社會網絡分析之研究–以網際網路搜尋爲例

王朝煌、彭議霆

中央警察大學資訊管理研究所

摘要：二十世紀以來資訊科技突飛猛進，尤其在電腦與網路結合後，網際網路漸漸成爲人類資訊活動最主要的傳輸媒體。人類行爲涉及網際網路的比例日漸增加，生活方式也隨著改變，以網路活動爲主的虛擬社會應運而生。類似於傳統的犯罪者會在犯罪現場留下物理跡證，如指紋、血跡、毛髮等等，虛擬社會的犯罪者往往也會在電腦和網路中留下電子跡證。如何分析電子跡證，有效掌握人類的資訊活動，以輔助資訊時代的犯罪偵查工作，實爲資訊時代極重要的課題。本研究運用功能強大的 Google 搜尋引擎進行特定人物相關資料的搜集，再藉由中央研究院的中文斷詞系統萃取與該特人物相關之人物名稱，然後以社會網絡圖形介面顯示相關人物間的關係。實驗顯示本研究之分析技術可以發現犯罪份子的關聯及其社會網絡關係。

關鍵詞：關聯分析、社會網路分析、網際網路搜尋、斷詞技術

綱要

一、緒論

根據洛卡得物質交換理論(Locard’s Exchange Principle)[1]，犯罪者往往會在犯罪現場留下物理跡證，如指紋、血跡、毛髮等等，偵查人員因而可依據犯罪現場的痕跡、物證的位置和狀態的分析及物證的檢驗，從而確定或排除在犯罪現場發生的事件及行爲過程，以推斷犯罪者與犯罪現場的關連[2]。資訊科技發明以前，人類無形的資訊活動只能透過面對面的言語交換，面對面的資訊交換行爲雖然可能間接造成和留下物理證據狀態的改變，但資訊活動只能在當事人腦海中留下記憶或印象，因此這方面的證據只能透過詰問審訊的方式獲得，輔以物理證據的佐證與邏輯推論，證明其真實性。因此傳統的犯罪偵查工作主要以犯罪現場的物理證據作爲重建犯罪事實的基礎。

在電信科技發明以後，人類可以借由通信科技進行點對點的遠距通訊，面對面的言語交換不再是(無形)資訊唯一的交換方式，通訊活動除了在當事人腦海中留下記憶外，其間接在通信設施所留下的通聯紀錄也成爲犯罪偵查重要輔助證據，因此運用通聯記錄的關聯分析(link analysis)成爲犯罪偵查最重要的工具之一。此外犯罪偵查單位更進一步運用電信監聽技術，直接掌握當事人的通信內容，犯罪偵查的證據蒐集從而超越了物理證據的範圍。

二十世紀以來資訊科技突飛猛進，尤其在電腦與網路結合後，網際網路漸漸成爲人類資訊活動最主要的傳輸媒體。人類不但可以透過資訊高速公路，悠遊於網際網路所建構的虛擬世界，更可以互動式的電腦網路媒體作爲橋樑，形成虛擬社群[1]。然而由於電腦驚人的處理效率以及網路活動具跨越時空、隱密、及匿名等特性，使得有心人或犯罪者，不但可以利用網際網路作爲作案管道，並且可以利用其隱密及匿名等特性，避免執法機關的偵察與逮捕，致使網路世界逐漸淪爲犯罪淵藪。近年來網路詐欺、網路援交、網路金融犯罪、以及侵犯智慧財產權等犯罪問題不斷地成長，即爲明顯的實例。

類似於傳統的犯罪者會在犯罪現場留下物理跡證，如指紋、血跡、毛髮等等，電腦相關犯罪者往往也會在電腦和網路中留下電子跡證(electronic trail)。如何運用電子跡證的分析技術，有效掌握人類的資訊活動，以輔助資訊時代的犯罪偵查工作，實爲資訊時代極爲重要的課題。例如在傳統社會，偵查人員可經由巡邏查察及透過各種監視設施，窺知犯罪徵候，進而加以監控查察。然而傳統的巡邏查察及監視設施，在社會漸漸轉型爲虛擬社群的過程中，將漸漸失去其效用，亟待研擬新的監控查察方法。雖然聊天室、電子郵件、及網路（視訊）電話等通訊乃憲法秘密通訊自由所保障的私領域行爲，必須在法律的授權下，才可進行監控。但網站、論壇（部落格）、及網路遊戲等公共的空間，調查人員可視工作需要進行必要的資料蒐集與監控。

