國立交通大學

資訊學院　資訊學程

碩 士 論 文

基於電子郵件通聯之社交網路探勘

Mining Social Relationships from E-mail Logs

研 究 生：詹大偉

指導教授：彭文志　教授

中 華 民 國　九 十 九 年 一 月

基於電子郵件通聯之社交網路探勘

Mining Social Relationships from E-mail Logs

研 究 生：詹大偉　　Student：Ta-Wei Chan

指導教授：彭文志　　Advisor：Wen-Chih Peng

國 立 交 通 大 學

資訊學院　資訊學程

碩 士 論 文

A Thesis

Submitted to College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements for the Degree of

in

Computer Science

January 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年一月

基於電子郵件通聯之社交網路探勘

學生：詹大偉　　指導教授：彭文志

國 立 交 通 大 學　　資 訊 學 院　　資 訊 學 程 碩 士 班

摘　要

隨著全球資訊網路（WWW） 的發展，電子郵件（E-Mail）已成為一種不可或缺的、流行的通訊方式。現今，電子郵件不但是人與人溝通最重要的工具，對各企業更已是商務往來最主要的溝通橋樑。企業內部溝通、重要會議資料、商務往來書信、甚至年節祝賀等，也都已經缺少不了電子郵件。

E-Mail內含可觀的企業資源，E-Mail的歷史資料與內容可以匯聚成為企業知識庫，E-Mail附加檔案的再利用更可以避免資源浪費，從E-Mail延伸的行為，如收送時間，地點，單位，日流量等等，加上E-Mail行為與資料的比對分析得到的重要情報。經過分析比較產生的E-Mail資訊，相信對企業重要的決策議題與管理，具備相當程度的參考價值。另外透過E-Mail構建之社會關係網路，是Mining E-Mail的一個新的應用。基於以上的觀察，本篇論文實作一基於電子郵件之社群網路探勘系統，在本系統中， 透過所開發的模組，可以分析出公司哪一群人(寄件者)與外部人員群組(收件者)的對應關係。在實作驗證上， 我們透過真實電子郵件通聯的紀錄， 找出可以分析出公司中部門與部門的關係程度、部門與供應鏈的關係。進而尋找出每個部門對營業額的影響程度、SALES部門對供應鏈與營業額的關係。

Mining Social Relationships from E-mail Logs

Student：Ta-Wei Chan Advisor：Dr. Wen-Chih Peng

Degree Program of Computer Science
National Chiao Tung University

Abstract

With the popularity of Internet, E-Mail has become one important communication media. It can be seen that e-mail is used for personal and enterprise business purpose. Not only the important interpersonal communication tool, but also the most important communication bridge for enterprise business. For example, e-mails play an important role in information exchange, commercial business and holiday greetings.

With the above observation, in this paper, we intend to discover community structures via email logs. E-mail of the historical data and content can be converged into a corporate knowledge data base. Since E-mail logs contain e-mail usage information, such as sending and receiving time, location, organize department, daily flow and so on, community structures mined from e-mail logs have considerable social behaviors for decision making and management. Therefore, this study implements a framework of mining community structures from e-mail logs which consists of pre-processing module, mining module and community module. Through the developed modules in this system, we can analyze the relationship between senders and recipients. Moreover, we could derive the relationship among departments and the relationship between departments and suppliers. Furthermore, we try to find out the impact on sales turnover of each department, and the influence on the revenue from the relationship between sales department and supply chain.

誌　　謝

首先我要誠摯地感謝我的指導教授彭文志博士對我的啟蒙與指導。彭教授從論文題目的確定、研究的方法到比較的方式，都非常用心指導，因此我才能順利完成我的碩士論文。在論文撰寫的這段期間，非常感謝 ADSL 實驗室的學長、學弟們，謝謝你們不厭其煩的指出我研究中的缺失，且總能在我迷惘時為我解惑，提供寶貴的意見，使得本論文能夠更完整而嚴謹。

另外，我要感謝幾位和我一起努力奮鬥的同學 Kevin、Yummy、Fuwin...，回想這三年來，修課時總是互相解決疑問，每當考試期間，我們總是相約在圖書館 k 書，偶而還會相約去打牙祭，有你們這群好朋友，讓三年多研究所的生活增添不少歡樂回憶！

最後，我要感謝我的母親，從小無怨無悔的給我支持。您的支持及鼓勵，是趨動我完成學業的動力之一！同時，我特別要謝謝在背後默默支持我的親愛老婆和我可愛的三個女兒，每當我最無力、最需要有人依靠時，妳們總是在我身邊，耐心地安慰我，鼓勵我。在我學業、論文忙得焦頭爛額之際給予我最大的包容，讓我能無後顧之憂專心在學業及論文寫作上。

如今我終於完成了這份論文了，謝謝那些曾經幫助我的師長、同學及家人們，這份榮耀應該是屬於你們的。

目　　錄

表目錄

圖目錄

第一章 緒論

1.1 研究動機

在現今資訊科技發達的環境中，網際網路的普及化，數以百萬計的資料庫被應用在商業管理、企業管理、工程管理、科學領域以及一些其他的應用上，儲存在其中的資料數目以驚人的速度成長中，大量的資料使得我們需要新的技術與工具來處理與分析存在其中的有價資訊，於是產生了資料探勘（Data Mining）這一門重要的學術。所謂的資料探勘就是處理、分析大量的資料，將隱藏在其中的有用資訊挖掘出來的過程。資料探勘的技術在近幾年來已經逐漸成熟，並且廣泛被應用在各種不同的領域上，但是如何從日益遽增的大量資料中，以更有效率的方式粹取出更符合需求的有用資訊，便是資料探勘所面臨的最大挑戰。

隨著全球資訊網路（WWW） 的發展，電子郵件(Electronic mail, E-mail)已成為一種重要的、流行的通訊方式。對多數的企業來說，電子郵件已經是商務往來的主要溝通工具，無論是重要的會議資料或是公司的商務書信，也多採用電子郵件作為資訊傳遞的媒介。

E-Mail 在人們的生活中越來越風行，因為E-Mail具有成本低，容易使用的優點。而且，與傳統的語音通訊比較，是一種更為進步的通訊方式，不需要通訊雙方同時在線上。因此E-Mail越來越受到人們的歡迎。隨著網際網路的發展，E-Mail在經濟和社會的發揮的作用越來越明顯，但同時也帶來很多的問題。Mining E-Mail 技術為這些問題提供了解決方案，並取得較好的效果。

電子郵件是企業現在不可或缺的工具。企業越來越依賴電子郵件的溝通，各個部門與部門的溝通、部門的關係與公司營運是否有關聯，相信這是每個老闆所關心的。人資部門主管，對於透過 E-Mail 人與人的溝通，是否能有效了解員工想什麼，員工的需求是什麼，找到員工福利與公司利益之最大平衡點，進而激勵員工士氣，達到公司持續成長的目的。IT 部門主管，是否能透過這樣的資訊，提早預警企業機密資訊的外洩，防止不肖之員工危害公司；分析對公司有利之資訊，提供老闆決策的參考。我們想透過 Mining E-Mail Log 技術為這些問題找到相關的方法，進而提供解決方案，並取得對企業有貢獻的效果。

圖 1　企業內部對外溝通 E-Mail 示意圖

我們希望使用一種資料探勘的方法，找出 E-Mail Log 之間的存在關係，加上我們提出的研究，找出公司中部門與部門的關係程度、部門與供應鏈的關係。進而尋找出每個部門與公司營業額之間的關係；SALES 部門對供應鏈與營業額之間的關係。判斷這些關係，是否可以提供參考資訊，應用於公司的管理與發展。

透過 E-Mail 進行的通訊在一定程度上反映了人們之間的關係，即發送者與接收者的聯繫。因此，透過資料探勘大量的 E-Mail，可以構建出一定範圍內的社會關係網路(Social Network)。以 E-Mail 作為探勘社會關係網路的資訊，具有以下的優勢：

