從購買意願資料中挖掘高度相關性的關聯規則

翁政雄
中臺科技大學資訊管理系

摘要

關聯規則探勘技術是一項重要的資料挖掘技術，這項技術可以從交易資料庫中挖掘消費者購買行為之間的關聯性。現今的行銷策略皆視顧客為公司重要的獲利來源。因此，公司應該積極尋找潛在的顧客，並發展合適的行銷策略以吸引他們。為達上述目的，許多公司已經開始積極收集相關資料庫，並嘗試從這些資料庫中找出有意義的規則，藉以發展合適的行銷策略以吸引這些潛在顧客。本研究探討如何利用關聯規則分析消費者購買手機的決策考量因素，利用關聯規則之支持度與信心度分析消費者基本資料與手機產品特性之間的關聯性，以提供給行銷部門及產品設計部門分別作為行銷策略制定之參考與設計出更符合消費者的產品。然而，使用Σ-count方式累計過多具有低支持度的項目集時，卻容易產生不具關聯性的高頻項目集。因此，本研究發展新的方法嘗試從消費者的購買意願中挖掘有意義且有關聯性的規則。此方法乃是運用α-cut的概念過濾不具關聯性的低支持度項目，並且利用相關係數（*lift*）進一步強化現有挖掘關聯規則的基本機制（支持度-信心度），嘗試從消費者購買意願資料中找出有意義且相關的規則。實驗結果顯示本研究所提出的方法可以找出有價值且具有高度相關的關聯規則。

關鍵字：資料探勘、關聯規則、關聯分析

Mining association rules with high correlation from the purchasing intension data

Cheng-Hsiung Weng
Department of Management Information Systems,
Central Taiwan University of Science and Technology

Abstract

Association rule mining is an important data analysis method that can discover associations within data. This technique can mine the associations between the consumer's behaviors. The current marketing strategies perceive customers as important resources to a company for making more profit. Therefore, it is essential to companies to successfully discover potential customers and then develop new marketing strategies to attract them. To achieve these aims, many companies have gathered significant numbers of large databases to discover meaningful patterns and then develop new marketing strategies to attract the potential customers. However, using the Σ-count, the summation of a large number of itemsets with very small support may induce irrelevant associations. To this end, this study proposes a new approach to discover interesting and relevant patterns from consumer's purchasing intension. This approach is based on the α -cut method to filter out the irrelevant patterns with small support. Furthermore, a correlation measure, also known as lift, is used to augment the support-confidence framework for association rules. Next, we develop an algorithm to discover relevant and interesting association rules from purchasing intensions. Experimental results from the survey data show that the proposed approach can help to discover interesting and valuable patterns with high correlation.

Key words: data mining; association rule; corelation analysis

壹、緒論

關聯規則探勘技術是一項重要的資料挖掘技術，這項技術可以從交易資料庫中挖掘消費者的購買行為的關聯性（Berry and Linoff, 1997; Han and Kamber, 2001）。Agrawal et al.（1993）首先提出關聯規則的定義：所有的關聯規則必須符合兩項門檻值，分別是最小支持度（minimum support）及最小信心度（minimum confidence）。基本上，關聯規則探勘技術主要分成2個步驟：（1）以支持度為基準，找出大於最小支持度的高頻項目集，以及（2）以信心度為基準，從高頻項目集中，找出大於最小信心度的關聯規則。

現代企業在競爭激烈的環境中，除了必須收集大量的客戶、產品…等資料之外，還必須將資料做有效的分析與處理，以產生對企業有用的資訊，因此企業將收集來的資料儲存在資料庫中，以便直接從資料庫中找出有用的資訊。這些資訊可以幫助決策者做決策，加強與客戶之間的關係、鞏固客源，更可為企業提供行銷策略的重要依據（Kotler and Keller, 2001）。關聯規則的相關技術主要是用來找出資料庫中項目（item）之間的關聯性，而依據項目之間的關聯性，可建立關聯規則。由於關聯規則能描述項目不同種類之間的關聯性，因此，可以找出並顯示商品銷售的關聯性和客戶消費傾向等資訊，如：某類客戶在購買新手機時希望具備的加值服務為「遊戲功能」，則其新手機款式會選擇「一體成型」的信心度是73%。手機的製造公司可運用這些資訊在產品的設計上。除此之外，手機行銷部門也可以針對手機價格部分，推出更吸引消費者的促銷活動，以期能更有效地將商品賣出和增加公司的獲利。

傳統上，關聯規則探勘的資料種類可以區分為：類別型資料（categorical data）（Agrawal et al, 1993）、數值型資料（numerical data）（Hong et al, 1999）以及序數型資料（ordinal data）。其中，序數型資料可視為一種特殊的類別型資料，即為有順序性的類別型資料，Chen and Weng（2008）從不精確的序數資料挖掘關聯規則，其研究顯示：若是能考量序數資料間存在著相似度，將可以找出更多有意義的規則。然而，考慮項目之間的相似度固然可能產生更多有意義的高頻項目集，卻也同時產生相似度太低的項目集。根據Djouadi et al.（2007）的研究顯示：計算高頻項目集時，若是累計支持度過低的項目集，可能會產生不具關聯性的規則。

為解決上述的問題，本研究將修改Apriori演算法，考慮項目之間的相似度，並應用α-cut過濾支持度過低的項目集，期能挖掘出有意義且具有高度關聯性的關聯規則。第二章回顧關聯規則與其在行銷應用上相關的文獻，第三章說明問題定義。第四章為演算法設計，將詳細說明本研究所提出的α-Fuzzy-Apriori演算法，並且比較α-Fuzzy-Apriori演算法與傳統Apriori演算法的差異。第五章為實驗部份，本研究將以消費者購買行為為例，驗證本研究的可行性。第六章將本研究的成果加以探討，並說明結論與未來研究。

貳、文獻探討

本章節分為兩個部分進行討論，第2.1 節中討論關聯規則探勘的相關研究，第2.2節

則探討關聯規則探勘在行銷上應用的相關研究。

一、關聯規則

關聯規則是最常被使用來表示產品項目之間關聯性的方法之一。挖掘關聯規則的目的：在大量的交易資料中，找出不同項目之間的關聯性，從關聯規則所顯示出的消費傾向，對企業在從事行銷組合及市場預測等活動時，可提供非常有價值的資訊。由於Apriori演算法已經廣泛且成功應用在許多領域，因此許多改良的演算法先後經被提出（Agrawal et al, 1993; Lian et al, 2005; Tseng et al, 2008; Srikant et al., 1996）。

Agrawal and Srikant（1994）提出Apriori演算法，用以處理類別型資料（categorical data），例如：消費者所購買的物品。然而，在交易資料庫中的每一筆交易資料除了包含所購買的物品之外，也包含物品被購買的數量。物品與被購買的數量這一層關係對於行銷策略的決定也有重要的影響。處理數值型資料的方法可以區分成兩類：離散化分割（crisp partitions）以及模糊分割（fuzzy partitions）。

