| | |
|---|---|
| P.O.Box 8718, Beijing 100080, China | Journal of Software, Jan. 2004,15(1):94-102 |
| E-mail: jos@iscas.ac.cn | ISSN 1000-9825, CODEN RUXUEW, CN 11-2560/TP |
| http://www.jos.org.cn | Copyright © 2004 by The Editorial Department of Journal of Software |

# 挖掘闭合模式的高性能算法

刘君强, 孙晓莹, 庄越挺, 潘云鹤

Full-Text PDF Submission Back

刘君强1， 孙晓莹1， 庄越挺2， 潘云鹤2 1(杭州商学院 计算机信息工程学院,浙江 杭州 310035)2(浙江大学 人工智能研究所,浙江 杭州 310027)

作者简介: 刘君强(1962－),男,浙江杭州人,博士,教授,主要研究领域为人工智能,数据挖掘;孙晓莹(1963－),女,讲师,主要研究领域为管理信息系统;庄越挺(1965－),男,博士,教授,博士生导师,主要研究领域为计算机图形学,人工智能;潘云鹤(1946－),男,教授,博士生导师,主要研究领域为人工智能,计算机辅助设计.

联系人: 刘君强 Phn: +86-571-88835193, E-mail: liujunq@mail.hzic.edu.cn

Received 2002-10-24; Accepted 2003-09-05

## Abstract

The set of frequent closed patterns determines exactly the complete set of all frequent patterns and is usually much smaller than the laster. Yet mining frequent closed patterns remains to be a memory and time consuming task. This paper tries to develop an efficient algorithm to solve this problem. The compound frequent item set tree is employed to organize the set of frequent patterns, which consumes much less memory than other structures. The tree is grown quickly by integrating depth first and breadth first search strategies, opportunistically choosing between two different structures to represent projected transaction subsets, and heuristically deciding to build unfiltered pseudo or filtered projections. Efficient pruning methods are used to reduce the search space. The balance of the efficiency and scalability of tree growth and pruning maximizes the performance. The experimental results show that the algorithm is a factor of five to three orders of magnitude more time efficient than several recently proposed algorithms, and is also the most scalable one. It can be used in the discovery of non-redundant association rules, sequence analysis, and many other data mining problems.

Liu JQ, Sun XY, Zhuang YT, Pan YH. Mining frequent closed patterns by adaptive pruning. *Journal of Software*, 2004,15 (1):94~102.
http://www.jos.org.cn/1000-9825/15/94.htm

## 摘要

频繁闭合模式集惟一确定频繁模式完全集并且尺寸小得多,然而挖掘频繁闭合模式仍然是时间与存储开销很大的任务.提出一种高性能算法来解决这一难题.采用复合型频繁模式树来组织频繁模式集,存储开销较小.通过集成深度与宽度优先策略,伺机选择基于数组或基于树的模式支持子集表示形式,启发式运用非过滤虚拟投影或过滤型投影,实现复合型频繁模式树的快速生成.局部和全局剪裁方法有效地缩小了搜索空间.通过树生成与剪裁代价的平衡实现时间效率与可伸缩性最大化.实验表明,该算法时间效率比其他算法高5倍到3个数量级,空间可伸缩性最佳.它可以

进一步应用到无冗余关联规则发现、序列分析等许多数据挖掘问题.

基金项目：Supported by Zhejiang Provincial Natural Science Foundation of China under Grant No.602140 (浙江省自然科学基金); the National Natural Science Foundation of China under Grant No.60272031 (国家自然科学基金); the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No.20010335049 (国家教育部博士点基金); the Key Science-Technology Project of the National "Ninth Five-Year-Plan" of State Development Planning Commission of China under Grant No.JGJ-1998-2077 (国家计委"九五"重大科技攻关项目)

## References:

[1] Pasquier N, Bastide Y, Taouil R, Lakhal L. Discovering frequent closed itemsets for association rules. In: Beeri C, et al, eds. Proc. of the 7th Int'l. Conf. on Database Theory. Jerusalem: Springer-Verlag, 1999. 398~416.

[2] Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Beeri C, et al, eds. Proc. of the 20th Int'l. Conf. on Very Large Databases. Santiago: Morgan Kaufmann Publishers, 1994. 487~499.

[3] Pei J, Han J, Mao R. CLOSET: An efficient algorithm for mining frequent closed itemsets. In: Gunopulos D, et al, eds. Proc. of the 2000 ACM SIGMOD Int'l. Workshop on Data Mining and Knowledge Discovery. Dallas: ACM Press, 2000. 21~30.

[4] Burdick D, Calimlim M, Gehrke J. MAFIA: A maximal frequent itemset algorithm for transactional databases. In: Georgakopoulos D, et al, eds. Proc. of the 17th Int'l. Conf. on Data Engineering. Heidelberg: IEEE Press, 2001. 443~452.

[5] Zaki MJ, Hsiao CJ. CHARM: An efficient algorithm for closed itemset mining. In: Grossman R, et al, eds. Proc. of the 2nd SIAM Int'l. Conf. on Data Mining. Arlington: SIAM, 2002. 12~28.

[6] Liu JQ, Pan YH, Wang K, Han J. Mining frequent item sets by opportunistic projection. In: Hand D, et al, eds. Proc. of the 8th ACM SIGKDD Int'l. Conf. on Knowledge Discovery and Data Mining. Alberta: ACM Press, 2002. 229~238.

[7] Srikant R. Quest synthetic data generation code. San Jose: IBM Almaden Research Center, 1994. http://www.almaden.ibm.com/software/quest/Resources/index.shtml

[8] Blake C, Merz C. UCI Repository of machine learning. Irvine: University of California, Department of Information and Computer Science, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html