挖掘重要關聯規則-以搭配銷售為例

翁政雄
中臺科技大學資訊管理系
chweng@mgt.ncu.edu.tw

摘要

將產品或服務組合銷售在實務上已是極為普遍的行銷手法。近年來，此種行銷手法已經應用在網路商店中，藉以提升產品的銷售數量。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資料之間的關聯性。大部分的關聯規則探勘技術乃是以支持度-信心度架構為基礎。雖然，支持度-信心度的架構可以過濾大多數無意義的規則，然而，對使用者而言仍有多數的無意義的規則依舊被找到。除此之外，對行銷人員而言，哪一項關聯規則較為重要呢？可以協助行銷人員制定適當行銷策略(如搭配銷售)，以提供較佳購買計畫以吸引消費者呢？為了解決上述問題，本研究提出一種新的方法，嘗試從相關項目集中(relevant itemsets)找到有意義且重要的關聯規則。首先，本研究運用相關性指標(correlation measure)找出具有相關性的關聯規則，進而運用重要性指標(importance measure)篩選出重要的關聯規則。本研究所提出的重要性指標可以強化現有支持度-信心度架構，讓使用者透過重要性指標的高低衡量關聯規則的重要性。實驗結果顯示本研究所提出的方法可以篩選重要且具有高度相關的關聯規則，以協助行銷人員制定合適的行銷策略(如搭配銷售)。

關鍵詞：搭配銷售、資料探勘、關聯規則、關聯分析、重要性

挖掘重要關聯規則-以搭配銷售為例

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(Han and Kamber, 2006)。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。

Apriori 演算法已經被廣泛用來從交易資料庫中尋找高頻項目集。由於該演算法的成功與廣泛使用，諸多不同改良型的演算法已經先後被提出，而這些演算法所處理的資料型態可分為：種類型資料型態(Agrawal et al., 1993; Agrawal and Srikant, 1994)、序數型資料型態(Chen and Weng, 2008)以及數值型資料型態(Delgado et al., 2003; Lian et al., 2005)。

為了尋找更多的關聯規則，我們可以將最小支持度門檻值設的更低。然而，當高頻項目集(符合最小支持度門檻值的項目集)數量增加時，則關聯規則的數量所會隨之大量增加。除此之外，並非所有以支持度-信心度的架構為基礎所產生的關聯規則都是有意義的。因此，Aggarwal and Yu (1998)以統計學上的相關性分析(correlation analysis)做為另一項衡量指標，用以強化支持度-信心度架構，期能找出有意義的關聯規則。一項有意義的關聯規則表示成：$X \Rightarrow Y$ [support, confidence, correlation]。從此之後，各種相關性的衡量指標先後被提出(Han et al., 2007)，包括：*lift*, χ^2, cosine 與 all-confidence。其中，Brin et al. (1997)提出 *lift* 及 χ^2 相關性衡量指標，而 all-confidence 相關性衡量指標則由 Omiecinski (2003)提出。

將產品或服務組合銷售在實務上已是極為普遍的行銷手法，不僅賣方可從商品組合中增加產品的銷售數量，而買方的消費者常可因購買商品組合而享有優惠的價格或是額外的贈品(Garfinke et al., 2006)。「商品組合」(Bundle)，根據 Guiltinan (1987)的定義，係以一個單一的包裝，行銷兩種或多種的產品或服務。例如，旅行業者推出的旅遊套餐包括航空機票、飯店住宿、租車服務等。

為了增加產品之銷售量，廠商往往會以商品組合方式進行目標商品的促銷活動，促銷活動的方式林林總總，例如相同品牌的產品組合、性質相近的產品組合甚至是其他表面上看來似乎不太相關的商品組合，以獲取銷售機會並達成降低成本的目的。然而，如何決定商品組合中「何種商品」可以用來促銷「另一種商品」，並且能夠成功刺激消費者的購買慾望是一個值得研究的議題。

首先，如果該「商品組合」能夠成功刺激消費者的購買慾望，意味著該「商品組合」經常被消費者「一起購買」。因此，我們可以從交易資料庫尋找「高頻項目集」，即可以找出「何種商品」經常被消費者「一起購買」。然而，「高頻項目集」無法辨識出「何種

商品」可以用來促銷「另一種商品」。因此，本研究提出重要性指標(importance index)用以決定關聯規則的重要性，進而辨識出「何種商品」可以用來促銷「另一種商品」。

以購物籃分析為例，情況一：假設我們得知一項規則($bread \Rightarrow milk$)具有 80%的信心度，即顧客購買麵包的情況下，會同時購買牛奶的機率為 80%。假設我們得知另一項規則($\overline{bread} \Rightarrow milk$)具有 70%的信心度，即顧客不購買麵包的情況下，會同時購買牛奶的機率為 70%。比較上述兩項規則($bread \Rightarrow milk$)與($\overline{bread} \Rightarrow milk$)，我們得知：顧客買麵包的情況下，同時買牛奶的頻率更高。即「買麵包」的消費者比「不買麵包」的消費者有更高的機率「買牛奶」。因此，規則($bread \Rightarrow milk$)對行銷決策者是一項重要的規則。因為，可以運用此規則制定相關的行銷活動。例如：將「麵包與牛奶」搭配銷售，運用「麵包」刺激「牛奶」的銷售量。

情況二：倘若，另一項規則($\overline{bread} \Rightarrow milk$)具有 90%的信心度，即顧客「不購買麵包」的情況下會同時「購買牛奶」的機率為 90%。則規則($bread \Rightarrow milk$)將不再具有重要性，因為顧客「不購買麵包」的情況下會同時「購買牛奶」的機率更高。即「不買麵包」的消費者比「買麵包」的消費者有更高的機率「買牛奶」。因此，規則($bread \Rightarrow milk$)對行銷決策者不再是一項重要的規則。