關於離散化分割（crisp partitions），Srikant et al.（1996）運用離散化分割（crisp partitions）的方式將連續的數值型資料進行離散化成數個區間（interval），每一個區間即是特定的類別型資料。因此，傳統處理類別型資料的演算法便可以繼續沿用。而模糊分割（fuzzy partitions）乃是利用模糊理論（fuzzy theory）（Zadeh, 1971）將連續的數值型資料進行離散化成各種語意變數（linguistic variables），此項技術已經被廣泛應用，例如：Hong et al.（1999）利用模糊分割（fuzzy partitions）的方法從數值型資料庫中挖掘模糊關聯規則。Hu et al.（2003） 則利用模糊分割的方法分別將類別型資料及數值型資料轉換成各種語意變數，進而挖掘有意義的模糊關聯規則。

除了類別型資料與數值型資料之外，另一種經常出現在交易資料庫的資料型態為序數型資料。序數型資料則是一種特殊的類別型資料。不同於類別型資料，序數型資料間彼此具有順序關係。例如：彩虹的顏色。關於顏色的相似程度，紅色與橙色的相似度明顯大於相對於紅色與藍色之間的相似度。因此，若能考量序數型資料屬性之間的相似度，將能挖掘出更多有意義的關聯規則。Chen and Weng（2008）從不精確的序數資料挖掘關聯規則，其研究顯示：若是能考量序數資料的相似程度，將可以找出更多有意義的規則。然而，考慮項目之間的相似度固然可能產生更多有意義的高頻項目集，卻也同時產生相似度太低的項目集。根據Djouadi et al.（2007）的研究顯示：計算高頻項目集時，若是累計支持度過低的項目集，可能會產生不具關聯性的規則。

二、關聯規則在行銷上的應用

現今的行銷策略深刻體認到顧客是公司最重要的資源，因此如何發掘新的顧客及留住高價值的舊顧客對公司而言非常重要。為了達到這項目的，許多公司收集大量的資料庫資料加以分析，期能發掘新的機會與商業策略。Shepard（1998）認為資料庫行銷是一種資訊導向（information driven）的行銷過程，其結合了資料庫技術來輔助關係行銷的實行，能夠讓行銷人員發展、測試、實行、衡量與適當修正顧客化的行銷策略。

為了提升某些產品的銷售和增加利潤，公司或組織會運用資料探勘技術，針對客戶的歷史資料庫進行分析與研究，以建立新客戶的選擇模式。利用關聯規則演算法對交易資料進行探勘，經過探勘所得到的資訊，我們便可發掘出一些潛在產品間的關聯規則，並可得知規則是否成立，如果成立我們就提供這些關聯規則給企業經營者制定決策時的參考，並藉由規則加入至知識庫中，增進顧客服務品質及減少銷售資源的浪費。

在零售市場中，了解顧客購買行為的改變，將有助於擬定更有效的促銷活動，Chen et al.（2005）整合顧客行為變數、人口統計變數以及交易式資料庫，用以挖掘顧客購買行為的改變。Lin and Hong（2008）整合顧客的各種行為變數，並利用關聯規則探勘技術，以偵測顧客購買行為的改變，進而建立有效的交互銷售（cross-selling）活動。除此之外，搭配銷售（bundle）是常見的促銷方式，其作法乃是將不同的產品搭配銷售，以改善其中一項產品的銷售量。然而，如何從眾多的搭配銷售組合中，找出最佳的方案非常困難。Yang and Lai（2006）運用關聯規則探勘技術從顧客的網頁瀏覽記錄中及顧客訂單中挖掘顧客線上購物行為，以找出瀏覽資料與顧客訂單之間的關聯性，進而協助行銷人員制定更佳的搭配銷售策略。 從上述的研究中，我們得知關聯規則探勘已經普遍應用於行銷領域，以協助行銷人員制定更佳的行銷策略。

從上述的研究中，關聯規則演算法所處理的資料種類可以區分為：類別型資料（Agrawal et al, 1993; Agrawal and Srikant, 1994）、數值型資料（Hong et al, 1999）以及序數型資料（Chen and Weng, 2008）。處理類別型資料時，只會有出現與否的情況（即支持度為1或0）。因此，不會有支持度過低的項目集。然而，運用模糊分割處理數值型資料時，則可能出現部份隸屬的情況（0<支持度<1）。因此，可能產生支持度過低的項目集。另外，運用相似度矩陣處理序數型資料時，也會部份隸屬的情況（0<支持度<1）。因此，可能產生支持度過低的項目集。

由於購買意願資料的資料型態包含：類別型資料、數值型資料以及序數型資料。Chen and Weng（2008）的研究顯示：若是能考量序數資料的相似程度，將可以找出更多有意義的規則。然而，（1）考慮項目之間的相似度固然可能產生更多有意義的高頻項目集，卻也同時產生相似度太低的項目集。（2）運用模糊分割處理數值型資料時，則可能產生相似度太低的項目集。為解決上述的問題，本研究將（1）考慮序數型項目之間的相似度，用以挖掘出更多有意義的項目集，並且（2）應用α-cut過濾支持度過低的項目集，以避免產生不具關聯性的規則。

參、問題定義

從消費者的問卷訪談中，我們可以得知許多有關消費者購買時的考量因素（或稱屬性），透過這些屬性的關聯性分析，我們可以了解消費傾向等資訊，善用這些重要的資訊，設計部門可以設計出更符合消費者的產品。另外，行銷部門也可以針對價格部分，推出更吸引消費者的促銷活動，以期能更有效地將商品賣出和增加公司的營業獲利。本章將詳細定義：如何從消費者的問卷資料中挖掘關聯規則的問題。首先，我們定義演算

法中所使用的項目（item），進而定義不同資料型態的資料如何計算支持度，包含：類別型資料、數值型資料及序數型資料。

定義1. 令$IT=\{it_1, it_2, \ldots, it_m\}$為所有項目的集合，$S=\{s_1, s_2, \ldots, s_m\}$為所有值的集合且$s$-item表示成$(it_i, s_i)$，其中，$1 \le i \le m$，$it_i \in IT$為第$i$個項目的名稱，而且$s_i \in S$為項目$it_i$的值。語意上，$(it_i, s_i)$用以表示消費者在某項購買屬性上的選擇。

範例1. 假設我們有一資料集合包含五筆資料，如表1所示。在此範例中，共有6種不同屬性，這些屬性可以區分為手機屬性與消費者屬性。其中，手機屬性包含現有手機的廠牌、新手機的廠牌、新手機的價位、選擇該廠牌手機最主要原因、新手機款式。而消費者資料屬性包含年級、性別。就屬性資料型別而言，現有手機的廠牌、新手機的廠牌、選擇該廠牌手機最主要原因、新手機款式屬性為類別型資料（categoriacal data），例如：（現有手機的廠牌, Sony）表示消費者現有手機的廠牌是Sony；（新手機款式, 一體成型）則表示消費者希望其新手機款式為一體成型；而價格屬性則屬於數值型資料（numerical data），例如：（新手機的價位, 4000）則表示消費者所希望的手機價格為4000元。除此之外，年級屬於序數型資料（ordinal data），例如：（年級, 大學部四年級）則表示消費者現在就讀大學部四年級。

表1：手機購買決策資料集合

No	現有手機的廠牌	新手機的廠牌	新手機的價位	選擇該廠牌手機最主要原因	新手機款式	年級
1	SONY	SONY	4000	款式	折疊	大學部四年級
2	SONY	SONY	1500	功能	滑蓋	大學部一年級
3	SONY	Nokia	5000	廠牌	一體成型	大學部二年級
4	Nokia	HTC	5000	功能	滑蓋	大學部三年級
5	SONY	SONY	5000	功能	一體成型	大學部三年級

定義2. 令$IC=\{ic_1, ic_2, \ldots, ic_n\}$為所有規則項目的集合，$F=\{f_1, f_2, \ldots, f_n\}$為所有規則項目值的集合，其中，$n \le m$，$ic_j \in IC$為第$j$個規則項目的名稱，而且$f_j$為規則項目$ic_j$的值，且關聯規則項目$r$-item表示成$(ic_j, f_j)$，$1 \le j \le n$。本研究中所定義的關聯規則項目$r$-item可以是類別型（categorical）規則項目及語意型（linguistic）規則項目。本研究將使用$b_j=(ic_j, f_j)$表示規則項目r-item。規則項目集r-itemset B 則表示規則項目r-item的集合，其中所有的規則項目r-item必須有不同的項目名稱。本研究使用$B=\{(ic_1, f_1), (ic_2, f_2), \ldots, (ic_n, f_n)\}$表示規則項目集$r$-itemset，$n \le m$。

範例2. 例如：{(現有手機的廠牌, SONY), (新手機的價位, *high*)}是一個規則項目集r-itemset.