網際網路的資料雖然甚爲龐雜，但其分析應用可區分爲兩類:當事人主動參與的社會活動，如共同選課、同列榜單、同學錄、共同著作出版…等資料；當事人相關報導，即由第三者(如記者)對當事人的報導，如新聞報導、網路日誌、傳記、法院判決（例）書…等資料。前項可作爲社會網絡分析資料，而後者則可以關聯分析技術加以應用。本研究運用功能強大的 Google 搜尋引擎，搜尋網際網路公共領域(public domain)的資料，搜集特定人物的相關資料，並藉由中央研究院之中文斷詞系統萃取與該特人物相關的人物名稱，再以圖形介面顯示相關人物間的關係，期能作爲監控查察之參考。

[1] 根據學者 Wally Bock 的定義，形成社群一般具共同僻好（common interests）、互動頻繁(frequent interaction)、及互相認識(identification)等特色[3]。虛擬社群乃以網站、論壇（部落格）、聊天室、電子郵件、網路（視訊）電話、網路遊戲等電腦媒體管道，所促成的社群組織。社群成員間通常透過網站或論壇發表文章，以交換心得，討論共同主題，及尋求同儕的支持，並以別名或電子郵件作爲彼此的識別，經由互動漸漸形成以電腦網路爲媒體的虛擬社群。

二、文獻探討

（一）網際網路搜尋（Internet Search）

網際網路搜尋引擎一般以網路蜘蛛(web spider)，定期蒐集網際網路上的資料，建立資料庫與索引資料及提供搜尋服務。索引最基本方法有兩種，分別是以「關鍵字」及以「概念」爲索引[4]。關鍵字索引是目前最常使用的方法；以概念爲基礎的索引，則透過人工彙整定義概念，逐一檢查文件，以確定文件所屬的範圍做成索引。如使用者輸入：「全世界最高的山」，搜尋引擎會由資料庫定義中，找出符合該定義的山爲「喜馬拉雅山」。

Google 搜尋引擎是由二名史丹福大學博士生，拉里·佩奇和謝爾蓋·布林在 1996 年所創立，除資料的索引分析外，網站間關係的分析與運用，也是 Google 搜尋引擎的重要特色。Google 搜尋引擎是目前網際網路上最大、影響最廣泛的搜索引擎。Google 搜尋引擎每日處理超過 2 億次查詢。除了網頁外，Google 搜索引擎也提供搜尋圖像、新聞網頁、影片、地圖、部落格等服務，Google 搜索引擎其他服務有：Google 網上論壇和 Google 圖片搜索服務、Google 新聞、Google 網頁目錄、Google Answers、Froogle、Google Web API、Google Book Search、Picasa、Google、Notebook、Google Maps、Google Earth、Google SketchUp、Google Moon、Google Local、Google Mars 等等[5]。

（二）斷詞系統

字是最基本的語義單位，字與字的組合及構成詞，可以延伸字的表達範圍。例如「蜻」與「蜓」雖各具基本語意，組合成「蜻蜓」一詞，便可以描述另一新的動物。人類善於組合既有的概念，以描述新的現象或事物。詞與詞或詞與字又可加以組合，例如「竹」與「蜻蜓」組合成「竹蜻蜓」，及「電子」與「計算機」組合成「電子計算機」等等，可以描述嶄新的事物。詞可分爲單字詞、多字詞，如「紙」、「小說」[6]。文件資料的自動化處理非常複雜，爲簡化處理程序，通常先將文件資料進行斷詞，再接續語言分析、資訊抽取、資訊檢索等工作。因此中文自動分詞的工作成了語言處理不可或缺的技術。當處理不同領域的文件時，領域相關的特殊詞彙或專有名詞，常常造成分詞系統因爲參考詞彙的不足而產生錯誤的切分。爲了解決這個問題，最有效的方法是補充領域詞典加強詞彙的搜集[7]。斷詞系統可分爲五大單元，分別爲前置處理單元、構詞單元、斷詞單元、後置構詞單元及詞類標記產生單元。前置處理單元將輸入的字串做處理，如字串中有英文、標點符號或其他符號時，將字元轉換爲全形，再輸入構詞單元。斷詞單元利用斷詞規則，經詞庫比對原字串後而結合的詞，再交由後置構詞單元檢查詞句中是否可以互相結合，最後將斷詞結果產生所要的詞與詞類標記[8]。目前中文斷詞，主要有詞庫斷詞法、統計斷詞法及混合式斷詞法[9,10]，彙整如表 2.1。

表 2.1 中文斷詞法

方法	說明
詞庫斷詞法	利用詞庫和文件中的句子做字詞比，以找出文件內字詞。爲保持詞庫斷詞正確性，詞庫內容須時常維護與更新。
統計斷詞法	從大量領域的文件資料庫中分析統計，以找出鄰近字元共同出現的頻率及前後字之分佈情形作爲斷詞依據。例如某一系列的字 ABCD 共同出現的頻率頗高，且 A 之前及 D 之後出現的字則較爲多樣化，則 ABCD 爲一詞的可能性較高。
混合式斷詞法	利用詞庫斷出不同的組合字詞，再利用字詞的統計資訊，找出最佳的斷詞組合。