從上述的範例中，我們得知：藉由正規則($bread \Rightarrow milk$)與負規則($\overline{bread} \Rightarrow milk$)的信心度(即條件機率)的比較，可以判斷出正規則($bread \Rightarrow milk$)的 LHS 項目(*bread*)是否為導致消費者會購買 RHS 項目(*milk*)的主要原因。以情況一為例：規則($bread \Rightarrow milk$)具有 80%的信心度，且得知另一項規則($\overline{bread} \Rightarrow milk$)具有 70%的信心度。即「買麵包」的消費者比「不買麵包」的消費者有更高的機率「買牛奶」。因此，關聯規則($bread \Rightarrow milk$)對行銷決策者是一項重要的規則。因為，行銷管理者可以運用此規則制定相關的行銷活動。例如：將「麵包與牛奶」搭配銷售，運用「麵包」刺激「牛奶」的銷售量。故本研究提出新的衡量指標(*importance* measure)，用以篩選出對行銷策略重要的關聯規則，進而過濾不重要的規則。

雖然，支持度-信心度架構可以過濾許多無意義的關聯規則。然而，對使用者而言仍有多數的無意義的規則依舊被找到。一項規則是否有意義可以客觀地評估。換言之，我們可以運用客觀的衡量指標過濾無意義的關聯規則。Aggarwal and Yu (1998) 以統計學上的相關性分析(correlation analysis)做為另一項衡量指標，用以強化支持度-信心度架構，期能找出具有相關性的關聯規則。然而，藉由相關性分析僅能了解高頻項目集(frequent itemsets)之間的相關性，無法得知關聯規則的重要性(importance)。

從上述的討論，我們得知：(1)相關性分析(correlation analysis)，可以過濾相關性較低的關聯規則，而只保留具有高度相關性的關聯規則。(2)負關聯規則可能與正關聯規則具有重要性。然而，如何從眾多的正關聯規則與負關聯規則中，篩選出重要的關聯規則是非常值得研究的議題。從相關性分析(correlation analysis)，僅能得知關聯規則中項目集之間的相關程度，並無法辨識出「何種商品」可以用來促銷「另一種商品」。為了解決上述問題，本研究提出重要性指標(*importance*)，用以強化現有支持度-信心度架構，讓使用者運用重要性指標的高低，藉以衡量關聯規則的重要性，進而辨識出「何種商品」可以用來促銷「另一種商品」。

本研究其他章節規劃如下：第二章說明問題定義。第三章為演算法設計，將詳細說明本研究所提出的演算法，並且比較本研究演算法與傳統 Apriori 演算法的差異。第四

章為實驗部份，將以實際案例驗證本研究的可行性。第五章將本研究的成果加以探討，並說明結論與未來研究。

貳、問題定義

本章將詳細定義：重要性指標(*importance*)以及如何決定關聯規則的重要性。首先，我們先回顧關聯規則的定義，並且說明支持度-信心度為何無法過濾掉所有無意義的關聯規則。次之，我們簡介相關性指標(*lift*)的定義，以及如何運用相關性指標(*lift*)篩選出具有相關性的關聯規則。最後，我們將介紹重要性指標(*importance*)，以及如何運用重要性指標(*importance*)篩選出重要的關聯規則。

定義 1. (Agrawal et al., 1993) 令 $I=\{i_1, i_2, \ldots, i_m\}$為所有項目的集合且資料庫 D 為交易資料 T 的集合，而交易資料 T 為項目的集合，其中 $T \subseteq I$。令 X 為項目的集合，交易資料 T 包含 X，若且為若 $X \subseteq T$。關聯規則表示成 $X \Rightarrow Y$，其中 $X \subset I$, $Y \subset I$ 且 $X \cap Y = \varnothing$。支持度(*support*, *s*)用以表示關聯規則$\{X \Rightarrow Y\}$在資料庫 D 中出現 $X \cup Y$ 的百分比。信心度(*confidence*, *c*)用以表示關聯規則$\{X \Rightarrow Y\}$在資料庫 D 中出現 X 的情況下同時也出現 Y 的百分比。因此，$Support(X \cup Y)$, $Confidence(X \Rightarrow Y)$的定義如下所示：

$$Support(X \cup Y) = \frac{|X \cup Y|}{|D|} \tag{1}$$

$$Confidence(X \Rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)} \tag{2}$$

其中，$|D|$表示資料庫 D 的總交易資料筆數，$|X \cup Y|$表示資料庫 D 同時出現 $X \cup Y$ 的交易資料筆數。

範例 1. 假設我們有一資料集合包含 1,000 筆交易資料。令 $\overline{coffee}$ 表示該筆交易資料不包含 *coffee*，而 $\overline{milk}$ 表示該筆交易資料不包含 *milk*。則 1,000 筆交易資料的分類彙總數據，如表 1 所示。從表 1 我們得知：$Support(coffee \cup milk) = 400/1000 = 0.40$ 和 $Confidence(coffee \Rightarrow milk) = 400/600 = 0.67$. 倘若，最小支持度與最小信心度分別為 40% 及 60%。則我們得到下列關聯規則：
$coffee \Rightarrow milk$ [*support*=40%, *confidence*=67%] .