1、E-Mail 的使用範圍極其廣泛，因此透過 E-Mail Log 是探勘社會關係網路的好途徑。
2、E-Mail 具有相對標準格式，這些資訊便於電腦處理。
3、E-Mail 不但記錄了人們之間 ” 互動” 的關係，而且還提供了通訊頻率、通訊時間等特性，可以利用這些特性構建有權重的社會關係網路。
4、E-Mail 帶有時間戳記訊息，這有利於發現社會關係網路的動態特性。

1.2 研究背景與目的

本人於業界工作任職資訊部門，認為對於 E-Mail 的資料管理，是一項對 IT 人員極其重要的資訊及管控，郵件進出之內容、大小、是否與公司為競爭對手，都可透露出相當的資訊。所任職公司亦曾經因為不肖高階員工離職，要求之前所屬員工，藉由 E-Mail 傳送公司機密資料，給其參閱使用，造成公司不白之損失。IT 部門對於公司的營業額成長與衰退都很清楚，我們也希望在 E-Mail Logs 中，找到公司各個部門之密切關係，部門與供應鏈之關係，比較這些關係與營業額成長與否有關係。並希望找到公司興衰是否於 E-Mail 中透露出端倪，進而未來於就任之公司，提供主管這樣的分析資訊。本研究既因為這些因素而產生。

本研究主要是運用資料探勘(Data Mining)與社會網路分析(Social Network Analysis)支援 IT 服務管理中心的問題管理(Problem Management)。藉由資料探勘(Data Mining)來找出每封 E-Mail 他們的互相關聯性，以提供解決問題的參考，進而對相關的 E-Mail 做分群(Clustering)；另外，透過社會網路分析(Social Network Analysis)方法，找出工作性質相似的工作者，架構出彼此的社會網路關係圖。問題管理(Problem Management)，主要是為了降低因 IT 基礎架構錯誤而引發的事件與問題對於企業的負面衝擊，比免相關錯誤所引發的事件重複發生，故將分析是件資料庫中的問題資料，試著從可能的原因清單中，找出根本的主因，為完整的解決方法。

1.3 論文主軸

基於以上的研究目的，在本篇論文中，我們首先在第二章進行文獻探討。在文獻探討中主要可以分成二小節。第一小節對於現行資料庫，資料大量產生時，所使用資料分析研究，多使用資料探勘方式，對其基本的做法與觀念加以介紹。第二節則是介紹社會網路分析，其是研究行為者於團體中一言一行，加以嘗試去瞭解這些行為者的人際關係狀況、尋找人際關係的特徵、及發覺這些關係對個人或是組織的影響。

經由第二章的詳細介紹資料探勘，在本論文第三章中會介紹用來探勘的方法，LCM-freq演算法，這個演算法架構於backtracking方法之下，以及運用有效率的頻繁項目次數計算方法，並加以介紹其真實的資料探勘做法。並在第二小節中，舉一個簡單的例子來說明LCM-freq 演算法執行所有的程序過程，並將一步一步的解釋其步驟。在第三小節中，使用真實資料來套用LCM-freq 演算法執行程序過程，並將2007年每個月的E-mail Log做出結果。而第四小節，我們提出一些方法，對每個群組來做分析，群組與群組的比較等，最後針對我們做的群組算出其純度與亂度。

在實驗與討論部份，藉由公司獲得之實際E-Mail進出資料，用來加強佐證本篇

論文的動機及價值。第一節為資料處理與實驗結果；第二節為部門與部門關係實驗結果與細節；第三節為部門與供應鏈實驗結果與細節；第四節為其他實驗結果與細節。

透過以上的討論，我們對整篇論文作最後總結並且提出個具體的方法供許多企業參考使用。此外，亦描述未來可能遇到的情況與未來後續研究的建議。

第二章 文獻探討

2.1 資料探勘

2.1.1 資料探勘的定義

資料探勘(Data Mining)就是從資料中發掘出資訊或知識，有人稱為「知識發掘」(Knowledge Discovery in Database，KDD)，也有人稱為「資料考古學」(Data Archaeology)、「資料型態分析」(Data Pattern Analysis)、「功能相依分析」(Functional Dependency Analysis)、「資料庫知識探勘」(Knowledge Mining from Database)、「知識萃取」(Knowledge Extraction)、「資料分析」(Data Analysis) 等等，均意指對資料庫中所隱含資訊（如知識法則或資料的含義等)，所做的探勘程序。

資料探勘是一種自動模擬、偵測出資料庫中的相關樣式（Relational Pattern）的技術。依據使用者需求從資料庫中選擇合適資料，並加以處理、轉換、發掘至評估的一連串過程，期能找出真實世界運行時隱含在其內的運作現象，以輔助解決相關問題，這個技術將可以幫助分析人員發現隱含在企業趨勢和企業資料的特徵，甚至用來預測未來趨勢以達成預估目標。

表 1 為資料探勘定義的相關文獻整理

學者、專家或廠商	資料探勘（DM）定義
Frawley（1991）	資料探勘在資料庫中發掘出飛顯然的、前所未有的及潛在的可能有用資訊的過程。
Group and Owrang（1995）	資料探勘是由已存在的資料中，發掘新事實即發現專家尚未知曉的新關係。
Hall（1995）	資料探勘是一種結合資訊視覺化、機器學習、統計方法及資料庫等多種技術，以便從龐大資料量中，萃取法則形式或其他模式所表達的知識。
Fayyad（1996）	資料庫知識發現是種辨別有效的、新奇的、前在有用的以及最終能被瞭解的模式（Pattern）的重要過程。
Cabena（1997）	資料探勘是將先前所未知的隱藏資料，從大型資料庫中有效地抽出以提供給高階主管作為決策參考。

2.1.2 資料探勘的功能

一般而言，資料探勘功能能包含下列五項功能，將這些功能的意義及可能使用的技巧簡述如下：

1. 分類(Classification)
   按分析對象的屬性分門別類加以定義，建立類組(Class)。例如：信用卡區分為白金卡、金卡、普卡。
   最常使用的技巧：決策樹(Decision Tree)、記憶基礎推理(Memory-based Reasoning)等。

2. 推估(Estimation):
   根據既有連續性數值之相關屬性資料，以獲致某一屬性為知之值。例如：依性用卡的申請者之職業、教育程度、消費行為來推估其信用卡消費額。
   最常使用的技巧：統計方法上之相關分析、迴歸分析及類神經網路方法。

3. 預測(Prediction)
   根據對象屬性之過去觀察值來推估該屬性未來之值。例如：根據客戶過去的刷卡消費金額預測其未來之刷卡消費金額。
   最常使用的技巧：迴歸分析、時間數列分析及類神經網路方法。

4. 關聯分組(Affinity Grouping)：
   從所有物件決定那些相關物件應該放在一起。例如：超市中相關之美妝用品(化妝品、保養品)，放在同一間貨架上。在客戶行銷系統上，此種功能用來確認交叉銷售(Cross Selling)的機會以設計出吸引人的產品群組。

5. 同質分組 (Clustering)：
   將異質母體中區隔為較具同質性之群組(Clusters)。事先未對於區隔加以定義，而資料中自然產生區隔。同質分組相當於行銷術語中的區隔化(Segmentation)。
   最常使用的技巧：K-means 法及 Agglomeration 法。

