

基于改进关联规则的 Web 使用挖掘方法研究

张 翔¹, 陈胜勇²

(1 金华职业技术学院, 浙江 金华 321007; 2 浙江工业大学, 浙江 杭州 310014)

摘 要: 本文提出了改进关联挖掘算法的数据挖掘方式. 在自然连接产生候选集以前, 先进行一个修剪过程, 减少参加连接的项集数量, 进而减小生成的候选项集规模, 减少了循环迭代次数和运行时间, 同时在连接判断步骤中减少多余的判断次数, 避免传统方法的弊端. 仿真实验证明, 这种改进算法能够大幅提高数据的查询速度, 取得很好的效果.

关键词: 数据挖掘; 关联规则; 改进算法

中图分类号: TP393

文献标识码: A

文章编号: 1000—7180(2012)03—0110—03

Based on the Improvement of Web Association Rules Method Research

ZHANG Xiang¹, CHEN Sheng-yong²

(1 Jinhua College of Vocation and Technology, Jinhua 321007, China;

2 Zhejiang University of Technology, Hangzhou 310014, China)

Abstract: This paper puts forward the improved algorithm for mining association of data mining. In the natural connection way to generate candidate set before you a clip in the process, reduce the connection, and reduce the number of sets of the generation of candidate itemsets scale, reduce the iterative times and running time, at the same time in connection to reduce redundant judgement steps the judgment of the traditional method, avoid times the drawbacks. Simulation experiments show that the improved algorithm can greatly improve data inquires, speed got very good effect.

Key words: data mining; association rules; improved algorithm

1 引言

使用传统方式进行数据的相互传递, 有很多缺陷, 首先大范围获取候选集内容时, 需要巨额花费, 利用了可以排列顺序的谓词, 这样会导致投入非常大; 其次利用 I/O 频率过高, 没有合理使用索引, 让查询更加复杂, 降低了查询效率, 从而降低了系统性能; 再次, 表连接时不合理, 引起集合中有很多数据冗余, 导致查询时间延长, 影响检测效率; 最后, 算法在收敛到某一个部分的最佳取值时停止运算, 并没有获取最终结果^[1].

为了避免上述缺陷, 本文需要对相关数据挖掘

算法进行优化处理, 缩短数据库查询时间, 使算法能够更加便捷、高效. 通过实践可以看出, 优化处理后的算法效率有了明显提高, 使用价值也更加明显.

2 相关数据挖掘的原则、理论及缺点

Agrawal 等人在上世纪 90 年代提出相关数据挖掘方法. 此方法先统计各个项目在数据库中的相关数据量, 然后将统计结果与相关数据量极小值对比, 获取 1-使用频率较高的数据集合 M_1 , 然后对此数据集合进行连接, 获取 2-备选数据集合, 从此集合中得到 2-使用频率较高的数据集合 M_2 , 不断重复上述过程, 假设使用频率较高的数据集合中数据

收稿日期: 2011—07—20; 修回日期: 2011—08—15

基金项目: 金华市科技计划项目(2007-1-067)

量不再增多,则停止运算.在第 l 次重复上述过程时,得到第 $(l-1)$ 次数据检测过程中获取的出现频率较高的数据集合 M_{l-1} 和通过Apriori-gen表达式获取的备选数据集合 D_l .继续对数据库中的数据进行检测,通过运算获取 D_l 中备选数据集合相关数据量和出现频率较高的数据集合 M_l .通过上述步骤,能够获取相关数据挖掘方法^[2-3].但是这个过程也可以体现出此算法的诸多缺陷:

(1)不断检测数据库中相关数据必须利用比较大的I/O承载量,极大的降低了数据挖掘效率,从而降低了系统性能.

(2)获取数目较多的备选集合.因此查询过程比较复杂,不利于简单、便捷的获取查询结果.

(3)当数据量较大时,在连接过程中会产生大量备选数据集合,而在确定备选数据集合中是否是出现频率较高的数据集合时,需要反复检测数据库,运算步骤很多,耗费时间很长,算法效率会大大降低.

3 以相互联系规律为基础的数据挖掘算法

3.1 改进算法

(1)降低备选数据集合数量.首先,检测数据库获取备选数据集合 D_1 ,统计备选数据集合相关数据量,得到一阶出现频率较高的数据集合 M_1 ,其次,对数据重复进行迭代处理,连接 M_{l-1} 获取 D_l .优化处理算法需要先统计 M_{l-1} 里所有项目出现频率,删除 M_{l-1} 中项目集合数量小于 $l-1$ 的项目集合,将其余项目集合连接后获取 D_l . D_l 里备选项目集合数量减少,因为获取每个备选项目集合都需要对数据库进行一次全面检测,所以减少备选数据集合数目就可以减少运算步骤,缩短运算时间,从而提高查询效率,最终实现算法优化.