表 1　A 2×2 contingency table for two items.

	coffee	$\overline{coffee}$	$\sum_{row}$
milk	400	350	750
$\overline{milk}$	200	50	250
$\sum_{col}$	600	400	1,000

然而，關聯規則{ $coffee \Rightarrow milk$ [*support*=40%, *confidence*=67%]}可能誤導使用者。因為從表 1 得知：顧客購買牛奶(*milk*)的機率為 75%，高於關聯規則($coffee \Rightarrow milk$)的信心度 67% 。換言之，買咖啡(*coffee*)並不會導致買牛奶(*milk*)的機率提升，反而下降。

從上述的範例中得知：支持度-信心度架構並無法完全過濾掉無意義的關聯規則。Piatetsky-Shapiro (1991)認為：倘若 $Support(X \cup Y) = Support(X) \times Support(Y)$ ，則關聯規則 $X \Rightarrow Y$ 是一項無意義的關聯規則。為了解決上述問題，相關性指標(correlation measure)已經被用來強化支持度-信心度架構，期能找出有意義的關聯規則。

定義 2. (Brin et al., 1997) 令 $P(X)$ 代表資料庫 D 中，每一筆交易資料包含項目 X 的機率且 $P(\overline{X}) = 1 - P(X)$ 代表資料庫 D 中，每一筆交易資料不包含項目 X 的機率。相同地，$P(XY)$ 代表資料庫 D 中，每一筆交易資料同時包含項目 X 及項目 Y 的機率，而 $P(\overline{X}Y)$ 代表資料庫 D 中，每一筆交易資料不包含項目 X，但是含項目 Y 的機率。倘若 $P(X \cup Y) = P(X) \times P(Y)$，則表示項目 X 與項目 Y 之間互相獨立；否則，項目 X 與項目 Y 之間具有相關性。*lift* 指標是一種相關性指標，可以用來衡量項目之間的相關性，其定義如下：

$$lift(X,Y) = \frac{P(X \cup Y)}{P(X)P(Y)} = \frac{Support(X \cup Y)}{Support(X) \times Support(Y)} = \frac{Confidence(X \Rightarrow Y)}{Support(Y)} \tag{3}$$

範例 2. 承範例 1，關聯規則{*coffee* ⇒ *milk*}的相關性為：*lift*(*coffee*, *milk*)=(0.40)/(0.60×0.75)=0.89. 因為關聯規則{*coffee* ⇒ *milk*}的相關小於 1，故項目(*coffee*)與項目(*milk*)之間存在負相關性(negative correlation)。然而，這種負相關性(negative correlation)無法從支持度-信心度架構中得知。藉由相關性指標，可以了解關聯規則中項目與項目之間的相關性，因此，運用相關性指標，關聯規則可以進一步表示成{$X \Rightarrow Y$ [support, confidence, correlation]}。除此之外，其他的相關性指標尚有 χ^2、cosine 及 all-confidence 等(Han et al., 2007)。

除了上述討論的關聯規則為正關聯(positive association)形式，而關聯規則也存在負關聯(negative association)的形式，這類型的關聯規則稱為負關聯規則(Savasere et al., 1998)。其中，負關聯規則用以描述「出現項目」與「未出現項目」之間的關係，例如：買咖啡，但不買牛奶。

隨著資料探勘技術與工具的發展，相關的研究開始聚焦於負關聯(negative associations)的研究(Savasere et al. 1998)。不同於傳統的正關聯規則(positive association rules)，Wu et al. (2004)將關聯規則的研究延伸到其他形式的關聯規則，如{ $\overline{X} \Rightarrow Y$ 、$X \Rightarrow \overline{Y}$ 與 $\overline{X} \Rightarrow \overline{Y}$ }。

定義 3. (Negative association rule; Yuan et al., 2002)負關聯規則(negative association rule)表示成 $\overline{X} \Rightarrow Y$ (或 $X \Rightarrow \overline{Y}$)，其中 $X \subset I, Y \subset I$ 且 $X \cap Y = \varnothing$；然而，關聯規則 $\overline{X} \Rightarrow \overline{Y}$ 中，左項目($\overline{X}$)與右項目($\overline{Y}$)皆為「未出現項目」，其意義等同於關聯規則 $Y \Rightarrow X$。因此，關聯規則 $\overline{X} \Rightarrow \overline{Y}$ ，將不視為負關聯規則。故本研究所定義的負關聯規則為：$\overline{X} \Rightarrow Y$ 與 $X \Rightarrow \overline{Y}$。其中，負關聯規則{ $\overline{X} \Rightarrow Y$ }的支持度與信心度定義如下：

$$Support(\overline{X} \cup Y) = \frac{Support(Y) - Support(X \cup Y)}{|D|} \tag{4}$$

$$Confidence(\overline{X} \Rightarrow Y) = \frac{Support(Y)}{1 - Support(X)} \times (1 - Confidence(Y \Rightarrow X)) \tag{5}$$

範例 3. 令 $\overline{coffee}$ 表示該筆交易資料不包含 *coffee*，而 $\overline{milk}$ 表示該筆交易資料不包含

milk，而$\overline{bread}$表示該筆交易資料不包含 *bread*。假設有交易資料統計數據如表 2 所示。從表 2 得知：買 *bread* 共有 650 筆、不買 *bread* 共有 350 筆、買 *milk* 共有 450 筆、買 *coffee* 共有 300 筆以及買 others(其他商品)共有 350 筆。因此，得知：$Support(\overline{milk} \cup bread) =$ (650-300)/1000 = 0.35, $Support(milk)$=0.45, $Support(\overline{milk}) = 1 - Support(milk) = 1 - 0.45 =$ 0.55, $Support(bread)$=0.65 且 $Confidence(bread \Rightarrow milk) = 0.30/0.65 = 0.46$. 則 $Confidence(\overline{milk} \Rightarrow bread) =$ (0.65/0.55)×(1-0.46)=0.64. 最後，我們得到負關聯規則 $\{\overline{milk} \Rightarrow bread\}$，如下：

$\overline{milk} \Rightarrow bread$ [*support*=35%, *confidence*=64%] .

