從交易關係之觀點設計網拍異常帳號分類方法

蕭志豪
元智大學資訊管理學系
博士生
s1019208@mail.yzu.edu.tw

蕭心旻
元智大學資訊管理學系
碩士生
s1016206@mail.yzu.edu.tw

連于萱
元智大學資訊管理學系
碩士生
s1016230@mail.yzu.edu.tw

古永昌
元智大學資訊管理學系
博士生
ycku1230@gmail.com

邱昭彰
元智大學資訊管理學系
教授
imchiu@saturn.yzu.edu.tw

摘要

由於網路拍賣平台具有高匿名程度、不完善的法律規範、低度的市場進入障礙等因素，使得詐騙者可以輕易的進行詐騙行為。在臺灣，拍賣詐財更在官方財務損失排行榜中名列前茅。故本研究以臺灣 Yahoo! 奇摩拍賣為資料來源，利用社會網路分析 (Social Network Analysis, SNA) 的概念提出創新的相似度衡量指標，根據帳號之間的交易關係評估帳號的相似程度。並且以常見的資料探勘分類方法建立預測模型。根據實驗結果顯示，異常類帳號之召回率 ($Recall_{abnormal}$) 最高可達 75.01%，證明研究方法確實有效。

關鍵詞：社會網路分析、資料探勘、網路拍賣詐欺偵測

Abstract

Because of the anonymity, loose constraints, and low entry costs associated with Internet auctions, fraudsters can easily set up Internet auction scams. In Taiwan, the auction fraud also caused high losses. This paper uses experimental data gathered from Yahoo! Auction Web site in Taiwan, and designs four similarity indicators to evaluate the transaction behavior similarity between whole accounts based on social network analysis concepts. Also, this study uses Classification approach to construct abnormal accounts predict models that focus on the type of failure to ship. The testing results based on 5-fold average show the best recall rate of abnormal reached 75.01%. So this methodology can achieve superior classification performance.

Keywords: Social Network Analysis, Data Mining, Internet Auction Fraud Detection.

1. 前言

根據東方快線市調研究部於 2011 年 10 月 26 日所發布關於網友們最常逛的網路購物平台前三名依序為: Yahoo! 奇摩拍賣 (66%) 、露天拍賣 (49%) 及 Yahoo! 購物中心 (48%)，證明 Yahoo! 奇摩拍賣更是當前最常被使用的購物網站。儘管評價系統已經成為保障交易安全的重要關鍵 [10]，網路拍賣詐欺事件卻持續的發生，許多知名、眾多用戶所使用的網拍網站提供的評價機制仍舊極為簡易，經營者也無意願去建立或提供一個完善的評價機制 [9]。因此，設計有效的網路拍賣異常帳號偵測方法是急迫的。

因此，本研究期望提出網路拍賣異常帳號之偵測方法，故以臺灣 Yahoo! 奇摩拍賣為資料來源，以自行開發的網頁爬蟲程式 (Crawler) 從網站中擷取帳號相關資訊後，利用社會網路分析 (Social Network Analysis, SNA) 的概念為基礎，根據帳號間的交易關係設計創新的相似度衡量指標，帳號間交易行為之相似程度，並常見分類方法各自建立預測模型。透過異常類的召回率 ($Recall_{abnormal}$) 進行預測模型評估。實驗結果顯示，平均 $Recall_{abnormal}$ 可以達到 75.01%。結果證明本研究提出之相似度指標 A 結合分類方法用於網路拍賣異常帳號偵測上有一定的成效。

本研究之文章結構如下: 第二章針對網路拍賣詐欺作介紹，並整理近年來於相關的網路拍賣詐欺偵測、社會網路分析等相關研究，從中獲取研究方向與方法的基礎。第三章針對研究方法中各個步驟所使用的方法進行詳細的說明。第四章呈現實際實驗後所得之結果，並針對預測結果進行評估，找出最佳的預測模型。第五章以相似度指標與演算法的角度剖析預測的結果。第六章則提出研究方法之貢獻以及未來改善與研究方向。