(2)算法中项目集合依据字母排列顺序,从而获取有序序列,根据 M_{l-1} 获取 D_l 时,先对比分析、进行连接,然后使用项目集合有序序列减少判定次数,简化运算步骤,从而缩短运算时间.

(3) D_l 减少后,经过运算获取 M_l 的 l 项目集合数量,假设项目集合数量小于 $l+1$,那么不用计算 $l+1$ 项目集合,停止迭代处理后,算法随即停止.

3.2 改进算法理论基础

下面阐述优化处理该算法的相关理论:针对出现频率较高的项目集合 M_l ,假设 $|M_l| < l-1$,那么停止对 $(l-1)$ -项目数据集合进行运算,在这个表达式里, $|M_l|$ 代表出现频率较高的 l -项目数据集合数量.在 $(l-1)$ -项目出现频率较高的数据集合中, $l-$

子集个数是 $l-1$.假设 M_l 里的数据量小于 $l-1$,那么数据库里没有 $(l-1)$ -项出现频率较高的数据集合.通过上述算法可以最终获取全部出现频率较高的数据集合,为下一步计算奠定基础^[3].

在数据挖掘算法项目集合按顺序排列时,假设 $(l-1)$ -出现频率较高的项目集合是 $M_b = \{M_b[1], M_b[2], \dots, M_b[l+1]\}$ 和 $M_c = \{M_c[1], M_c[2], \dots, M_c[l+1]\}$, M_b 和 M_c 符合下述条件:

$$M_b[1] = M_c[1], M_b[3] = M_c[2], M_b[l+3] = M_c[l+2], M_b[l+1] = M_c[l+2]$$

假设 $(l+1)$ -出现频率较高的项目集合不与 M_b 连接,那么这个集合不能连接 M_c 及其后的全部 $(l+1)$ -项目集合.

根据上述理论,算法用下面步骤阐述:

(1)检测数据库获取备选项目集合 D_1 ,通过运算获取 D_1 里所有备选项目集合相关数据量,这个数据量的值直接关系到运算步骤和运算时间,影响查询效率.

(2)删除小于相关数据量极小值的项目集合,获取一阶出现频率较高的项目集合 M_1 .

(3)将 M_{l-1} 自连接获取备选项目集合 D_l 之前,通过运算获取 M_{l-1} 里全部项目出现频率.

(4)删除 M_{l-1} 里频率小于 $l-1$ 次的项目集合, M_{l-1} 处理后记为 M'_{l-1} .

(5)将 M'_{l-1} 自连接获取 D_l ,根据推论对算法进行优化,使其连接判断方法更加准确、便捷.

(6)经过运算缩小 D_l 的取值范围,获取 D_l ,统计其中 l -项目集合数量.假设数量小于 $l+1$,那么停止计算,否则返回步骤(3)重复计算,则能够获取最终需要查询的数据,在数据库中数据量较大时,极大的缩短了查询时间,提高了查询效率.

3.3 改进算法的应用

将优化处理后的数据挖掘方法使用到实践中,需要先检测数据库里所有项目集合的相关数据量, $D_1 = \{(A, 3), (B, 5), (C, 4), (D, 3), (E, 3)\}$;抽取相关数据量大于或者等于2的项目集合构造1-出现频率较高的项目集合 $M_1 = \{A, B, C, D, E\}$.然后连接并统计项目集合相关数据量,获取2-出现频率较高的项目集合 $M_2 = \{AB, AC, AD, BC, BD, BE, CD, CE\}$.

算法依据 L_{k-1} 连接获取 D_l 时,首先,需要通过运算获取每一项在 l -出现频率较高的项目集合 M_l 里的相关数据量,缩小其取值获取 M'_l .其次,根据 M'_l 连接获取 D_{l+1} 时,更改 M'_l 里出现频率较高的项

目集合判定连接方法. 连接 M'_2 获取 D_3 时, 以出现频率较高的项目集合 AB 为例, 因为 AB 和 BC 不可以连接, 所以不判定 AB 是否可以与其后的 BD 、 BE 连接, 所以省略了 2 个判断步骤. 缩短判定需要的时间, 提高了查询效率. 在获取 D_3 时, 利用传统方式需要进行 $7 * (7+1)/2 = 28$ 次检测, 而优化处理后的算法只需要进行 $(4+3+2+3+2+1+1) = 16$ 次检测. 优化处理后, 检测次数减少了很多, 缩短了运算时间, 提高了检测效率.