表 2　A 3×2 contingency table for three items.

	bread	$\overline{bread}$	$\sum_{row}$
milk	300	150	450
coffee	250	150	400
others	100	50	150
$\sum_{col}$	650	350	1,000

從表 2，我們可以找到其他的負關聯規則：$\{\overline{coffee} \Rightarrow bread$ [*support*=40%, *confidence*=67%]}。依據定義 1，我們也可以找到 2 項關聯規則：{*milk* ⇒ *bread* [*support*=30%, *confidence*=67%]} 與{*coffee* ⇒ *bread* [*support*=25%, *confidence*=63%]}。對決策制定而言，負關聯規則也可以與正關聯規則一樣扮演重要的角色。以 $\{\overline{coffee} \Rightarrow bread$ [*support*=40%, *confidence*=67%]}為例，我們得知：顧客不購買 *coffee*，但是購買 *bread* 的信心度(機率)為 67%。

制定決策時，決策者不知如何從眾多的關聯規則中(正關聯規則與負關聯規則)，篩選出重要規則以利決策之制定。例如：從上述的規則中，我們可以得到下列規則：規則(1){*milk* ⇒ *bread* [*support*=30%, *confidence*=67%]} 、規則 (2){ $\overline{milk} \Rightarrow bread$ [*support*=35%, *confidence*=64%]}、規則(3){*coffee* ⇒ *bread* [*support*=25%, *confidence*=63%]} 以及規則(4){ $\overline{coffee} \Rightarrow bread$ [*support*=40%, *confidence*=67%]}。就規則(1){*milk* ⇒ *bread* [*support*=30%, *confidence*=67%]} 與規則 (3){*coffee* ⇒ *bread* [*support*=25%, *confidence*=63%]}而言，其支持度與信心度皆大於 25%與 60%。然而，這兩項規則對決策者而言都很重要嗎？在產品銷售上(如搭配銷售策略)何者比較重要呢？

就搭配銷售策略而言，倘若產品 *X* 的銷售能夠帶動產品 *Y* 的銷售，即「消費者購買 *X* 會強化購買 *Y* 的機率」，則關聯規則{*X*⇒*Y*}對搭配銷售策略的制定是非常重要的。反之，倘若消費者「不購買 *X*」的情況下，反而購買 *Y* 的機率更高，則關聯規則{*X*⇒*Y*}對搭配銷售策略而言，就不再重要了。因為，「產品 *X* 的銷售能並不會帶動產品 *Y* 的銷售」。

根據關聯規則的定義：信心度為條件機率的概念，即關聯規則{*X*⇒*Y*}的信心度為「購買 *X* 的條件下，購買 *Y* 的機率」。故信心度高僅代表條件機率高，即「消費者在**購買 *X***的條件下，大部分會購買 *Y*」。然而，有可能絕大多數的消費者是傾向「消費者在**不購買 *X***的條件下，更會購買 *Y*」。因此，透過關聯規則的相關性指標：{*X*⇒*Y* [support, confidence, correlation]}，我們僅知道關聯規則的項目之間的相關性程度，而無法得知關聯規則的重要性。

綜合上述討論，本研究提出關聯規則的重要性指標(*Importance*)：對於關聯規則{$X \Rightarrow Y$}，比對正關聯規則{$X \Rightarrow Y$}與其所對應的負關聯規則{$\overline{X} \Rightarrow Y$}之信心度比值。藉以判定關聯規則{$X \Rightarrow Y$}中 LHS 項目{ X }是否會導致 RHS 項目{ Y }出現的機率。就購物籃分析為而言，本研究希望透過重要性指標(*Importance*)了解：關聯規則{$X \Rightarrow Y$}中 LHS 項目{ X }是否會強化購買 RHS 項目{ Y }出現的機率。除此之外，由於人們熟悉運用正負值判斷重要與否(正值表示重要；負值表示不重要)，故本研究將信心度比值取對數(log)以便利決策者判斷上之方便，換言之，正值代表重要，負值代表不重要，而且數值越大代表越重要。關於本研究所提出的重要性指標(*Importance*)定義如下：

定義 4. (*Importance*; 本研究) 假設我們有正關聯規則{$X \Rightarrow Y$}以及其所對應的負關聯規則{$\overline{X} \Rightarrow Y$}。則正關聯規則{$X \Rightarrow Y$}的重要性定義如下：

$$Importance(X \Rightarrow Y) = \log \frac{Confidence(X \Rightarrow Y)}{Confidence(\overline{X} \Rightarrow Y)} \tag{6}$$

除此之外，根據定義 2：$lift(X,Y) = \dfrac{Confidence(X \Rightarrow Y)}{Support(Y)}$,

因此，正關聯規則{$X \Rightarrow Y$}的重要性指標(*Importance*)亦可定義如下：

$$Importance(X \Rightarrow Y) = \log \frac{lift(X,Y) \times Support(Y)}{lift(\overline{X},Y) \times Support(Y)} = \log \frac{lift(X,Y)}{lift(\overline{X},Y)} \tag{7}$$

範例 4. 承範例 3，從表 2 我們可以找到 2 項正關聯規則{*milk* ⇒ *bread* [*support*=30%, *confidence*=67%]} 以及 {*coffee* ⇒ *bread* [*support*=25%, *confidence*=63%]}。除此之外，我們也可以找這兩項正關聯規則所對應的負關聯規則：{$\overline{milk} \Rightarrow bread$ [*support*=35%, *confidence*=64%]}以及{$\overline{coffee} \Rightarrow bread$ [*support*=40%, *confidence*=67%]}。因此，我們可以計算出：*Importance*(*milk* ⇒ *bread*) = log(0.67/0.64)=log(1.05)= 1.99% 以及 *Importance*(*coffee* ⇒ *bread*) = log(0.63/0.67)=log(0.94)= -2.67%。由上述的重要性指標(*Importance*)，我們得知：正關聯規則{$milk \Rightarrow bread$}是重要的關聯規則，因為它的重要性(*Importance*)大於 0。然而，正關聯規則{$coffee \Rightarrow bread$}不是重要的關聯規則，因為它的重要性(*Importance*)小於 0。除此之外，關聯規則{$milk \Rightarrow bread$}的重要性(1.99%)大於關聯規則{$coffee \Rightarrow bread$}的重要性(-2.67%)，因此，關聯規則{$milk \Rightarrow bread$}的重要性比較高。

