針對重要稀少性資料之一種有效率關聯式探勘方法設計

龔旭陽
國立屏東科技大學資訊管理學系

林美賢
國立中山大學資訊工程學系

林靖祐
國立屏東科技大學資訊管理學系

賴威光
國立中山大學資訊工程學系

摘要

關聯法則(Association Rules)廣泛應用於資料探勘研究方法，於過往研究中，大都針對支持度(Support)較高之高頻項目集(Frequent ItemSets)進行探勘，然而卻無法迅速且有效探勘出支持度小但卻擁有重要關聯性之重要稀少性資料(Significant Rare Data)，亦即所謂之半高頻項目集(Semi-frequent ItemSets)。現今有部份研究針對具備重要關連法則之稀少性資料，進行相關探勘方法設計，其方法大都採用由下而上(Bottom-Up)搜尋方式，但往往無法有效率探勘出最大半高頻項目集(Maximal Semi-frequent ItemSets)。針對上述問題，本研究提出與設計專門針對重要稀少性資料之最大半高頻項目集探勘演算法(Maximum Semi-frequent Itemsets Algorithm, MSIA)，MSIA可有效整合分群(Cluster)與分解(Decomposition)探勘概念，並結合篩選法(Filter)與相對支持度(Relative Support)分析方法，採由上而下(Top-Down)之搜尋機制進行高效率最大半高頻項目集探勘。由效能實驗結果可知，MSIA於探勘過程中可以有效降低原始來源資料庫(Source Database)讀取掃描次數，提升探勘效能以節省探勘時所花費之時間成本，進而有效且快速取得重要稀少性資料中之最大半高頻項目集。

關鍵字：關聯法則、重要稀少性資料、最大半高頻項目集、分群、相對支持度

An Efficient Method for Mining Association Rules on Significant Rare Data

Hsu-Yang Kung
Department of Management Information Systems,
National Pingtung University of Science and Technology

Mei-Hsien Lin
Department of Computer Science and Engineering, National Sun Yat-sen University

Ching-Yu Lin
Department of Management Information Systems, National Pingtung University of Science and Technology

Wei-Kuang Lai
Department of Computer Science and Engineering, National Sun Yat-sen University

Abstract

Mining out the association rules is the popular research issue in data mining research. In recent years, many studies have focused on discovering the important association rules based on the criteria of maximum support and confidence for frequent itemsets. The significant rare data, i.e., the semi-frequently itemsets, are not easily to mine out the important association rules using traditional mining methods. Some mining methods based on the bottom-up policy can not efficiently mine out association rules from longer length of semi-frequent itemsets. The time complexity of mining process is very high due to the generation of large candidates by repeatedly scanning source database. This research proposed the maximum semi-frequent itemsets algorithm (MSIA), which quickly and efficiently mining out the association rules on the significant rare data. MSIA is a top-down approach by combining the techniques of clustering, decomposition, filtering, and relative supports to efficiently search the source database. From the performance of experiment results, the MSIA can decrease the time complexity of scanning database and thus significantly reduce the number of candidate itemsets. MSIA efficiently mines out the useful association rules from the maximum semi-frequent itemsets.

Key words : Association rule; Significant rare data; Semi-frequent ItemSets; Cluster; Decomposition

壹、導論

透過資料探勘(Data Mining)技術，可由龐大的資料庫中萃取出具有價值的隱藏資訊與知識(Chen et al. 1996; Han et al.1997)，所採用之相關資料探勘技術可大略分為關聯法則(Association Rules) (Agrawal et al. 1993; Han et al. 1997; Park et al. 1997)、分類式法則(Classification Rule) (Ali et al. 1997; Tsay & Chang-Chien 2004)、組群式法則(Clustering Rule) (Kaufman & Rousseeuw 1990)以及序列型資料樣型分析(Sequential Pattern Analysis) (Agrawal & Srikant 1995; 魏志平 & 董和昇 2000)等，其中關聯法則為使用較為廣泛之技術。關聯法則主要為尋找出資料庫中各項目集(ItemSet)間的關聯性，其利用支持度(Support)與信心度(Confidence)進行篩選，發掘出相依關聯性高且具備重要關聯規則之項目集合，亦即所謂高頻項目集(Frequent ItemSets)或較大項目集(Large ItemSets)，例如Apriori Algorithm與Direct Hashing and Pruning Algorithm，即屬這一類別之關聯法則探勘演算法(Agrawal & Srikant 1994; Park et al. 1997)。

關聯法則探勘技術大都針對出現頻率較高即支持度高之項目集進行探勘，但於某些案例上，部份出現頻率較低之稀少性項目集卻仍有可能與其它項目集有著重要且具價值的關聯規則存在。就以咖啡機與咖啡豆之範例而言，咖啡機其購買比率極低，咖啡豆則佔有較高的購買比率，然而大多數購買咖啡機時大都會伴隨購買咖啡豆，因此這樣的購買關聯特性，確實存在著一定的利用價值，對於此類稀少但卻擁有重要關聯法則之項目集可將其歸類為重要稀少性資料(Significant Rare Data)(Yun et al. 2003)。這些資料表示該項目集合擁有較低之支持度，但與其它項目集有著相當高之信心度，對於重要稀少性資料這部份於 Yun et al.(2003)學者其研究中有所定義，將這類的關聯規則稱為半高頻項目集(Semi-frequent ItemSets)，並且更進一步提出(Relative Support Apriori Algorithm, RSAA)進行重要稀少性資料之關聯法則探勘(Yun et al. 2003)。然而由於RSAA演算法是依據Apriori演算法進行改良設計，加入相對支持度(Relative Support)來輔助分析篩選重要稀少性資料，但無論RSAA或Apriori於探勘過程中需一直重覆掃描讀取資料庫，因此相當浪費處理時間，且由於RSAA演算法與Apriori演算法同樣採用由下而上(Bottom-Up)的探勘搜尋方式，因此若使用者欲探勘最大半高頻項目集(Maximal Semi-frequent ItemSets)時，仍必須依序由最小項目集往上探勘，過程中必定會產生許多候選項目集(Candidate ItemSets)，如此並無法確切符合使用者需求。因此針對上述問題，部份學者即著重於減少讀取資料庫次數與降低候選項目集做為主要之研究發展方向，並改採由上而下(Top-Down)的探勘程序，改善較長高頻項目集之搜尋探勘效能，如 D. Lin與Y. J. Tsay等多位學者，即分別提出Pincer-Search演算法與CDAR演算法，此兩種探勘技術皆可利用由上而下的方式進行探勘，以快速發掘出較大之高頻項目集(Maximal Frequent ItemSets)(Lin & Kedem 1998; Tsay & Chang-Chien 2004)。其中CDAR演算法更利用分群(Cluster)與分解(Decomposition)技術，以降低探勘時讀取資料庫次數減少比對候選項目次數，進而減少記憶體需求與降低探勘過程所花費的時間成本，但上述兩項技術卻無法有效適用於重要稀少性資料之探勘。

綜觀上述問題，本研究提出一最大半高頻項目集探勘演算法(Maximum Semi-frequent Itemsets Algorithm, MSIA)為專用於探勘稀少性資料中最大半高頻項目集之演算法。MSIA採用CDAR演算法之分群(Cluster)與分解(Decomposition)探勘概念，減少比對之候選項目集，並結合相對支持度(Relative Support)分析方法，進行重要稀少性資料之探勘，透過由上而下(Top-Down)的搜尋機制，進行高效率的稀少性資料最大半高頻項目集之探勘，以達到快速且精確找尋出重要稀少性資料關聯法則。

本論文於第二章節為文獻回顧，描述MSIA所引用之相關技術與文獻，並且進行分析與評估比較，第三章節則針對MSIA之設計方法流程與演算法詳細描述，第四章節則進行MSIA與RSAA之範例探勘與兩種方法之比較，第五章節則描述說明本研究之MSIA效能分析，最後於第六章節將對本研究作一結論。