2.1.3 資料探勘的步驟

資料探勘可視為知識發掘的一部分，因此，資料探勘的步驟亦可視為知識發掘簡單而基本的程序，至於要如何進行資料探勘？它的步驟有哪些？經整理分述如下圖：

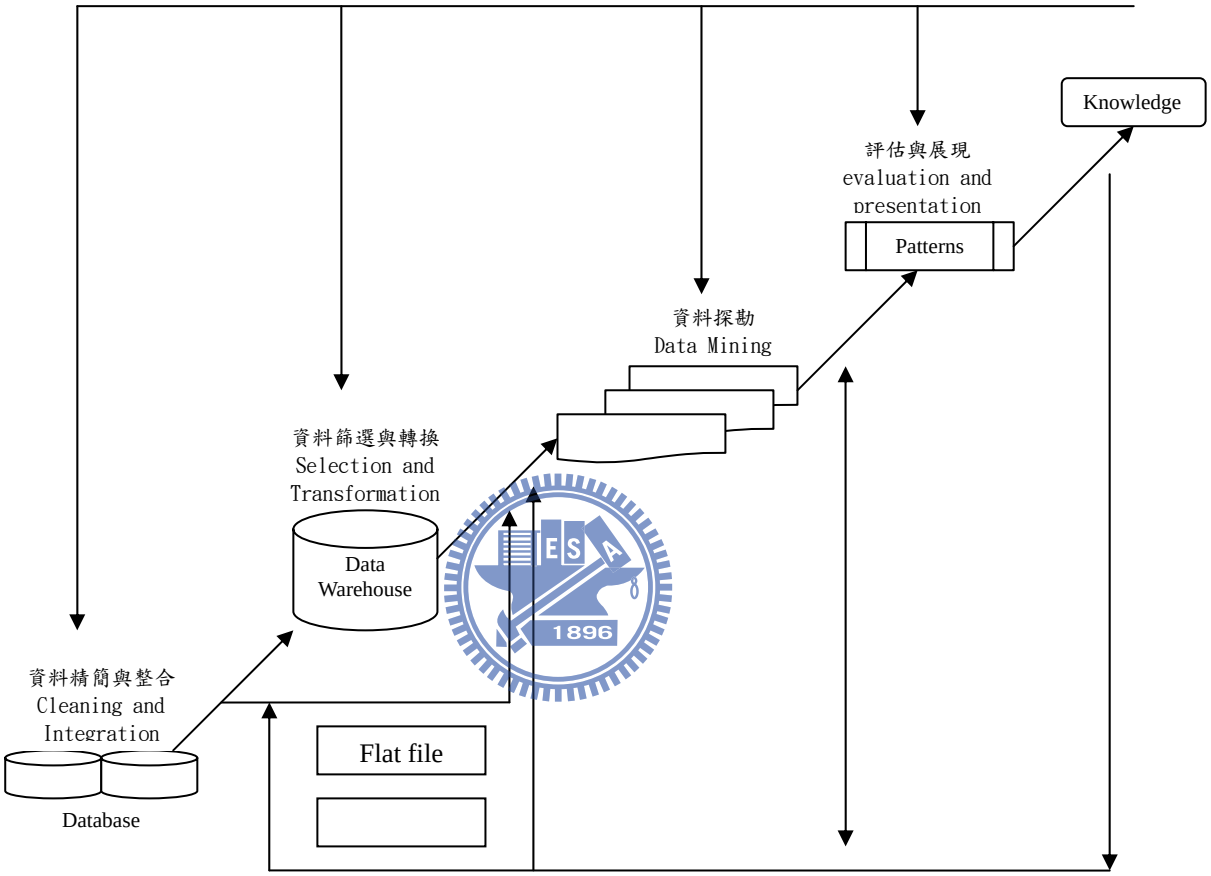


圖 2　資料探勘的步驟

1. 資料精簡(Data Cleaning)：去除重覆資料、錯誤資料或不一致的資料。
2. 資料整合(Data Integration)：將不同來源的資料加以整合。
3. 資料篩選(Data Selection)：由原資料庫中將要操作的資料抽出另存，可以加速 KDD 的處理程序。
4. 資料轉換(Data Transformation)：透過資料轉換的過程，可以增加要描述主題的資訊或去除多餘的資料。
5. 資料探勘(Data Mining)：實際的資料探勘工作。
6. 樣本型態的評估(Pattern Evaluation)：針對某些有興趣的問題去定義實際的樣本型態，以便描述知識。

7. 知識展現(Knowledge Presentation)：將探勘出的資訊或知識，透過視覺化的工具表現在使用者前。

2.2 社會網路分析

社會網路分析是研究行為者（actor）彼此之間的關係〔18〕，所謂的行為者可以是個人、組織或是家庭。社會網路分析嘗試去瞭解這些行為者的人際關係狀況、尋找人際關係的特徵、及發覺這些關係對個人或是組織的影響。社會網路分析總括來說是一門整合的行為科學，其中包含了社會理論、實體觀察研究、數學、統計、圖學等學科。社會網路分析最基本的成分就是點與線(Node and Link)，點代表行動者，而線代表行動者之間的關係或是聯繫。社會網路分析著重在行動者之間的關係，而獨自的行動者或是行動者的個人屬性，因此社會網路分析的資料收集方式是有別於傳統的方法，傳統的資料收集是以機率獨立取樣為主，而社會網路分析是將所有與研究相關的資料都收集起來。例如研究是探討教室中友誼的關係時，社會網路分析會先找尋一個起始點 A 同學，請 A 同學列出自己所熟識的同學，進一步在找尋 A 同學所列出的熟識同學名單，請他們列出所熟識的同學，如此反覆追尋下去將所有相關的行為者資料收集起來後，再進一步探討其行為者之間的關係。

2.2.1 社會網路分析的種類

探討關係或是聯繫的連結情況，將能顯露出行為者的社會網路資訊，藉由觀察社會網路的連結情況，將能瞭解行為者的社會網路特徵。一般來說社會網路分析依照『資料蒐集』方法分成自我中心網路（Ego-centered Networks）分析及完整網路（Whole Networks）分析兩種。

（1） 自我中心網路分析

自我中心網路分析只考慮與焦點（Focal）行為者相關的聯繫，以特定的行為者為探討中心，探討與其中心相關的行為者之間的社會網路情況，其探討的議題包含了自我中心網路的大小、差異性、屬性是否同質等。 自我中心網路分析的資料收集方法，首先要訂定出研究的焦點行為者是誰，請焦點行為者列舉他所認識的人及其關係，接著進一步詢問焦點行為者所列舉的相關行為者，瞭解他們彼此的關係。 自我中心網路分析可以清楚顯示個人的社會網路特徵，包含其相關的行為者為誰、關係內容為何及各行為者之間彼此的連結情況；而這種分析方法適合應用在研究母體（Population）非常大或是研究範圍不易訂定時。

（2） 完整網路

所謂的完整網路分析方法是指在某種特定的範圍下，研究範圍內所有行為者的關係，其範圍可以是正式的組織、系所部門、俱樂部、親屬團體等。這種方法考慮範圍內所有行為者的關係，也就是需要所有行為者彼此之間的關係資料，如果研究範圍下有N 個行為者，則必頇要考慮到所有行為者彼此之間的N＊（N-1）條關係。另外完整網路分析可以探討的議題，除了自我中心網路分析方法的討論議題外，還加上子群組（Subgroup）、中心性（Centrality）分析等，以下將分別針對子群組之分群方法與中心性做介紹。

2.2.2 社會網路分析分群方法

辨識出網路中行為者形成的聚合子群體（cohesive subgroup）是社會網路分析的主要議題之一。聚合子群體是網路中的行為者子集合，相對於其他行為者子集合擁有較強、直接、頻繁的連結關係。許多學者根據社會網路的特性公式化cohesive subgroup的定義與方法，以下介紹常見的幾種定義。

（1） Clique

一個Clique是網路的子集，此子集包含三或三個以上的節點，每一節點直接連接同一Clique中的所有其他節點。Clique必頇是一個極大完全子圖（maximal complete sub-graph），亦即每個Clique會包含所有可能包含的節點，在某特定Clique之外不存在任一節點與此Clique之所有成員皆有直接連接關係。Clique對於cohesive subgroup的定義相當嚴格，在鬆散的網路中只能找到極少量的Clique，在分析真實資料時並不實用，真實資料中找出的Clique通常很小，而且和其他Clique有所重疊，因此學者們提出了幾個改良的方法以寬鬆Clique的嚴格限制。

（2） N-Clique

N-Clique為一網路的子集Ns，此子集中任兩節點之相連距離不得大於門檻值n， N-Clique必須是一個極大完全子圖，其定義公式如下：

$d(i, j) \leq n$ for all ni, nj$\in$Ns

d(i, j)表節點i與節點j之最短路徑所需經過的edge數。由於N-Clique沒有限制節點相連時經過的路徑必頇是N-Clique集合內部的路徑，因此可能造成兩個問題：N-Clique中的節點相連距離可能會超過門檻值n、N-Clique中的節點彼此可能不相連。為了改善這些問題發展出N-Clan與N-Club。

