高效率之高頻資料項目型樣挖掘演算法

羅美榮、程仲勝

E-mail: 9124932@mail.dyu.edu.tw

摘 要

資料挖掘(DATA MINING)是現今一個相當熱門的研究領域，尤其在探討如何從龐大資料庫中快速找出 所有的高頻資料項目型樣(FREQUENT PATTERNS OR FREQUENT ITEMSETS)的議題，更是被廣泛討論與研 究。針對此一議題的文獻中，多半採取APRIORI-BASED的方式，此方式需經過頻繁的資料庫存取，並且 要產生龐大的候選物項集，因此在效率上很差。而以高頻資料項目型樣樹(FP-TREE)為主的演算法雖不 需產生候選物項集並且只需要掃描資料庫兩次，但其資料結構相當複雜。因此最近提出的高頻資料項目 型樣串列(FPL)演算法利用簡單線性串列資料結構來儲存所有的資料，以改進FP-TREE演算法複雜的資料 結構。然而FPL演算法仍需掃描資料庫兩次。本論文中，我們設計一具有高效率之高頻資料項目型樣挖 掘演算法FPLI，此演算法主要針對FPL演算法加以改進。FPLI只需掃描交易資料庫一次，資料結構如同 FPL利用簡單線性串列資料結構來儲存資料庫中的所有交易資料，且在此串列上執行較FPL簡單的運算。 另外，FPLI演算法亦可適用於交易資料庫異動或最小支持度變動時之高頻資料項目型樣挖掘而不需要重 新掃描資料庫、重建資料結構，實驗結果顯示我們的方法有很好的效能。

關鍵詞 : 資料挖掘、高頻資料項目型樣、候選物項集

目錄

參考文獻

[1]R. AGARWAL,C.AGGARWAL,ANDV.V.V.PRASAD."A TREE PROJECTION ALGORITHM FOR GENERATION OF FREQUENT ITEMSETS,"IN J. PARALLEL AND DISTRIBUTED COMPUTING, 2000.

[2]R.AGRAWAL,T.IMILIENSKI,AND A.SWAMI,"MINING ASSOCIATION RULES BETWEEN SETS OF ITEMS IN LARGE DATABASES",PROC.OF THE ACM SIGMOD CONFERENCE ON MANAGEMENT OF DATA,MAY 1993 PAGE (S): 207-216.

[3]R.AGRAWAL AND R.SRIKANT,"FAST ALGORITHMS FOR MINING ASSOCIATION RULES IN LARGE DATABA -SES",PROC.OF THE 20TH INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASE,SEPTEMBER 1994 PAGE(S): 478-499.

[4]R.J.BAYARDO JR.,"EFFICIENTLY MINING LONG PATTERNS FROM DATABASES,"PROC.OF THE ACM SIGM -OD CONFERENCE ON MANAGEMENT OF DATA,JUNE 1998 PAGE(S): 85-93.

[5]S.BRIN,R.MOTWANI,J.D.ULLMAN,AND S.TSUR,"DYNAMIC ITEMSET COUNTING AND IMPLICATION RULES FOR MARKET BASKET DATA",1997 ACM SIGMOD CONFERENCE ON MANAGEMENT OF DATA,1997 PAGE(S): 255-264.

[6]M.S.CHEN,J.HAN,AND P.S.YU,"DATA MINING: AN OVERVIEW FROM A DATABASE PERSPECTIVE",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,VOL.8,NO.6,DECEMBER 1996.

[7]J.HAN,J.PEI,AND Y.YIN,"MINING FREQUENT PATTERNS WITHOUT CANDIDATE GENERATION," PROC. ACM SIGMOD,2000 PAGE(S): 1-12.

[8]H.TOIVONEN,"SAMPLING LARGE DATABASES FOR ASSOCIATION RULES ",VLDB,1996 PAGE(S):134.145.

[9]L.KAUFMAN AND P.J.ROUSSEEUW,"FINDING GROUPS IN DATA:AN INTRODUCTION TO CLUSTER ANALYSIS ",JOHN WILEY & SONS, 1990.

[10]D.LIN AND Z.M.KEDEM,"PINCER-SEARCH:A NEW ALGORITHM FOR DISCOVERING THE MAXIMUM FREQUEN -T SET,"PROC.OF THE 6TH INTL.CONF.ON EXTENDING DATABASE TECHNOLOGY,1998.

[11]J.S.PARK,M.S.CHEN,AND P.S.YU,"AN EFFECTIVE HASH BASED ALGORITHM FOR MINING ASSOCIATION RULES",PROC.OF THE ACM SIGMOD CONFERENCE ON MANAGEMENT OF DATA,MAY 1995 PAGE(S):175-186.
[12]F.C.TSENG AND C.C.HSU."CREATING FREQUENT PATTERNS WITH THE FREQUENT PATTERN LIST,"PROC. OF THE ASIA PACIFIC CONFERENCE OF DATA MINING AND KNOWLEDGE DISCOVERY, HONG KONG,2001 PAGE(S): 376-386.
[13]S.M.WEISS AND C.A.KULIKOWSKI,"COMPUTER SYSTEM THAT LEARN:CLASSIFICATION AND PREDICTION METHODS FROM STATISTICS,NEURAL NETS,MACHINE LEARNING,AND EXPERT SYSTEM",MORGAN KAUFMAN, 1991. S.Y.WUR AND Y.LEU,"AN EFFECTIVE BOOLEAN ALGORITHM FOR MINING ASSOCIATION RULES IN LARGE DATABASES",DATABASE SYSTEMS FOR ADVANCED APPLICATIONS,1999.PROCEEDINGS.,6TH INTERNATIONAL CONFERENCE ON , 1999 PAGE(S):179-186.