貳、文獻回顧

本研究所設計之MSIA主要為整合RSAA與CDAR兩項資料探勘演算法之特色優點，並加以改良以符合重要稀少性資料中最大半高頻項目集之探勘，本章節即針對重要稀少性資料之定義、國內資料探勘文獻探討、Yun et al. (2003)所提出之RSAA和Tsay與Chang-Chien(2004)學者所提出之探勘技術進行相關介紹，並將針對RSAA與CDAR作分析比較(Tsay & Chang-Chien 2004; Yun et al. 2003)。

一、重要稀少性資料關聯法則

支持度即代表項目集於資料庫中所出現頻率高低，擁有高支持度之項目集即代表於資料庫中重覆筆數較多，而擁有支持度較低者，則代表筆數較少，而這即為稀少性資料(Rare Data)。而在以往的關聯法則探勘技術中，大多數是把這些稀少性資料排除不加以考慮，雖然稀少性資料都是無法滿足一般演算法中所設定之最小支持度(Minimum Support)門檻值，但事實上稀少性資料並非全部都是無意義之項目集；在稀少性資料中有可能一部份的項目集是與其它項目集存在著相當高的信心度，這些特別的項目集即稱為重要稀少性資料(Significant Rare Data)，而此類項目集合亦可稱為半高頻項目集(Semi-frequent ItemSets)或類高頻項目集(Quasi-frequent ItemSets) (Yun et al. 2003)。舉例假設某一資料庫中包含*{X}*、*{Y}*與*{Z}*等三個資料項目，且其各別支持度分別為10%、50%與60%，然而每當項目集合中出現*{X}*項目時，即有80%的機率會同時會伴隨出現*{Y, Z}*此項目集合，由此可知*{X}*雖然其支持度僅10%，但與*{Y, Z}*項目集合間卻有著極高信心度，類似上述之特殊關聯性，在某些案例上必定擁有相當的利用價值如預測或防災，然而現今大多數的探勘技術中並無法適用於這類重要稀少性資料關聯法則上，例如當以Apriori演算法來發掘有具有這樣特性之重要稀少性資料關聯法則時，使用者將把探勘篩選之最小支持度門檻降低，然而降低最小支持度門檻值雖可以考慮到支持度較小之項目集，但如此將導致大多數資料項目集皆能滿足最小支持度門檻值，進而被歸類至高頻項目集合中，結果產生巨大數目之多餘且不重要的關聯法則(Liu et al. 1999)。

二、國內資料探勘文獻探討

近年來國內外有多數學者著力於關聯法則之高頻項目集探勘技術發展研究與改良，以下即針對林淑菁(2004)學者所提出之MFSA 演算法(Maximum Frequent Itemsets Algorithm, MFSA)和蔡玉娟等(2003)學者所提出之快速反向關聯法則(Fast-Backward Association Rule, FBAR)演算法進行探討說明(林淑菁 2004; 蔡玉娟等 2003)。

（一）MFSA演算法

林淑菁(2004)學者所提出之MFSA其主要整合Pincer-Search 演算法的由上而下(Top-down)搜尋機制與Parameterised演算法利用參數設定使之可由下向上(Bottom-up)一次搜尋多層的基礎概念，設計出以雙向搜尋(Two-way Search)之探勘模式，以更快速的取得最大高頻項目集或較長之高頻項目集，並可大量減少高頻候選項目集數目，以有效降低資料庫掃描次數，提升最大高頻項目集之探勘處理效率(Denwattana & Getta 2001; Lin & Kedem 1998; 林淑菁 2004)。

（二）FBAR演算法

由於一般關聯法則演算法須受限由單一組合之高頻項目集開始，透過不斷的重複組合與運算步驟，才可找出較大之高頻項目集，進而導致其探勘效率相當低落，因此蔡玉娟等(2003)學者針對此項問題所提出之快速反向關聯法則演算法(FBAR)以進行改善，FBAR採以反向於Apriori演算法之程序，透過事先的一次資料庫掃描，將不符合欲探勘目標之交易項目集剔除，再由最長之項目集逐層分解候選項目集，以減少大量的交易記錄比對時間，提供一快速發掘適合長度之高頻項目集，因此FBAR之探勘程序屬於由上而下(Top-Down)的方式，對於最大之高頻項目集擁有較佳之探勘效率(Lin & Kedem 1998; 林淑菁 2004)。

針對上述國內學者所提出之MFSA與FBAR等演算法，皆可有效率的探勘取得最大或較大之高頻項目集，然而對於半高頻項目集之探勘，卻無法提供有效且合適之重要性稀少資料探勘模式，因此雖可大幅降低對於較大高頻項目集之探勘花費時間成本，然而卻無法有效適用於重要稀少性資料之探勘。

三、RSAA演算法

Yun et al.(2003)學者提出(Relative Support Apriori Algorithm, RSAA)之演算法，其作法主要為改良Apriori演算法，以運用於重要稀少性資料之探勘，其相關流程步驟敘述如下(Liu et al. 1999)。

（一）RSAA演算法支持度定義

RSAA演算法主要之概念為使用相對支持度(Relative Support)來判斷候選項目集合是否為半高頻項目集，進而探勘出重要稀少性資料，於RSAA探勘程序中，必須定義二個支持度門檻值，分別為第一最小支持度(1^{st} Support)和第二最小支持度(2^{nd} Support)，1^{st}

Support主要是將項目集合分類為稀少資料項目集與高頻項目集，而2nd Support則主要將不滿足1st Support門檻值之稀少項目集進行再次篩選，因此在設定上1st Support和2nd Support必需滿足1st Support > 2nd Support之條件，否則篩選結果一般高頻項目集與稀少性資料之高頻項目集，將發生重覆之現象，如此將造成發掘過程與結果產生問題(Liu et al. 1999)。

除了1st Support和2nd Support之外，RSAA還有使用另一個最小相對支持度(Least Relative Support)門檻值進行篩選，主要因素為含有稀少資料項目之候選項目集合其支持度必定較低，因此RSAA設計一公式用以計算每個含有稀少性資料之候選項目集合的相對支持度(Relative Support)，假設欲探勘之資料庫中，含有n個資料項目(Item)，而欲計算相對支持度之項目集合為$\{I_1, I_2, ..., I_K\}$，首先必需先掃描資料庫，以取得$\{I_1, I_2, ..., I_K\}$之支持度，再將該$\{I_1, I_2, ..., I_K\}$項目集支持度分別除以該集合中各個別資料項目之支持度，所有計算結果中最大值即為該候選項目集合之相對支持度，因此相對支持度亦可視為候選項目集合與稀少資料項目間的信心度計算(Liu et al. 1999)。

（二）RSAA演算法探勘流程

RSAA演算法探勘流程如圖1所示，其步驟如下(1) 分別篩選出滿足1st Support之高頻項目集與不滿足1st Support之稀少項目集，(2) 利用2nd Support門檻值進行稀少項目集之第二次篩選，將過於稀少之資料項目集去除，(3) 將滿足2nd Support之稀少項目集與滿足1st Support之高頻項目集進行Join/Pruning組合處理，(4) 將Join/Pruning後之候選項目集(Candidate ItemSet)進行相對支持度(Relative Support)計算，若該候選項目集之相對支持度高於最小相對支持度門檻值，則代表該侯選項目集為一項重要稀少性資料，亦即所謂之半高頻項目集，(5) 將所取得之半高頻項目集彼此進行Join/Pruning，再將產生之候選項目集進行相對支持度評估計算，並依最小相對支持度門檻值進行半高頻項目集篩選。如此重覆步驟(5)至無法產生再任何候選項目集為止。