定義3. 給定序數o_i 與序數o_j，令$sim\ (o_i, o_j)$表示序數o_i與序數o_j的相似度。本研究運用相似度矩陣Sim表示序數型態項目之間的相似度。

範例3. 年級的相似度矩陣Sim，如表2所示。從相似度矩陣Sim，我們知道sim(大一, 大一)=1.00, sim(大一, 大二)=0.30 以及 sim(大一, 大三)=0.10.

表2：年級相似度矩陣Sim

	大一	大二	大三	大四
大一	1.00	0.30	0.10	0.00
大二	0.30	1.00	0.30	0.10
大三	0.10	0.30	1.00	0.30
大四	0.00	0.10	0.30	1.00

定義4. 假設有一序數型資料項目s-item $a_i=(it_i, s_i)$以及一序數型規則項目r-item $b_j=(ic_j, f_j)$。令$sup\ (a_i, b_j)$表示資料項目a_i與規則項目b_j的支持度，其定義如下：

$$sup(a_i,b_j)=\begin{cases} sim(s_i,f_j) & , \text{ if } it_i = ic_j \text{ and } s_i \in \text{ordinal data} \\ 0 & , \text{ otherwise} \end{cases}.$$

如果類別型資料項目之間不存在相似度，則上述公式仍然適用，因為只有完全相似的項目其相似度才是100%，否則，其相似度為0%。

範例4. 假設有年級相似度矩陣如表2所示，並且有一序數型資料項目s-item $a_1=$(年級, 大一)以及一序數型規則項目r-item $b_1=$(年級, 大二)。則支持度$sup\ (a_1, b_1)=sim$(大一, 大二)$=0.30$。

語意變數（Linguistic Variable）是以自然語詞中的語詞為值，例如：可以用詞組（低、中、高）來表達評估者對評估值價值程度的感受。語意變數的概念可以適當的表達這些主觀的判斷，用於處理不明確或模糊的資訊（Zadeh, 1971）。本研究利用三角模糊數來表示語意變數的隸屬函數。

定義5. 一個模糊集合F可以透過隸屬函數$m_F\ (x)$加以表示x，而其隸屬程度則介於區間值 [0, 1]。

範例5. 假設有三種語意變數（低、中、高）表示價格的高低，分別由三個隸屬函數為：P_{low}, P_{middle}及P_{high}等加以表示。從這三個隸屬函數，我們得知：$P_{low}(1500)=0.75$, $P_{low}(2000)=0.50$, $P_{middle}(4000)=0.50$及$P_{high}(5000)=1.00$.

$$P_{low}(p)=\begin{cases} 1 & , \text{ if } p \le 1000 \\ \dfrac{3000-p}{3000-1000} & , \text{ if } 1000 \le p \le 3000 \end{cases}; \quad (1)$$

$$P_{middle}(p)=\begin{cases} \dfrac{p-1000}{3000-1000} & , \text{ if } 1000 \le p \le 3000 \\ 1 & , \text{ if } p = 3000 \\ \dfrac{5000-p}{5000-3000} & , \text{ if } 3000 \le p \le 5000 \end{cases}; \quad (2)$$

$$P_{high}(p)=\begin{cases} \dfrac{p-3000}{5000-3000} & , \text{ if } 3000 \le p \le 5000 \\ 1 & , \text{ if } p \ge 5000 \end{cases}. \quad (3)$$

定義6. 假設我們有一數值型資料項目s-item表示為$a_i=(it_i, s_i)$、一語意型規則項目r-item 表示為$b_j=(ic_j, f_j)$ 以及隸屬函數(FS_{f_j})，其中隸屬函數$FS_{f_j}(s_i)$用以表示數值s_i隸屬於f_j的程度。因此，項目a_i相對於規則項目b_j的支持度的計算如下所示：

$$sup(a_i, b_j) = \begin{cases} FS_{f_j}(s_i) & , \text{ if } it_i = ic_j \text{ and } s_i \in \text{numerical data} \\ 0 & , \text{ otherwise} \end{cases}.$$

範例6. 假設我們有一數值型資料項目s-item $a_2=(it_2, 5000)$、一語意型規則項目r-item $b_2=(ic_2, high)$以及隸屬函數(P_{high})，如範例5所示。則支持度$sup\ (a_2, b_2)=sup((it_2, 5000), (ic_2, high))=1.00$。從上述的範例中，我們運用隸屬函數將數值型的資料轉換成語意變數（屬於種類型資料的一種），進而計算手機價格5000元，其價格屬於$high$的支持程度為1.00。

定義7. α-cut係將模糊集合轉為明確集合的工具，在模糊集合論中佔有相當重要的地位（Bodjanova, 2002），茲定義α-cut如下：

對模糊集合A而言，若給定一實數值α，$\alpha \in (0, 1]$，則對模糊集合A取α-cut所形成的明確集合$A_\alpha=\{x|\mu(x) \geq \alpha\}$，區間範圍$[A_l^\alpha, A_u^\alpha]$。其中，我們稱$\alpha$為門檻值（threshold value），當$\alpha$值越大，表示門檻值越高，則其所對應的區間值越小；同理，當α值越小，表示門檻值越低，則其所對應的區間值越大。當$\alpha=1$時，即成為單一的實數值。

然而，根據Djouadi et al.（2007）的研究顯示：計算高頻項目集時，若是累計太多支持度過低的項目集，可能會產生不具關聯性的規則。因此，本研究將運用α-cut的概念，將支持度偏低的項目集過濾掉，以免產生不具關聯性的規則。因此，運用α-cut的支持度定義如下所示：

$$sup^\alpha(a_i, b_j) = \{sup(a_i, b_j) \geq \alpha\}, \alpha \in (0,1]$$

範例7. 假設我們有一數值型資料項目s-item $a_2=(it_2, 1500)$、一語意型規則項目r-item $b_2=(ic_2, middle)$以及隸屬函數(P_{middle})，如範例5所示。則支持度$sup(a_2, b_2)=sup((it_2, 1500), (ic^2, middle))=0.25$。然而，從上述的範例中我們得知0.25的支持度屬於偏低的支持度。因此，若給定α-cut的門檻值$\alpha=0.30$，則$sup(a_2, b_2)=0.25<0.30$將被過濾掉。

