系统的 PH 值传感器故障诊断的研究.电子测量与仪器学报,2001,15(3):30–35,40.

7 何鹏,楚艳红.基于数据挖掘的温室多参数控制算法的研究.农机化研究,2012(10),180–183.

8 Han JW, Micheline Kamber,Jian Pei.Data Mining: Conceptsand Techniques. 3rd Ed. USA: Morgan Kaufmann Publications, 2011.

9 Malgaokar S, Surve S, Hirave T. Use of minging techniques to improve the effecitiveness of marketing of marketing and sales. Mumbai, India. IEEE. 2013. 1–5.

10 Park JS, Chen MS, Yu PS. An effective hash-based algorithm for mining association rules. Proc. of ACM SIGMOD International Conference on Management of Data. San Jose,CA . May 1995. 175–186.

11 Mannila H, Toivonen H, Verkamo A. Efficient algorithm for discovering association rules. AAAI Workshop on Knowledge Discovery in Databases. 1994. 181–192.

12 Han J, Fu Y. Diseovery of multiple level assoeiation rules from large databases. Proc. of the 21st Int. Conf. on Very Large Databases(VLDB'95). Zurich, Switzerland. Morgan Kaufinann Publisher. 1995. 420–431.

13 王丹,张浩,陆剑峰.针对高项频繁集的关联规则改进算法.计算机工程,2006,32(40),29–30,80.