由Apriori所改良之RSAA，會有以下兩項問題，(1) 資料庫掃描次數過高：RSAA之比對方式亦與Apriori之反覆比對方式雷同，因此將造成資料庫必需一直大量重覆讀取；(2) 無法快速取得較大之半高頻項目集：由於RSAA之探勘過程是由下而上(Bottom-Up)搜尋方式，候選項目須由項目個數最少之項目集依序向上探勘，一旦使用者僅欲取得較大之半高頻項目集時，亦需由最短之候選項目集開始依序探勘，因此尋找最大半高頻項目集時，RSAA之搜尋時間成本相當高。

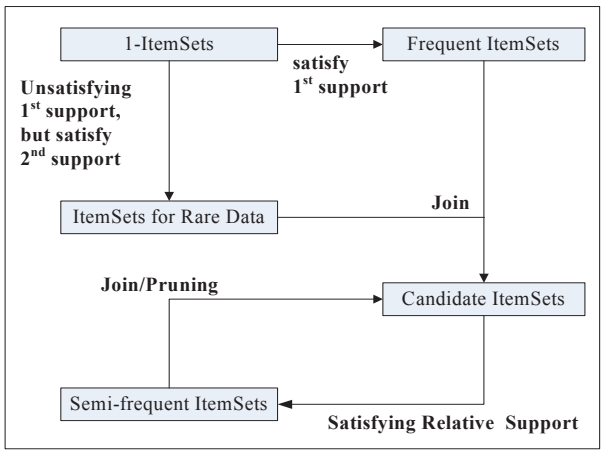

圖1：RSAA探勘流程

四、CDAR演算法

CDAR演算法是由Tsay 與 Chang-Chien(2004)學者所提出，其全名為分群分解關聯法則(Cluster-Decomposition Association Rule, CDAR) (Tsay & Chang-Chien 2004)，CDAR利用分群與分解之探勘概念，有效降低資料庫搜尋次數，並透過由上而下(Top-Down)之方式，快速探勘出較大之高頻項目集合。

（一）CDAR演算法基本介紹

CDAR演算法可有效降低探勘流程中資料庫掃描次數，其研究中顯示僅需讀取資料庫一次，便能利用有效的分群(Cluster)機制將資料庫進行分群(Tsay & Chang-Chien 2004)。其分群機制主要為依據資料庫中每項紀錄集合之長度做為分群依據，分群後隨即針對最大候選項目集(Maxiaml Candidate ItemSets)進行支持度評估，直接快速求得最大高頻項目集。CDAR與傳統由下而上之探勘方式其相異之處為，傳統探勘技術大都由較小候選項目集篩選出較小高頻項目集後，再將所得之高頻項目集透過由連結(Join)以及修整(Prune)兩項程序產生較大之候選項目集，再繼續向上探勘。而CDAR則採用由上而下之探勘方式，一開始即直接探勘取得最大高頻項目集，若欲再探勘較小之高頻項目集時，則必需使用分解(Decomposition)技術，將最大高頻項目集進行分解為多個較小之候選項目集合，但由於CDAR演算法於第一次掃描資料庫時即一併評估紀錄分群中每個項目集合之支持度，因此各較小候選項目集之支持度亦無須藉由重覆掃描資料庫來取得，直接由第一次掃描資料庫後所得分群紀錄中比對獲得。

（二）CDAR演算法探勘流程

CDAR演算探勘法流程如圖2所示，其主要探勘步驟如下說明，假設某資料庫中最長之項目集長度為M，步驟(1) 首先須掃描資料庫一次，依據各項目集合長度分類為M個群組，並於分群過程中記錄該項目集出現頻率，例如某項目集合長度為k，則該項目集便分類儲存於第k個群組中，該群組名稱命名為Cluster(k)，於分群完畢後，即進入下一步驟。

步驟(2) 是由項目集長度最長之Cluster(*M*)中進行高頻項目集之探勘，Cluster(*M*)所含之項目集即為最大候選項目集，而其支持度若滿足最低支持度門檻值，即屬於最大高頻項目集合，而不滿足最低支持度門檻值之項目集則先歸類為臨時項目集(Temp ItemSets)，當該Cluster(*M*)內所有項目集皆篩選過後，則進入分解程序，將臨時項目集分解為多個長度為*M-1*之較小項目集。分解後之項目集於Cluster(*M-1*)群組中發生重覆情況，則該較小項目集之支持度為兩者之總合，若Cluster(*M-1*)群組中之項目集與分解後之項目集無重覆之情形，則直接將Cluster(*M-1*)群組中之項目集一併列為長度*M-1*之候選項目集，倘若分解後結果為已取得高頻項目集之子集合，則不加以考慮，待所有長度*M-1*之候選項目集分群確認完畢後，即再度進行最小支持度門檻值之篩選。若高於最小支持度門檻值之集合，則屬於長度為*M-1*之高頻項目集，而不滿足最小支持度門檻值之集合，則繼續進行分解程式，重覆進行步驟(2)，直至項目集長度過小無法再進行分解或已達使用者欲探勘之項目集合最小長度(Min Length)為止；因此由上述之CDAR探勘流程，可明顯發現CDAR是一套採由上而下之探勘機制，不僅可快速探勘出最大高頻項目集，更可有效降低資料庫搜尋次數(Tsay & Chang-Chien 2004)。

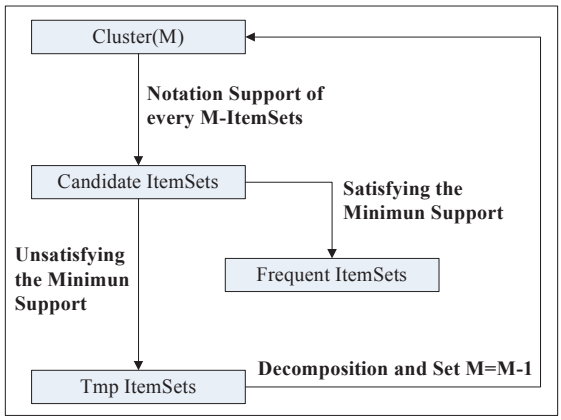

圖2：CDAR探勘流程

CDAR其特色優點為可快速探勘取得較大之高頻項目集，且有效降低資料庫掃描次數，提升探勘效率並節省花費之時間成本，然而CDAR雖具備了上述之特性與優點，但其僅適合應用於一般高頻項目集之探勘，無法適用於重要稀少性資料之關聯法則探勘，此外由於探勘重要稀少性資料之半高頻項目集時，其考慮之候選項目集與探勘一般高頻項目集之候選項目集有所不同，因此並無法直接套用於半高頻項目集之探勘。

五、比較分析

針對RSAA與CDAR資料探勘演算法之特色與優缺點整體分析比較整理如表1所示。RSAA雖可有效探勘取得重要稀少性資料，但使用者若欲探勘取得較大半高頻項目集時，其探勘效率將明顯下降，而CDAR採以由上而下的探勘程序，雖可大幅提升對於較

大高頻項目集之探勘效率，然而卻無法有效適用於重要稀少性資料之探勘。因此本研究針對RSAA與CDAR缺點與不足之處進行改善，設計MSIA演算法以提供高效率的較大半高頻項目集探勘模式。

表1：RSAA與CDAR比較分析表

方法名稱	RSAA	CDAR
探勘目的	可有效探勘重要稀少性資料 (Significant Rare Date)	可快速探勘較大之高頻項目集 (Maximal Frequent ItemSets)
探勘程序	由下而上(Bottom-Up)	由上而下(Top-Down)
核心技術概念	採用 Apriori 進行改良	利用分群(Cluster)與分解 (Decomposition)技術
資料庫掃描次數	多	少
候選項目數量	多	少
探勘效率	差	佳
缺　　點	探勘過程產生大量候選項目集，造成效率低落	探勘效率較佳，但無法適用於重要稀少性資料之探勘

參、最大半高頻項目集探勘演算法設計

本研究所設計之最大半高頻項目集探勘演算法(Maximum Semi-frequent Itemsets Algorithm, MSIA)將整合並利用分群、分解與相對支持度(Relative Support)等相關技術概念，提供較有效率之重要稀少性資料中最大半高頻項目集探勘技術，以下分別MSIA所使用之相關支持度篩選門檻定義與探勘流程步驟做一詳述說明。