定義8. 給定使用者定義的α-cut門檻值α，假設有一資料項目集s-itemset $A=\{(it_1, s_1), (it_2, s_2), \cdots, (it_m, s_m)\}$以及規則項目集 r-itemset $B=\{(ic_1, f_1), (ic_2, f_2), \cdots, (ic_n, f_n)\}$（$n \leq m$）. 分別可以找到$i_1, i_2, \cdots, i_n$使得$a_{i_j}$對應$b_j$，令支持度$sup^\alpha(A, B)$表示$A$對應到$B$的程度，其中所有的子項目集的支持度皆大於或等於門檻值α，則$sup^\alpha(A, B)$的定義如下：

$$sup^\alpha(A, B) = \text{Min}_{j=1}^{n} sup^\alpha(a_{i_j}, b_j).$$

範例8. 給定使用者定義的α-cut門檻值$\alpha=0.3$，假設有一資料項目集s-itemset $A=\{$(新手機的價位, 5000), (年級, 大學部三年級)$\}$、規則項目集r-itemset $B=\{$(新手機的價位, $high$), (年級, 大學部三年級)$\}$以及隸屬函數(P_{high})。則支持度$sup^{0.3}(A, B)=min(1.00, 1.00)=1.00$。

定義9. 給定使用者定義的α-cut門檻值(α)，假設有一資料庫DB，令資料項目集s-itemset A_i 為資料庫中的第i筆資料，表示成$A_i=\{(it_1, s_1), (it_2, s_2), \cdots, (it_m, s_m)\}$，其中項目

$a_i=(it_i, s_i)$可以是類別型資料項目s-item或是數值型資料項目s-item。假設有一規則項目集r-itemset $B=\{(ic_1, f_1), (ic_2, f_2), \cdots, (ic_n, f_n)\}(n \le m)$，其中規則項目$b_j=(ic_j, f_j)$是規則項目。則$DB$資料庫中，規則項目集$r$-itemset B的α-cut支持度可表示成$sup^{\alpha}_{DB}(B)$，其定義如下所示：

$$sup^{\alpha}_{DB}(B) = (\sum_{sid \subseteq DB} sup^{\alpha}(A_{sid}, B)) / |DB|,$$

其中，$|DB|$表示資料庫中的資料筆數

範例9. 給定使用者定義的α-cut的門檻值$\alpha=0.3$，假設有一資料庫（如表1所示）、有一規則項目集r-itemset $B=\{$(新手機的價位, *high*), (年級, 大學部三年級)$\}$。第2筆資料{(新手機的價位, 1500), (年級, 大學部一年級)}，因為*sup* ((年級, 大學部一年級), (年級, 大學部三年級))$=0.1 <$ 門檻值$\alpha(0.3)$，故此筆資料將過濾掉。換言之，因為低於門檻值$\alpha(0.3)$的資料支持度可視為0。規則項目集r-itemset B的支持度$sup^{\alpha}_{DB}(B)$=(min (0.50, 0.30) + min (0, 0) + min (1.00, 0.30) + min (1.00, 1.00) + min (1.00, 1.00) /5=(0.30+0+0.30+1+1) /5=2.6/5=0.52。

表3：資料項目集的支持度

TID	sup^{α}_{DB}		
	新手機的價位（high）	年級（大學部三年級）	手機的價位（high）∪年級（大學部三年級）
1	0.50	0.30	0.30
2		0（~~0.10~~）	
3	1.00	0.30	0.30
4	1.00	1.00	1.00
5	1.00	1.00	1.00
AVG	0.70	0.52	0.52

傳統上，支持度與信心度為衡量關聯規則是否成立的基本機制。然而 Chen et al.（1996）認為上述兩項基本門檻仍不足以過濾過無意義的關聯規則。為了解決上述問題，統計學上的相關係數（*lift*）概念用以衡量項目之間的相關程度，已經被用來強化以支持度及信心度為基礎的架構（International Business Machines, 1996）。既然，挖掘高度相關的項目集是本研究的目標，因此本研究亦使用相關係數（*lift*）篩選出高度相關的項目集，進而產生有意義且具高度相關的關聯規則。

定義10. 給定4個使用者定義的門檻值，分別為α-cut(α)、支持度(σ_{sup})、信心度(σ_{conf})以及相關係數(σ_{lift})，倘若一規則項目集r-itemset B 的支持度$sup^{\alpha}_{DB}(B)$不小於支持度門檻值(σ_{sup})，則規則項目集r-itemset B是一個高頻項目集，其中$B=X \cup Y$ 且 $X \cap Y=\phi$。規則$X \Rightarrow Y$的信心度表示成*conf* ($X \Rightarrow Y$)，其定義為$sup^{\alpha}_{DB}(B) / sup^{\alpha}_{DB}(X)$。倘若規則$X \Rightarrow Y$的信心度不小於信心度門檻值($\sigma_{conf}$)，則關聯規則$X \Rightarrow Y$成立。除此之外，規則$X \Rightarrow Y$的相關係數表示成*lift* ($X \Rightarrow Y$)，其定義為*conf* ($X \Rightarrow Y$) / $sup^{\alpha}_{DB}(Y)$。倘若規則$X \Rightarrow Y$的相關係數不小於相關係數門檻值(σ_{lift})，則關聯規則$X \Rightarrow Y$的項目集之間具有高度關聯性。

範例10. 假設有一資料庫DB（如表1所示）以及4個使用者定義的門檻值，分別為α

-cut(α=0.3)、支持度(σ_{sup}=0.3)、信心度(σ_{conf}=70%)以及相關係數(σ_{lift}=100%)，令規則項目集 r-itemset X={(年級, 大學部三年級)}，Y={(新手機的價位, *high*)}且$B=X\cup Y$。從表3我們得知$sup^{\alpha}_{DB}(B)$=0.52且$sup^{\alpha}_{DB}(X)$=0.52皆不小於支持度門檻值(σ_{sup}=0.3)。因此，規則項目集 B 和 X 都是高頻項目集，關聯規則$X\Rightarrow Y$的信心度$conf\ (X\Rightarrow Y)$=0.52/0.52=100%不小於信心度門檻值(σ_{conf})，則關聯規則$X\Rightarrow Y$成立。其規則的意義為：假如消費者是大學部三年級學生，則其買新手機的價位為*high*。除此之外，此規則的相關係數為$conf\ (X\Rightarrow Y)/sup^{\alpha}_{DB}(Y)$=142.86%。

肆、演算法設計

本章節將詳細說明本研究所提出的α-Fuzzy-Apriori演算法，並且比較α-Fuzzy-Apriori演算法與傳統Apriori演算法的差異。由於α-Fuzzy-Apriori演算法乃是以Apriori演算法為基礎所發展出來，其基本概念仍然遵循Apriori演算法的精神。因此，4.1節將先介紹Apriori演算法，4.2節則說明α-Fuzzy-Apriori演算法。最後4.3節舉例說明α-Fuzzy-Apriori演算法的計算過程。

一、Apriori演算法的基本概念

Apriori 演算法為主的關聯規則探勘步驟主要分成兩大步驟，分別為：(1)找高頻項目集，以及(2)產生關聯規則。以下說明Apriori 演算法擷取高頻k-項目集(k>1)並找出關聯規則的步驟：