2. 文獻探討

本章首先探討網路拍賣詐欺之定義、類型以及相關偵測方法研究，接著根據社會網路分析相關概念進行剖析。

2.1 網拍異常帳號偵測相關研究

過去有研究者 [6] [8] 考量以 SNA 找出群體中由詐欺帳號所組成的犯罪群體具可行性，故利用 k-core 指標配合資料探勘分類演算法，應用於 Yahoo! 奇摩拍賣異常帳號偵測中，偵測正確率至少高於 90%，證明以 SNA 中的 k-core 指標作為網路結構中異常節點的辨識指標能夠有效改善預測結果。另外有學者 [11] 改良 NetProbe 以 Markov Random Field (MRF) 模型為基礎，將帳號以交易關係建構成圖型網路，並且藉由 Belief Propagation algorithm 觀察傳播矩陣 (Propagation Matrix) 以及鄰近節點 (帳號) 的狀態 (詐欺、共犯、正常) 推斷出該帳號詐欺的機率，將其用於 66,000 個擷取自 eBay 的帳號進行實驗，除了執行速度比原本更加快速外，亦保留了 99%的精確度。Almendra & Schwabe (2009) 曾以網拍網站上賣家的大頭照提供給問卷受訪者辨識，期望透過問卷調查可以定義詐欺者共同的臉部特徵，作為未來詐欺偵測時的參考依據 [2]。Chang & Chang (2010) 亦利用實例學習為基礎 (Instance based learning) 的分類方法於 Yahoo! 奇摩拍賣的帳號資料上，其偵測潛在詐欺者的召回率 (Recall rate) 達 84% [4]。Chang & Chang (2011) 利用過去網拍詐欺者的行為週期建立一個兩階段的模型架構，用於台灣 Yahoo! 奇摩拍賣的帳號資料上其 True positive rate 平均達 93% [7]。根據近年來對於偵測網拍詐欺的相關研究顯示，大多數的研究者會採用資料探勘的分類或分群演算法作為主要偵測方法的基礎，而較少研究是應用社會網路分析在異常帳號偵測上。

2.2 社會網路分析 (Social Network Analysis, SNA)

SNA 的關鍵主軸是關聯 (Relation)，主要分為: 群體關聯 (Group relation) 與網路關聯 (Network relation)。群體關聯中的成員互動密切、範圍狹窄、流動性低；網路關聯中的成員互動較少、範圍較廣、具有多變的關聯性、流動性高 [16]。透過關聯性的衡量，可以得知群體的內成員的距離、密度、互動頻率、中心性等 [16]。

Qi et al. (2010) 提出三個以網頁結構中超連結間雙向 (bi-directional) 關聯為基礎的相似度衡量指標，計算網站的相似程度後，應用 Quasi Clique Merge (QCM) 階層式分群演算法於極端政治運動的網站之分群，實驗結果證明新的相似度計算方法可有效幫助網站分群 [13]。其網站之間的相似度計算方式為三個指標值相加總後的值，分別為: 是否直接相連 (direct linking)、是否共同連向 (co-linking)、是否共同被連 (co-linked) 等。其中 direct linking 代表兩個網站之間的是否有互相建立超連結，一個網站通常會建立友站連結以供使用者點選，而那些連結幾乎都會連結至具相同性質的網站，因此，若兩者互有建立超連結連結到對方網站，則兩個網站的 direct linking 指標值越高；co-linking 代表兩個網站建立超連結的相似程度，若兩個網站都曾經建立超連結至同一個網站，則兩者 co-linking 指標值越高；co-linked 代表兩個網站被連結的相似程度，若兩個網站都曾經被同一個網站建立超連結，則兩者 co-linked 指標值越高。計算各個網站之間的三個相似度指標後將三個值進行加總，即成為兩個網站之間的相似程度。該研究考量 SNA 中節點與節點的關聯，利用網站之間超連結的關聯，以及網站間是否存在共同的行為，提出創新的相似度衡量指標，因此，本研究將其計算相似度的概念應用於網路拍賣的交易網路裡，從帳號的交易關係當中，計算帳號之間的交易行為相似程度。