（3） N-Clan

N-Clan必定是一個N-Clique，並且限制計算d(i, j)時所經過的edge必頇是Ns所涵蓋的路徑。實作N-Clan的方法是先找出N-Clique，再過濾掉內部節點連結距離超過門檻值N的N-Clique。

（4） N-Club

N-Club是一個極大完全子圖，和N-Clan同樣限制計算d(i, j)時所經過的edge必頇是Ns所涵蓋的路徑。和N-Clan的不同點在於，N-Club不一定會是一個N-Clique，但一個N-Club必定被某個N-Clan所包含。

（5） K-Plex

K-Plex是一個極大完全子圖，包含gs個節點，每節點至少要連結到同一K-Plex的其他gs-k個節點。K-Plex要求節點成員對大部分的成員有連結，傾向找出許多成員較少的小群組。其定義公式如下：

d(i)≧(gs-k) for all ni ∈Ns

（6） K-Core

K-Plex是定義每個節點可以省略的連結數，而K-Core與K-Plex相反，是定義每個節點最少必頇擁有的連結數，每一個節點都必頇至少連結到群內的k的節點。其定義公式如下：

d(i)≧k for all ni ∈Ns

利用社會網路分析軟體如UCINET，可以找出符合以上幾種定義的聚合子群體。使用以上社會網路分析的分群方法搜尋群組時，除了設定該分群方法的定義門檻值（如N-Clique的距離門檻值N或是K-Core的連結數門檻值K）外，還可以設定欲搜尋群組所包含成員數目的下限。成員數下限必須搭配門檻值的設定，若設定不佳將使搜尋所得結果失去代表性，如使用N-Clique、N-Clan或N-Club演算法時，若將設定群組成員數下限為三且距離門檻值N設定為二，由於條件過於寬鬆，則所搜尋到的群組將會較無意義。而當條件如Clique過於嚴格，又會因不易找到符合條件的群體而失去實用性。因此在決定使用的分群方法後，還必須審慎考量搜尋的設定，設定最適合的搜尋門檻值。[Social Network Analysis: Methods and Applications](Wasserman & Faust, 1994)

2.2.3 評估節點重要性

在社會網路中權力（power）是基本且重要的研究議題，可藉由權力來瞭解行為者的重要性。權力乃是由關係構成，一個擁有較高權力的行為者，代表其與其他節點擁有較為緊密的關係，該行為者會擁有較高的影響力、較多可選擇的機會與資源。社會網路分析利用中心性的計算來衡量行為者的權力大小。中心性的探討分為三種類型。

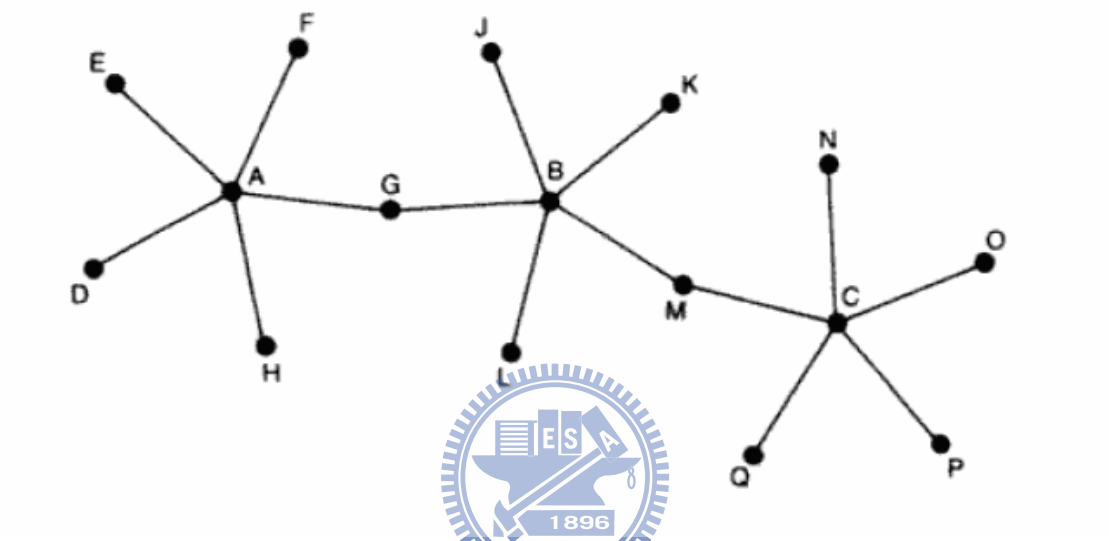


圖 3 社會網路分析之中心性圖例

（1） 程度中心性（Degree centrality）

用以衡量區域中心性（Local Centrality），觀察某一節點與周圍節點的連結關係。一個節點連結到越多節點，就越有權力，如圖中的點A連結到六個節點，而節點E只連結到一個節點，點A明顯有較高的權力，因為當A與E之間的連結被移除，E將失去與其他節點的聯繫而變得孤立，而A仍然可以正常連結至其他節點。圖中的A、B、C擁有較高且相同的程度中心性，皆是區域中心。提出程度中心性為相鄰節點數目除以所有點的數目，將區域中心性轉換為比例的形式。

（2） 接近中心性（Closeness centrality）

用以衡量全域中心，對整體網路的所有點進行衡量。節點可以用比較短的距離接觸到其他節點，或是以較短的距離被其他節點接觸，這種架構

優勢可以被轉換成權力。計算每一節點與其他節點之總和，可以發現圖中的節點B擁有最短的距離總和，不但是區域中心，還是全域中心，比貧A、C來得更重要。

（3） 中介中心性（Betweenness centrality）

衡量某節點在任兩點之間的程度。當節點A要連結節點B，一定要通過中間的節點G，節點G擔任A、B之間的中介，擁有收費的權力。 圖中的節點G與節點M擁有最高的中介中心性。中介中心性的概念簡單，但計算複雜且耗時。

第三章 探勘方法

3.1 E-Mail 社群探勘之流程

在本論文中，我們提出了一個電子郵件社群探勘系統，在本系統中，我們首先從 E-mail log，做 Pre-process 的步驟。接著使用 LCM-freq 演算法做分群，得到寄件者與收件者相同之群組，再針對群組的特性，找出裡面之關連性。首先找出寄件者群組中部門與部門之關係、部門與營業額之關係；接著針對群組中寄件者與收件者之供應鏈找出關係、SALES 部門與下游客戶關係與營業額之關係；針對我們提出之關係值去分析純度與亂度、再把我們算出之關係值以社會網路的架構來顯現。