一、MSIA相關支持度門檻定義

MSIA主要包含三個支持度門檻定義，分別為（一）高頻項目最小支持度(Frequent Itemsets Min Support, FIMSup)、（二）稀少項目最小支持度(Rare Itemsets Min Support, RIMSup)與（三）相對最小支持度(Relative Min Support, RMSup)，當資料庫之項目集合其支持度滿足FIMSup時，即代表該項目集合為高頻項目集(Frequent Itemset, FI)。若項目集合其支持度不滿足FIMSup但滿足RIMSup，該項目集合則為稀少項目集(Rare Itemset, RI)，而未滿足RIMSup則將其定義無意義性之稀少性資料。因此根據上述特性，FIMSup必需大於RIMSup，否則稀少性項目集與高頻項目集將重覆出現相同項目集，而RMSup則為MSIA發掘過程中判斷稀少性資料候選項目集是否為半高頻項目集之主要評估標準。

二、MSIA之FIMSup與RIMSup門檻值定義

MSIA探勘流程中FIMSup與RIMSup為決定高頻項目與稀少項目個數之主要篩選依據，FIMSup與RIMSup門檻值之不當定義，將導致無法取得有用之重要稀少性資料探勘

結果，於本研究中分別針對FIMSup與RIMSup門檻值設定設計一標準定義規則，其中FIMSup之定義如公式(1)所示，而RIMSup則如公式(2)所示，資料庫中所有個別項目集分別為$I_1, I_2, \ldots, I_n$，$\Delta 1$為個別項目最大支持度以下，欲分類為高頻項目之支持度偏移比率(Offset of Support Ratio)，而$\Delta 1$則為不滿足FIMSup條件下欲求稀少項目之支持度偏移比率。

$$FIMSup = Max(Sup(I_1), Sup(I_2), \ldots, Sup(I_n)) - \Delta_1 \quad \text{公式(1)}$$

$$RIMSup = FIMSup - \Delta_2 \quad \text{公式(2)}$$

三、MSIA相對支持度計算

探勘重要稀少性資料過程中，較無法利用一般之支持度來判斷項目集合是否為半高頻項目集，須藉由候選項目集對於稀少性資料之信心度來進行判斷，即為所謂的相對支持度(Relative Support)，假設欲求相對支持度之候選項目集合為$C=\{I_1, I_2, \ldots, I_m\}$，而$C$集合中屬於稀少性資料之項目集合為$R=\{I_j, \ldots, I_k\}$，以$Sup(I)$表示各資料項目集之支持度，則該$C$候選項目集相對支持度$RSup(I_1, I_2, \ldots, I_m)$計算將如公式(3)所示。在此以範例進行說明本MSIA，假設某候選項目集合為$\{W, X, Y, Z\}$，其支持度為20%，$\{X\}$與$\{Y\}$屬稀少性資料，支持度分別為25%與30%，則該候選項目集相對支持度為 $Max(\frac{20}{25}, \frac{20}{30}) = 80\%$ ，由此可知$\{W, X, Y, Z\}$與稀少性資料具有明顯關聯規則；MSIA所設計之相對支持度計算公式，是由RSAA之概念進行改良，差異處為MSIA於計算過程中，僅考慮稀少性資料之信心度，並不會如RSAA考慮多餘不必要資訊，如此將有助於提升探勘效率。

$$RSup(I_1, I_2, \ldots, I_m) = Max(\frac{Sup(I_1, I_2, \ldots, I_m)}{Sup(I_j)}, \ldots, \frac{Sup(I_1, I_2, \ldots, I_m)}{Sup(I_k)}) \quad \text{公式(3)}$$

四、MSIA探勘流程步驟設計

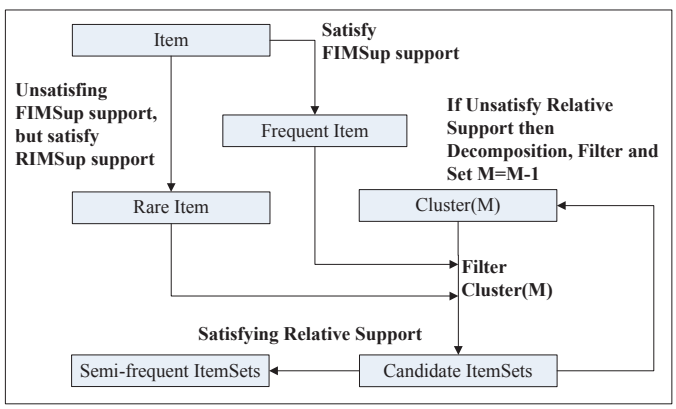

圖3：MISA探勘流程

MSIA探勘方法可分為下列6個步驟，(一) 產生高頻項目 (Frequent Item, FI)：由資料庫中選取出滿足FIMSup之項目，並將其定義為高頻項目，(二) 產生稀少性項目 (Rare Item, RI)：則將剩於無法滿足FIMSup之項目，進行支持度評估，若其支持度高於RIMSup，則將其歸類為稀少性項目，(三) 項目集分群與篩選(Cluster and Filter)：當完成上述步驟後即依據各別項目集合之長度*i*進行分群*Cluster(i)*，若探勘資料庫中最大之項目集合長度為*M*，排除長度為1之群組不加以考慮，則所有項目集最多將分類為*M-1*個群組，例如*{A, B, C}*集合屬於Cluster(3)之群組，此外於分群過程中，並須一併紀錄項目集合重覆出現頻率(Frequency)，例如*{A, B, C}*於資料庫中重覆出現*t*次相同記錄集合，則*Frequency{ A, B, C}=t*，然而完成分群後MSIA須再執行篩選動作，將項目集合中未包含任何稀少性資料之項目集合排除，主要因為於探勘半高頻項目集過程中此類項目集對於發掘結果將不會造成任何影響，因此對其加以排除不加以考慮，(四) 計算候選項目集相對支持度：MSIA對於候選項目集之計算公式如公式(3)所示，主要將候選項目集支持度個別除以該集合中稀少性項目之支持度，並於所有計算結果中取出最大值，即設定此值為該候選項目集之相對支持度；當候選項目集所屬相對支持度高於RMSup門檻值時，(五) 取得重要稀少性資料，完成長度*M*之半高頻項目集探勘後，若欲向下繼續探勘，則須針對剩餘未滿足RMSup之候選項目集進行，(六) 分解與篩選(Decomposition and Filter)，假設前一階段為針對長度為*M*之群組進行探勘，則分解程序必須將長度為*M*之剩餘未滿足RMSup候選項目集分解為數個長度為*M-1*之子項目集，完成分解後即進行篩選，篩選過程將未包含任何RI與FI之子項目集剔除，而包含稀少性資料之子項目集則列為長度為*M-1*候選項目集，長度*M-1*之候選項目集其Frequency計算方式為逐一比對Cluster(*M-1*)內所有項目集，並將發生重覆之候選項目集出現頻率與Cluster(*M-1*)內該相同項目集出現頻率進行加總，加總所得結果即為該候選項目集頻率，求得後即可利用該頻率進行候選項目集之相對支持度計算，完成上述步驟後再重覆進行步驟(四)，直至無法進行分解或達使用者欲求項目集最小長度為止，本研究所提之MSIA探勘流程圖如圖3所示，而演算法如表2所示。