從上述的範例得知：透過重要性(*Importance*)指標，我們可以將關聯規則的形式表示成{$X \Rightarrow Y$ [support, confidence, correlation, importance]}，用以呈現關聯規則的重要性。透過重要性的數值大小，我們可以了解關聯規則的重要性程度。故本研究所提出的重要性指標將有助於使用者從眾多的關聯規則中，篩選出重要的關聯規則，以協助管理者行銷策略之制定。

參、演算法設計

本章節將詳細說明本研究所提出的 **IARM (Important Association Rule Mining)**演算法，並且比較 **IARM** 演算法與傳統 Apriori 演算法(Agrawal and Srikant, 1994)的差異。

由於**IARM**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，3.1 節將先介紹 Apriori 演算法，3.2 節則說明 **IARM** 演算法。

一、 Apriori 演算法

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

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

(7). 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。

雖然，Apriori 演算法所使用的支持度-信心度架構可以過濾大多數無意義的規則，然而，對使用者而言仍有多數的無意義的規則依舊被找到。為了解決上述問題，本研究運用相關性指標(correlation measure)選出具有相關性的關聯規則，進而運用重要性指標(importance measure)找出有重要的關聯規則。本研究所提出的重要性指標可以強化現有支持度-信心度架構，讓使用者透過重要性指標的高低衡量關聯規則的重要性。

二、 **IARM** 演算法

本章節將介紹本研究所提出之 **IARM** 演算法的運作方式，在介紹之前，我們首先比較 Apriori 演算法與 **IARM** 演算法的差異。兩者的差異歸納如下：

(1) 負關聯規則(Negative association rules)：傳統的 Apriori 演算法用於尋找正關聯規則(positive association rules)，而不是負關聯規則。然而，負關聯規則也如同正關聯規則一樣具有實用價值。**IARM** 演算法則運用相關的正關聯規則，計算負關聯規則的支持度與信心度，如定義 3 所示。

(2) 相關係數(*lift*)：**IARM** 演算法應用統計學上的相關係數(*lift*)概念用以衡量關聯規則左項目(LHS)與右項目(RHS)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。

(3) 重要性指標(*importance*)：**IARM** 演算法運用重要性指標(importance measure) 衡量關聯規則的重要性，進而強化支持度及信心度為基礎的 Apriori 演算法架構。

如圖 1 所示，**IARM** 演算法分成 4 大步驟，分別為： (1)利用反覆的方式，逐一統計高頻項目集的支持度。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，進而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，再利用高頻項目集 L_k，合併產生項目集長度為(k+1)的候選項目集 C_{k+1}。(2)利用所找到的高度關聯性的高頻項目

集 L_k 產生正關聯規則(*PARs*)，規則是否成立則以信心度門檻值(σ_{conf})為衡量依據。除此之外，應用統計學上的相關係數概念用以衡量關聯規則左項目(LHS)與右項目(RHS)之間的相關程度，並以門檻值(σ_{lift})為衡量依據，僅保留具有高度關聯性高頻項目集的正關聯規則。(3)產生負關聯規則(*NARs*)以做為後續篩選重要的關聯規則之用。(4)產生重要的關聯規則：比對正關聯規則與負關聯規則的信心度，以決定關聯規則的重要性。進而篩選出重要的關聯規則。

Input: A database, D_B; a predefined minimum support σ_{sup}; a predefined minimum confidence σ_{conf}; a predefined minimum correlation σ_{lift}; a predefined minimum correlation $\sigma_{importance}$.
Output: A set of association rules
Method:
// Phase 1 Call the *FreqItemsets_gen* Subroutine
(1). For each item it_i, calculate its support.
(2). Check whether the support of each item it_i is no less than the minimum support σ_{sup}. If it is, put it into the set of frequent one-itemsets (L_1).
(3). Generate candidate set C_{k+1} from L_k.
(4). Compute the supports of all itemsets in C_k and determine L_k.
(5). If L_{k+1} is null, go to phase 2; otherwise, set $k = k + 1$ and repeat steps (3)–(5).
// Phase 2 Call the *PAR_gen* Subroutine
(1). Generate positive association rules from all frequent and relevant itemsets with the two thresholds σ_{conf} and σ_{lift}.
// Phase 3 Call the *NAR_gen* Subroutine
(1). Generate negative association rules based on both the positive association rule and its relevant-frequent itemsets.
// Phase 4 Call the *IAR_gen* Subroutine
(1). Determine important association rules by calculating the *Importance* value for from the positive association rule and its negative association rule.

圖 1　IARM 演算法

肆、實驗結果

本研究利用數項實驗數據衡量，用以驗證 **IARM** 演算法的成效。本研究以松青超市某週的銷售紀錄資料集為例。其中，有效資料總共有 77506 筆交易資料。本研究使用筆記型電腦進行實驗，電腦配備如下：(1)CPU 為 Intel Centrino 1400 MHz processor、(2)記憶體有 512MB 及 (3)使用 Windows XP 作業系統。

本研究所進行的實驗，包含：(1)測試 **IARM** 演算法在不同支持度(σ_{sup})情況下的執行時間；(2)統計 **IARM** 演算法在不同支持度(σ_{sup})情況下的高頻項目集個數；(3)統計 **IARM** 演算法在不同相關係數(σ_{lift})情況下的關聯規則數，以及(4)統計 **IARM** 演算法在不同重要性($\sigma_{importance}$)情況下的關聯規則數。

在第一個實驗中，將測試 **IARM** 演算法在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 77506 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 Apriori 演算法相同(Agrawal and Srikant, 1994)。

在第二個實驗中，將統計 **IARM** 演算法在不同支持度情況下的高頻項目集個數。如圖 3 所示，與傳統的 Apriori 演算法相同，當支持度越小時將產生越多的高頻項目集(Agrawal and Srikant, 1994)。從第一個實驗與第二個實驗的數據中，得知本研究所提出 **IARM** 演算法正確無誤，後續的實驗中，將進一步探討不同相關係數(σ_{lift})門檻值對 **IARM** 演算法產生關聯規則數的影響。