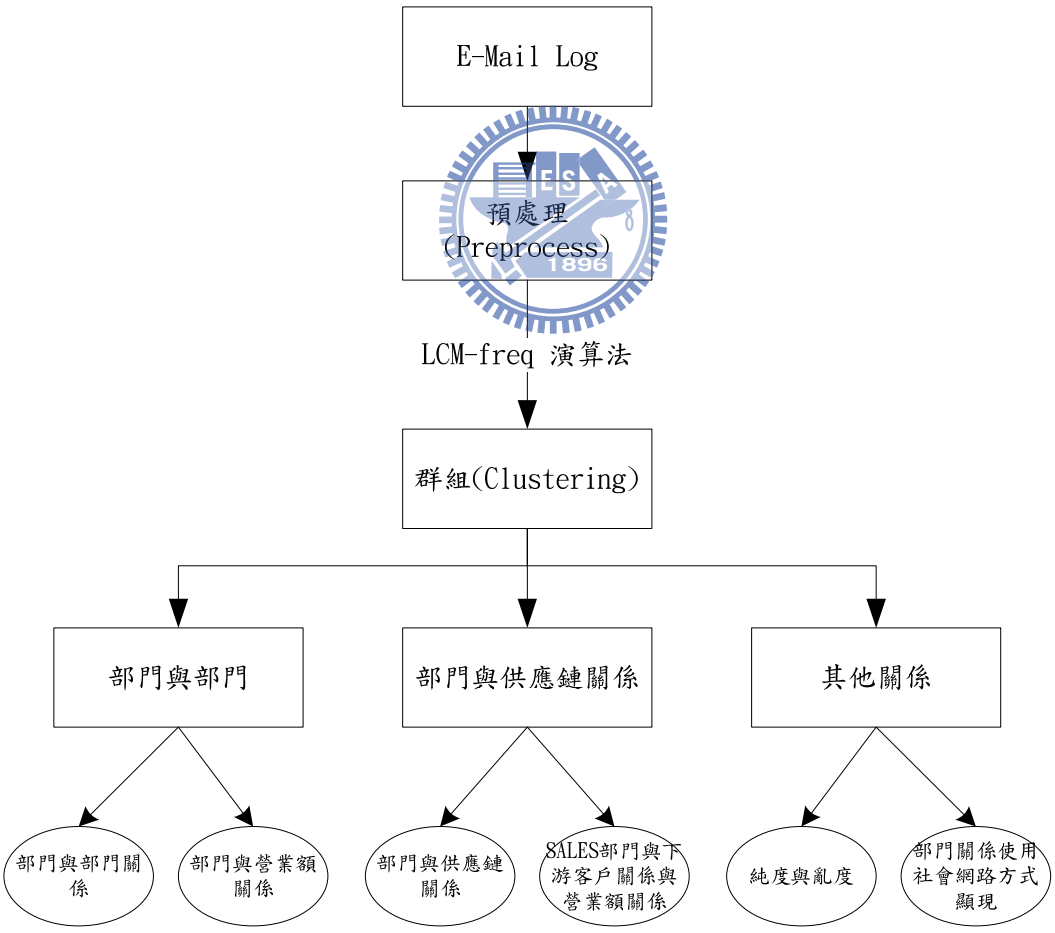


圖 4 探勘流程圖

3.2 LCM-freq 演算法

我們要Mining E-Mail Log有大量的收件者與寄件者之資料，LCM-freq演算法，使用parent-child關係去定義頻繁項目集合；若child為頻繁的項目集合，則parent必為頻繁的項目集合；反之，parent為頻繁的項目集合，則child可能為頻繁的項目集合；因此利用此定理可以依線性時間搜尋頻繁項目。LCM-freq演算法利用陣列方式儲存資料庫，降低主記憶體耗費，但是卻要耗費較多暫存空間，來進行遞迴搜尋頻繁項目集合。記憶體足夠的情況下，LCM-freq演算法將在密集資料庫是個不錯選擇。綜合以上種種，符合對Mining E-Mail Log的特點，大量的記憶體對現今的硬體來說不是大問題，因而我們選用LCM-freq演算法來做為我們資料探勘的工具。

由Takeaki Uno, Taisuya Asai, Yuzo Uchida, Hiroki Arimura [1] 學者提出，他們在技術上其使用簡單陣列技術壓縮資料庫，不像其他演算法壓縮資料庫亦使用binary tree 之方法，例如FP-growth 演算法；這個演算法架構於backtracking 方法之下，以及運用有效率的頻繁項目次數計算方法。

Backtracking 方法是由depth-first 變化衍生而來的，主要在於遞迴方式搜尋頻繁項目集合(Frequent Itemset)，其重覆置入頻繁項目集合P，而產生尾端為P 之延伸集合，因此延伸集合若為頻繁的，將重覆產生延伸集合於遞迴方式。其定理，若一個延伸集合為頻繁的，其延伸前集合一定為頻繁的，例如{ABC}為頻繁項目集合，則它的{AB}一定為頻繁的，這根據Apriori 演算法定理之頻繁項目集合其子項目集合也為頻繁項目集合。有效率的頻繁項目次數計算方法是將水平資料轉換成垂直資料並以陣列儲存，然而將兩兩的項目之交易記錄陣列進行交集動作，只要有共同的交易記錄，將這筆交易記錄之陣列位置顯示為1 且做支持度計算；因此，搜尋頻繁項目集合就以線性的發展，可以快速獲得頻繁項目集合。

Backtracking 方法主要是以parent-child 關係去定義頻繁項目集合，因此，它的架構顯示如下圖4，若有一個transaction dataset 且有5 個項目{a,b,c,d,e}與最小支持度為3，開始進行backtracking 演算法；首先置入一個空集合而且是頻繁的，並開始置入頻繁項目{a}，{a}為頻繁項目集合，進行遞迴方式加入{b}項目集合，產生延伸集合為{ab}，但是{ab}非頻繁的，根據定理，所以有關{ab}之延伸集合皆為非頻繁的項目集合；{a}之尾端置入{c}，產生延伸集合為{ac}，從資料庫得知{ac}為頻繁的，因此{ac}為新的集合；依{ac}之尾端置入{d}，產生新的延伸集合{acd}，但從資料庫得知為非頻繁的，所以回到{ac}；後續一直進行遞迴方式逐一尋找頻繁項目集合，直到尾端沒有任何的項目可延

伸；因此，最後頻繁項目集合為{{a}、{ac}、{ace}、{ad}、{ae}、{b}、{bc}、{bce}、{bd}、{be}、{c}、{cd}、{cde}、{ce}、{d}、{de}、{e}}。

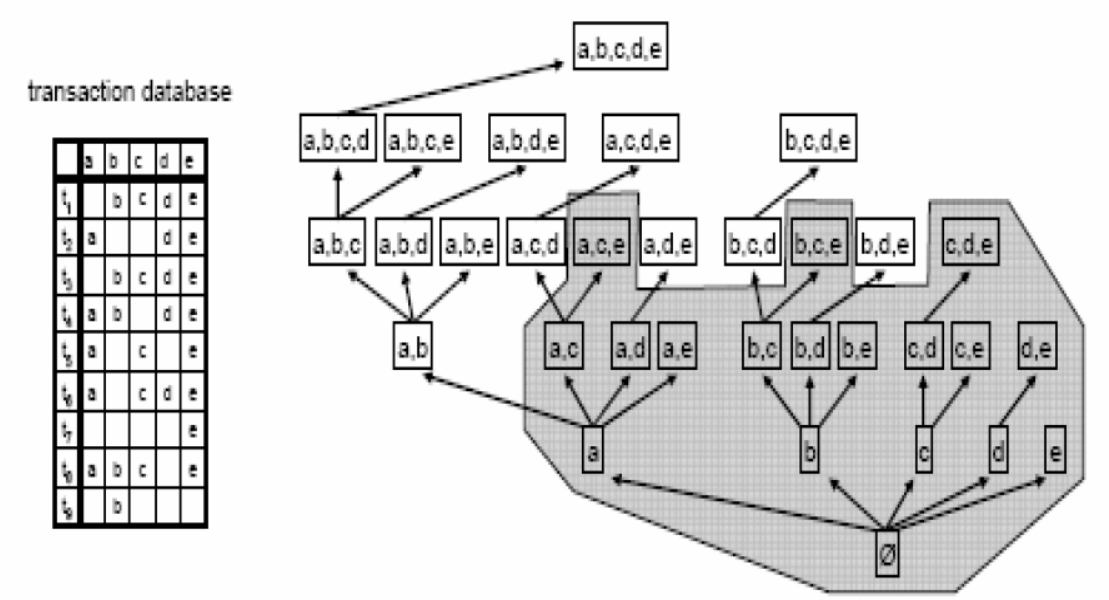


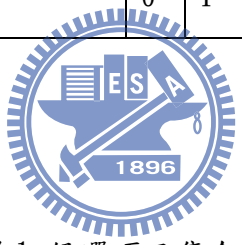
圖 5 樹狀模式描述 backtrack algorithm 使用 depth-first 方法

3.3 LCM-freq 演算法實例說明

舉一個簡單的例子來說明LCM-freq 演算法執行所有的程序過程。假設有一4 筆交易記錄的資料庫D 並且以交易中的項目按字典次序存放表示(如表2)，再進行探勘最小交易支持度為2(出現次數為2)的頻繁項目集合

表 2 LCM-freq 演算法之資料庫 D

交易項目 / 交易紀錄編號	A	B	C	D	E
T001	1	0	1	1	0
T002	0	1	1	0	1
T003	1	1	1	0	1
T004	0	1	0	0	1