第一步驟：找高頻項目集

（1）找出高頻項目集$_{k-1}$，若為∅，則停止執行；

（2）由(1)中找出任兩個有(k-2)項目相同的項目集$_{k-1}$，組合成項目集$_k$；

（3）判斷由(2)所找出的項目集$_k$，其所有包括的項目集$_{k-1}$之子集合是否都出現在(1)中，假如成立就保留此項目集$_k$；否則就刪除。

（4）再檢查由(3)所擷取的項目集$_k$是否滿足最小支持度，假如符合就成為高頻項目集$_k$；否則就刪除。

（5）跳至(1)繼續找高頻項目集$_{k+1}$，直到無法產生高頻項目集為止。

第二步驟：產生關聯規則

（1）將所有高頻k-項目集(k>1)拆解成$X\rightarrow Y$，$X,Y\in I$且$X\cap Y=\varnothing$。

（2）判斷所有的規則是否符合最小信心度，若符合則成為關聯規則。

傳統的Apriori演算法僅適用於類別型的資料，而且其計算支持度的方式僅考慮出現與否，項目集有出現，其隸屬程度為1，故次數加1。換言之，不允許出現部份隸屬的情況。然而，現實世界的交易資料通常包含序數型資料（如一年級、二年級等）以及數值型資料（如1000）。除此之外，由於資料間可能存在相似度，因而有部份相似（部份隸屬）的情況。此時，支持度將不再0或是1等整數值，而是0.5或0.7等小數的情況。因此，

傳統的Apriori演算法不適用於數值型資料以及部份隸屬的情況。為了解決上述問題，本研究利用模糊理論中的隸屬函數將數值型資料轉換成語意變數，並改良傳統傳統的Apriori演算法的缺點。

二、α-Fuzzy-Apriori演算法

本章節將介紹α-Fuzzy-Apriori演算法的運作方式，在介紹之前，我們首先比較Apriori演算法與α-Fuzzy-Apriori演算法的差異。兩者的差異歸納如下：

（1）資料型態：傳統的Apriori演算法僅適用於類別型的資料，而α-Fuzzy-Apriori演算法則可以處理類別型的資料以及數值型的資料。

（2）隸屬函數：Apriori演算法計算項目之間的相似度（支持度）時，僅允許100%或0%相同的情況。換言之，不允許部份相似（部份隸屬）的情況，例如：80%相似。為了解決上述問題，α-Fuzzy-Apriori演算法則允許部份相似（部份隸屬）的情況。

（3）α-cut門檻值：α-Fuzzy-Apriori演算法僅考慮支持度皆大於或等於門檻值α的項目集。計算高頻項目集時，過濾支持度過低的項目集，以避免產生不具關聯性的規則。

（4）計算候選項目集：傳統的Apriori演算法僅適用於類別型的資料，而且其計算支持度的方式僅考慮出現與否，項目集有出現，其隸屬程度為100%，故次數加1。換言之，不允許出現部份隸屬的情況。考量資料間可能存在相似度，因而有部份相似（部份隸屬）的情況，因此計算候選項目集的支持度時，α-Fuzzy-Apriori演算法則允許0或1，甚至0.5或0.7等小數的情況。

（5）相關係數(*lift*)：α-Fuzzy-Apriori演算法應用統計學上的相關係數(*lift*)概念用以衡量項目之間的相關程度，用以強化支持度及信心度為基礎的Apriori演算法架構。

如圖1所示，α-Fuzzy-Apriori演算法分成3大步驟，分別為：（1）應用隸屬函數將原始資料轉換成新的資料庫，其中每一筆資料都將轉換成（項目集, 支持度）的格式，如此將有助於累計每一項目集的支持度，並且運用α-cut門檻值(α)過濾支持度過低的項目集，以避免產生不具關聯性的規則。（2）利用反覆的方式，逐一統計項目集的支持度，進而找出高頻項目集。其主要概念為：首先產生項目集長度為k的候選項目集C_k^α，在篩選出支持度不小於支持度門檻值(σ_{sup})的高頻項目集L_k^α，再利用高頻項目集L_k^α，合併產生項目集長度為($k+1$)的候選項目集C_{k+1}^α，關於產生候選項目集的子程式，請參考附錄A。除此之外，α-Fuzzy-Apriori演算法應用統計學上的相關係數概念，用以衡量項目之間的相關程度，並以門檻值(*lift*)為衡量依據，僅保留具有高度關聯性的高頻項目集。（3）利用所找到的高度關聯性的高頻項目集L_k^α產生關聯規則，規則是否成立則以信心度門檻值(σ_{conf})為衡量依據。

由上述的說明中得知：α-Fuzzy-Apriori演算法仍是以Apriori演算法為基礎，利用反覆

的方式，逐一統計項目集的支持度，進而使用最小支持度做為判斷是否成為高頻項目集的標準，倘若某項目集的支持度不小於最小支持度，則該項目集即為高頻項目集。不同的是：運用α-cut門檻值(α)過濾支持度過低的項目集，以避免產生不具關聯性的規則。因此，α-Fuzzy-Apriori演算法所產生的高頻項目集數量將少於Apriori演算法的高頻項目集數量，因為支持度過低的項目集將被過濾掉。由於α-Fuzzy-Apriori演算法仍是利用反覆的方式，逐一統計項目集的支持度，進而使用最小支持度做為判斷是否成為高頻項目集的標準，故α-Fuzzy-Apriori演算法所產生的高頻項目集仍具有反單調性，即高頻項目集的子集合一定是高頻項目集。

Input: A database, D_B; membership functions (FS_{f_j}); a predefined α-cut threshold α; a predefined minimum support σ_{sup}; a predefined minimum confidence σ_{conf}; a predefined minimum *lift* σ_{lift}.
Output: A set of fuzzy association rules
Method:
// **Phase 1 Call the** *Sup_Transform* Subroutine
(1). For each transaction
Transform each *s*-item data into *r*-items;
Remove the itemsets with support smaller than the α-cut threshold;
Store these results as a new transaction in a new database D_T^{α}.
// **Phase 2 Call the** *Itemsets_gen* Subroutine
(1). For each *r*-item $ic_{j,r}^{\alpha}$, calculate its support.
(2). Check whether the support of each *r*-item $ic_{j,r}^{\alpha}$ is no less than the minimum support σ_{sup}. If it is, put it into the set of frequent one-itemsets (L_1^{α}).
(3). Generate candidate set C_{k+1}^{α} from L_k^{α}.
(4). Compute the supports of all *r*-itemsets in C_{k+1}^{α} and determine L_{k+1}^{α}.
(5). Compute the *lift*s of all *r*-itemsets in L_{k+1}^{α} and determine releavant L_{k+1}^{α}.
(6). If L_{k+1} is null, go to phase 3; otherwise, set $k = k + 1$ and repeat steps (3)–(5).
// **Phase 3 Call the** *FAR_gen* Subroutine
(1). Generate fuzzy association rules from all *r*-itemsets with the two thresholds σ_{conf} and σ_{lift}.