Chiu et al. (2011) 曾利用 2-core 從網路拍賣交易網路中過濾出異常的帳號群，並且經由多種資料探勘分類演算法建立預測模型，其決策樹所建立的預測模型測試結果證明該屬性能夠有效偵測詐欺帳號 [6]。因此，本研究採用帳號間是否存在於同一個 2-core 結構中作為相似度指標。另外，亦會衡量帳號間的買、賣之交易行為相似性也可能顯示出共犯關係或是否進行詐欺行為的跡象。

3. 研究方法

本研究主要分為帳號資訊取得、交易行為相似性衡量、異常帳號偵測等三大步驟。首先以 Yahoo! 奇摩拍賣為資料來源，透過資訊擷取 (Information Retrieval) 技術取得研究所需帳號資訊。並且提出四個相似度指標，透過帳號間的交易關係衡量帳號間交易行為的相似程度產生相似度矩陣，並且配合三種常見的分類演算法各自進行預測模型建立。研究流程如

圖 1 所示。

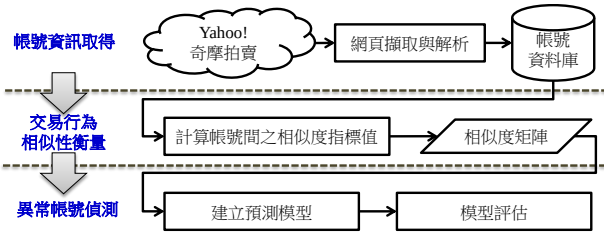

圖 1 研究流程

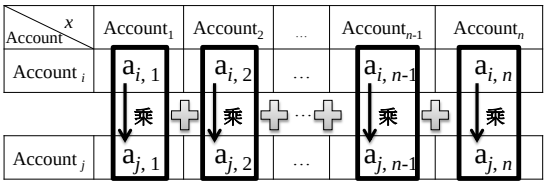

圖 2 購買行為相似性計算範例

3.1 帳號資訊取得

首先經由 Crawler 程式，選定根帳號後，向 Yahoo! 奇摩拍賣網站伺服器要求根帳號所擁有的會員評價資訊網頁。取得其網頁原始碼後，根據研究所需欄位進行解析 (Parsing) 的動作，從網頁原始碼當中，取得所需實驗所需欄位，其中包括有帳號、買賣紀錄之交易對象帳號、交易日期、會員狀態 (是否遭停權) 等資訊。根帳號相關資訊擷取完畢後，程式將向伺服器要求出現於根帳號會員評價資訊中的所有交易對象之會員評價資訊網頁，繼續擷取下一層的帳號相關資訊。重複此程序，直到由根帳號往下擷取交易網路中四層帳號為止。由於 Yahoo! 奇摩拍賣網站防火牆會阻擋同一個網路 IP 位址，當 IP 在沒有登入該網站時，每小時最多只能開啟 500 個網頁，每當超過流量限制，則會傳回網站錯誤訊息，並且封鎖 IP 三小時。故本研究於 Crawler 程式開發時，為了解決此問題，會在程式中加入自行更換 IP 位址的功能，以利實驗樣本收集的進行。

3.2 交易行為相似性衡量

本研究以 SNA 概念為基礎，考量交易歷史紀錄中的帳號與帳號其之間的交易關係，提出以下四個相似度指標:

1.直接連接性: 兩帳號間之相似程度隨著兩者曾經直接交易的次數而遞增。若帳號 i 與帳號 j 之間不曾進行過任何交易，則兩者之間不存在連結關係；反之，則兩者相似度隨之增加。若兩帳號不曾交易過，則指標值為 0；若帳號 i 曾和 j 買過商品或是帳號 i 曾賣過商品給 j，則指標值 θ 為 1；若帳號 i 曾和 j 買過商品而且帳號 i 曾賣過商品給 j，則指標值 θ 為 2。

2.購買行為相似性: 兩帳號間之相似程度隨著兩者曾經向同一個賣家購買過商品而改變。若帳號 i 與帳號 j 之間曾經共同向某特定帳號買過商品，則兩者交易行為相似度越高；反之越低。計算圖示如圖 2:

3.販賣行為相似性: 兩帳號間之相似程度隨著兩者曾經賣過商品給同一個買家而改變。若帳號 i 與帳號 j 之間曾經共同賣過商品給某特定帳號，則兩者交易行為相似度越高；反之越低。計算圖示如圖 3:

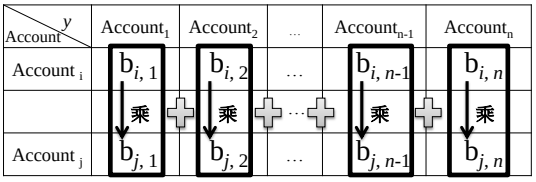

圖 3 販賣行為相似性計算範例

4.共存性: 兩帳號間之相似程度隨著兩者共同存在於一個 2-core 結構中而改變。存在於同一個 2-core 中。由於 SNA 具有辨識出社會網路中子群聚的分析指標 k-core，而 Chiu et al. (2011) 使用 2-core 作為異常帳號群體的過濾方法，其分類預測正確率高於 90%。故本研究採用是否存在於同一個 2-core 為帳號相似度衡量指標之一。若帳號 i 和 j 不存在於同一個 2-core 中，則指標值 c 為 0；反之為 1。

3.3 異常帳號偵測

本研究將透過相似度矩陣，使用 C4.5 決策樹、SVM、Naïve Bayes (NB) 等分類演算法分別進行預測模型之建立，並進行預測結果的比較。三種演算法介紹如下:

1.C4.5 決策樹:

C4.5 決策樹由 Quinlan (1993) 提出，是一種由上而下建立的決策樹 [12]。建立方式是從根節點為起始，透過統計方法從所有屬性當中決定哪些屬性對於將訓練資料集分類是有幫助的，其中最佳的屬性即為根節點。若屬性為離散型，則子節點會由所有可能的值創造；若屬性為連續型，則子節點會由所有可能的離散區間創造。訓練資料集將會經由適當的節點排序，然後不斷的重複測試每個節點是否在訓練

資料集中為最佳的節點 [15]。

2.支援向量機 (Support Vector Machine, SVM):

Cortes & Vapnik (1995) 提出 SVM，使用統計學習理論作為其理論基礎 [3]。傳統的 SVM 將輸入向量映射到高維度特徵空間中，然後在此空間找尋一個能夠將資料分割成兩邊的超平面，並求最大之邊際值 (Margin) 以最小化邊界的錯誤，並且將原始非線性的類別空間建構呈現性模型，以達到分類的目的。SVM 即從空間中找到一個超平面 (optimal hyperplane) 能夠把類別與類別區隔的最好。

3.貝氏分類器 (Naïve Bayes, NB):

本研究使用相似度矩陣進行分類預測模型建立，所有屬性皆為數值型屬性，而 Weka 資料探勘工具中，會透過核心密度估算子推估數值型屬性的機率。

NB 是以貝氏理論 (Bayes theorem) 為基礎的分類演算法，Sahami et al. (1998) 首先將其應用於垃圾郵件的過濾中 [14]。NB 認定訓練資料集中每一個樣本是由多個離散型屬性以及有限集合中的一個類別所組成。NB 可以利用現有的訓練資料集，計算出未知樣本最有可能的類別其出現的機率，並且假設每一個屬性與其類別是各自獨立的 [15]。

在某些領域的應用上，其分類效果優於類神經網路和決策樹。對榆類別型屬性值的資料，NB 會根據訓練樣本，計算每個屬性於各個類別出現的機率值，對於所給予測試資料的屬性值 (a1, a2, a3,……, an) 指派具有最高機率值的類別。

4.實驗結果

實驗數據由自行撰寫的 Crawler 程式於 2011 年 2 月 14 日起，從台灣 Yahoo! 奇摩拍賣網站中，由根帳號往下截取四層帳號相關資訊。而根帳號由 Yahoo! 拍賣中隨機決定。自 2012 年 02 月 14 日由 10 個根帳號往下擷取四層後，總共擷取 578 個帳號，其中 570 個為正常帳號，8 個遭停權。資料描述於表 1 中。

表 1 資料描述

項目	說明
商品類別	無限制
根帳號選擇方式	隨機選擇 10 個帳號
擷取方式	由根帳號往下擷取四層
日期	2012/02/14
樣本數量	8/570
（異常/正常）	

