以布林運算為基礎探勘網路拍賣異常之競標者

陳垂呈
南台科技大學資訊管理系
E-mail：ccchen@mail.stut.edu.tw

摘要

在眾多電子商務經營形態中，網路拍賣是最受消費者歡迎的交易方式之一，雖然競標者可以享受到自由喊價的趣味性，但也衍生了許多的交易問題。其中最常見的有競標者喊價之後，未能履行付款購買拍賣品的義務，或是某些競標者只是為了哄抬拍賣品的價格，而不是實際真正的競標者。因此，如何偵測出具有異常競標的競標者，使得網路拍賣交易能夠正常的進行，即成為網路拍賣經營者必須解決的問題之一。在本篇論文中，我們以競標者過去之交易資料為探勘的資料來源，其交易資料包含有競標過的產品項目及購買過的產品項目，利用探勘技術來發掘具有異常競標之行為特徵的競標者。我們從競標者過去購買過的產品項目中，以布林運算為基礎，找出產品項目之間的關聯規則，根據關聯規則來計算競標者過去競標過的產品項目中，彼此之間的關聯性，若未能滿足所設定的最小相似度，即稱之為異常的競標者。此探勘結果，對拍賣品避免被惡意哄抬價格，或事先預防無意購買拍賣品之競標者的喊價行為，將可提供非常有用的資訊。我們根據所提出的方法，設計與建置一個探勘異常競標者的系統。

關鍵詞：資料探勘、關聯規則、布林運算、網路拍賣、異常競標

一、簡介

隨著電子商務(e-commerce)的發展，消費者逐漸地習慣在網路上購買產品，在電子商務眾多的經營形態中，網路拍賣是最受消費者歡迎的交易方式之一，在拍賣網站上，消費者可以陳列寄賣的產品，而競標者可以享受到自由喊價的趣味性。隨著拍賣交易的熱絡，也隨之衍生了許多的交易問題，其中最常見的有競標者喊價之後，未能履行付款購買拍賣品的義務，或是某些競標者只是為了哄抬拍賣品的價格，而不是實際真正的競標者。因此如何發掘出異常的競標者，使得拍賣交易能夠正常的進行，即成為網路拍賣經營者必須解決的問題之一。

藉由資訊技術的支援，網路拍賣經營者可以很輕易地蒐集競標者的交易記錄，並且快速的累積，若能從這些大量的交易資料中，找出競標者的消費模式，以協助偵測競標者之競標是否異常，對拍賣品避免被惡意哄抬競價，或事先預防無意購買拍賣品之競標者的喊價行為，將可提供非常有用的資訊，以遏止異常競標行為之事件持續的發生。

資料探勘(data mining)是從大量資料中找出潛在有用的知識與資訊，以提供決策分析之參考資訊，目前資料探勘技術已普遍地應用在各領域中。在本篇論文中，我們利用資料探勘技術來發掘具有異常競標之行為特徵的競標者，並根據我們所提出的方法，設計與建置一個探勘異常競標者的系統。

在本篇論文中，我們以競標者過去之交易資料為探勘的資料來源，其交易資料包含有競

標過的產品項目及購買過的產品項目，利用探勘技術來發掘具有異常競標之行為特徵的競標者。我們從競標者過去購買過的產品項目中，以布林運算(Boolean computation)為基礎，找出產品項目之間的關聯規則(association rules)，根據關聯規則來計算競標者過去競標過的產品項目中，彼此之間的關聯性，若未能滿足所設定的最小相似度，即稱之為異常的競標者。例如，假設找出的關聯規則有 A→B 及 A→C，若一競標者過去競標過的產品項目有 ABCDE 等 5 項，因為競標過之產品包含有關聯規則的 AB 及 AC 產品項目，故其相似度為：ABC 的個數/ABCDE 的個數=3/5=60%，若未能達到所設定的最小相似支持度，即稱之為異常的競標者。

本篇論文的架構如下：下一節中，我們介紹資料探勘技術、及其在探勘異常交易的相關研究；在第三節中，我們以布林運算為基礎，利用關聯規則來發掘具有異常競標之行為特徵的競標者，並以一個實例來說明探勘的過程；第四節中，我們說明探勘系統的實作應用；最後，我們在第五節中做一結論。

二、相關研究

資料探勘是從大量資料中挖掘出潛在有用資訊與知識，發現專家尚且未知的新關係，以提供給企業專業人員參考。資料探勘可完成以下任務或是更多：關聯規則(association rules)、分群(clustering)、分類(classification)、次序相關分析(sequential pattern analysis)等[5]。目前已有許多利用資料探勘技術分析交易異常的相關研究[1-3]，其中[1]利用分群技術來偵測信用卡的交易是否異常，[2]利用資料探勘技術來分析個人消費行為，以預測信用卡之詐欺事件，[3]利用複合項關聯規則(association rules with composite items)，提出一個兩階段探勘的方法，來發掘一消費者目前之信用卡交易是否異常。