表2：MISA演算法

MISA Algorithm

```
D : Database (a set of transactions)
FI : Frequent Item, RI : Rare Item, ML : Min Length
I =I1; I2; . . . ; Im : a set of data items accessed by all transactions in D
T=T1;T2; . . . ; Tn: a set of all transactions in D
Begin MISA Algorithm
    For i=1 to m
        If Sup(Ii) > FIMSup then
            Put Ii into FI
        Else If Sup(Ii) > RIMSup then
                Put Ii into RI
    Next
    For i=1 to n
        L = Length(Ti)
        If (Ti∩FI) ≠θ and (Ti∩RI) ≠θ and (Ti∩Cluster(L)) =θ then
            Put Ti into Cluster(L)
    Next
    Set i=Max(Length of all Cluster)
    Do while (i=1) or (i>= ML)
        Set RS=Select Itemsets from Cluster(i)
        Do while RS.EOF
            If (Rsup(Itemset) > RMSup) then
                Get Semi-frequent ItemSet
            Else
                Cluster(i-1) = Decomposition and Filter (Itemset) ∪Cluster(i-1)
            RS.movenext
        Loop
        RS.close
        Set i=i-1
    Loop
End
```

五、以貝氏定理為基礎之重要稀少性資料驗證

透過探勘演算法所取得之重要稀少性資料(Significant Rare Data)對使用者而言，是否實為重要且具備高度價值之重要稀少性資料，仍需依使用者主觀的判斷與認定，因此針對上述重要稀少性資料無法明確認定之問題，本研究將貝式定理(Bayes Theorem)概念導入，建立一以貝氏定理為基礎之重要稀少性資料驗證模式，提供使用者可自行針對探勘結果進行檢視衡量，辨識該取得之項目集與關連是否確為具備價值之重要稀少項目，進而提供稀少性資料探勘過程中項目集合之判斷與參考。

於本稀少性資料驗證模式中，使用者必須事先將幾次的MSIA探勘結果，判斷其對於使用者而言是否實際具備高度價值與重要性，並將辨識結果進行記錄與儲存為歷史辨識資料，當系統擁有基礎之樣本記錄資訊後，即可進行稀少性資料驗證服務。由於MSIA所

探勘取得之重要稀少性資料項目集是由多個子項目所組成，因此若欲判斷該項目集合是否是具備人為主觀條件之重要稀少性資料，則需利用所包含之子項目與相關歷史辨識記錄資料，透過貝式分類法評估驗證。

驗證推論過程中首先假設h_j表示欲驗證辨識結果，而X為欲進行驗證辨識之項目集合，而x_i則代表X所包含相關項目，由上述參數設定，可利用$P(h_j|X)$來表示當項目集合為X時其驗證辨識結果為h_j之機率為多少，因此當$P(h_j|X)$計算取得機率越大時，則代表該X項目集合與該驗證辨識結果h_j間的關連性越高，由貝式定理(Bayes Theorem)得知，$P(h_j|X)$求算方式如公式(4)所示。

$$P(h_j \mid X) = \frac{P(X \mid h_j)P(h_j)}{P(X)} \qquad \text{公式(4)}$$

X 欲驗證辨識之項目集合
x_i 為X所包含之子項目
h_j 為欲驗證辨識之結果

由於項目集合驗證是依據各個$P(h_j|X)$機率值大小進行辨識，由於公式1其分母為$P(X)$，因此可將項目集合各驗證辨識結果$P(h_j|X)$機率求算過程同乘$P(X)$，以進行簡化的貝式定理假設，將公式(4)簡化如公式(5)所示。

$$P(h_j \mid X) = P(X \mid h_j)P(h_j) \qquad \text{公式(5)}$$

然而項目集合X之組合項目相當複雜，因此可能無法由歷史辨識記錄資料中取得完全相同等之記錄，因此無法直接計算其$P(h_j|X)$值，但因X 所包含之子項目屬性都是條件獨立(Conditionally Independent)，因此$P(h_j|X)$機率值計算可透過公式(6)進行推估，由於歷史辨識記錄資料中對於x_i與h_j的關連記錄較易取得，因此重要稀少性資料驗證模式首先針對個別$P(x_i|h_j)$機率進行求算，待所有子項目之$P(x_i|h_j)$機率皆取得後，再進行$P(X \mid h_j)$機率計算。

$$P(X \mid h_j) = \prod_{i=1}^{n} P(x_i \mid h_j) \qquad \text{公式(6)}$$

利用上述公式計算可以取得當驗證辨識結果h_j時，驗證辨識項目集合為X之$P(X \mid h_j)$機率為多少，再透過公式(7)進行$P(h_j|X)$機率之計算，取得欲驗證辨識項目集合為X驗證辨識結果為h_j之機率為多少，取其機率值最大者，即可判斷欲X其辨識結果為何，當取得稀少性資料驗證結果後，系統會提供結果給使用者進行參考，其後系統進一步將使用者對於驗證結果之正確性判斷進行記錄，用以提升重要稀少性資料驗證服務之準確性。

$$P(h_j \mid X) = P(X \mid h_j)P(h_j) = \prod_{i=1}^{n} P(x_i \mid h_j)P(h_j) \qquad \text{公式(7)}$$

肆、MSIA範例探勘

本研究將以表3之資料庫作為MSIA範例推導，針對MSIA之探勘流程作一詳細說明。此外本研究首先以RSAA進行探勘，用以輔助驗證本研究所提出之MSIA正確性。於RSAA探勘範例中1st Support設定為50%、2nd Support為30%，相對支持度(Relative Support)門檻值為70%，而欲求半高頻項目集之最小長度為2，由RSAA所發掘出之重要稀少性資料關聯法則(半高頻項目集)分別為*{B,E}*、*{C,D}*、*{C,E}*與*{C,D,E}*，於RSAA範例探勘過程中，判斷篩選出稀少項目與高頻項目需掃描資料庫1次，而由於範例資料庫中最長項目集合為5，因此針對各長度之候選項目集進行相對支持度計算時，將必需重覆掃描資料庫多次。

表3：範例資料庫

編號	項目集合內容	編號	項目集合內容
1	A、E、F	6	B、E、F
2	B、D、E	7	A、B
3	A、C、D、E、F	8	B、E、F
4	C、D、E、F	9	C、D、E
5	A、D、E、F	10	A、D、E、F

MSIA範例探勘中採用相同篩選門檻值與範例資料庫，其中FIMSup與RIMSup設與RSAA探勘範例中1st Support與2nd Support相同之值，分別為50%與30%，而相對支持度門檻值設為70%，欲求之半高頻項目集最小長度同為2，其探勘流程如圖4所示，詳細內容述說如下。

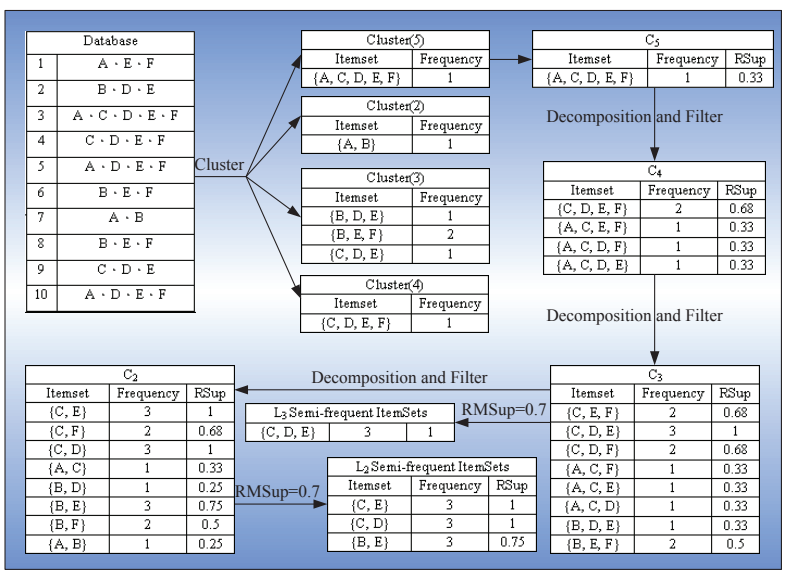

圖4：MSIA探勘範例

（一）選出高頻項目集(FI)與稀少性項目集(RI)：範列資料庫各項目之支持度如表4所示，因此經篩選過後，滿足FIMSup之FI分別為*{A}*、*{D}*、*{E}*與*{F}*，而其餘且支持度高於RIMSup之RI分別為*{B}*與*{C}*。