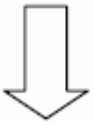


步驟1：

搜尋資料庫找出所有長度為1 候選項目集合，並把水平資料庫轉換成垂直資料庫，之後找出長度為1 頻繁項目集合。(如表3)

表 3 LCM-freq 演算法之資料庫 D 轉換

交易項目 / 交易紀錄編號	A	B	C	D	E
T001	1	0	1	1	0
T002	0	1	1	0	1
T003	1	1	1	0	1
T004	0	1	0	0	1



交易紀錄編號 / 交易項目	T001	T002	T003	T004
A	1	0	1	0
B	0	1	1	1
C	1	1	1	0
D	1	0	0	0
E	0	1	1	1

Li	Cont
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

步驟2：

進行backtracking 方法置入空集合，並把{A}置入與交易記錄陣列暫存，之後置入{B}於{A}尾端，作兩兩陣列交集與支持度計算，得到{AB}與交易記錄之陣列{0010}，然而不符合最小支持度，返回至{A}。（如圖6）

A	1	0	1	0
B	0	1	1	1

支持度1

AB	0	0	1	0

圖 6 {A}與{B}兩兩互相合產生{AB}其支持度為 1

步驟3：

把{A}之交易記錄陣列繼續暫存，置入{C}於{A}尾端，作兩兩陣列交集與支持度計算，得到{AC}與交易記錄之陣列{1010}，然而符合最小支持度，得到{AC}

A	1	0	1	0
C	1	1	1	0

支持度2

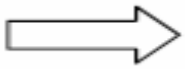
AC	1	0	1	0

圖 7 {A}與{C}兩兩互相結合產生{AC} 其支持度為 2

步驟4：

進入{AC}並且把交易記錄陣列暫存，置入{E}於{AC}尾端，作兩兩陣列交集與支持度計算，得到{ACE}與交易記錄之陣列{0010}，然而不符合最小支持度，所以返回{AC}，但是{E}為最後一個項目，所以返回至{A}。（如圖8）

AC	1	0	1	0
E	0	1	1	1



支持度1

ACE	0	0	1	0

圖 8 {AC}與{E}兩兩互相結合產生{ACE} 其支持度為 1

步驟5：

返回至{A}之交易記錄陣列暫存，置入{E}於{A}尾端，作兩兩陣列交集與支持度計算，得到{AE}與交易記錄之陣列{0010}，然而不符合最小支持度，所以返回{A}，但是{E}為最後一個項目，所以返回至空集合。（如圖9）

A	1	0	1	0
E	0	1	1	1

支持度1

AE	0	0	1	0

圖 9 {A}與{E}兩兩互相結合產生{AE}其支持度為 1

步驟6：

因返回至空集合，所以置入{B}與交易記錄陣列暫存，置入{C}於{B}尾端，作兩兩陣列交集與支持度計算，得到{BC}與交易記錄之陣列{0110}，然而符合最小支持度，得到{AC}頻繁項目集合與交易記錄陣列。（如圖10）

B	0	1	1	1
C	1	1	1	0

支持度2

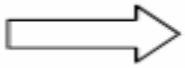
BC	0	1	1	0

圖 10 {B}與{C}兩兩互相結合產生{BC}其支持度為 2

步驟7：

進入{BC}並且把交易記錄陣列暫存，置入{E}於{BC}尾端，作兩兩陣列交集與支持度計算，得到{BCE}與交易記錄之陣列{0110}，然而符合最小支持度，得到{BCE}頻繁項目集合與交易記錄陣列，但是{E}為最後一個項目，所以返回至{BC}。（如圖11）

BC	0	1	1	0
E	0	1	1	1



支持度2

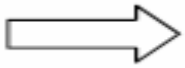
BCE	0	1	1	0

圖 11 {AC}與{E}兩兩互相結合產生{BCE}其支持度為 2

步驟8：

返回至{BC}之交易記錄陣列暫存，但是{BC}尾端並沒有項目可以置入，所以返回{B}之交易記錄陣列暫存，置入{E}於{B}尾端，作兩兩陣列交集與支持度計算，得到{BE}與交易記錄之陣列{0111}，然而符合最小支持度，得到{BCE}頻繁項目集合與交易記錄陣列，但是{E}為最後一個項目，所以返回至空集合。(如圖12)

B	0	1	1	0
E	0	1	1	1



支持度3

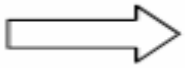
BE	0	1	1	1

圖 12 {B}與{E}兩兩互相結合產生{BE}其支持度為 3

步驟9：

因返回至空集合，所以置入{C}與交易記錄陣列暫存，置入{E}於{C}尾端，作兩兩陣列交集與支持度計算，得到{CE}與交易記錄之陣列{0110}，然而符合最小支持度，得到{CE}頻繁項目集合與交易記錄陣列，但是{E}為最後一個項目，所以返回至空集合。(如圖13)

C	1	1	1	0
E	0	1	1	1



支持度2

CE	0	1	1	0

圖 13 {C}與{E}兩兩互相結合產生{CE}其支持度為 2

步驟10：

因返回至空集合，所以置入{E}與交易記錄陣列暫存，但是{E}為最後一個項目，所以返回空集合並停止演算法。所以找出所有頻繁項目有{{A}、{AC}、{BC}、{BCE}、{BE}、{CE}、{E}}。

3.4 LCM-freq 演算法實作說明

3.4.1 E-Mail 資料前置處理

原始資料格式如下圖所示(如圖14)，由公司E-Mail Log機器擷取我們需要的資訊，我們每個檔案中包含有欄位有日期、時間、寄件者、收件者、主旨、附件檔、郵件大小、寄送IP、處理狀態和說明。

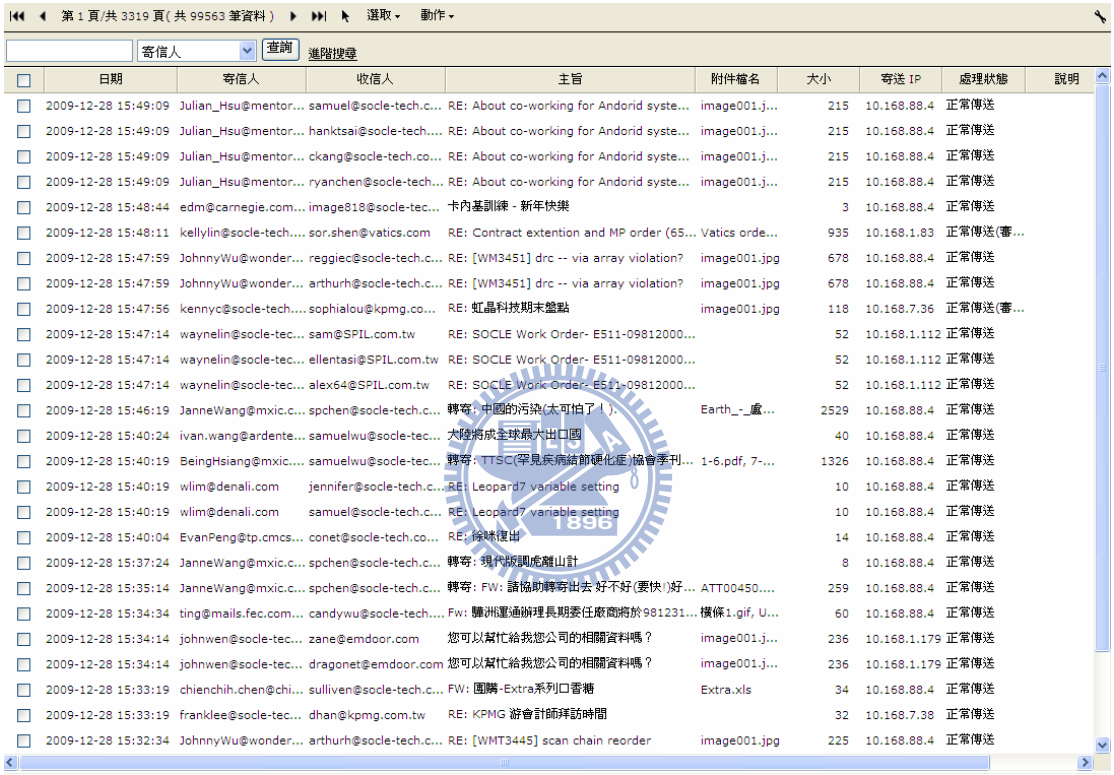


圖 14 E-Mail Log

由公司E-Mail Log機器擷取下我們需要的資訊，由於我們將分析2007一整年之資料，我們將每個月資料儲存成一個檔案，我們將會有十二個檔案，每個檔案中包含有這幾個欄位：日期、時間、寄件者、收件者。