Agrawal 等人[6]首先提出擷取關聯規則來顯示出項目之間的關聯性，關聯規則的定義說明如下：假設 I 是所有項目的集合，T 是全部交易資料的集合，一筆交易資料 T_j，$T_j \in T$，是由一些項目所形成的集合，稱之為項目組(itemsets)，若一個項目組包含有 k 個項目，稱之為 k-項目組(k-itemsets)，$k \geq 1$，以 $itemset_k$ 表示之。在項目組 X 與 Y 之間有一關聯規則被表示成 $X \rightarrow Y$，X、$Y \subseteq I$ 且 $X \cap Y = \varnothing$，其中 X 稱之為前項目組，Y 稱之為後項目組。有兩個參數 s 與 c 分別為支持度(support)與信賴度(confidence)，用來決定關聯規則是否成立；支持度 s 的定義為：在所有的交易集合中，同時包含有 $X \cup Y$ 的比率值，即 s=(同時包含有 $X \cup Y$ 的交易數量)/(總交易數量)；信賴度 c 的定義為：在包含有 X 的交易集合中，也同時包括有 Y 的比率值，即 c=(同時包含有 $X \cup Y$ 的交易數量)/(包含有 X 的交易數量)。擷取出來的關聯規則，其支持度與信賴度必須大於或等於所指定的最小支持度與最小信賴度，這樣的關聯規則才成立。

關聯規則的探勘過程，主要分成以下兩個階段：首先，找出滿足最小支持度的所有項目組，這些滿足最小支持數量的項目組就稱之為高頻項目組(frequent itemsets)，若某 k-項目組滿足最小支持數量，即稱之為高頻 k-項目組(frequent k-itemsets)，以 $frequent_k$ 表示之；然後，就根據前階段所找出的高頻項目組及以最小信賴度為條件，計算出所有符合的關聯規則。例如 ABC 為高頻 3 項目組，假如關聯規則 AB→C 滿足最小信賴度，則此關聯規則成立，擷取關聯規則的相關研究可參考[4, 7-11]。

在眾多擷取關聯規則的方法中，Apriori 演算法[7]是最具代表性的方法之一，以下我們

說明 Apriori 演算法擷取關聯規則的過程：

(1) 找出高頻(k-1)-項目組，k>1，若為∅，則停止執行。
(2) 由(1)中找出任兩個有 k-2 項目相同的高頻(k-1)-項目組，組合成 k-項目組。
(3) 判斷由(2)所找出的 k-項目組，其所有包括的(k-1)-項目組之子集合是否都出現在(1)中，假如成立就保留此 k-項目組，否則就刪除。
(4) 再檢查由(3)所擷取的 k-項目組是否滿足最小支持度，假如符合就成為高頻 k-項目組，否則就刪除。
(5) 計算高頻 k-項目組所形成的關聯規則，若滿足最小信賴度，則關聯規則成立。
(6) 跳至(1)找高頻(k+1)-項目組，直到無法產生高頻項目組為止。

在本篇論文中，我們從競標者曾經購買過的產品項目中，擷取出項目之間的關聯規則，然後根據關聯規則的消費傾向，從競標者過去競標過的產品項目中，計算項目彼此之間的關聯性，以發掘具有異常競標之行為特徵的競標者。

三、以布林運算為基礎探勘異常之競標者

在探勘關聯規則的方法中，[10]曾經提出一個布林演算法，並證明其執行效率優於 Apriori 演算法，而[4]根據 Apriori 演算法的執行步驟，提出一個以布林運算為基礎的方法，將可提升[10]所描述之演算法的執行效能。在此章節中，我們以競標者之交易資料為探勘的資料來源，其交易資料包含有競標過的產品項目及購買過的產品項目。我們首先從購買過的產品項目中，以布林運算為基礎，找出項目之間的關聯規則，然後再根據關聯規則的消費傾向，計算競標過的產品項目中，項目彼此之間的關聯性，以發掘具有異常競標之行為特徵的競標者。此章節共分為兩小節如下：第一小節中，我們從競標者購買過的產品項目中，找出項目之間的關聯規則，並發掘具有異常競標之行為特徵的競標者；第二小節中，我們以一實例來說明探勘的過程。