进行连接后, 通过运算获取所有项目集合相关数据量, 得到出现频率较高的 3-项目集合, 统计所有项目在其相关数据量, 删除其中频率低于 2 的集合, 从而使运算更加简单.

利用上述方式优化该算法后, 运算量减小了, 运算时间缩短, 提高了数据挖掘能力.

4 实验结果及其分析

在实践中, 通过 Web 试验数据模拟试验, 下面阐述一下具体方法: 检测数据库中数据, 统计所有项目集合相关数据量, 并将其与相关数据量极小值对比, 获取出现频率较高的 1-项目集合 M_1 , 依次进行连接, 获取备选数据集合 D_2 , 使用相同方式可以获取出现频率较高的 2-项目集合 M_2 , 不断重复, 最终获取备选数据集合 D_l 和出现频率较高的 l -项目集合 M_l . 最终获取数据库中需要查询的相关数据.

优化处理后的优势: 连接 M_{l-1} 与获取 D_l 之间进行了修剪, 删除频率低于 $l-1$ 的项目数据集合, 获取 M'_{l-1} , 使 D_l 中备选数据集合数量降低; 当 M'_{l-1} 连接获取 D_l 时更新判定方法, 提高判定效率, 通过统计出现频率较高的项目集合数目, 确定是否停止迭代处理. 减少了检测次数, 缩短了检测时间, 提高了查询效率. 表 1 提供了试验环境相关参数.

表 1 实验环境

CPU	AMD Athlon(tm) X2 Dual-Core QL-60 1.90GHz
内存	1GB
操作系统	Windows 7 旗舰版
开发环境	Java JDK1.6
数据库	Oracle 10g

获取两种算法在不同数据库中查询时间的区别, 其中相关数据量系数是 0.3, 超过这个临界值的集合是出现频率较高的项目集合, 抽取 400 个到 2000 个不同数据库中的数据, 使用传统算法和优化处理后的算法进行运算, 获取出现频率较高的项目集合. 这

两种算法的运算时间如图 1 所示.

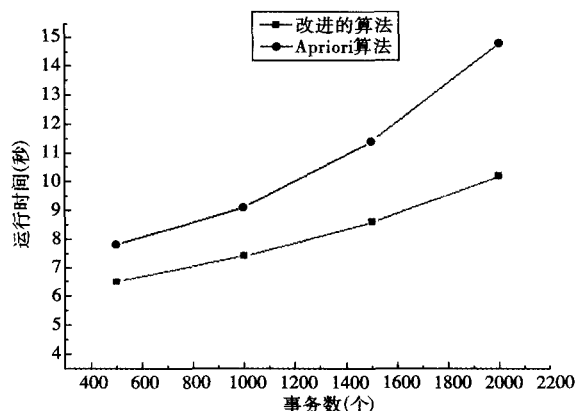


图 1 事务数不同时算法的运行时间对比

通过图 1 可以看出, 假设相关数据量系数是确定的, 那么优化处理后的算法与传统算法对比, 数据库中样本含量越大, 效率优势越大. 因为优化处理主要是降低了需要检测数据库中所有数据的次数, 如果数据量越大, 则检测一次数据库中所有数据的耗时越长, 查询效率越低, 相反, 优化处理后的算法, 降低了检测次数, 提高了查询效率.

检测同一个数据库中的数据时, 两种算法在相关数据量系数极小值不同时, 进行运算需要的时间是有区别的, 其中假设数据库中数据量是 2000 个. 传统算法和优化处理后的算法在相关数据量系数是 0.1~0.4 时检测需要时间的区别如图 2 所示.

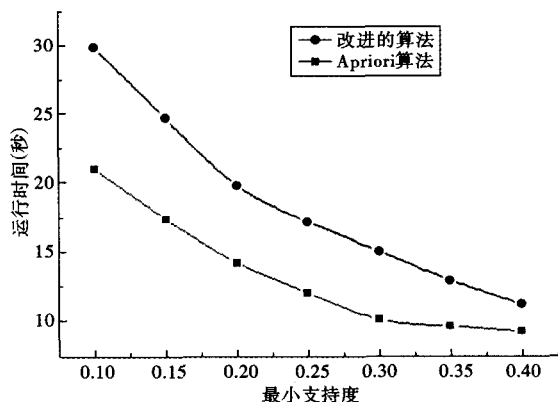


图 2 支持度不同时运行时间对比

从图 2 可以看出, 改进的算法在执行时间上明显优于传统算法, 特别是在支持度越小的情况下这种优势更加明显, 这是因为在支持度越小的情况下, 候选集一定的情况下, 频繁集规模越大, 在以后每次迭代过程中, 连接的次数也更多, 相应产生候选集的规模也越大. 在支持度较大的情况下, 相应的频繁集