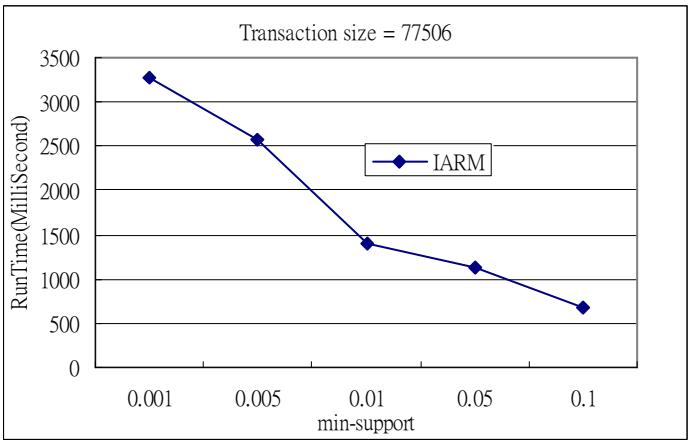

圖 2 執行時間 vs. 最小支持度

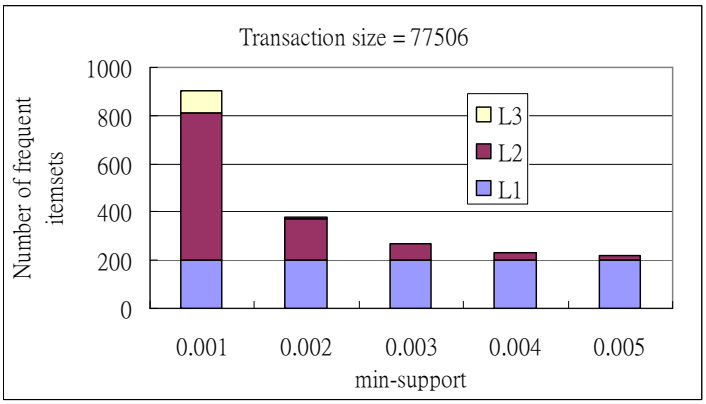

圖 3 項目集個數 vs. 最小支持度

在第三個實驗中，將測試 **IARM** 演算法在固定支持度，但是不同相關係數(σ_{lift})門檻值情況下的關聯規則數(由 L2 所產生的關聯規則)。本研究所使用的資料樣本大小依舊為 77506 筆有效資料。如圖 4 所示，關聯規則數隨著相關係數(σ_{lift})門檻值增加而減少，因為相關係數(σ_{lift})門檻值越小將產生更多不具相關性的關聯規則。

在第四個實驗中，將統計 **IARM** 演算法在固定支持度，但是不同重要性($\sigma_{importance}$)門檻值情況下的關聯規則數。本研究所使用的資料樣本大小依舊為 77506 筆有效資料。如圖 5 所示，重要關聯規則數隨著重要性($\sigma_{importance}$)門檻值增加而減少，因為重要性($\sigma_{importance}$)門檻值越小將產生更多不重要關聯規則。

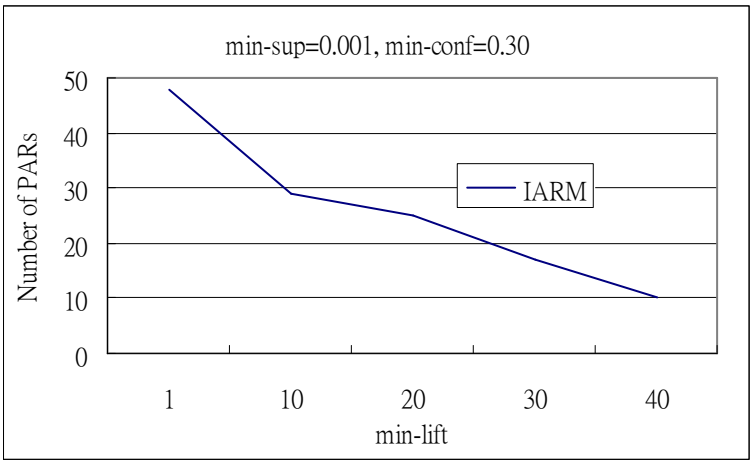

圖 4　Rules vs. *lift*.

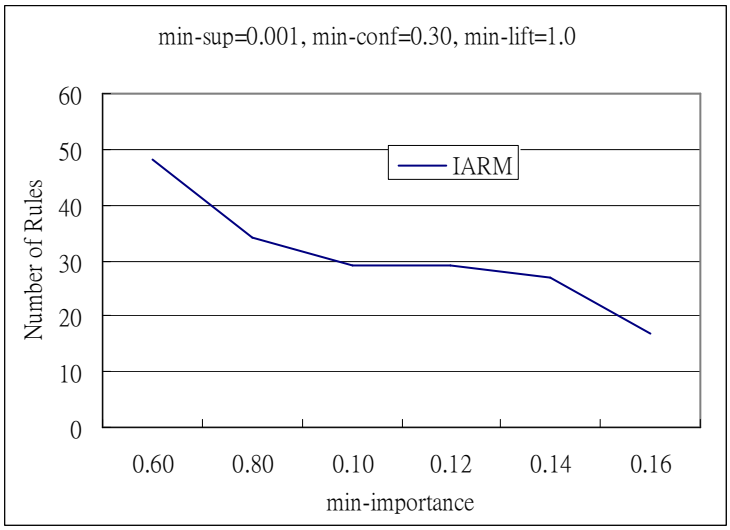

圖 5　Rules vs. *importance*.