表4：範列資料庫基項目資訊

No	Item	Support	FIMSup	RIMSup
1	{A}	50%	Satisfy	Satisfy
2	{B}	40%	Unsatisfy	Satisfy
3	{C}	30%	Unsatisfy	Satisfy
4	{D}	60%	Satisfy	Satisfy
5	{E}	90%	Satisfy	Satisfy
1	{F}	70%	Satisfy	Satisfy

（二）項目集分群與篩選：依據項目集合長度進行分群為4個群組，分群過程中須同時紀錄項目集出現頻率，以便往後進行計算候選項目集相對支持度之用。

（三）計算候選項目集相對支持度：*Cluster(5)*中*{A,C,D,E,F}*之相對支持度為 0.33無高於RMSup門檻值，因此該項目集並非半高頻項目集。

（四）分解與篩選：由於未達欲求之項目集合最小長度2，因此必需將*{A,C,D,E,F}*分解並分群於原有之*Cluster(4)*中，*{A,C,D,E,F}*可分別分解為*{C,D,E,F}*、*{A,D,E,F}*、*{A,C,E,F}*、*{A,C,D,F}*與*{A,C,D,E}*，其中*{A,D,E,F}*未包含任何RI值，因此將其剔除，而*Cluster(4)*中*{C,D,E,F}*其頻率更改為2。

（五）由於未達最欲求半高頻項目集最小長度，因此依序針對*Cluster(4)*、*Cluster(3)*與*Cluster(2)*進行重覆流程(3) 候選項目集相對支持度計算與流程(4)分解與篩選兩項步驟，最後探勘所得之半高頻項目集結果分別為*{B,E}*、*{C,D}*、*{C,E}*與*{C,D,E}*。

由結果得知採用RSAA與MSIA皆能獲得相同結果，因此足以驗證本演算法之正確性，然而MSIA於探勘過程中，卻可大幅降低掃描資料庫次數，其效能將明顯提升，時間複雜度為*O(n)*。

伍、效率分析與實作

本研究所設計之最大半高頻項目集探勘演算法，能有效整合分群、分解與相對支持度(Relative Support)等相關技術概念，提供較有效率之最大半高頻項目集探勘技術。本章節將針對MSIA進行系統實作與效能分析，而MSIA相關實作系統畫面如圖5所示，本MSIA實作系統可提供使用者透過網頁瀏覽器直接進行操作，本研究之實作系統探勘流程操作設計中，使用者可依據其欲探勘需求，自行調整FIMSup、RIMSup與RMSup等篩選門檻參數值，來探勘半高頻項目集，此外探勘過程中亦可由系統畫面快速取得探勘相關訊息，其中包含分群狀況、分解與篩選結果、探勘原始資料筆數、各候選項目集相對支

持度、探勘花費時間成本與最後所得之探勘結果，將有助於使用者了解MSIA之運作過程。

效率分析實驗中將MSIA與現有針對半高頻項目集之探勘演算法RSAA進行相關探勘執行效能比較分析，透過效能評估分析，用以佐證本研究所設計之MSIA對於較大半高頻項目集擁有較佳的探勘效率，並驗證MSIA之可行性。效能分析實驗環境採用Pentium M 1.6GHz中央處理器、728MB記憶體與Microsoft Windows 2000作業系統之電腦主機進行相關測試，實驗開發程式語言為VB Script，資料庫管理系統(DBMS)則採用Microsoft Office Access 2003，本研究所探勘之資料交易項目集合為採用隨機方式，實驗過程總共產生1,500筆假想交易資料。

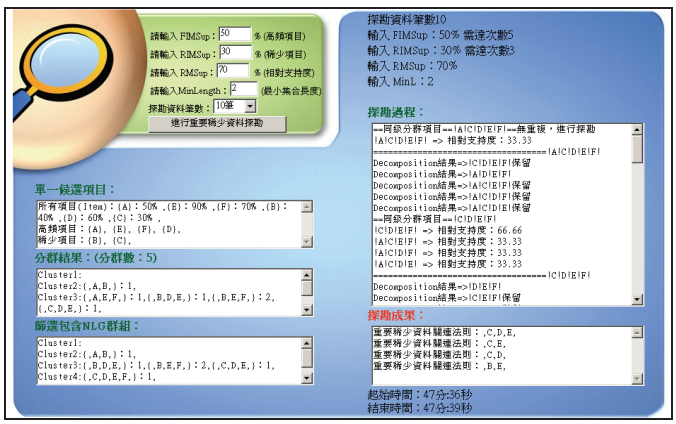

圖5：MSIA系統實作畫面

效能評估分析過程中，於實驗軟硬體設備與探勘資料交易項目集合等環境條件皆相等情況下，本研究將分別針對下列四項分析議題進行效能評估測試，四項分析議題分別為（一）資料交易項目筆數、（二）最小相對支持度、（三）稀少項目最小支持度、與（四）高頻項目最小支持度等門檻值調整對於探勘效能結果之影響分析，以下即針對各項效能分析實驗進行詳述說明。

一、調整資料交易項目筆數之影響

為驗證本研究所提出之MSIA具有較佳的半高頻項目集探勘效率，本實驗章節中，分別採用五組不同之交易項目集資料庫以MSIA與RSAA進行重要稀少性資料探勘，於實驗過程中MSIA與RSAA之RMSup設為9%、FIMSup設為11%、RIMSup亦固定設定為9%，而欲求之半高頻項目集最小長度門檻值同設定為5，匯入資料庫交易項目集筆數分別為500筆、750筆、1000筆、1250筆與1500筆，經過實驗後所得結果如下表5與圖6所示，由實驗結果可以得知MSIA於探勘較大半高頻項目集之探勘過程中，其處理所花費時間與效能皆較佳於RSAA，因此可以驗證MSIA於不同的資料庫大小情況下，皆可獲得較佳之探勘效果。

表5：資料交易項目筆數調整狀態與效能比較

Number of transactions	RMSup	FIMSup (1st Support)	RIMSup (2nd Support)	MSIA Time (ms)	RSAA Time (ms)
500	10%	11%	9%	5015	20136
750	10%	11%	9%	6109	23171
1000	10%	11%	9%	7293	24642
1250	10%	11%	9%	8671	25134
1500	10%	11%	9%	9525	25847

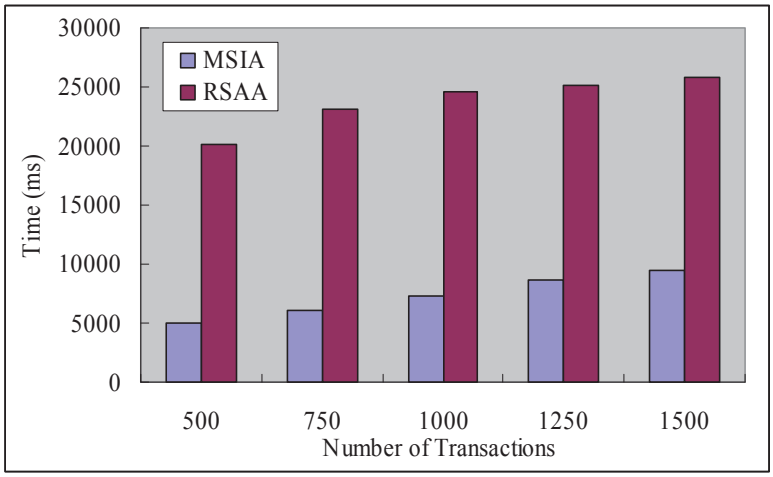

圖6：資料交易項目筆數調整之效能比較

二、調整最小相對支持度之影響

最小相對支持度為探勘半高頻項目集過程中重要的篩選門檻依據，本實驗章節將MSIA與RSAA中之稀少項目最小支持度門檻值皆設定為9%，高頻項目最小支持度門檻值則皆設為10%，而欲求之半高頻項目集最小長度門檻值同設定為5，於RIMSup與FIMSup兩項門檻值為固定之前提下，進行最小相對支持度調整實驗測試，實驗中以取得最大半高頻項目集為主要目標，MSIA與RSAA所得實驗結果如表6與圖7所示。