(一) 擷取購買產品項目之間的關聯規則

以布林運算為基礎來擷取關聯規則，已經被證明可以有效地提升探勘關聯規則的執行效率[4, 10]。在此一小節中，我們從競標者購買過的產品項目中，首先利用[4]所提出的演算法，擷取產品項目之間的關聯規則。我們說明一些名詞定義如下：

- $I=\{i_1, i_2, \ldots, i_n\}$，是全部項目(items)的集合，共有 n 項。
- $T=\{T_1, T_2, \ldots, T_j, \ldots, T_m\}$，是全部交易資料的集合，共有 m 筆，其中 T_j 為第 j 筆交易資料，$1 \le j \le m$。
- TB_j 為在 T_j 中購買過的產品項目，由 n 位元(bits)所組成，其格式表示成 $TB_j=[b_1, b_2, b_3, \ldots, b_f, \ldots, b_n]$，$b_f \in \{0, 1\}$，$1 \le f \le n$，若有出現第 f 項的項目，則 $b_f=1$，否則 $b_f=0$。
- TA_j 為在 T_j 中競標過的產品項目，由 n 位元(bits)所組成，其格式如同 TB_j。
- $itemset_k$ 表示包含有 k 個項目的項目組，其資料格式如同 TB_j。
- $frequent_k$ 表示包含有 k 個項目的高頻項目組，其資料格式如同 TB_j。
- $Aitemset_j$ 表示若關聯規則之項目組有出現在 TA_j 中，則這些關聯規則之項目組的聯集，其資料格式如同 TB_j。

我們分別使用 *or*(圖 1)、*xor*(圖 2)、及 *and*(圖 3)布林運算(如圖 1)，可以很有效率地分別計算出兩項目組之間位元的聯集、相異的位元、及值為“1”的相同位元。

or	0	1
0	0	1
1	1	1

圖 1

xor	0	1
0	0	1
1	1	0

圖 2

and	0	1
0	0	0
1	0	1

圖 3

我們分別將交易資料中購買過的產品項目及競標過的產品項目，轉換成位元的資料型態，若項目有出現在購買過的產品項目中，或出現在競標過的產品項目中，則對應位元設定為“1”，否則設定為“0”，在每一筆交易資料中，購買過的產品項目及競標過的產品項目，都是各以 n 位元的格式表示之。我們從購買過的產品項目中，以[4]所提出的演算法，來探勘項目之間的關聯規則，其過程說明如下：

(1) 找出 $frequent_{k-1}$，k>1，若為∅，則停止執行。

(2) 任意兩個 $frequent_{k-1}$ 做 *or* 布林運算，假如結果為 $itemset_k$，即有 k 個項目其值為“1”，且非重複者，就保留此 $itemset_k$，否則就刪除[10]。

(3) 判斷由(2)所找出的 $itemset_k$，其包含的 $itemset_{k-1}$ 之子集合，是否都出現在(1)中，可將 $itemset_k$ 與(1)中所有各 $frequent_{k-1}$ 做 *xor* 布林運算，計數結果為 $itemset_1$ 的數目是否等於 k，假如成立就保留此 $itemset_k$，否則就刪除。

(4) 檢查由(3)所擷取出的 $itemset_k$，是否滿足最小支持度，可將 $itemset_k$ 與交易資料 T_j 中的 TB_j 執行以下的運算：

$itemset_k$ *or* TB_j *xor* TB_j，$1 \le j \le m$，

若結果為 $itemset_0$，則表示 $itemset_k \subseteq TB_j$，掃瞄所有 TB_j 之後，判斷出現的次數是否滿足最小支持度，假如符合就成為 $frequent_k$，否則就刪除。

(5) 對 $frequent_k$ 計算可能形成的關聯規則，若前項目組設定為 X，則後項目組 Y 可由以下布林運算找出[10]：

$$Y = frequent_k \text{ xor } X \quad \text{.....(a)}$$

若關聯規則 $X \rightarrow Y$ 滿足最小信賴度，則關聯規則成立。

(6) 跳至(1)找 $frequent_{k+1}$，直到無法產生高頻項目組為止。

根據關聯規則所顯示出的消費傾向，從競標者曾經競標過的產品項目中，計算項目彼此之間的關聯性。我們可執行以下公式，來計算由 $frequent_k$ 所形成的關聯規則是否被包含於 TA_j 中：

$$frequent_k \text{ or } TA_j \text{ xor } TA_j \quad \text{.....(b)}$$

若結果為 $itemset_0$，則表示關聯規則的項目組 $\subseteq TA_j$ 中。我們對每一關聯規則執行公式(b)的運算，並將包含於 TA_j 中之關聯規則的項目組執行 *or* 布林運算，即可計算出有出現在 TA_j 中之關聯規則的項目組的位元聯集，以 $Aitemset_j$ 表示之。例如，{A, B, C, D, E}為全部項目所形成的集合，若關聯規則 A→B 及 A→C 成立，某一競標者 T_j 曾經競標過的產品項目為 ABCD，因為 AB=[11000]⊆[11110]，AC=[10100]⊆[11110]，則 $Aitemset_j$=[11000] *or* [10100]=[11100]，為 ABC。在此我們定義競標者 T_j 之相似度為：