實驗首先以 5-fold 交叉驗證 (Cross validation) 的方式進行。並且將測試資料之預測結果以 5-fold 之平均的異常類召回率 ($Recall_{abnormal}$) 進行評估。將預測結果進行比較。原始測試資料 5-fold 平均測試結果如表 2。以測試資料之 5-fold 平均異常類召回率 ($Recall_{abnormal}$) 進行模型評估，$Recall_{abnormal}$ 可檢視實際為異常的帳號被偵測為異常的比例。實驗結果顯示，預測模型皆無法偵測到異常帳號。無法有效偵測的原因，是因為異常帳號的數量過少，導致訓練模型對於異常帳號的模式訓練不足。

表 2 實驗結果 (原始)

Classifier	$Recall_{abnormal}$
C4.5	0.00%
SVM	0.00%
NB	0.00%

因此，實驗採用 Upsampling 的方式，將原始資料中 8 個異常帳號數量乘以 2 倍，使用 570 個正常帳號與 16 個異常帳號重新進行 5-fold 交叉驗證實驗。實驗結果顯示，所有預測模型皆可偵測到異常帳號，最佳 $Recall_{abnormal}$ 為 NB 的 75.01%。5-fold 平均測試結果如表 3 所示。

表 3 實驗結果 (Upsampling)

Classifier	$Recall_{abnormal}$
C4.5	6.67%
SVM	73.35%
NB	**75.01%**

5.討論

在拍賣市場中，雖然不同帳號的下標或購物習慣都不盡相同，帳號之間的互動方式可能也不一樣，根據實驗結果發現，正常帳號與異常帳號在交易行為上仍然是有所區別的。

本研究創新利用帳號之間的交易行為設計四個相似度計算指標，其中包括帳號之間的直接連接性、購買行為相似性、販賣行為相似性、共存性等四個。這四個指標不僅考量兩個帳號之間的互動行為，也考慮到兩個帳號與其他帳號之間的購買與販賣行為會影響到兩個帳號的行為相似性，另外也加入社會網路分析的概念，辨識兩個帳號是否共同存在於一個小群聚

當中。

然而從實驗結果中最佳 $Recall_{abnormal}$ 達75.01%可得知，透過本研究提出的四個衡量不同面向的相似度指標，可以成功幫助分類方法將大部分的正常、異常帳號區隔，此即代表這兩種型態的帳號在行為上確實不同。以資料集 A 為例，經由實驗結果發現，異常帳號與異常帳號之間共通特性除了交易數量、對象少之外，在實驗所擷取的樣本中，所有的異常帳號皆不共同存在於一個 2-core 中，使得他們的共存性為 0，而正常帳號因其交易範圍廣、交易對象較多，與異常帳號的行為完全不同，使得分類方法可正確偵測出異常帳號。

但如果帳號是網拍平台的新手、鮮少進行交易者或是只跟異常帳號交易過的會員，則會因為與異常帳號的行為相似而遭到預測模型誤判。所以不論是進行交叉驗證實驗或模型驗證實驗，都有少數的正常帳號會被判斷為異常。

6. 結論

本研究以資訊擷取的技術從臺灣 Yahoo! 奇摩拍賣中，擷取帳號的名稱、交易紀錄等資訊，並且提出創新的相似度計算方法，根據帳號之間的交易關係為基礎，考量兩帳號間是否曾經進行過交易、兩帳號是否曾經共同向某特定帳號購買過商品、兩帳號是否曾經共同賣過商品給某特定帳號，加上社會網路分析中的 k-core 指標，設計出四個相似度計算指標衡量帳號間的交易行為相似性。計算完帳號間交易行為相似性產生相似度矩陣之後，再配合 Weka 資料探勘工具中 C4.5, SVM, NB 等三種分類演算法建立預測模型，並且以異常類的召回率 ($Recall_{abnormal}$) 針對異常帳號預測模型進行評估。其交叉驗證 5-fold 平均的 $Recall_{abnormal}$ 可以達到 75.01%。未來將設計新的相似度指標以及相似度計算方式，以減少正常帳號誤判為異常情況。另外，將擷取國內外不同拍賣網站的帳號資訊進行實驗，評估研究方法於實務上之可行性與有效性。最後期望提供給拍賣平台經營者一套完整的異常帳號偵測機制，降低網路拍賣詐欺事件的發生次數，減少平台使用者金錢損失。