本研究運用中央研究院中文斷詞系統部份功能，該系統包含約拾萬詞的詞彙庫及附加詞類、詞頻、詞類頻率、雙連詞類頻率等資料。根據中研院斷詞系統說明指出，每篇文章中約有 3%~5%的詞彙是未知詞，文章經過斷詞後，所取出的詞性有數十種，部份詞性說明如表 2.2 所示，例如專有名稱（Nb），「巨蛋棒球場」、「貓空纜車」皆屬之[7]。

表 2.2 中央研究院中文斷詞系統部份詞性列表

標記	對應詞類	
Na	Naa, Nab, Nac, Nad, Naea, Naeb	普通名詞
Nb	Nba, Nbc	專有名詞
Nc	Nca, Ncb, Ncc, Nce	地方詞
Ncd	Ncda, Ncdb	位置詞
Nd	Ndaa, Ndab, Ndc, Ndd	時間詞

（三）社會網路分析（Social Network Analysis）

社會網路指將人們連結在一起的社會關係網路，並利用社會圖（sociogram），以點表示成員，以線表示成員間的關係，呈現這些社會組態的屬性，衡量社會凝聚力和社會壓力[11]。社會網絡分析方法包括核心措施、鑑定小組、分析角色、圖論基礎、排列型

統計分析等等。傳統社會網路分析法通常透過蒐集問卷、訪談、觀察蒐集社會網路資料，並運用圖論（graph theory）作進一步分析與解釋。社會網路分析是一種強而有力的分析方法，依節點、單雙向關係，以及關係的強度(如共同出現次數)等，可分析在特殊領域成員間彼此的關係。

1.節點「Vertex」，以 V 表示，圖形爲「●」，本文中以一個節點代表某一特定人物名稱，各節點間可能存在直接或間接關係，也可能不存在任何關係。

2.社會網路分析通常以矩陣 A 來表示，以 A_{ij} 表示點 i 和點 j 間的關係，A_{ij} 值爲 0 表示點 i 和點 j 間沒有關係，否則表示點 i 和點 j 間有關聯。若以圖形顯示，點(人物)與點間的關係「Edge」，以 E 表示：$A \rightarrow B$ 代表 A 與 B 爲單向關係，$A \leftrightarrow B$ 代表 A 與 B 爲雙向關係。

本研究社會網路分析使用軟體爲 UCINET。UCINET 可讀寫各種不同的檔案格式，主要用來分析所搜集的數據，及以一維和二維的方式分析資料。UCINET 包括 NetDraw、Spreadsheet、Mage、Pajek 等四大項功能，本研究分別運用其 Spreadsheet 及 NetDraw 兩功能以建立關係矩陣及繪製社會網路分析圖形。

三、實驗設計

本研究以 Google 搜尋引擎進行特定人物(或稱關鍵人)相關資料，並藉由中央研究院的中文斷詞系統萃取與該特人物相關之人物名稱，所蒐集的相關之人物名稱再運用 Google 搜尋引擎進行作第二層的搜集，之後彙整關係矩陣資料，最後以圖形介面顯示相關人物間的關係，本研究的實驗設計與流程如圖 3.1 所示：