(下转第 116 页)

本文的方法把大区域分成 5 个小区, 在无线网络监测区域内, 建立 5 组具有相同结构的节点. 设定 10 个随机感应器节点为入侵节点, 分别用传统的集中式定位搜索算法和本文的算法对这些被入侵的异常节点进行检测, 记录两种方法的耗时, 并利用下面公式计算平均误差:

$$E_{\text{error}} = \frac{\sum_{i=1}^m \|x_{\text{real}}^i - x_{\text{est}}^i\|}{mR} \quad (9)$$

平均误差是用估计待检测位置信息与被入侵节点实际的位置信息之间的欧式距离来表示的, 距离越大误差越大. 计算两种节点异常检测方法的平均误差, 并记录两种方法完成被入侵节点检测的耗时, 两组数据的对比结果如表 1 所示. 图 2 为两种算法的平均误差趋势图.

表 1 两种方法定位数据对比

定位方法	定位耗时	平均误差
传统集中式	9.1 s	26.2%
本文算法	4.2 s	6.3%

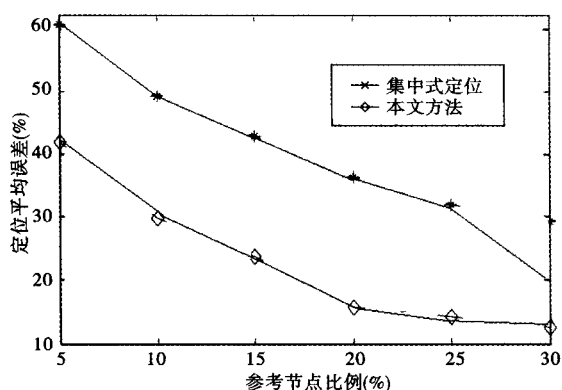


图 2 平均误差趋势图

(上接第 112 页)

以及连接产生的候选集规模也就较小, 所以改进的算法效率优势就不那么明显.

总体上, 可以得出, 在候选集相同的情况下, 改进的算法的执行效率明显优于传统算法.

5 结束语

本文的主要内容是对基于关联规则数据挖掘算法进行了分析研究, 先详细描述了算法的步骤, 再结合实际例子进行应用, 然后分析了算法的缺点, 针对它的缺点, 提出了改进的算法. 改进的算法是通过减少连接判断次数, 以及在生成候选集前对 L_k 进行了裁枝, 产生新的频繁集 L'_k , 然后连接产生新的候选

由上述实验结果可以看出, 本文的方法不仅缩短了定位耗时而且具有较高的定位准确率, 满足了无线传感网络中对异常节点定位实时性的要求.

5 结束语

本文提出了一种区域联系的异常节点定位检测方法. 通过把大的无线网络分成小的区域, 进而运用寻优算法, 把定位过程分布到各个小网络中相关节点上, 利用小区域相关节点中的正常信息, 与局部节点信息进行比较, 完成一次多节点比较, 避免了传统集中式方法大规模比较的弊端, 最终实现了无线传感网络入侵的快速定位.

参考文献:

- [1] 孙佩刚, 赵海, 张文波, 等. 普适计算中定位服务的参考点布置及选择算法[J]. 电子学报, 2006, 34: 1456-1463.
- [2] Koutsonikolas D, Das S M, Hu Y C. Path planning of mobile landmarks for localization in wireless sensor networks[C]//Proceedings of the 26th IEEE International Conference on Distributed Computing Systems Workshops. Amsterdam, The Netherlands: IEEE, 2006: 80-86.
- [3] 姜华, 袁晓兵, 付耀先, 等. 无线传感器网络中信道仿真模型的研究[J]. 计算机仿真, 2006(11): 129-132.
- [4] 王伟. 多信息源的实时入侵检测方法研究[M]. 西安: 西安交通大学出版社, 2005.

作者简介:

范荣真 男, (1974—), 硕士, 高级工程师. 研究方向为网络安全、电子商务安全.

集. 最后实验也证实了改进的算法的优越性.

参考文献:

- [1] 朱元忠. 入侵检测系统发展研究[J]. 北京工业职业技术学院学报, 2006(1): 8-11.
- [2] 陶剑文. 一种分布式 Web 日志挖掘系统的设计与实现[J]. 计算机仿真, 2006(10): 122-123.
- [3] 刘兵. Web 数据挖掘[M]. 北京: 清华大学出版社, 2009.

作者简介:

张翔 女, (1973—), 硕士, 讲师. 研究方向为图像处理、数据库技术.

陈胜勇 男, (1973—), 博士, 教授, 博士生导师. 研究方向为计算机视觉、图像分析与处理、机器人智能技术.