圖1：The α-Fuzzy-Apriori Algorithm

三、範例

本章節將使用前一章節的表1資料為例，說明本研究所提出α-Fuzzy-Apriori 演算法的運作過程，其中包含下列3個主要的步驟：（1）資料轉換。（2）尋找高頻項目集。（3）產生關聯規則。詳細描述如下：

步驟1. 根據第三章節所定義的支持度計算方式，分別計算出類別型資料項目以及數值型資料項目所對應的規則項目集及其支持度。其中，α-cut門檻值設為0.3，而所需的3個隸屬函數(FS_{fj}) 如前一章節所示。其計算後的資料儲存於暫時資料庫D^T中，如表4所示。

表4：暫時資料庫D^T

No	現有手機的廠牌	新手機的廠牌	新手機的價位	選擇該廠牌手機最主要原因	新手機款式	年級
1	（SONY, 1.0）	（SONY, 1.0）	（middle, 0.5） （high, 0.5）	（款式, 1.0）	（折疊, 1.0）	（大學部三年級, 0.3） （大學部四年級, 1.0）
2	（SONY, 1.0）	（SONY, 1.0）	（low, 0.75） ~~（middle, 0.25）~~	（功能, 1.0）	（滑蓋, 1.0）	（大學部一年級, 1.0） （大學部二年級, 0.3） ~~（大學部三年級, 0.1）~~
3	（SONY, 1.0）	（Nokia, 1.0）	（high, 1.0）	（廠牌, 1.0）	（一體成型, 1.0）	（大學部一年級, 0.3） （大學部二年級, 1.0） （大學部三年級, 0.3） ~~（大學部四年級, 0.1）~~
4	（Nokia, 1.0）	（HTC, 1.0）	（high, 1.0）	（功能, 1.0）	（滑蓋, 1.0）	~~（大學部一年級, 0.1）~~ （大學部二年級, 0.3） （大學部三年級, 1.0） （大學部四年級, 0.3）
5	（SONY, 1.0）	（SONY, 1.0）	（high, 1.0）	（功能, 1.0）	（一體成型, 1.0）	~~（大學部一年級, 0.1）~~ （大學部二年級, 0.3） （大學部三年級, 1.0） （大學部四年級, 0.3）

步驟2.1 針對暫時資料庫D^T中的每一規則項目r-item計算其支持度，並且判斷是否不小於支持度門檻值($\sigma_{sup}=0.4$)，若是，則存入高頻項目集L_1，如表5所示。

表5：L_1高頻項目集

項目集	支持度
現有手機的廠牌.SONY	0.80
買新手機的廠牌.SONY	0.60
新手機的價位.high	0.70
選擇該廠牌手機最大原因.功能	0.60
新手機的款式.一體成型	0.40
新手機的款式.滑蓋	0.40
年級.大學部三年級	0.52

步驟2.2. 運用組合（join）的方式，從L_1高頻項目集產生候選項目集C_2，例如：（現有手機的廠牌.SONY, 買新手機的廠牌.SONY）,（現有手機的廠牌.SONY, 新手機的價位.high）…（新手機的款式.滑蓋, 年級.大學部三年級）。計算C_2中每個候選項目集的支持度，則可以進一步判斷項目集長度為2的高頻項目集L_2，如表6所示。

表6：L_2高頻項目集

No	項目集	支持度
1	（現有手機的廠牌.SONY, 買新手機的廠牌.SONY）	0.60
2	（現有手機的廠牌.SONY, 新手機的價位.high）	0.50
3	（現有手機的廠牌.SONY, 選擇該廠牌手機最大原因.功能）	0.40
4	（現有手機的廠牌.SONY, 新手機的款式.一體成型）	0.40
5	（買新手機的廠牌.SONY, 選擇該廠牌手機最大原因.功能）	0.40
6	（新手機的價位.high, 選擇該廠牌手機最大原因.功能）	0.40
7	（新手機的價位.high, 新手機的款式.一體成型）	0.40
8	（新手機的價位.high, 年級.大學部三年級）	0.52
9	（選擇該廠牌手機最大原因.功能, 新手機的款式.滑蓋）	0.40
10	（選擇該廠牌手機最大原因.功能, 年級.大學部三年級）	0.40

步驟2.3. 既然高頻項目集L_2不是空集合，則重複進行步驟2.2的組合（join）的程序，進而產生項目集長度為3的候選項目集C_3。既然高頻項目集L_3不是空集合，則重複進行步驟2.2的組合（join）的程序，進而產生項目集長度為4的候選項目集C_4。然而，計算所有C_4中每個候選項目集的支持度之後，發現並不存在項目集長度為4的高頻項目集L_4，則停止步驟2.2的組合（join）的程序。

以下將說明如何由L_k產生C_{k+1}的詳細過程：在組合（join）的過程中，必須先判斷2個L_k高頻項目集的前$(k-1)$個項目集是否相同，若是，則由相同的前$(k-1)$個項目集組合（join）第1個L_k高頻項目集的第k個項目與第2個L_k高頻項目集的第k個項目組合（join）成長度為$(k+1)$個候選項目集。以表6為例，有2個長度為2高頻項目集(L_2)分別為（現有手機的廠牌.SONY, 買新手機的廠牌.SONY）以及（現有手機的廠牌.SONY, 新手機的價位.high）。由於2個高頻項目集(L_2)的前1(2-1＝1)個項目集皆為“現有手機的廠牌.SONY”，則由“現有手機的廠牌.SONY”組合（join）“買新手機的廠牌.SONY”及“新手機的價位.high”組合成C_3（現有手機的廠牌.SONY, 買新手機的廠牌.SONY, 新手機的價位.high）。經計算候選項目集C_3的支持度之後所得到高頻項目集L_3，如表7所示。

表7：L_3高頻項目集

No	項目集	支持度
1	（現有手機的廠牌.SONY, 買新手機的廠牌.SONY, 選擇該廠牌手機最大原因.功能）	0.40
2	（現有手機的廠牌.SONY, 新手機的價位.high, 新手機的款式.一體成型）	0.40
3	（新手機的價位.high, 選擇該廠牌手機最大原因.功能, 年級.大學部三年級）	0.40

步驟3. 從所有高頻項目集中產生關聯規則。在信心度$(\sigma_{conf}=100\%)$以及相關係數$(\sigma_{lift}=100\%)$的前提下，由表7中的L_3高頻項目集所產生的規則，如表8所示。

表8：從高頻項目集（L_3）產生的關聯規則

No	關聯規則
1	若（現有手機的廠牌.SONY且 選擇該廠牌手機最大原因.功能），則（買新手機的廠牌.SONY）；（信心度＝100%；相關係數＝166.67%）
2	若（買新手機的廠牌.SONY且 選擇該廠牌手機最大原因.功能），則（現有手機的廠牌.SONY）；（信心度＝100%；相關係數＝125.00%）
3	若（現有手機的廠牌.SONY且 新手機的款式.一體成型），則（新手機的價位.high）；（信心度＝100%；相關係數＝142.86%）
4	若（新手機的價位.high 且 新手機的款式.一體成型），則（現有手機的廠牌.SONY）；（信心度＝100%；相關係數＝125.00%）
5	若（新手機的價位.high且 選擇該廠牌手機最大原因.功能），則（年級.大學部三年級）；（信心度＝100%；相關係數＝ 192.31%）
6	若（選擇該廠牌手機最大原因.功能 且 年級.大學部三年級），則（新手機的價位.high）；（信心度＝100%；相關係數＝142.86%）
7	若（新手機的款式.一體成型），則（現有手機的廠牌.SONY 且新手機的價位.high）；（信心度＝100%；相關係數＝200.00%）