從上述的實驗中，我們得知：(1)相關係數(*lift*)能過濾關聯規則中 LHS 與 RHS 相關性不高者，進而篩選出 LHS 與 RHS 具高度相關性的關聯規則。(2)本研究運用所提出的重要性指標(*importance*)，更進一步確認具高度相關性的關聯規則中哪些規則的 LHS 的確會高度導致 RHS 的發生，藉以判斷出該規則的重要性。其中，重要性指標(*importance*)運用取對數(log)方式計算正負關聯規則的信心度比值，運用正負值的方式判斷關聯規則的重要性，即無意義的規則其重要性為負值。

因為運用傳統的 Apriori 演算法(最小支持度與最小信心度)即可找出所有規則集合(包含重要的規則)。故本研究並不會找出新奇的規則，而是協助決策者過濾不重要規則，聚焦於重要的規則。規則的重要與否除了決策者人為主觀之判斷外，本研究提出客觀的重要性指標(*importance*)，藉以判斷出「哪些規則的 LHS 的確會高度導致 RHS 的發生」。而篩選出具有高度重要性的規則後，決策者可以運用行銷活動之制定，例如：搭配銷售。

最後，本研究運用 4 個使用者定義的門檻值，分別為支持度(σ_{sup}=0.2%)、信心度(σ_{conf}=30%)、相關係數(σ_{lift}=100%)以重要性($\sigma_{importance}$=160%)及所找出的部分關聯規則，如表 3 所示。表 3 的所有重要規則的重要性指標值皆大於 0，故表示每一項規則的 LHS

產品皆能帶動 RHS 產品的銷售。例如，規則#1 與規則#2 的 RHS 產品為“桂冠魚卵卷-80g”，而 LHS 產品分別為“桂冠蟳味棒-100g”及“桂冠蝦球-150g”皆為促銷 RHS 產品(桂冠魚卵卷-80g)合適的產品後選項目。除此之外，經由比較重要性指標值，我們知道產品組合{桂冠蝦球-150g、桂冠魚卵卷-80g }較佳，因為其重要性指標值較高(imp=**1.891**)略高於產品組合{桂冠蟳味棒-100g、桂冠魚卵卷-80g}的重要性指標值(imp=**1.804**)。除此之外，以規則#17 為例，LHS 產品與 RHS 產品皆是“**小泡芙**”，然而“**義美草莓小泡芙-65g**”適合用於搭配“**義美牛奶小泡芙-65g**”銷售，以增加“**義美牛奶小泡芙-65g**”的銷售量。因為根據重要性的定義，「消費者**購買**“義美草莓小泡芙-65g”，而且同時購買“義美牛奶小泡芙-65g”」**機率高於**「消費者**不買**“義美草莓小泡芙-65g”，而且同時購買“義美牛奶小泡芙-65g”」

關於管理意涵部分：為了增加產品之銷售量，廠商往往會以商品組合方式進行目標商品的促銷活動，促銷活動的方式林林總總，例如相同品牌的產品組合、性質相近的產品組合甚至是其他表面上看來似乎不太相關的商品組合，以獲取銷售機會並達成降低成本的目的。傳統關聯規則以支持度-信心度架構為基礎，透過「高頻項目集(支持度)」商家僅能了解消費者購哪些商品一起購買以及透過「條件機率(信心度)」商家僅能了解消費者購哪些商品的條件下，一起購買其他商品的機率。然而，單從「高頻項目集(支持度)與條件機率(信心度)」無法辨識出「何種商品」可以用來促銷何種「另一種商品」。因此，本研究提出重要性指標(importance index)用以決定關聯規則的重要性，藉由正關聯規則(positive association rules)與其所對應的負關聯規則(negative association rules)信心度之比較，進而辨識出「何種商品」可以用來促銷「另一種商品」。因此，本研究運用重要性指標(importance index)所篩選出的重要規則，將可以提供給管理者做為制定「產品組合」行銷策略之依據。

表 3 關聯規則

No	*LHS*	*RHS*	*sup*	*conf*	*lift*	*imp*
1	**桂冠蟳味棒-100g**	**桂冠魚卵卷-80g**	**0.002**	**0.326**	**43.183**	**1.804**
2	**桂冠蝦球-150g**	**桂冠魚卵卷-80g**	**0.003**	**0.343**	**45.388**	**1.891**
3	桂冠蝦球-150g	桂冠黃金魚蛋-150g	0.004	0.380	31.368	1.642
4	桂冠蟳味棒-100g	桂冠蝦球-150g	0.003	0.346	37.239	1.711
5	桂冠魚卵卷-80g	桂冠蝦球-150g	0.003	0.422	45.388	1.836
6	桂冠花枝餃	桂冠蝦餃	0.003	0.326	31.871	1.644
7	桂冠魚卵卷-80g	桂冠蟳味棒-100g	0.002	0.326	43.183	1.803
8	海霸王蝦餃-140g	海霸王花枝餃-140g	0.003	0.345	50.626	1.933
9	海霸王燕餃-140g	海霸王花枝餃-140g	0.003	0.324	47.565	1.893
10	海霸王燕餃-140g	海霸王魚餃-140g	0.003	0.343	31.133	1.620
11	海霸王蝦餃-140g	海霸王魚餃-140g	0.003	0.398	36.202	1.708
12	海霸王花枝餃-140g	海霸王魚餃-140g	0.002	0.358	32.525	1.618
13	海霸王花枝餃-140g	海霸王蝦餃-140g	0.003	0.415	50.626	1.885
14	海霸王燕餃-140g	海霸王蝦餃-140g	0.003	0.367	44.849	1.851

No	*LHS*	*RHS*	*sup*	*conf*	*lift*	*imp*
15	海霸王蝦餃-140g	海霸王燕餃-140g	0.003	0.373	44.849	1.847
16	海霸王花枝餃-140g	海霸王燕餃-140g	0.003	0.396	47.565	1.844
17	**義美草莓小泡芙-65g**	**義美牛奶小泡芙-65g**	**0.003**	**0.306**	**30.154**	**1.611**