參考文獻

[1] 東方快線研究部，「Yahoo! 奇摩拍賣」讓網友百訪不厭，2011。

[2] Almendra, V. and Schwabe, D., “Fraud Detection by Human Agents-A Pilot Study,” ***E-Commerce and Web Technologies***, pp.300-311, 2009.

[3] Cortes, C. and Vapnik, V., “Support-Vector Networks,” ***Machine Learning***, Vol. 20, Issue 3, pp.273-297, 1995.

[4] Chang, W.H. and Chang, J.S., “Using Clustering Techniques to Analyze Fraudulent Behavior Changes in Online Auctions,” ***International Conference Networking and Information Technology (ICNIT)***, 2010 on, pp.34-38, 2010.

[5] Chang, W.H. and Chang, J.S., “An Online Auction Fraud Screening Mechanism for Choosing Trading Partners,” ***International Conference on Education Technology and Computer (ICETC)***, V5-56-V5-60, 2010.

[6] Chiu, C., Ku, Y., Lie, T. and Chen, Y., “Internet Auction Fraud Detection Using Social Network Analysis Classification Tree Approaches,” ***International Journal of Electronic Commerce***, Vol. 15, No. 3, pp.123-147, 2011.

[7] Chang, W.H. and Chang, J.S., “A novel two-stage phased modeling framework for early fraud detection in online auctions,” ***Expert Systems with Applications***, Vol. 38, Issue 9, pp.11244-11260, 2011.

[8] Ku, Y., Chen, Y.C. and Chiu, C., “A Proposed Data Mining Approach for Internet Auction Fraud Detection,” ***Proceedings of the 2007 Pacific Asia conference on Intelligence and security informatics***, pp.238-243, 2007.

[9] Liu, X., Kaszuba, T., Nielek, R., Datta, A. and Wierzbicki, A., “Using stereotypes to identify risky transactions in Internet auctions,” ***Social Computing (SocialCom), 2010 IEEE Second International Conference on***, pp.513-520, 2010.

[10] Morzy, M., Wierzbicki, A. and Papadopoulos, A.N., “Mining online auction social networks for reputation and recommendation,” ***Control and Cybernetics***, Vol. 38, No. 1, pp.87-106, 2009.

[11] Pandit, S., Chau, D.H., Wang, S. and Faloutsos, C., “NetProbe A Fast and Scalable System for Fraud Detection in Online Auction Networks,” ***International World Wide Web Conference Committee (IW3C2)***, pp.201-210, 2007.

[12] Quinlan, J.R., “C4.5: Programs for machine learning.,” *Morgan Kaufmann Publishers Inc.*, San Francisco, 1993.

[13] Qi, X., Christensen, K., Duval, R., Fuller, E., Spahiu, A., Wu, Q. and Zhang, C.Q., "A Hierarchical Algorithm for Clustering Extremist Web Pages," ***International Conference on Advances in Social Networks Analysis and Mining (ASONAM***, pp.458-463, 2010.

[14] Sahami, M., Dumais, S., Heckerman, D. and Horvitz, E., "A Bayesian Approach to Filtering Junk E-Mail," ***AAAI Workshop on Learning for Text Categorization***, Wisconsin. AAAI Technical Report WS-98-05, 1998.

[15] Setsirichok, D., Piroonratana, T., Wongseree, W., Usavanarong, T., Paulkhaolarn, N., Kanjanakorn, C., Sirikong, M., Limwongse, C. and Chaiyaratana, N., "Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening," ***Biomedical Signal Processing and Control***, Vol. 7, Issue 2, pp.202-212, 2012.

[16] Wang, J.C. and Chiu, C.C., "Recommending trusted online auction sellers using social network analysis," ***Expert Systems with Applications***, Vol. 34, Issue 3, pp.1666-1679, 2008.