表6：相對支持度調整狀態與效能比較

RMSup	FIMSup (1st Support)	RIMSup (2nd Support)	MSIA Time (ms)	RSAA Time (ms)
12%	10%	9%	3156	4578
11.5%	10%	9%	3281	4078
11%	10%	9%	3140	4046
10.5%	10%	9%	4015	33281
10%	10%	9%	3875	34265
9.5%	10%	9%	3953	37750
9%	10%	9%	4234	37328
8.5%	10%	9%	3875	41031
8%	10%	9%	3703	41015

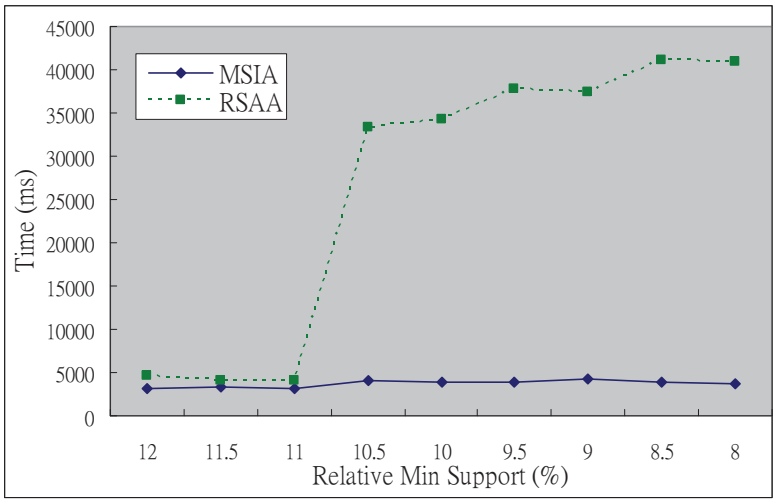

圖7：MSIA與RSAA相對支持度調整之效能比較

由實驗結果中各相異RMSup門檻條件下，MSIA探勘效能皆優於RSAA，因此可證實MSIA於探勘重要稀少性資料上有著較好的探勘執行效率；此外由實驗結果中推測得知，隨機產生之資料交易項目集合中半高頻候選項目集相對支持度(Relative Support)大多為小於11%，因此於RSAA探勘過程中若最小相對支持度門檻值設定小於11%時，將導致產生巨量的半高頻候選項目集，且RSAA演算法對於個別半高頻候選項目集之相對支持度求算，需分別重覆掃描資料交易項目集合，進而造成探勘所花費之時間成本大幅增加，此外若利用RSAA進行長度較長之半高頻項目集探勘時，則產生之半高頻候選項目集數量將更為龐大。

MSIA探勘流程中透過了事先的分群與篩選動作，資料交易項目集合中所包含之半高頻候選項目集支持度大都於此步驟中定義計算完成，因此於計算各候選項目集所屬相對支持度(Relative Support)時，可減少資料庫掃描次數，提升相對支持度計算效率，綜觀上述結果得知，MSIA較不受RMSup調整影響，能有效降低資料庫掃描次數，並且獲得較佳之探勘效能。

三、調整稀少項目最小支持度之影響

稀少項目最小支持度門檻值為決定高頻項目(FI)與稀少項目(RI)總合數量之主要判斷依據之一，當RIMSup提升時高頻項目與稀少項目之總合數量將減少，反之降低RIMSup時將造成高頻項目與稀少項目總合數量增加，而高頻項目與稀少項目則為產生半高頻候選項目集之重要因子，因此本實驗於高頻項目最小支持度、相對支持度與資料交易項目集筆數皆為相等條件下，進行稀少項目最小支持度之調整與效能比較，詳細門檻值調整數據如表7所示，而所得之相關實驗結果如圖8所示。

表7：稀少項目最小支持度調整狀態與效能比較

RMSup	FIMSup (1st Support)	RIMSup (2nd Support)	MSIA Time (ms)	RSAA Time (ms)
10%	11%	10.4%	4750	843
10%	11%	10.2%	4781	828
10%	11%	10%	5015	859
10%	11%	9.8%	8609	24171
10%	11%	9.6%	8593	25390
10%	11%	9.4%	8671	24234
10%	11%	9.2%	8843	23984
10%	11%	9%	9500	25953
10%	11%	8.8%	10578	169156
10%	11%	8.6%	10781	171437

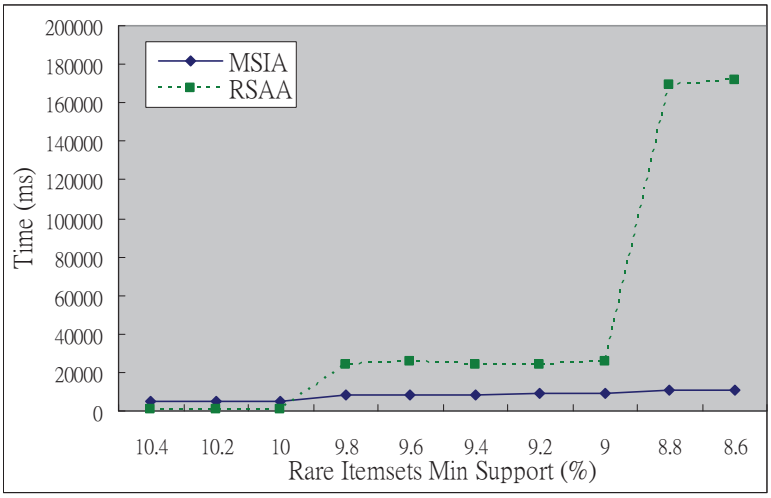

圖8：MSIA與RSAA稀少項目最小支持度調整之效能比較

本實驗中由於FIMSup固定為11%，因此當RIMSup進行調降時，所產生之稀少項目個數將向上調升，而RSAA探勘流程中須將稀少項目與高頻項目進行Join/Pruning處理來產生半高頻候選項目集，因此當稀少項目個數增加時，將導致半高頻候選項目集數量同時提高，而實驗中所探勘之資料交易項目集合中個別項目支持度大多小於9%，因此當RIMSup設定小於9%時，RSAA將產生大量的半高頻候選項目集，進而造成探勘效能降低，導致探勘時間大幅度提升，反觀本研究所設計之MSIA，於RIMSup門檻值調降時雖也造成了探勘時間的提升，但MSIA所採用之分解與篩選機制，卻可有效避免產生過多非必要之半高頻候選項目集，因此於本實驗中MSIA其探勘花費時間皆能有效控制於15000ms以內完成，因此MSIA於不同RIMSup條件下大都能獲得較佳之探勘效果。

四、調整高頻項目最小支持度之影響

高頻項目最小支持度之調整可直接影響高頻項目(FI)與稀少項目(RI)之數量，本實驗測試章節中，將於稀少項目最小支持度、相對支持度與資料交易項目集合筆數皆為相等之情況下，依據不同高頻項目最小支持度之調整，進行MSIA與RSAA兩項演算法的探勘效能測試，高頻項目最小支持度調整狀態與所取得之MSIA與RSAA之效能比較結果如表8與圖9所示。