擷取的資料如下圖所示(如圖15)，

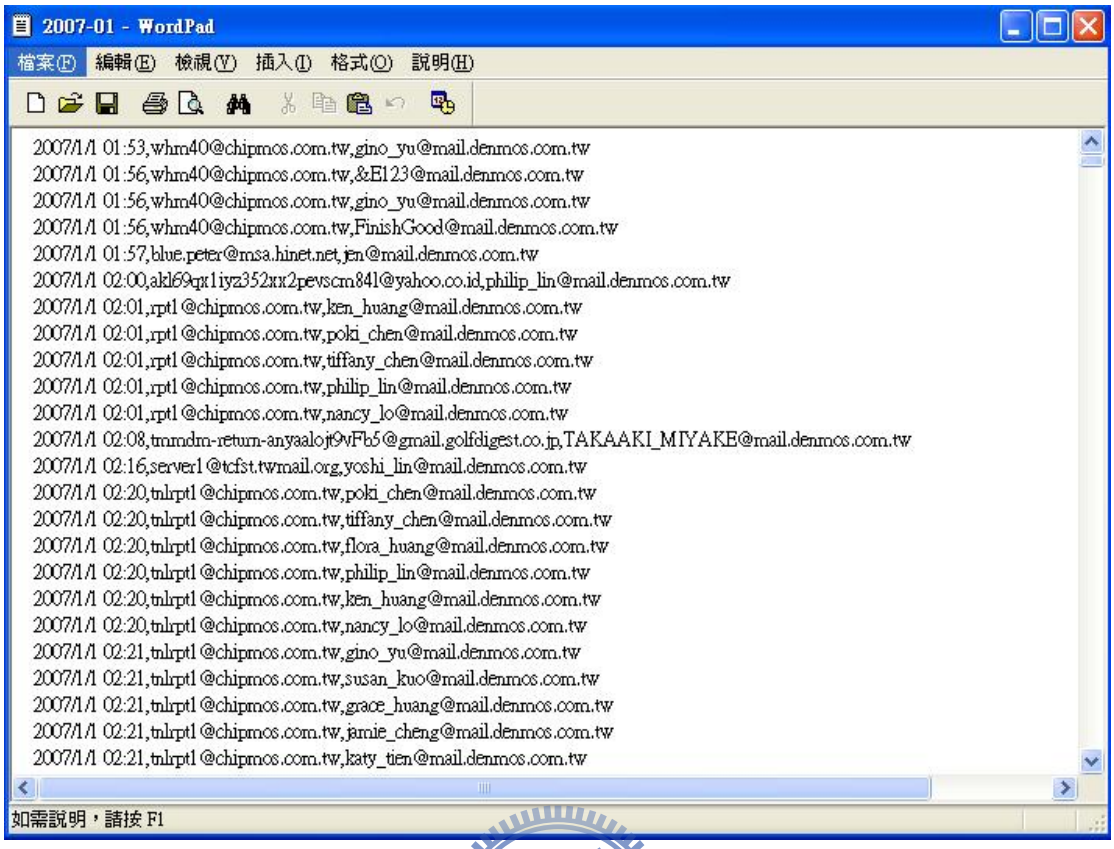


圖 15 2007 年 1 月之 E-Mail 之進出 Log

3.4.2 執行程式與獲取結果

由2007年每個月的資料，分別由以下指令執行，產生出我們需要的轉換格式。如下圖所示(如圖16)，

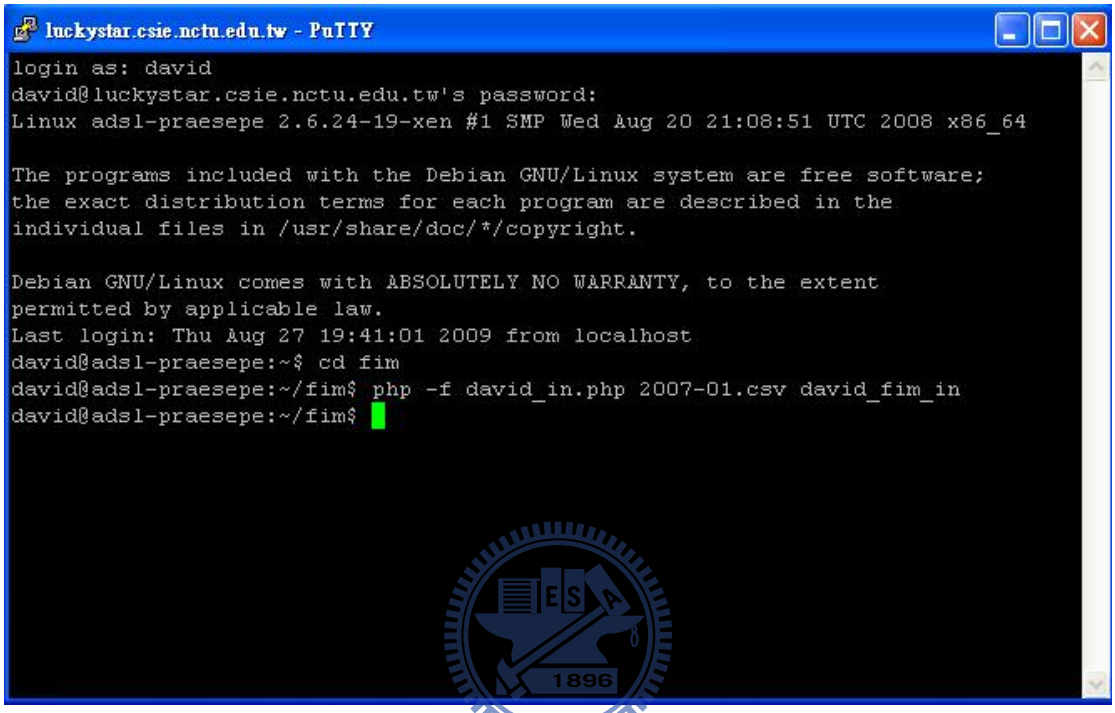


圖 16 執行轉檔成為需要之格式

由第一步驟產生的資料，執行LCM演算法程式，參數設定為4個收信者，12個發送者，產生出我們參數設定值的每個群組代碼。如下圖所示(如圖17)，其為2007年一月共有31組符合條件，其下圖分別為顯示4個收信者有六組，5個收信者有十二組，6個收信者有九組，7個收信者有四組，所以總共有三十一組。

