

基于改进 Apriori 算法的关联规则挖掘研究

朱其祥, 徐 勇, 张 林

(安徽财经大学 信息工程学院, 安徽 蚌埠 233041)

摘 要: 关联规则挖掘研究是数据挖掘研究的一项重要内容。经典的关联规则提取算法——Apriori 算法及其改进算法存在着一些不足, 一是会产生大量的候选项目集, 二是在扫描数据库时需要很大的 I/O 负载。通过对关联规则产生过程的实际实验分析发现, 可以采取利用频繁 $k-1$ 项集 L_{k-1} 对候选 k 项集 C_k 进行预先剪枝、及在扫描数据库过程中忽略对频繁项集的产生无贡献的交易记录的方法来改进关联规则提取的效率。

关键词: 数据挖掘; 关联规则; 频繁项集; Apriori

中图分类号: TP301.6

文献标识码: A

文章编号: 1673-629X(2006)07-0102-03

Research on Mining Association Rule Based on Improved Apriori Algorithm

ZHU Qi-xiang, XU Yong, ZHANG Lin

(School of Information Engineering, Anhui University of Finance & Economics, Bengbu 233041, China)

Abstract: Mining association rule is one of the most important topics of data mining. There are some shortcomings in mining association rules via Apriori algorithm or its improved algorithms. The first is many candidate items are generated; the second is more disk I/O needed. It's been found the efficiency of algorithms can be improved by pruning the candidate items C_k based on frequent items L_{k-1} , and ignoring the transactions which is useless for frequent items generated.

Key words: data mining; association rules; frequent items; Apriori

0 引 言

计算机科学与技术的迅速发展和广泛应用为人类的日常生活、工作和科学研究提供了极大的便利, 在计算机的帮助下人们可以将传统的事务做的更好。然而随着大规模、海量的数据被有效地存储起来之后, 人们却发现没有足够多的时间去理解这些数据, 人们处在被各种各样的数据——科学数据、医疗数据、市场数据等——淹没的境地, 理解这些数据已经远远超出了人的能力。因此面对这样一种“数据丰富, 但信息贫乏”^[1]的严峻形势, 寻找新的数据分析方法和工具, 以便从海量数据中提取有用知识已成为世界范围内的热门研究课题。数据库中知识发现 (Knowledge Discovery in Databases, KDD)^[2~4]正是在这样一种背景下产生并蓬勃发展起来的。数据库中知识发现, 也称为知识发现、数据挖掘 (Data Mining, DM), 出现于 20 世纪 80 年代后期, 是在数据库技术的基础上, 结合人工智能、机器学习、统计学、神经网络等多种学科技术产生的一

个具有很强生命力的新研究领域。90 年代开始有了突飞猛进的发展, 成为当前涉及人工智能、数据库理论与技术、电子商务等学科的一个非常活跃的研究领域, 在商务管理、生产控制、市场分析、工程设计和科学探索等方面表现出很好的应用前景, 将会成为未来几年内对社会产生深远影响的关键技术之一。其中关联规则挖掘研究^[5~7]是知识发现研究的一项重要内容, 其目的是发现大规模数据集中项集之间有趣的关联关系或模式。

关联规则挖掘问题首先是由 R. Agrawal 等于 1993 年提出, 其后许多的研究人员对该问题进行了广泛的研究, 主要集中在改进关联规则挖掘算法以提高挖掘的效率^[8~10]等和推广关联规则挖掘应用两个方面。至今, 最经典的关联规则挖掘算法仍是由 R. Agrawal 等提出的 Apriori 算法, 该算法的主要思想是首先寻找给定大数据集中的频繁项集, 然后通过频繁项集生成强关联规则。寻找频繁项集步骤的核心思想是用前一次扫描数据库的结果产生本次扫描的候选项目集, 从而提高搜索的效率。频繁项集和强关联规则的概念是通过支持度和可信度两个指标来确定的。满足支持度要求的项集被称为频繁项集, 同时又满足可信度要求的规则被称为强关联规则^[5]。关联规则的一个例子是“90% 的顾客在购买面包和黄油的同时也会购买牛奶”, 表示顾客在购买某些商品时很可能会