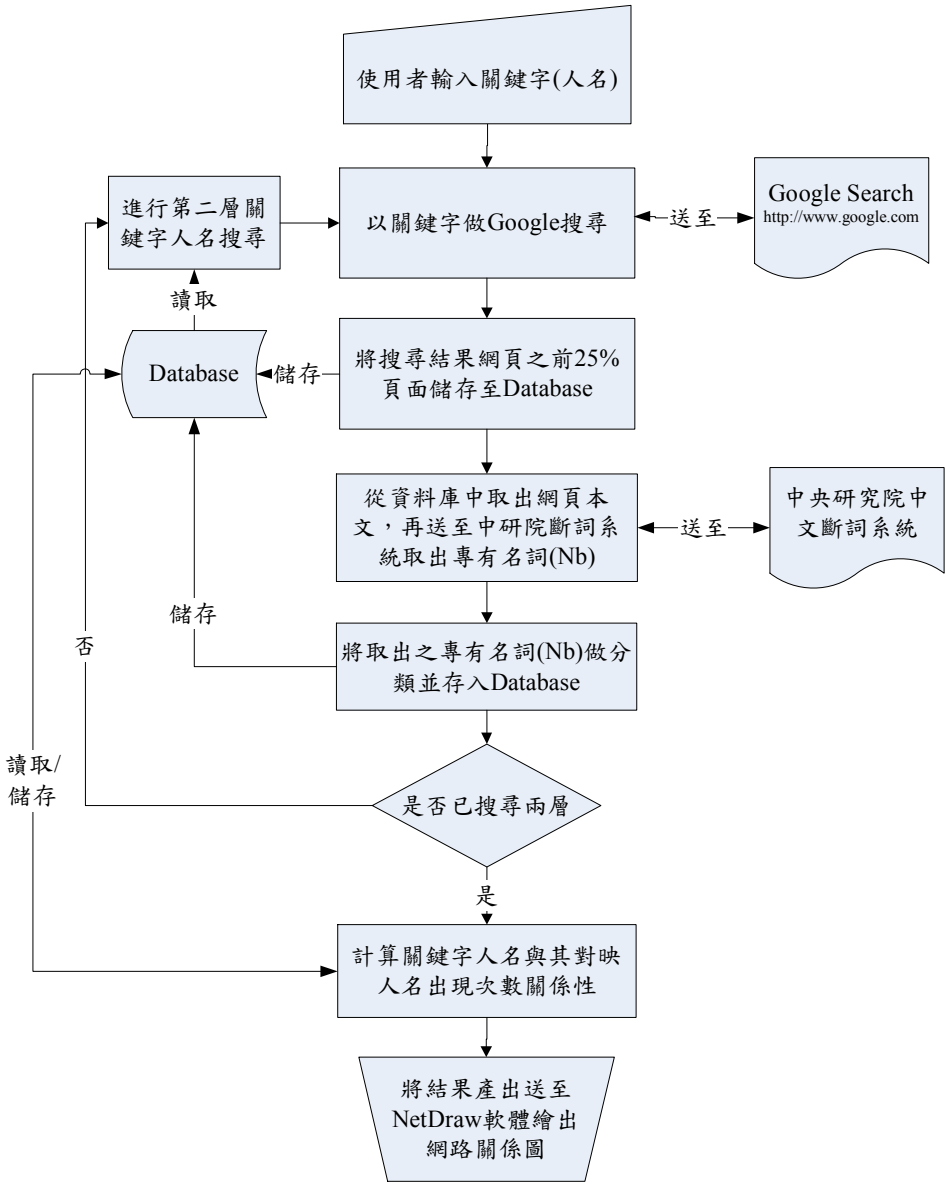

圖 3.1 實驗設計與流程

本研究使用自行開發之代理人程式，使用者輸入欲分析特定人物之名稱後，將其送至 Google 搜尋相關的網頁，取出前 25%[2]熱門網頁之 html、htm、txt 檔，並將內文儲存至資料庫，同時將資料送到中央研究院中文斷詞系統進行斷詞，萃取專有名詞（Nb）部份，再將它回存至資料庫中。第一層的搜尋與分析，找出與欲分析特定人物有關之人物名稱。接著進行第二層之搜尋，即將第一層搜尋所得之相關人物名稱，一一送至 Google 搜尋相關的網頁，並進行相關人物名稱的萃取與統計共同出現的次數。第二層的搜尋與分析可進一步找出特定人物與相關之人物相互間共同出現的關係。以張錫銘爲欲分析特定人物爲例進行第一層搜尋，所得資料經萃取人物名稱之後，與「張錫銘」重要相關（排名前 25%）的人物名稱有「段樹文」、「林慶益」、「洪俊彥」、「鄭進富」等五人。此五人物名稱再一一送至 Google 搜尋引擎，作第二層搜尋及萃取資料人物名稱，例如

[2] 依資料內提及欲分析特定人物之名稱的次數多寡排序。

與「林慶益」重要相關的人物名稱爲「張錫銘」、「鄭進富」、「林泰亨」等三人，如表 3.1 所示。

表 3.1 資料庫部份內容

資料編號	來源	欲分析特定人物名稱	相關人物名稱	詞性
1	自由電子報-社會新聞	張錫銘	段樹文	(Nb)
2	自由電子報-社會新聞	張錫銘	林慶益	(Nb)
3	自由電子報-社會新聞	張錫銘	林慶益	(Nb)
4	自由電子報-社會新聞	張錫銘	洪俊彥	(Nb)
5	自由電子報-社會新聞	張錫銘	鄭進富	(Nb)
6	自由電子報-社會新聞	張錫銘	王靖壹	(Nb)
7	自由電子報-社會新聞	張錫銘	林慶益	(Nb)
8	自由電子報-社會新聞	張錫銘	林慶益	(Nb)
9	臺頭號悍匪張錫銘被判三個無期及 55 年徒刑	林慶益	張錫銘	(Nb)
10	臺頭號悍匪張錫銘被判三個無期及 55 年徒刑	林慶益	張錫銘	(Nb)
11	臺頭號悍匪張錫銘被判三個無期及 55 年徒刑	林慶益	鄭進富	(Nb)
12	臺頭號悍匪張錫銘被判三個無期及 55 年徒刑	林慶益	林泰亨	(Nb)
13	臺頭號悍匪張錫銘被判三個無期及 55 年徒刑	林慶益	鄭進富	(Nb)