伍、實驗結果

本研究利用數項實驗數據衡量α-Fuzzy-Apriori演算法的成效。為了快速收集資料，本研究運用網頁問卷收集中部某大學學生的手機購買意願。其中，關於手機購買意願調查的問項，包括：（1）消費者基本資料：性別、大學年級、現有手機廠牌、最喜歡的休閒活動、平均多久換新手機；（2）新手機產品特性資料：廠牌、價格、最需要的功能、加值服務、款式、廣告訊息來源、選擇該廠牌的原因、購買地點以及選擇該購買地點的原因等因素。總共收集到531筆資料，剔除無效資料後，剩下473筆有效資料。本研究使用Sun Java language（J2SDK 1.3.1）開發α-Fuzzy-Apriori演算法，並利用筆記型電腦進行實驗，電腦配備如下：CPU為Intel Centrino 1400 MHz processor、記憶體有512MB及使用Windows XP 作業系統。

為了驗證本研究的可行性所進行的實驗，包含：（1）測試演算法在不同支持度情況下的執行時間；（2）統計演算法在不同支持度情況下的高頻項目集個數；（3）測試在不同α-cut門檻值情況下的執行時間，以及（4）統計演算法在不同α-cut門檻值情況下的高頻項目集個數。為了專注α-Fuzzy-Apriori演算法的設計與測試，本研究所使用的隸屬函數以及相似度矩陣，將聘請專業人士指定。

在第一個實驗中，將測試α-Fuzzy-Apriori演算法在固定的α-cut門檻值（0.3），但是不同支持度情況下的執行時間，本研究所使用的資料樣本大小為473筆有效資料。如圖2所示，執行時間隨著支持度增加而減少，當支持度越小時執行時間大幅增加，因為支持度越小將產生大量的高頻項目集。這樣的實驗結果與先前的Fuzzy-Apriori演算法相同（Hong et al., 1999）。但是，α-Fuzzy-Apriori演算法執行時間優於Fuzzy-Apriori演算法，其主要原因為：α-Fuzzy-Apriori演算法已經運用α-cut門檻值(α)過濾支持度過低的項目集，故在項目集個數較少的情況下，α-Fuzzy-Apriori演算法的執行時間將更短，這是非常合理的。在第二個實驗中，將統計α-Fuzzy-Apriori演算法在固定的α-cut門檻值，但是

不同支持度情況下的高頻項目集（L_1, L_2, L_3）個數。本研究所產生的高頻項目集（L_1, L_2, L_3），如表9所示，當支持度越小時將產生越多的高頻項目集，此結果與傳統的Apriori演算法相同（Agrawal and Srikant, 1994; Hong et al., 1999）。

從第一個實驗與第二個實驗的數據中，得知本研究所提出α-Fuzzy-Apriori演算法正確無誤，後續的實驗中，將進一步探討不同α-cut門檻值對演算法執行時間與高頻項目集個數的影響。

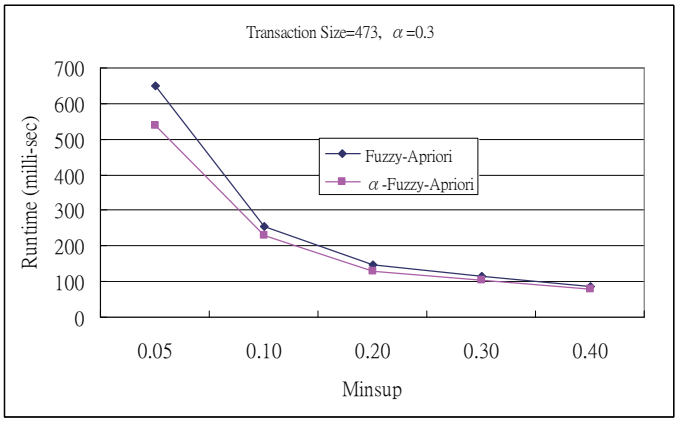

圖2：執行時間 vs. 最小支持度

表9：項目集個數 vs. 最小支持度

Min-sup	0.10	0.15	0.20	0.25	0.30
L1	42	34	30	24	21
L2	313	187	100	43	19
L3	520	102	14	2	0
Total	875	323	144	69	40

在第三個實驗中，將測試α-Fuzzy-Apriori演算法在固定支持度，但是不同α-cut門檻值情況下的執行時間，本研究所使用的資料樣本大小依舊為473筆有效資料。如圖3所示，執行時間隨著α-cut門檻值增加而減少。換言之，當α-cut門檻值越小時執行時間增加，因為α-cut門檻值越小將累計更多低支持度的項目集，而導致產生更多的高頻項目集。除此之外，倘若α-Fuzzy-Apriori演算法將α-cut門檻值設為0，即不運用α-cut門檻值(α)過濾支持度過低的項目集，則α-Fuzzy-Apriori演算法的執行時間，將與先前的Fuzzy-Apriori演算法相同（Hong et al., 1999）。在第四個實驗中，將統計α-Fuzzy-Apriori演算法在固定支持度，但是不同α-cut門檻值情況下的高頻項目集（L_1, L_2, L_3）個數。本研究所產生的高頻項目集（L_1, L_2, L_3），如表10所示，當α-cut門檻值越小時將產生越多的高頻項目集，其原因與第三個實驗相同。除此之外，倘若α-Fuzzy-Apriori演算法將α-cut門檻值設為0，即不運用α-cut門檻值(α)過濾支持度過低的項目集，則α-Fuzzy-Apriori演算法所產生的高頻項目集（L_1, L_2, L_3）數量為154個，將與先前的Fuzzy-Apriori演算法相同（Hong et al., 1999）。

從第三個實驗與第四個實驗的數據中，我們得知α-cut門檻值的確會對α-Fuzzy-Apriori演算法產生影響，而這樣的研究結果也印證Djouadi et al.（2007）的研究。從上述的討論中，得知本研究運用α-cut的概念的確可以將支持度偏低的項目集過濾掉，進而只保留有意義且具高度關聯性的規則。

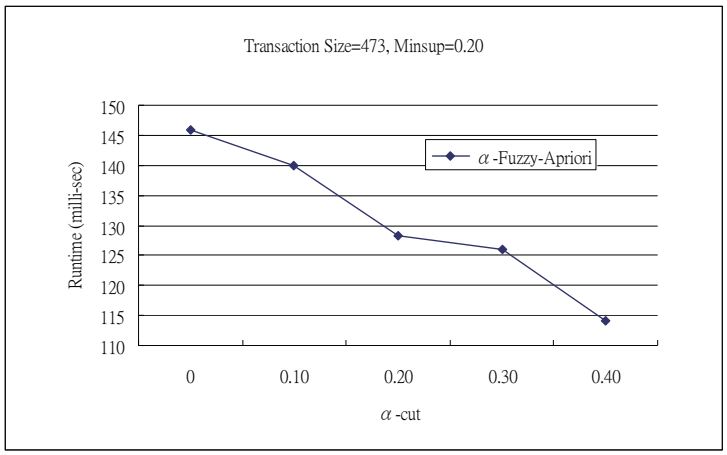

圖3：執行時間 vs. α-cut

表10：項目集個數 vs. α-cut

α-cut	0.0	0.1	0.2	0.3	0.4
L1	30	30	30	30	29
L2	109	109	100	100	81
L3	15	15	14	14	12
Total	154	154	144	144	122

最後，本研究運用4個使用者定義的門檻值，分別為α-cut(α=0.3)、支持度(α_{sup}=0.25)、信心度(α_{conf}=65%)以及相關係數(α_{lift}=100%)所找出的關聯規則，如表11及表12所示。