由上一實驗得知進行探勘之資料交易項目集合中各別項目支持度大多數小於9%，因此在RIMSup門檻值固定設為9%的情況下，FIMSup的調整對於所產生之高頻項目與稀少項目之總合數量並無影響，而主要影響為高頻項目與稀少項目佔其總合數量之分配比率，當FIMSup調降時將造成高頻項目個數提升，而稀少項目所佔比率下降，反之當FIMSup調升時，則高頻項目佔總合數量之比率將會下降，而稀少項目所佔比率則提升。於本章節實驗結果中得知，當FIMSup門檻值小於12%時，RSAA所花費之探勘時間大幅度上升，其主要原因為RSAA透過將高頻項目與稀少項目進行Join/Pruning來產生半高頻候選項目集，因此於高頻項目與稀少項目總合數量不變之情況下，倘若高頻項目與稀少項目所佔比率

相差較為懸殊時，則Join/Pruning所產生半高頻候選項目集較少，而若所佔比率較為接近時，則Join/Pruning產生半高頻候選項目集較多，由於資料交易項目集合內滿足RIMSup門檻值之項目大多小於12%，當FIMSup門檻值調降小於12%時，將造成高頻項目個數之提升，由於高頻項目所佔比率的提升，因此RSAA透過Join/Pruning產生半高頻候選項目集亦會大幅提升，進而導致探勘時間增長。

表8：高頻項目最小支持度調整狀態與效能比較

RMSup	FIMSup (1^{st} Support)	RIMSup (2^{nd} Support)	MSIA Time (ms)	RSAA Time (ms)
10%	14.5%	9%	1282	296
10%	14%	9%	1358	295
10%	13.5%	9%	4094	1453
10%	13%	9%	4358	1530
10%	12.5%	9%	6295	3030
10%	12%	9%	6203	2375
10%	11.5%	9%	9750	27344
10%	11%	9%	9594	29516
10%	10.5%	9%	3702	31077
10%	10%	9%	3890	32625

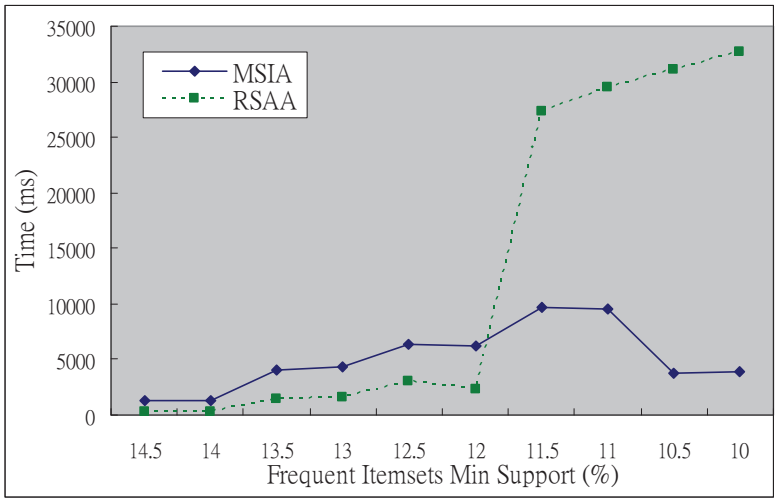

圖9：MSIA與RSAA高頻項目最小支持度調整之效能比較

本研究所設計之MSIA主要採以分群篩選制度，半高頻候選項目集是由包含高頻項目與稀少項目之分群項目集合中所產生，所以MSIA對於高頻項目與稀少項目之個數差距幅度之反應較不直接，因此當高頻項目與稀少項目個數相差過巨之特殊情況下，則可能發生MSIA探勘效率較差於RSAA之情形，但若於較正常之門檻值條件下，MSIA其探勘所花費的時間成本皆較低於RSAA，且本實驗測試中MSIA探勘時間皆有效控制於10000ms以內，反觀RSAA當FIMSup調整至小於12%以下時，其探勘時間卻大幅提升至25000ms以上，因此由上述結果得知，MSIA探勘效能較佳，並且較不受高頻項目與稀少項目分配比率影響。

陸、結論

本研究所設計之MSIA主要為提供較大半高頻項目集之探勘演算法。然而以往關聯法則探勘技術中，大都採取由下而上的探勘程序，於探勘最大半高頻項目集過程中將造成過多候選項目集，並佔用大量記憶體空間，導致執行效率不佳，因此本論文針對上述問題進行改善，所設計之MSIA主要包含下列特色與優點。

1. 採用相對支持度(Relative Support)作為客觀之評估值，藉以篩選出較具意義之重要稀少性資料，亦即半高頻項目集。
2. 利用資料分群技術降低資料庫掃描讀取次數，MSIA於探勘過程中僅需掃描讀取資料庫一次即可完成探勘。
3. 採用由上而下之探勘程序，因此MSIA可快速取得最大半高頻項目集，並避免產生多餘重覆之探勘結果。

4. 透過分解與篩選機制，解決由上而下探勘程序容易產生過多不必要候選項目集的問題，進而提升探勘處理效率。

綜合上述MSIA所具備之特色優點，可知MSIA為一效率極佳之較大半高頻項目集探勘技術，未來將持續著重於候選項目集之篩選研究，以再提升本探勘技術之處理效率。

致謝

本論文承國科會專題研究計劃支援。計劃編號：NSC 98-2220-E-020-001和NSC 98-2220-E-020-002。

參考文獻

1. 林淑菁，民93，有效率尋找最大高頻項目組的方法，逢甲大學資訊工程學系碩士班碩士論。
2. 蔡玉娟、張簡雅文、黃彥文，2003，『快速反向關聯法則與調整緊密規則-促銷商品組合之應用』，資訊管理學報，第十卷．第一期：181～204頁。
3. 魏志平、董和昇，民89，『資料管理與分析』，*收錄於電子商務理論與實務*，梁定澎（編），華泰書局。
4. Agrawal, R., Imilienski, T., and Swami, A.“Mining Association Rules between Sets of Items in Large Databases,”*ACM SIGMOD International Conference on Management of Data*, 1993, pp. 207-216.
5. Agrawal, R. and Srikant, R.“Fast algorithms for mining association rules,”*VLDB Conference*, 1994.
6. Agrawal, R. and Srikant, R.“Mining Sequential Patterns,”*IEEE ICDE*, 1995, pp. 3-14.
7. Ali, K., Manganaris, S., and Srikant, R.“Partial Classification using Association Rules,”*3rd International Conference on Knowledge Discovery in Databases and Data Mining*, 1997, Newport Beach, California.
8. Chen, M.S., Han, J., and Yu, P.S.“Data Mining: An Overview from a Database Perspective,”*IEEE Transactions on Knowledge and Data Engineering* (8:6), 1996.
9. Denwattana, N. and Getta, J. R.“A Parameterised Algorithm for Mining Association Rules,”*Proceedings of the 12th Australasian Database Conference*, 2001, pp. 45-51.
10. Han, E.H., Karypis, G., and Kumar, V.“Scalable Parallel Data Mining for Association Rules,”*ACM SIGMOD*, 1997, pp. 277-288.
11. Kaufman, L. and Rousseeuw, P.J. *Finding Groups in Data: an Introduction to Cluster Analysis*, John Wiley & Sons, 1990.
12. Liu, B., Hsu, W., and Ma, Y.“Mining association rules with multiple minimum supports,”

5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, USA, 1999.

13. Lin, D. and Kedem, Z.M. "Pincer-Search: A New Algorithm for Discovering the Maximum Frequent Set," *VI International Conference on Extending Database Technology*, 1998.
14. Park, J.S., Chen, M.S., and Yu, P.S. "Using a Hash-Based Method with Transaction Trimming for Mining Association Rules," *IEEE Transaction on Knowledge and Data Engineering* (9:5), 1997.
15. Tsay, Y.J. and Chang-Chien, Y.W. "An efficient cluster and decomposition algorithm for mining association rules," *Information Sciences* (160), 2004, pp. 161–171.
16. Yun, H., Ha, D., Hwang, B., and Ryu, K.H. "Mining association rules on significant rare data using relative support," *Journal of Systems and Software* (67:3), 2003, pp.181-191.