

EFFICIENT ALGORITHMS FOR MINING FREQUENT PATTERNS

羅美榮、程仲勝

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

ABSTRACT

DATA MINING IS A VERY IMPORTANT DATABASE RESEARCH ISSUE. ESPECIALLY, THE GENERATION OF FREQUENT PATTERNS IN LARGE DATABASE HAS BEEN WIDELY STUDIED. MOST OF THE STUDIES TAKE THE APRIORI-BASED APPROACH, WHICH HAS GREAT EFFORT IN THE GENERATION OF CANDIDATE FREQUENT PATTERNS AND NEEDS MULTIPLE DATABASE ACCESSES. THE FP-TREE-BASED APPROACHES HAVE BEEN PROPOSED TO AVOID THE GENERATION OF CANDIDATE SET AND SCAN TRANSACTION DATABASE ONLY TWICE, BUT THEY WORK WITH MORE COMPLICATED DATA STRUCTURE. RECENTLY, A FREQUENT PATTERN LIST (FPL) ALGORITHM, USING A SIMPLE LINEAR LIST TO STORE ALL TRANSACTIONS, WAS PROPOSED TO IMPROVE THE FP-TREE ALGORITHM. HOWEVER, FPL ALGORITHM STILL NEEDS TO SCAN DATABASE TWICE. IN THIS PAPER, AN EFFICIENT FREQUENT PATTERN GENERATION ALGORITHM, CALLED FPLI, WAS PROPOSED TO IMPROVE THE FPL ALGORITHM. FPLI SCANS THE DATABASE ONLY ONCE AND USES A SIMPLE LINEAR LIST TO STORE ALL TRANSACTIONS LIKE FPL. BY PERFORMING SIMPLE OPERATIONS ON THE LIST, WE CAN DISCOVER THE FREQUENT PATTERNS QUICKLY. IT IS ALSO NOT NECESSARY FOR FPLI TO RESCAN DATABASE AND RECONSTRUCT DATA STRUCTURE WHEN TRANSACTION DATABASE IS UPDATED OR MINIMUM SUPPORT IS VARIED. EXPERIMENTAL RESULTS SHOW THAT THE FPLI ALGORITHM HAS MUCH BETTER PERFORMANCE THAN THE FPL ALGORITHM.

Keywords : DATA MINING、FREQUENT PATTERN、CANDIDATE ITEMSET

Table of Contents

第一章 緒論--P1 第二章 相關研究--P6 2.1 APRIORI演算法--P6 2.2 DHP演算法--P7 2.3 DIC演算法--P7 2.4 RANDOM SAMPLING演算法--P8 2.5 PINCER-SEARCH演算法--P9 2.6 FP-TREE演算法--P10 2.7 FPL演算法--P14 第三章 快速高頻資料項目型樣挖掘演算法--P19 3.1 問題定義與探討--P19 3.2 FPLI演算法--P20 3.2.1 辭彙說明--P20 3.2.2 演算法分析與描述--P21 3.2.3 複雜度分析--P35 第四章 適用於最小支持度改變以及資料庫異動之高頻資料項目型樣挖掘演算法--P37 4.1 最小支持度變動演算法--P38 4.2 新增部份交易演算法--P41 4.3 刪除部份交易演算法--P42 第五章 實驗及結果分析--P44 5.1 實驗測試例子說明--P44 5.2 實驗結果分析及效能評估--P45 5.2.1 FPLI演算法和FPL演算法的效能比較--P45 5.2.2 最小支持度變動演算法和FPL演算法的效能比較--P48 5.2.3 新增部份交易演算法和FPL演算法的效能比較--P50 5.2.4 刪除部份交易演算法和FPL演算法的效能比較--P52 第六章 結論與未來展望--P55 6.1 結論--P55 6.2 未來展望--P55 參考文獻--P57

REFERENCES

- [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 DATABASES",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 SIGMOD 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.