競標者 T_j 之相似度=($Aitemset_j$ *and* TA_j)的項目個數/TA_j 的項目個數。

若競標者之相似度未能滿足所設定的最小相似度，即稱之為異常的競標者。例如若關聯規

則 A→B 及 A→C 成立，則表示競標者購買 A 項目，也會有購買 B 項目或是 C 項目的傾向，若某一競標者曾經競標過的產品項目為 ABCDE，根據以上相似度的定義，因為 AB⊆ABCED、且 AC⊆ABCDE，故其相似度為：ABC 的個數/ABCDE 的個數=3/5=60%，若未能滿足所設定的最小相似度，即稱之為異常的競標者，否則為非異常的競標者。

(二) 實例說明

我們以表 1 之交易資料庫 *D* 來進行分析，*I*={A, B, C, D, E}為產品項目的集合，*T*={T_1, T_2, T_3, T_4}為 4 筆競標者之交易資料的集合，設定最小支持度為 40％(即最小支持數量為 1.6)，最小信賴度為 70%，最小相似度為 60％。探勘具有異常競標之行為特徵的競標者，其探勘的過程說明如下：

表 1、交易資料庫 *D*

交易資料編號	購買過之項目	競標過之項目
T_1	BC	ABCD
T_2	BDE	ABDE
T_3	A	ABC
T_4	BCE	BCE

首先將各競標者購買過的產品項目轉換成位元格式為：TB_1=[01100]、TB_2=[01011]、TB_3=[10000]、TB_4=[01101]。我們從競標者購買過的產品項目中，利用[4]所描述之演算法擷取高頻項目組的過程如下：

	TB_1	TB_2	TB_3	TB_4
A	0	0	1	0
B	1	1	0	1
C	1	0	0	1
D	0	1	0	0
E	0	1	0	1

步驟(4)

itemset$_1$

A	B	C	D	E
1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1
1	3	2	1	2

出現次數

≥1.6

frequent$_1$

B	C	E
0	0	0
1	0	0
0	1	0
0	0	0
0	0	1
3	2	2

出現次數

步驟(2)、(4)

itemset$_2$

BC	BE	CE
0	0	0
1	1	0
1	0	1
0	0	0
0	1	1
2	2	1

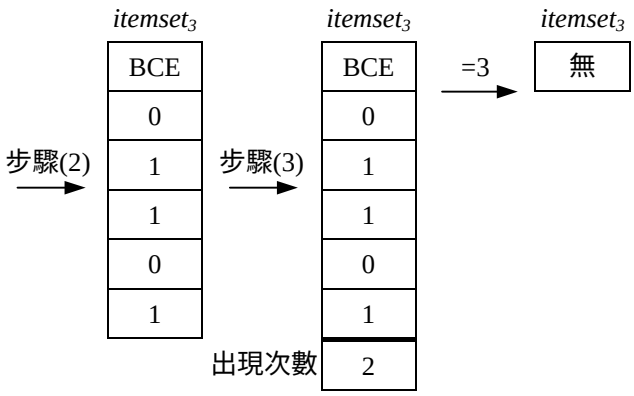
出現次數

≥1.6

frequent$_2$

BC	BE
0	0
1	1
1	0
0	0
0	1
2	2

出現次數



無 3-項目組。

擷取出的高頻 2-項目組有 BC、BE，我們利用公式(a)，分別計算其所形成的關聯規則 B→C、C→B、B→E 及 E→B 的信賴度，滿足最小信賴度者有 C→B 及 E→B。將各競標者曾經競標過的產品項目轉換成位元格式為：TA_1=[11110] 、 TA_2=[11011] 、 TA_3=[11100] 、 TA_4=[01101]。根據相似度的定義，計算各競標者的相似度為：