表11：從L_2所找出的關聯規則

No	關聯規則	支持度	信心度	相關係數
1	買新手機的廠牌.SONY⇒現有手機的廠牌.SONY	31.71%	69.44%	138.60%
2	新手機的價位.high⇒新手機的款式.一體成型	42.18%	74.03%	107.41%
3	選擇該廠牌手機最大原因.功能⇒新手機的款式.一體成型	36.58%	73.31%	106.36%
4	買新手機的地點.通訊行⇒新手機的款式.一體成型	28.96%	70.26%	101.94%
5	新手機具備哪種加值服務.遊戲⇒新手機的款式.一體成型	36.58%	73.62%	106.81%
6	廣告訊息來源.網路⇒新手機的款式.一體成型	38.90%	69.96%	101.51%
7	年級.大學部三年級⇒新手機的款式.一體成型	34.52%	68.99%	100.10%
8	年級.大學部四年級⇒新手機的價位.high	26.28%	66.01%	115.86%
9	年級.大學部四年級⇒新手機的款式.一體成型	27.74%	69.68%	101.09%
10	性別.男⇒新手機的價位.high	32.66%	65.47%	114.90%
11	性別.男⇒新手機的款式.一體成型	38.48%	77.12%	111.89%

表12：從L_3所找出的關聯規則

No	關聯規則	支持度	信心度	相關係數
1	（新手機的價位.high, 廣告訊息來源.網路）⇒新手機的款式.一體成型	25.37%	75.24%	109.16%
2	（新手機的款式.一體成型, 廣告訊息來源.網路）⇒新手機的價位.high	25.37%	65.22%	114.46%
3	（新手機的價位.high, 性別.男）⇒新手機的款式.一體成型	26.00%	79.61%	115.51%
4	（新手機的款式.一體成型, 性別.男）⇒新手機的價位.high	26.00%	67.58%	118.61%

陸、結論

關聯規則探勘技術是一項重要的資料挖掘技術，這項技術可以從交易資料庫中挖掘消費者的購買行為的關聯性。由於關聯規則能描述項目不同種類之間的關聯性，因此，可以找出並顯示商品銷售的關聯性和客戶消費傾向等資訊。雖然，考量序數資料間存在著相似度，將可以找出更多有意義的規則。然而，考慮項目之間的相似度固然可能產生更多有意義的高頻項目集，卻也同時產生相似度太低的項目集。為解決上述的問題，本研究以Apriori演算法為基礎，同時考慮項目之間的相似度，並應用α-cut過濾支持度過低的項目集，最後運用相關係數（*lift*）挖掘出有意義且具有高度關聯性的關聯規則。

本研究的貢獻歸納如下：（1）延續Chen and Weng（2008）的研究，將序數型的資料形態加入關聯規則探勘，以挖掘出更多有意義的規則；（2）應用α-cut過濾支持度過低的項目集，以避免累計支持度過低的項目集，避免產生不具關聯性的規則；（3）應用統計學上的相關係數（*lift*）篩選出具高度相關的關聯規則；以及（4）將上述的構想應用於手機購買意願調查資料中，並且挖掘出有意義的規則。

由於本研究所使用的相似度矩陣以及隸屬函數乃是由專家事先指定，因此未來研究期能由系統自動取得，以解決需專家事先指定的瓶頸。除此之外，未來仍希望繼續運用更有效率的衡量機制尋找有意義的關聯規則。

附錄A

Input: frequent itemset L_k.
Output: Candicate itemset C_{k+1}.
Method:
//self-joining L_k
insert into C_{k+1}
select p.item1, p.item2, …, p.item$_k$, q.item$_k$
from L_k p, L_k q
where p.item$_1$=q.item$_1$, …, p.item$_{k-1}$=q.item$_{k-1}$, p.item$_k$ < q.item$_k$

圖4：*The Candidate_gen Subroutine*

參考文獻

1. Agrawal, R., Imielinski, T. and Swami, A. "Mining association rules between sets of items in large databases," *Proceedings of ACM SIGMOD*, 1993, pp. 207-216.
2. Agrawal, R. and Srikant, R. "Fast algorithms for mining association rules," in: *Proceedings of the 20th International Conference on Very Large Data Bases*, Santiago, Chile, 1994, pp 487-499.
3. Berry, M. and Linoff, G. *Data Mining Techniques: For Marketing, Sales, and Customer Support*, Wiley, NY, 1997.
4. Bodjanova, S. "A generalized α-cut," *Fuzzy Sets and Systems* (126:2), 2002, pp. 157-176.
5. Chen, M.S., Han, J. and Yu, P.S. "Data Mining: An overview from a database perspective," *IEEE Trans Knowledge and Data Engineering* (8:6), 1996, pp. 866-883.
6. Chen, Y.L. and Weng, C.H. "Mining association rules from imprecise ordinal data," *Fuzzy sets and systems* (159:4), 2008, pp. 460-474.
7. Chen, Y.L., Tang, K., Shen, R.J. and Hu, Y.H. "Market basket analysis in a multiple store environment," *Decision Support Systems* (40:2), 2005, pp. 339-354.
8. Djouadi, Y., Redaoui, S. and Amroun, K. "Mining fuzzy association rules from uncertain data," in: *IEEE International Fuzzy Systems Conference*, London, 2007, pp. 1-6.
9. Han, J. and Kamber, W. *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, CA, 2001.
10. Hong, T.P., Kuo, C.S. and Chi, S.C. "Mining association rules from quantitative data," *Intelligent Data Analysis* (3:5), 1999, pp. 363-376.
11. Hu, Y.C., Chen, R.S. and Tzeng, G.H. "Discovering fuzzy association rules using fuzzy partition *methods*," Knowledge-Based System (16:3), 2003, pp. 137-147.
12. International Business Machines, *IBM Intelligent Miner User's Guide*, Version 1, Release 1, 1996.
13. Kotler, P. and Keller, K.L. A *framework for marketing management*, Pearson Education, Upper Saddle River, New Jersey, 2001.
14. Lian, W., Cheung, D.W. and Yiu, S.M. "An efficient algorithm for finding dense regions for mining quantitative association rules," *Computers and Mathematics with Applications* (50:3-4), 2005, pp. 471-490.
15. Lin, C. and Hong, C. "Using customer knowledge in designing electronic catalog," *Expert Systems with Applications* (34:1), 2008, pp. 119-127.
16. Shepard, D. *The New Direct Marketing: How to Implement a Profit-Driven Database Marketing Strategy* (3nd ed), David Shepard Associates, McGraw-Hill, 1998.
17. Srikant, R., Vu, Q. and Agrawal, R. "Mining Association Rules with Item Constraints," *SIGMOD International Conference on Management of Data*, 1996, pp. 1-12.

18. Tseng, M.C., Lin, W.Y. and Jeng, R. "Incremental maintenance of generalized association rules under taxonomy evolution," *Journal of Information Science* (34:2), 2008, pp. 174-195.
19. Yang, T.C. and Lai, H. "Comparison of product bundling strategies on different online shopping behaviors," *Electronic Commerce Research and Applications* (5:4), 2006, pp. 295-304.
20. Zadeh, L.A. "Quantitative fuzzy semantics," *Information Sciences* (3:2), 1971, pp. 159-176.