伍、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資料之間的關聯性。大部分的關聯規則探勘技術乃是以支持度-信心度架構為基礎。雖然，支持度-信心度的架構可以過濾大多數無意義的規則，然而，對使用者而言仍有多數的無意義的規則依舊被找到。

為了解決上述問題，本研究運用相關性指標(correlation measure)選出具有相關性的關聯規則，進而運用重要性指標(importance measure)找出有重要的關聯規則。實驗結果顯示本研究所提出的方法可以找出重要且具有高度相關的關聯規則。本研究的貢獻歸納如下：(1)應用統計學上的相關係數(*lift*)篩選出高度相關的關聯規則；(2)應用重要性指標(importance measure)篩選出重要的關聯規則以及(3)將上述的構想應用於超市銷售資料中，並且挖掘出重要且有意義的規則。

傳統關聯規則以支持度-信心度架構為基礎：(1)透過「高頻項目集(支持度)」商家僅能了解消費者購哪些商品一起購買，以及(2)透過「條件機率(信心度)」商家僅能了解消費者購哪些商品的條件下，一起購買其他商品的機率。然而，單從「高頻項目集(支持度)與條件機率(信心度)」無法辨識出「何種商品」可以用來促銷「另一種商品」。因此，本研究提出重要性指標(importance index)用以決定關聯規則的重要性，藉由正關聯規則(positive association rules)與其所對應的負關聯規則(negative association rules)信心度之比較，進而辨識出「何種商品」可以用來促銷「另一種商品」。因此，本研究所篩選出的重要規則將有助於管理者制定「產品組合」行銷策略之依據。

由於本研究所使用的 4 個門檻值：最小支持度、最小信心度、最小相關係數以最低重要性必須由專家事先指定。因此，未來研究希望能夠由其他技術自動取得，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。

致謝

本計畫為國科會專題研究計劃編號 NSC 100-2410-H-166-003 之部分研究結果，謹此致謝。

參考文獻

1. Aggarwal, C.C., and Yu, P.S. “A new framework for itemset generation,” in *Proceedings of the 1998 ACM Symp. Principles of database systems*, 1998, pp. 18-24.
2. Agrawal, R., Imieliński, T., and Swami, A. “Mining association rules between sets of items in large databases,” in *Proceedings of ACM SIGMOD*, 1993, pp. 207-216.
3. Agrawal, R., and Srikant, R. “Fast algorithms for mining association rules,” in *Proceedings of the VLDB Conference*, 1994, pp. 487-499.
4. Brin, S, Motwani, R., and Silverstein, C. “Beyond market baskets: generalizing association rules to correlations,” in *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, 1997, pp. 265-276.
5. Chen, Y.L. and Weng, C.H. “Mining association rules from imprecise ordinal data,” *Fuzzy sets and systems* (159:4), 2008, pp. 460-474.
6. Delgado, M., Marin, N., Sanchez, D., and Vila, M.A. “Fuzzy association rules: general model and applications,” *IEEE Transactions on Fuzzy Systems* (11:2), 2003, pp. 214-225.
7. Garfinke, R., Gopal, R., Tripathi, A., and Yin F. “Design of a shopbot and recommender system for bundle purchases,” *Decision Support Systems* (42:3), 2006, pp.1974-1986.
8. Guiltinan, J. “The price bundling of services: A normative framework,” *Journal of Marketing* (51:2), 1987, pp.74-85.
9. Han, J.W., and Kamber M. Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2006.
10. Han, J., Cheng, H., Xin D., and Yan, X. “Frequent pattern mining: current status and future directions,” *Data Mining and Knowledge Discovery* (15:1), 2007, pp. 55-86.
11. 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.
12. Omiecinski, E. “Alternative interest measures for mining associations in databases,” *IEEE Transactions on Knowledge and Data Engineering* (15:1), 2003, pp. 57-69.
13. Piatetsky-Shapiro, G. “Notes of AAAI’91 Workshop Knowledge Discovery in Databases,” in *Proceedings of KDD’91*, Anaheim, CA., 1991.
14. Savasere, A., Omiecinski, E., and Navathe, S. “Mining for strong negative associations in a large database of customer transactions,” in *Proceedings of the Fourteenth International Conference on Data Engineering*, Orlando, Florida, 1998, pp. 494-502.
15. Wu, X., Zhang, C., and Zhang, S. “Efficient mining of both positive and negative association rules,” *ACM Transactions on Information Systems* (22:3), 2004, pp. 381–405.
16. Yuan, X., Buckles, B., Yuan, Z., and Zhang, J. “Mining negative association rules,” in *Proceedings of the Seventh International Symposium on Computers and Communications*, Italy, 2002, pp. 623–629.

Discovering important association rules- A study for bundle promotion

Weng, Cheng Hsiung

Central Taiwan University of Science and Technology, Department of Management Information Systems

chweng@mgt.ncu.edu.tw

Abstract

The use of bundle pricing and promotions has been a common marketing practice for a long time. Recently, They have been used extensively and more frequently in online retailing to boost sales. Association rule mining is an important data analysis method that can be used to discover associations within data. The support-confidence framework is used by most association rule mining algorithms. Generally, a lot of interesting rules can be found by setting low support thresholds. Although, the support-confidence framework can filter out many uninteresting rules, many rules uninteresting to the users still remain. Besides, which rule is more important for decision maker for making marketing strategy, such as bundle promotion and provides the best possible purchase plan for attracting the consumers? For this reason, this study proposes a new approach to discover interesting and important association rules from relevant itemsets. First, a correlation measure is applied to augment the support-confidence framework for discovering correlation association rules. Then, a new criterion, named Importance, is applied to augment the support-confidence framework for discovering important association rules in advance. Experimental results from the survey data show that the proposed approach can help to discover interesting and valuable association rules with a high correlation for decision maker for making marketing strategy, such as bundle promotions.

Keywords: bundle; data mining; association rule; correlation analysis; importance