

論文名稱：預測式關聯規則演算法

總頁數：69

校院(所)組別：中國文化大學商學院資訊管理研究所

畢業時間及提要別：99 學年度第一學期碩士學位論文提要

研究生：李勃穎

指導教授：李長彥

李中彥

論文提要內容：

近年來關聯規則(association rules)技術已被廣泛的運用在資料探勘領域之中，關聯規則演算法分為兩個部份，首先從交易資料中找出購買次數高於支持度門檻的頻繁項目集，其次為這些頻繁項目集中找出商品之間購買的規則。

在關聯規則演算法中，時間的耗費主要在於找到頻繁項目集，而在以往的關聯規則演算法之中最常被使用的為 Apriori 演算法，雖然此演算法可以找出頻繁項目集，但是它存在著兩大缺點，第一點為產生過多的候選項目集，第二點為需多次掃瞄資料庫，而造成整體執行時間效率不佳。許多專家學者針對這兩個缺點提出改善的方式，然而皆未離開 Apriori 的架構，因此本論文提出預測式關聯規則演算法來提昇找到頻繁項目集的時間效率。

在預測式關聯規則演算法中，只需掃瞄資料庫兩次，第一次掃瞄完成長度項目分配表，接下來再利用使用者所輸入的長度誤差容許和頻繁誤差容許預測出所有的頻繁項目集，接著再一次掃瞄資料庫找出頻繁項目集，其優點為執行時間效率佳，缺點為可能產生誤差。

關鍵字:資料探勘(data mining)、關聯規則演算法(association rules)
apriori 演算法(apriori algorithm)

Predictive association rules algorithm

Student : Bo-Ying Li

Advisor : Prof.Tsang-Yean Lee
Prof.Chong-Yen Lee

Chinese Culture University

ABSTRACT

In past decades, the association rules technology has been applied in data mining domain. The association rules algorithm has two parts. The first part is finding the frequent item set where purchase of times over support threshold from transaction data. The second part is finding the association rules from frequent item set.

In the association rules algorithm, the first part is time-consuming. Apriori algorithm is the most often used association rules algorithm in former association rules algorithm. Although, this algorithm can find the frequent item set. But it has two shortcomings. The first shortcoming is generating candidate item set too much. The second shortcoming is scanning transaction data times without number. Therefore give occasion to time-consuming. Many experts propose the improvement ways in view of these two shortcomings. However the improvement ways are still using Apriori algorithm structure. In this paper we propose predictive association rules algorithm. This algorithm can find the frequent item set quickly.

Predictive association rules algorithm only need scan the database two times. First scan finish length-Item distribute total table. Then using the length error number and frequent error number predictive all frequent item set. Finally scan the transaction finding real frequent item set.

Keywords: data mining, association rules, apriori algorithm

誌 謝 辭

能完全這篇論文首要感謝的是我的指導教授李中彥教授與李長彥教授，讓我在許多瓶頸與挫折中打起精神記取經驗，並透過每週的討論給我正確的方向與針對我的問題給我建議與改善的方式，讓我受益良多。接著，我要感謝我的口試委員王美慈教授與李彥良教授，在我論文初審時給我許多寶貴的意見，讓我針對自己的缺點補強，讓論文更加完備。

接下來我要感謝我的同學林宇哲，在這些日子裡陪伴著我，陪我討論演算法的改善方式與程式上表現的方法，並耐心的指導我程式上的所有問題。另外，我要感謝鄭凱聰同學指導我在論文格式上的缺失與改正的方式，有了他們的協助我的論文才能順利完成與更加完善。

最後，我要感謝我的雙親，支持我、鼓勵我完成學業，有了他們支持讓我可以專心的研究，不必煩惱其他的事務。

內 容 目 錄

中文摘要	iii
英文摘要	iv
誌謝辭	v
內容目錄	vi
表目錄	viii
圖目錄	xi
第一章 緒論	1
第一節 研究背景與動機	1
第二節 研究目的	2
第三節 研究範圍與限制	3
第四節 研究流程	4
第五節 章節架構	5
第二章 文獻探討	7
第一節 資料挖掘介紹	7
第二節 購物籃分析	10
第三節 關聯規則演算法介紹	11
第四節 項目集拆解方式介紹	24
第三章 預測式關聯規則之頻繁項目集演算法	28
第一節 預測式演算法說明	28
第二節 誤差容許指數之設定	31
第三節 範例說明	33
第四章 頻繁項目集演算法實驗及討論	36
第一節 實驗設計及結果	36
第二節 不同演算法之優缺點	54
第五章 關聯規則的產生	57