接著將表 3.1 及統計資料轉換成共同出現矩陣，如表 3.2 所示，例如當以「張錫銘」

爲特定人物名稱進行搜尋時時，「林慶益」共同出現在與「張錫銘」相關資料前 25%中的次數爲 4，「段樹文」爲 1，餘依此類推。而當以「林慶益」爲特定人物名稱進行搜尋時時，「張錫銘」共同出現在與「林慶益」相關資料前 25%中的次數爲 2，「鄭進富」亦爲 2，餘依此類推。

表 3-2 轉換後部份關係矩陣

	張錫銘	林慶益	段樹文	洪俊彥	鄭進富	王靖壹	林泰亨
張錫銘	0	4	1	1	1	1	0
林慶益	2	0	0	0	2	0	1

最後，將轉換完之關係矩陣資料表匯入 UCINET 軟體加以描繪，並利用該軟體功能分析出各種可能存在的社會網路關係。

四、實驗結果與討論

將資料以 UCINET 軟體之 NetDraw 繪製社會網路關係圖如圖 4.1 所示。節點代表各個關係人物名稱，圖 4.1 中節點總數爲 131 個。點與點間的線段代表關係的存在，關係又可分爲「單向關係」與「雙向關係」，舉例來說：若搜尋人物 A 的相關資料前 25%包含人物 B，但搜尋人物 B 的相關資料前 25%未包含人物 A，則人物 A 與人物 B 之關係爲單向關係，反之亦同。但若搜尋人物 A 的相關資料前 25%包含人物 B，搜尋人物 B 的相關資料前 25%也包含人物 A，則人物 A 與人物 B 爲雙向關係。此外連接兩節點的線段，兩邊皆會有彼此共同出現次數，以表示相對的關係強度。