收稿日期: 2005-10-27

基金项目: 中华全国供销合作总社科研项目 (2006); 安徽财经大学青年科研资助项目 (ACKYQ06372C)

作者简介: 朱其祥 (1963-), 男, 安徽蚌埠人, 讲师, 硕士研究生, 研究方向为数据库技术、嵌入式系统。

同时购买其它的一些相关商品。发现这样的规则有助于商品货架设计、库存安排及根据购买模式对用户进行分类。

已有的关联规则挖掘算法及其相关改进算法^[9,11]主要是针对压缩候选项集、减少扫描数据库次数等方面的,而很少涉及从数据库本身的角度来考虑算法的改进。笔者在实际研究过程中发现,在扫描数据库的某些遍中一些记录对频繁项集的产生是无贡献的,也即在频繁项集产生的过程中数据库中的有些记录可以被忽略。因此,文中在此基础上提出了一种新的关联规则挖掘改进算法。

1 关联规则的有关定义和性质

定义1:设 $I = \{i_1, i_2, \dots, i_m\}$ 是一个项的集合,称之为项集; TD 是一个交易数据库, T 表示其中的一个交易,每个交易都有一个不为零的标识号(如交易号)TID,且有 $T \subseteq I$;设 $X \subseteq I$,当且仅当 $X \subseteq T$ 时称事务 T 包含或支持项集 X 。

定义2:关联规则是形如 $A \Rightarrow B$ 的蕴涵式,其中 $A \subset I, B \subset I$,且 $A \cap B = \emptyset$;交易数据库 TD 中的关联规则具有支持度 s 和可信度 c ;支持度是指 TD 中同时包含 A 和 B 的事务的百分比,可信度指 TD 中包含 A 的事务中同时也包含 B 的百分比。

定义3:对于一个给定的交易事务集 TD ,关联规则挖掘的任务是挖掘所有满足支持度和可信度约束的强规则。

根据上述定义,可以这样描述 Apriori 算法:Apriori 算法使用逐层搜索的迭代方法来产生频繁项集,设有频繁 k -项集 L_k ,通过 Galois 连接产生候选 $k+1$ 项集 C_{k+1} ,再通过扫描数据集产生频繁 $k+1$ 项集 L_{k+1} ,最后由产生的频繁项目集产生关联规则。

性质1:(Apriori 性质) 频繁项目集的所有非空子集都必须也是频繁的。

证明(用反证法):略。

推论1:一个非频繁项目集的任一超集必定也是非频繁的。

证明:根据定义若有 $k-1$ 项集 I_{k-1} ,不满足最小支持度阈值 \min_sup ,即 $P(I_{k-1}) < \min_sup$,则称 I_{k-1} 为非频繁的。若将任意一项(集) A 添加到 I_{k-1} 中,则必有 $P(I_{k-1} \cup A) < P(I_{k-1}) < \min_sup$,即 I_{k-1} 的任一超集($I_{k-1} \cup A$) 是非频繁的。得证。

推论2:若一个候选 k -项集的任一 $k-1$ 项子集不在 L_{k-1} 中,则该候选 k 项集是非频繁的。

证明:略。

性质2:若某交易记录 T 不支持频繁 $k-1$ 项集 L_{k-1} 中的每一元素,则 T 必不支持 C_k 中的任一元素。

证明:设频繁 $k-1$ 项集 L_{k-1} ,则必有 $\forall I_{k-1} \in T$;对于 $\forall X \in C_k$,则必有一 $k-1$ 项集 $I_{k-1} \subset X$,则有 $X \notin T$ 。得证。

推论3:若某记录 T 不包含候选项集 C_{k-1} 中的任一元

素,也必不包含 C_k 中的任一元素。

证明:略。

2 Apriori 算法改进

因为关联规则挖掘过程面对的是大规模数据库,导致挖掘算法效率较低的因素一是产生的候选项目集数量庞大,其次是数据库中的记录数很多导致过多的 I/O 开销。所以提高关联规则挖掘算法的效率可以从以下 3 个方面考虑:

(1) 减小候选项集 C_k 的规模。Apriori 算法及已有相关改进算法对从候选 k 项集 C_k 中产生频繁项目集 L_k ,是通过扫描数据库分别计算每个候选项的支持计数来完成的。通过上文的分析,在候选项集 C_k 产生后、扫描数据库计算每个候选项的支持计数之前,可以先判断 C_k 中每一元素 X 的 $k-1$ 项子集是否是 L_{k-1} 的子集。若是,则将 X 保留在 C_k 中,继续判断 C_k 中的下一个元素;否则将 X 从 C_k 中删除。实验证明,很多情况下这样先通过 L_{k-1} 对 C_k 进行预先剪枝可以大大减少 C_k 的规模。

(2) 在扫描数据库的过程中忽略对频繁项目集 L_k 中所产生的无贡献的交易记录,即减少扫描数据库时实际需要对其进行操作的交易数。在一次扫描数据库由候选项目集 C_k 确定频繁项目集 L_k 的过程中,若某条交易记录 T 不包含候选项集 C_k 中的任一元素,则可通过将该交易记录的标识号置为空(如 0)以在下一次扫描数据库时直接跳过该记录。因为若 T 不包含 C_k 中的任一元素,也必不包含 C_{k+1} 的任一元素。

(3) 通过程序优化提高算法效率。很多研究工作者在关联规则挖掘算法研究过程中,往往只注重对算法策略的考虑,而忽视对程序进行优化。良好的数据结构、编程风格及程序优化等对算法的效率是有影响的。例如,作者通过实验在 JAVA 语言中使用文件缓冲与未使用文件缓冲策略对 Windows 操作系统文件 Shell32.dll 进行读操作所耗时间比约为 3:100。

3 实验

采用一个实际超市交易数据集进行实验并与 Apriori 算法进行比较,数据集中共有交易记录 3000 条,交易商品种类(即项目数)为 88。

实验环境为 Pentium IV 2.6GHz,512M 内存,Windows XP 操作系统,编程语言为 JAVA。

图 1 显示了在最小可信度阈值为 0.6 时的不同支持度阈值下,分别用 Apriori 算法及改进后的算法对交易数据进行关联规则挖掘时所耗时间。

实验表明,改进后的算法基本上总是优于 Apriori 算法的;并且当支持度阈值较小时,改进算法的效率有更明显的提高。

通过实验也发现当数据库规模较小、数据库中包含的项数较小时,改进算法的效率没有明显的提高。这是因为

改进算法主要是针对减少候选项的数量、及跳过对于频繁项目集产生无贡献的记录的考虑而获得性能改进的。当数据库包含的项数较少时,通过连接操作产生的候选项数也较少;当数据库规模较小时,扫描数据库过程中忽略的记录也较少,所以在这种情况下,由于整个挖掘过程的所耗费的时间较少,从而改进算法的效率提高不明显。

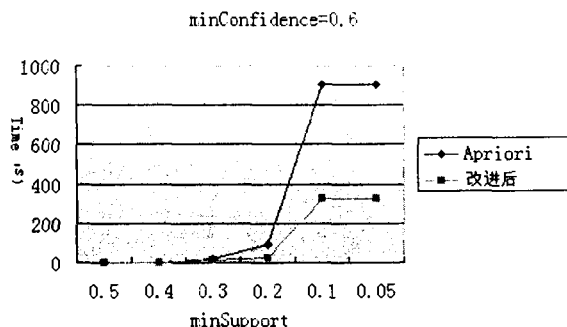


图 1 两种算法在不同支持度下的时间性能比较

4 小结

文中对关联规则挖掘的经典算法 Apriori 算法进行了讨论分析,研究了相关的性质,提出从减少候选项数及减少实际需要处理的交易记录数的角度改进挖掘算法的思想,并通过实验证明了改进方法的有效性与正确性。

关于关联规则挖掘研究的下一步研究方向可以是研究面向分布式环境下的关联规则挖掘问题,研究结合背景知识的关联规则挖掘问题。

参考文献:

- [1] Han Jia Wei, Kamber M. Data Mining - concepts and techniques[M]. San Francisco, CA: High Education Press, Morgan Kaufman Publishers, 2001.
- [2] Frawley W J, Piatetsky G, Shapiro C, et al. Knowledge Discovery in Databases: An Overview[A]. In: Piatetsky - Shapiro, Frawley W J. Knowledge Discovery in Databases[C]. Menlo Park, California: AAAI Press/The MIT Press, 1991. 1 - 27.
- [3] Fayyad U, Piatetsky - Shapim G, Smyth R. From Data Mining to Knowledge Discovery: An Overview[A]. In: Fayyad U. Advances in Knowledge Discovery and Data Mining[C]. Menlo Park, California: AAAI Press, 1996. 1 - 34.
- [4] Uthoramy R. From Data mining to Knowledge Discovery: Current Challenges and Future Directions[A]. In: Fayyad U. Advances in Knowledge Discovery and Data Mining[C]. Menlo Park, California: AAAI Press, 1996. 561 - 569.
- [5] Agrawal R, Imielinski T, Swami A. Mining Association Rules Between Sets of Items in Large Databases[A]. In: Proceedings of 1993 ACM - SIGMOD International Conference Management of Data(SIGMOD'93)[C]. Washington D. C. : [s. n.], 1993. 207 - 216.
- [6] Agrawal R, Srikant R. Fast algorithms for mining association rules in large database [R]. Technical Report FJ9893. San Jose, CA: IBM Almaden Research Center, 1994.
- [7] Agrawal R, Srikant R. Fast algorithms for mining association rules[A]. In: Proc of 20th Int Conf Very Large Databases (VLDB'94)[C]. CA: [s. n.], 1994. 487 - 499.
- [8] 徐 瑞, 乔志萍, 李伟华. 单维关联规则快速 Apriori 算法研究[J]. 微电子学与计算机, 2005, 22(2): 43 - 45.
- [9] 王创新. 关联规则提取中对 Apriori 算法的一种改进[J]. 计算机工程与应用, 2004(34): 183 - 185.
- [10] XU Yong, Zhou Sen - Xin, Gong Jin - Hua. Mining Association Rules with New Measure Criteria[A]. In Proc of the 4th Int Conf of ICMC[C]. [s. l.]: [s. n.], 2005. 2257 - 2260.
- [11] 李清峰, 杨路明, 张晓峰, 等. 数据挖掘中关联规则的一种高效的 Apriori 算法[J]. 计算机应用与软件, 2004, 21(12): 84 - 86.

(上接第 101 页)

问题。同时由于网络快速发展,新应用不断提出,网络正变得越来越复杂。为了应付日益复杂的网络,必须提出新型的网络测量方案,为整个网络稳健、可靠、高效运行提出依据。文中正是在这种背景下详细讨论了网络测量中的关键技术,为其进一步的发展打下基础。

参考文献:

- [1] Paxson V, Floyd S. Wide - area traffic: the failure of poisson modeling [J]. IEEE/ACM Transactions on Networking, 1995, 3(3): 226 - 244.
- [2] Real Time Flow Measurement Working Group [EB/OL]. <http://www.ietf.org/html.charters/rtfm-charter.html>, 2002.
- [3] Mahdavi J, Paxson V. IPPM Metrics for Measuring Connectivity, RFC2678 [EB/OL]. <http://www.ietf.org/rfc/rfc2678.txt>, 1999 - 09.
- [4] Paxson V. Measurement and Analysis of End - to - End Internet Dynamics[D]. Computer Science Division, University of California Berkeley, 1997.
- [5] Almes G, Kalidindi S, Zekauskas M. A One - way Delay Metrics for IPPM, RFC2680 [EB/OL]. <http://www.ietf.org/rfc/rfc2680.txt>, 1999 - 09.
- [6] Almes G, Kalidindi S, Zekauskas M. A Round - trip Delay Metrics for IPPM, RFC2681 [EB/OL]. <http://www.ietf.org/rfc/rfc2681.txt>, 1999 - 09.
- [7] Lai K, Baker M. Nettimer: a Tool for Measuring Bottleneck Link Bandwidth [DB/OL]. <http://mosquitonet.stanford.edu/laik/>, 2001.
- [8] Moon S B, Skely P, Towsley D. Estimation and Removal of Clock Skew from Network Delay Measurements[A]. proceedings of IEEE INFORCOM99[C]. New York, USA: IEEE, 1999. 227 - 234.