第一節	關聯規則運算	57
第二節	實驗設計及結果	58
第六章	結論及未來發展	61
參考文獻		63



表 目 錄

表 2-1	資料探勘的主要功能	8
表 2-2	雜貨店銷售表	11
表 2-3	名詞解釋表	11
表 2-4	符號解釋表	13
表 2-5	交易資料表	19
表 3-1	長度項目分佈累計示意表	29
表 3-2	符號解釋表	32
表 3-3	交易資料表	33
表 3-4	長度項目分佈累計表	34
表 3-5	排序後的長度一次數表	34
表 3-6	頻繁過濾清單	34
表 3-7	拆解名稱說明表	35
表 3-8	最終候選項目集清單	35
表 3-9	頻繁項目集	35
表 4-1	實驗平台	36
表 4-2	實驗目的表	37
表 4-3	初步實驗之資料設定	37
表 4-4	初步實驗之演算法細部設定表	38
表 4-5	初步實驗設定下各演算法的執行時間	39
表 4-6	初步實驗之頻繁項目集個數與支持度門檻關係表	39
表 4-7	實驗一之資料設定	40
表 4-8	實驗一之演算法細部設定表	40
表 4-9	實驗一設定下各演算法的執行時間表	41
表 4-10	實驗一頻繁項目集個數與支持度門檻關係表	42
表 4-11	實驗二資料設定	42

表 4-12	實驗二之演算法細部設定表	43
表 4-13	實驗二設定下各演算法的執行時間表	43
表 4-14	實驗二設定下預測式演算法的前置處理時間	43
表 4-15	頻繁項目集個數與資料量關係表	45
表 4-16	實驗三之資料設定	45
表 4-17	實驗三之演算法細部設定表	46
表 4-18	在實驗三設定下各演算法演算時間表	46
表 4-19	實驗三設定下預測式演算法的前置處理時間表	46
表 4-20	頻繁項目集個數與總項目數關係表	48
表 4-21	實驗四之資料設定表	48
表 4-22	實驗四之演算法細部設定表	49
表 4-23	在實驗四設定下各演算法演算時間表	49
表 4-24	頻繁項目集個數與分配方式關係表	50
表 4-25	實驗五之資料設定	51
表 4-26	實驗五演算法細部設定表	51
表 4-27	實驗五設定下預測式演算法演算時間表	51
表 4-28	頻繁項目集個數與頻繁誤差容許指數關係表	52
表 4-29	實驗六之資料設定表	53
表 4-30	實驗六之演算法細部設定表	53
表 4-31	實驗六設定下預測式演算法演算時間表	53
表 4-32	頻繁項目集個數與長度誤差容許指數關係表	54
表 4-33	各演算法比較表	56
表 5-1	頻繁項目集	57
表 5-2	關聯規則	58
表 5-3	實驗七之資料設定	58
表 5-4	實驗七關聯規則數量與演算法關係表	59
表 5-5	實驗八之資料設定	59

表 5-6 實驗八關聯規則數量與演算法關係表 60



圖 目 錄

圖 1- 1	研究流程圖	4
圖 2- 1	Apriori演算流程	13
圖 2- 2	重覆組合流程圖	14
圖 2- 3	Pascal演算流程圖	15
圖 2- 4	Partition演算流程圖	16
圖 2- 5	SWF解說圖	18
圖 2- 6	SWF檢驗流程圖	18
圖 2- 7	SWF演算流程圖	20
圖 2- 8	FP TREE演算流程圖	22
圖 2- 9	Counting base on matrix演算流程圖	23
圖 2-10	ICI拆解流程圖	25
圖 2-11	QSD建置二元樹流程圖	26
圖 2-12	IDA拆解流程圖	27
圖 3- 1	演算流程圖	28
圖 3- 2	前置作業流程圖	29
圖 3- 3	運算作業流程圖	30
圖 3- 4	運算作業演算法	31
圖 3- 5	誤差關係圖	32
圖 3- 6	誤差關係範例圖	32
圖 4- 1	初步實驗設定下執行時間與支持度門檻變化圖	39
圖 4- 2	實驗一設定下執行時間與支持度門檻變化圖	41
圖 4- 3	實驗二設定下執行時間與資料量變化圖	44
圖 4- 4	前置處理時間與資料量變化圖	44
圖 4- 5	實驗三設定下執行時間與總項目數量變化圖	47
圖 4- 6	前置處理時間與總項目數變化圖	47

圖 4-7	實驗四設定下執行時間與分配方式變化圖	50
圖 4-8	實驗五設定下執行時間與頻繁誤差容許變化圖	52
圖 4-9	實驗六設定下執行時間與長度誤差容許變化圖	54