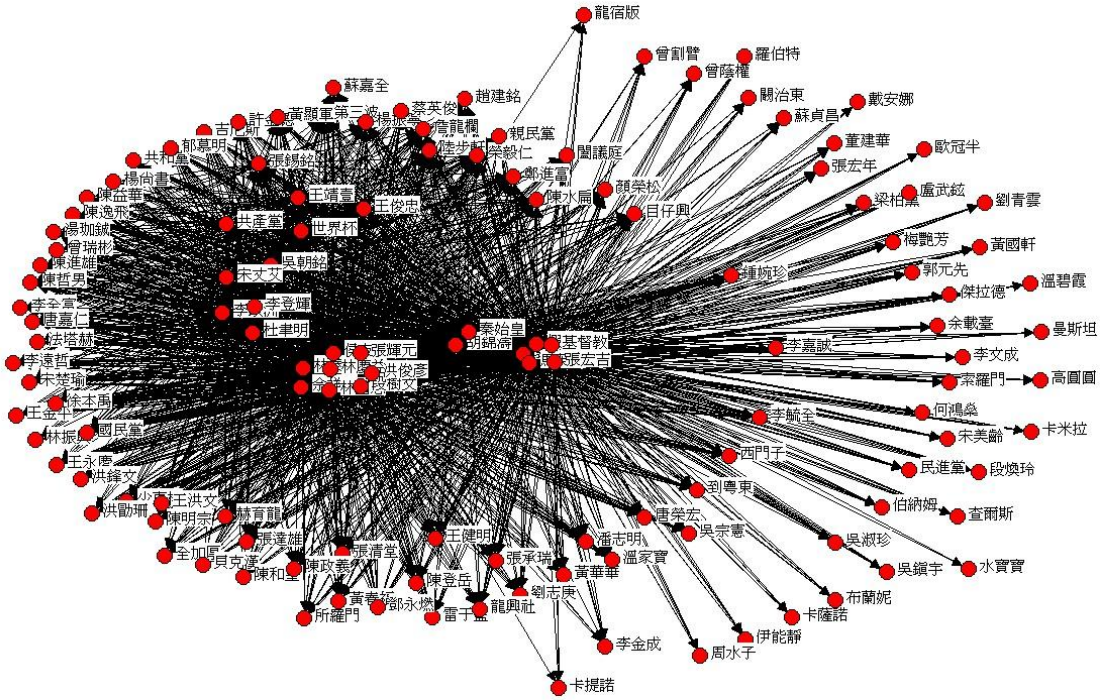

圖 4.1 張錫銘人際網絡關係分析圖

圖 4.1 中，「西門子」、「索羅門」、「到粵東」雖不是人物名稱，但屬於專有名詞，斷詞系統在斷詞過程中將該詞列爲專有名詞（Nb）中。這些較無相關的名稱可以調整顯示的關係強度門檻值篩去，例如將圖 4.1 關係圖中關係強度小於 5(約等於平均關係強度 4.84)的線段略去，可將關係圖簡化爲圖 4.2。此外圖 4.2 中線段以紅色表示雙向關係，黑色表示單向關係。另外，也以節點大小表示人物名稱出現次數之多寡。由圖 4.2 可明顯看出張錫銘與林國忠、段樹文、林泰亨有直接關聯，與侯友宜、楊尚書、蔡英俊則有間接關係。分析出現雙向關係原因如下：