T_1 之相似度為：[01100] and [11110] = [01100]，因此其相似度=2/4=50%。

T_2 之相似度為：[01001] and [11011] = [01001]，因此其相似度=2/4=50%。

T_3 之相似度為：[01100] and [11100] = [01100]，因此其相似度=2/3=67%。

T_4 之相似度為：[01101] and [01101] = [01101]，因此其相似度=3/3=100%。

經由以上相似度的計算，只有 T_3 與 T_4 滿足最小相似度，因此，具有異常競標之行為特徵的競標者有 T_1、T_2。

四、探勘異常競標者之系統實作

我們將前一章節所描述的演算法，應用到探勘具有異常競標之行為特徵的競標者的系統實作上，我們以C#為撰寫的程式語言，以下為此一系統的探勘過程。

圖4為此一系統的內部資料，包含「會員編號」、「身分證號碼」、「姓名」、「曾經購買產品」、「曾經競標產品」等欄位資料。

會員編號	身份證號碼	姓名	購買產品	競標產品
48890001	N123456789	李小姐	ABD	ABD
48890002	U182178494	張先生	B	AB
48890003	E210912164	洪先生	DE	DE
48890004	C123761256	鄭小姐	D	D
48890005	K224872352	陳先生	C	ACD
48890006	K154676115	吳先生	BC	ABC
48890007	G167373854	李小姐	E	E
48890008	N246479351	蕭先生	D	BD
48890009	E221568757	吳先生	A	AD
48890010	A248734932	洪小姐	BCE	BCE
48890011	H169872949	陳先生	AC	ABCD
48890012	R248628648	謝先生	B	AB
48890013	T167826859	謝小姐	BD	BDC
48890014	P193128956	董先生	CE	BCE
48890015	E119318698	林先生	A	ABC

圖 4、競標者的交易資料

圖 5 表示探勘畫面，可輸入「最小支持度」、「最小信賴度」、「最小相似度」等值。

圖 5、探勘的畫面

圖 6 表示探勘的結果，包含有找出的關聯

規則，經計算之後，發掘具有異常競標之行為特徵的競標者。

圖 6、探勘結果的畫面

五、結論

在電子商務中，網路拍賣是最受消費者歡迎的交易方式之一，隨著交易規模日益擴大與熱絡，也衍生了許多的交易問題，其中最常見的有競標者的欺騙競標行為，競標者只是為了哄抬拍賣品的價格，而不是實際真正的競標者，因此如何偵測出異常的競標者，即成為網路拍賣經營者必須解決的問題之一。在本篇論文中，我們利用關聯規則來發掘具有異常競標之行為特徵的競標者：我們先從競標者過去購買過的產品項目中，以布林運算為基礎，找出產品項目之間的關聯規則，然後從競標者過去競標過的產品項目中，根據關聯規則來計算競標者的相似度，若未能滿足所設定的最小相似度，即稱之為異常的競標者。此探勘結果，對無意購買拍賣品之競標者的喊價行為，將可提供非常有用的預警資訊。

六、參考文獻

[1] 汪昭緯，*應用分群技術偵測信用卡異常交易之研究*，國立中央大學資訊管理研究所，碩士論文，2002。

[2] 黃琮盛，*以個人消費行為預測信用卡詐欺事件之研究*，國立中央大學資訊管理研究所，碩士論文，2001。

[3] 陳垂呈、邱崇兼、黃昱銘，“探勘消費者信用卡之異常交易”，第三屆離島資訊與應用研討會，第 413-417 頁，2003。

[4] 陳垂呈，“以有效率的布林演算法來擷取關聯規則”，*2002 數位生活與網際網路科技研討會*，台南，成功大學，六月，2002。

[5] 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, pp. 866-883, 1996.

[6] R. Agrawal, T. Imielinski, and A. Swami, “Mining Association Rules between Sets of Items in Very Large Database,” *Proceedings of the ACM SIGMOD Conference on Management of Data*, pp. 207-216, 1993.

[7] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules,” *Proceedings of the 20th International Conference on Very Large Databases*, Santiago. Chile, September, pp. 487-499, 1994.

[8] J. S. Park, M. S. Chen, and P. S. Yu, “Using a Hash-Based Method with Transaction Trimming for Mining Association Rules,” *IEEE Transactions on Knowledge and Data Engineering*, Vol. 9, No. 5, pp. 813-825, 1997.

[9] R. Srikant and R. Agrawal, “Mining Generalized Association Rules,” *Proceedings of the 21th International Conference on Very Large Data Bases*, pp. 407-419, 1995.

[10] S. Y. Wur and Y. Leu, “An Effective Boolean Algorithm for Mining Association Rules in Large Databases,” *DASFAA*, 1999.

[11] X. Ye and J. A. Keane, “Mining Association Rules with Composite Items,” *Systems,*

Man, and Cybernetics, Computational Cybernetics and Simulation, IEEE Internaltional Conference, (2) , pp. 1367-1372, 1997.