1、 林國忠：爲張錫銘綁架集團黨羽。

2、 林泰亨：爲張錫銘綁架集團黨羽。

3、 段樹文：在 2004 年底要張錫銘向媒體投書放話。

出現與張錫銘有間接關係之姓名分析如下：

1、 侯友宜：案發當時擔任刑事局長，負責指揮專案小組成員偵辦。

2、 楊尚書：和欣客運小開，被張錫銘綁架。

3、 蔡英俊：承辦張錫銘案件檢察官。

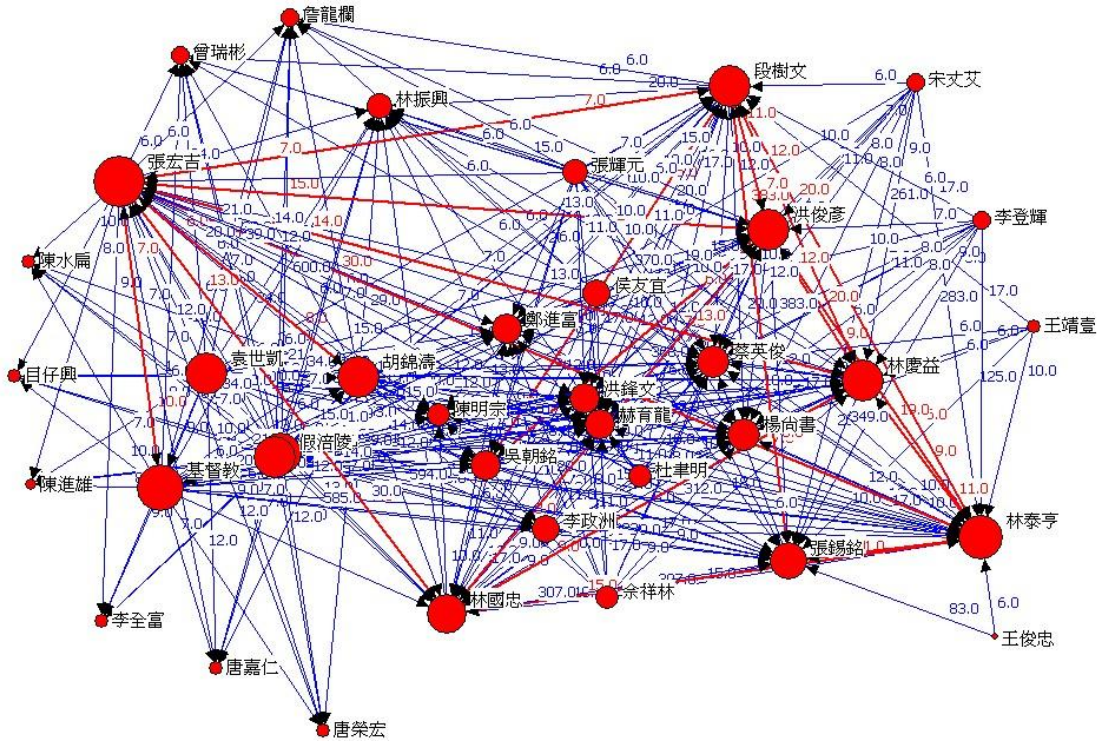

圖 4.2 簡化之張錫銘人際網絡關係圖

如將顯示關係強度門檻值調高爲 10，進一步簡化張錫銘之人際網絡關係，可得圖 4.3。節點明顯減少，可發現張錫銘犯罪集團黨羽、與較重要的受害者與偵辦人員間的關係。

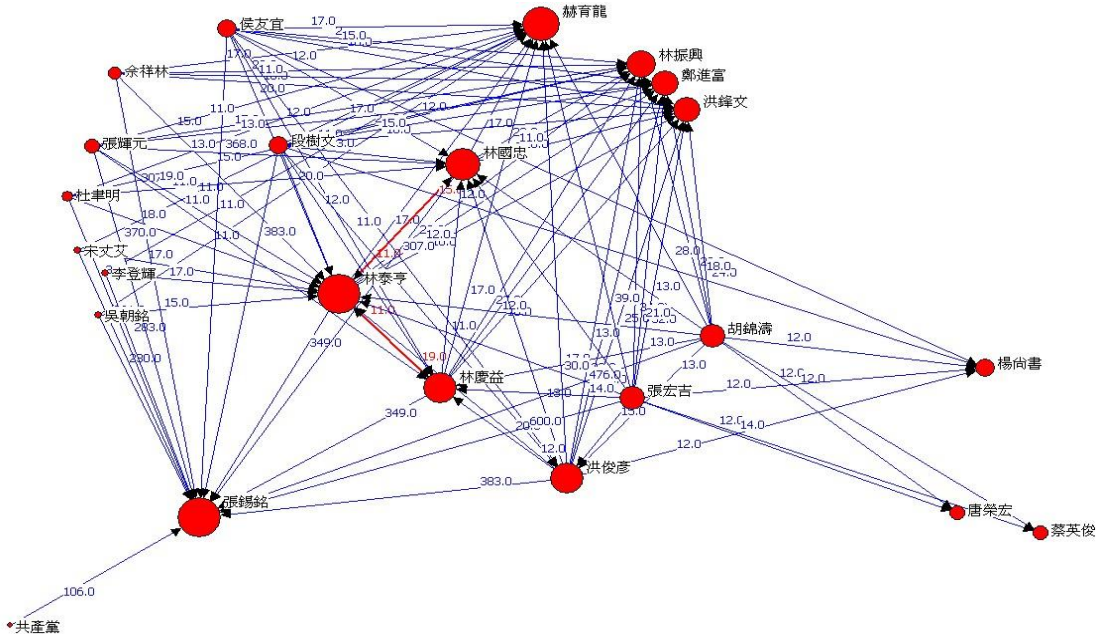

圖 4.3 關係強度大於 10 的張錫銘人際網絡關係圖

最後，具雙向關係且關係強度大於 5 的張錫銘人際網絡關係，如圖 4.4 所示。與張

錫銘具雙向關係且關係強度大於 5 的關係人有「洪俊彥」、「張宏吉」、「段樹文」、「林慶益」、「吳朝銘」、「段樹文」、「林國忠」等七名，由表 4.1 可知這七名人物與張錫銘間的特殊關係。彼等在資料庫中出現的次數都超過 100 次以上，關聯性相當高，此七名成員分別爲黨羽、受者者及主要的偵查人員。

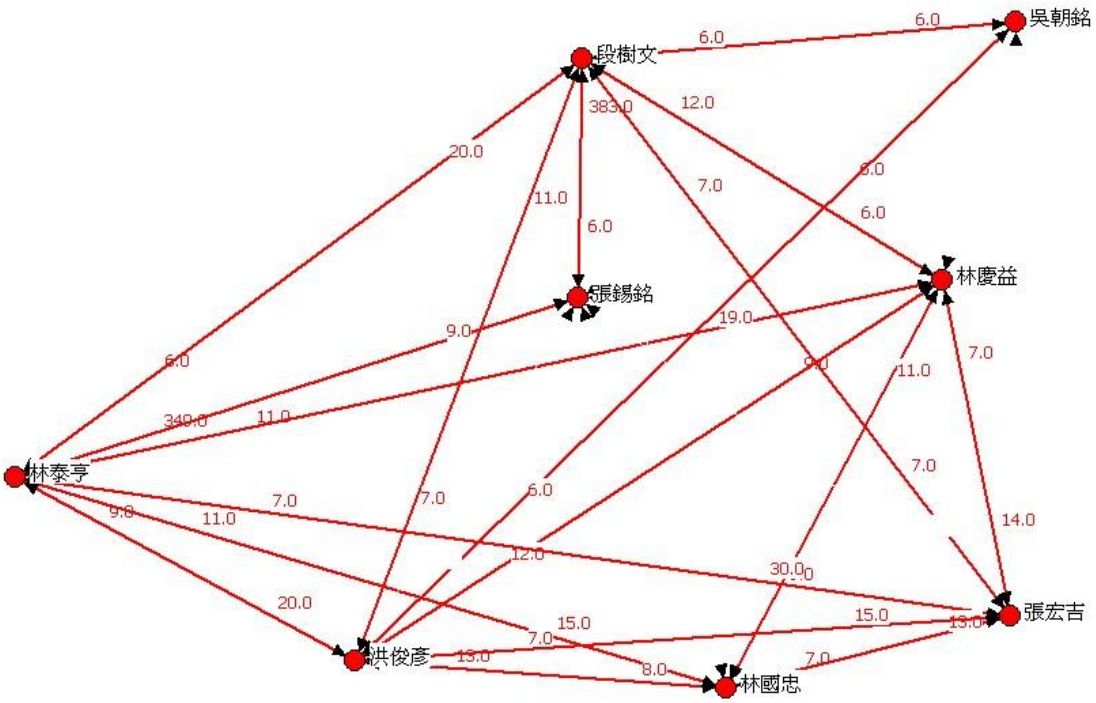

圖 4.4 雙向且關係強度大於 5 的張錫銘人際網絡關係圖

表 4.1 張錫銘綁架集團相關關係人說明

主要關係人	關係人背景說明	資料庫關鍵詞出現次數
林國忠	張錫銘綁架集團黨羽	227 次
陳進雄	張錫銘綁架集團黨羽	101 次
唐榮宏	張錫銘綁架集團黨羽	124 次
張宏吉	張錫銘綁架集團黨羽	140 次
吳朝銘	張錫銘綁架集團黨羽	123 次
曾瑞彬	張錫銘綁架集團黨羽	132 次

詹龍欄	詹龍欄綽號穿山甲，張錫銘爲其小弟	104 次
洪俊彥	南投電台主持人洪俊彥，張錫銘犯下之綁架案件	216 次
林振興	爲薛球及張錫銘等擄人勒贖犯罪集團幕後藏鏡人首腦的台中角頭	418 次
鄭進富	台南市當舖老闆鄭進富，張錫銘犯下之綁架案件	318 次
洪鋒文	張錫銘懷疑被高雄角頭洪鋒文出賣，投書媒體對其發出狙殺信	368 次
赫育龍	電玩老板于國柱遭張錫銘綁架，被認爲是幕後人物之一	436 次
蔡英俊	蔡英俊，承辦張錫銘綁架案件台南地檢署檢察官	242 次
楊尚書	和欣客運小開楊尚書，張錫銘犯下之綁架案件	236 次
林慶益	新營市東方酒店股東，被張錫銘於 84 年 2 月 20 日凌晨，在新營市東方酒店停車場，開槍殺害	243 次

五、結論

本研究運用功能強大的搜尋引擎，分層搜尋特定人物之網際網路資料，再藉由中央研究院中文斷詞系統萃取出網路資料中的人物名稱，統計分析特定人物與相關人物間的關係與強度，並運用社會網路分析軟體，繪製其社會網路關係圖。並以「張錫銘」綁架集團爲例，說明本研究的實驗結果。由張錫銘人際網絡關係圖發現與張錫銘有關的人物，包括黨羽、受害者及偵辦人員，都可從網際網路搜尋的資料分析獲得。可見經由網際網路搜尋與本研究所研擬的分析方法，可以發現犯罪份子間的關聯與其社會網絡關係。由於許多犯罪者爲累犯，警察機關可整合內部之資訊系統，如：犯罪資料庫、筆錄資料庫，再輔以本研究所提出之社會網路分析方法，以掌握更詳盡的相關資訊，以提高打擊不法犯罪份子之效能。

謝誌

本文感謝中央研究院詞庫小組、Google 搜尋引擎及 UCINET 軟體等提供的協助，使本研究實驗得以順利進行，特此致謝。

參考文獻

[1] Richard **Saferstein**, *Criminalistics—An introduction to Forensic Science*, 6th edition, Prentice Hall, 1998.

[2] H. C. Lee, *Crime Scene* **Investigation**, Central Police University Press, Taoyuan, Taiwan, ROC.

[3] Bock, W., 2001, "*Christmas*, Communities, and Cyberspace", http://www.bockinfo.com/docs/community.htm.

[4] 搜尋引擎的原理是什麼，http://blog.sina.com.tw/googleadsense/article.php?pbgid=22525&entryid=26，2006。

[5] http://zh.wikipedia.org.

[6] 黃居仁、陳克健，"中央研究院漢語料庫的內容與說明"，http://www.sinica.edu.tw/SinicaCorpus/98-04.pdf，1998。

[7] http://ckipsvr.iis.sinica.edu.tw.

[8] 唐大任，中文斷詞器之研究，國立交通大學電信工程學系碩士論文，1999。

[9] 賴錦慧，*新聞事件之變化探勘以支援決策制定*，國立交通大學資訊管理研究所碩士論文，2004。

[10] 王朝煌，2004，"資料分析技術與情報運用之探討"，*通識教育教學及研究方法學術研討會論文集*。

[11] 邱議德，以社會網路分析法評估工作團隊知識創造與分享，國立中正大學資訊管理所碩士論文，2003。