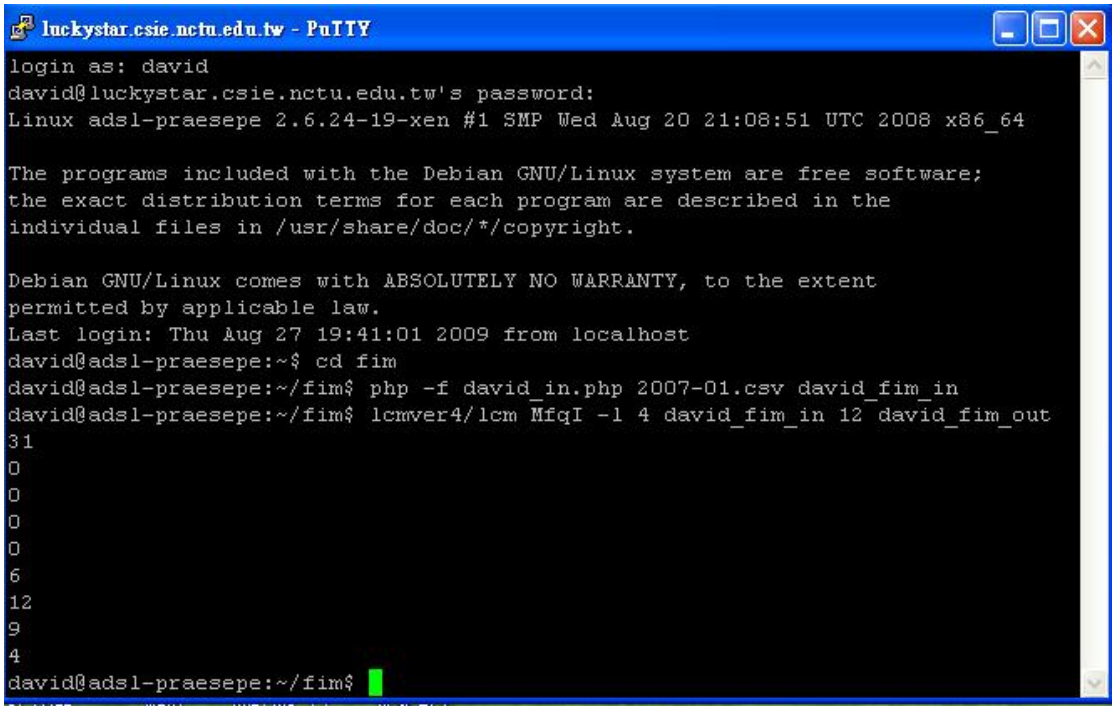


圖 17 執行 LCM 程式，得到當月之群組資料

由第二步驟產生的資料，執行代碼轉換回E-Mail Address，產生出每個群組，組號、收信者與發送者之資料。如下圖所示(如圖18)，

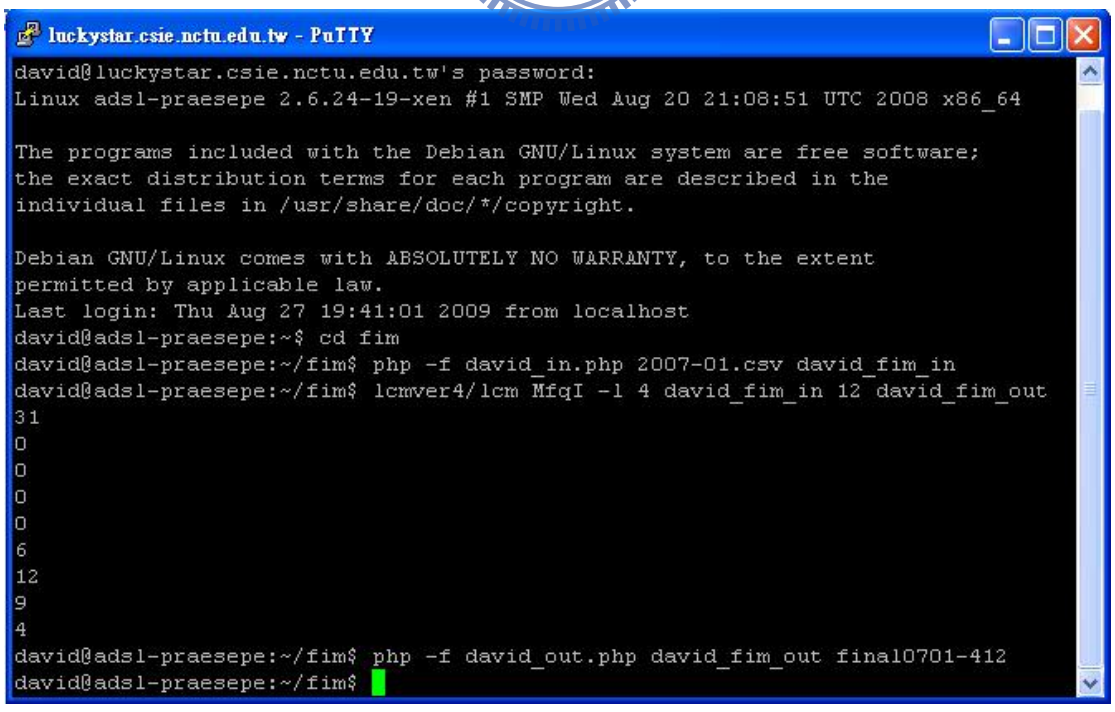


圖 18 執行轉檔程式將所得代號轉回 E-Mail 格式

以下為2007年一月之群組，組號、收信者與發送者之資料。如下圖所示(如圖19)，

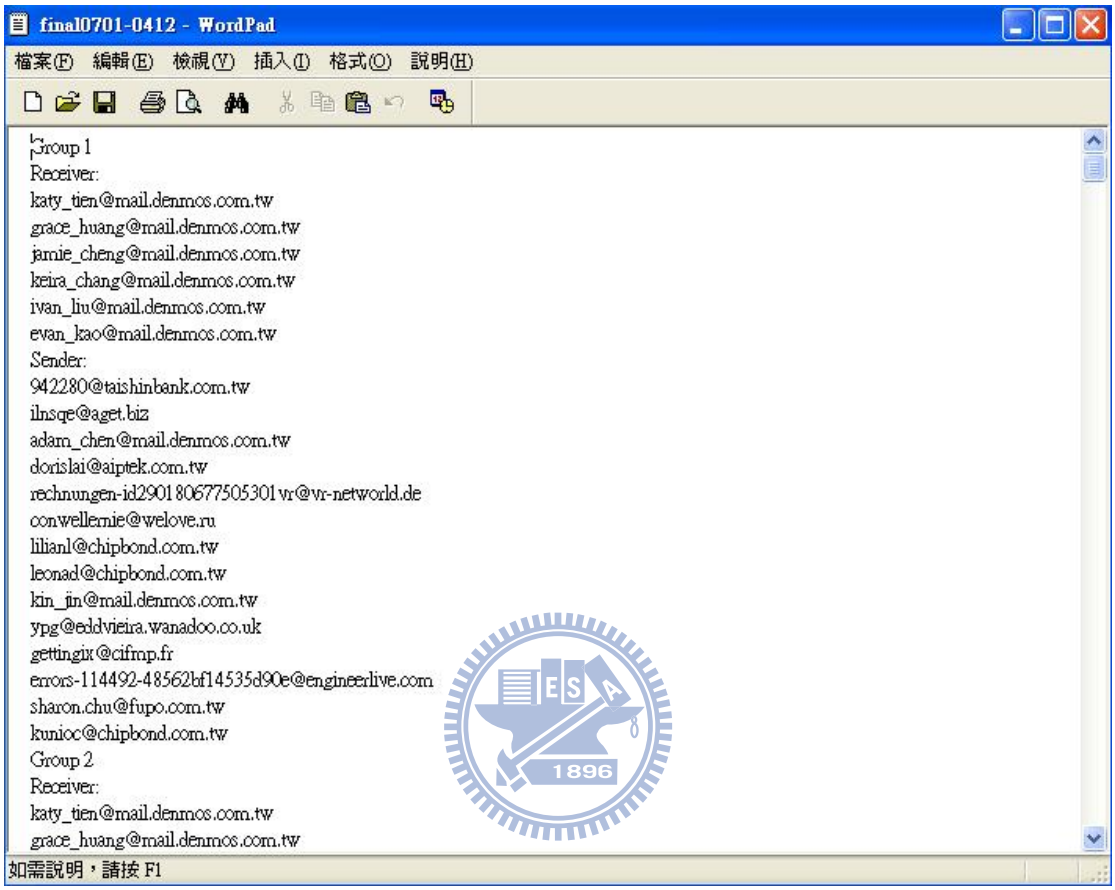


圖 19 執行完畢後之資料格式

3.5 群組間的關係研究

當我們使用LCM之方法資料探勘出的群組，收件者中，包含不同部門的人員資料，我們想將此資料進一步分析，研究部門與部門之間的關係。收件者、寄件者，也存在著不同程度的關係，由於產業別不同可以對寄件者之資料，按產業屬性分別去作群組，進而分析其結果。

3.5.1 群組內部門與部門關係

每個收件者隸屬於不同部門，我們將每個收件者在每個群組中每個部門之人員總數加總，其數值代表這個部門在這個群組之值，兩兩部門分別對其之值之乘積，既代表部門與部門之間的關係值，關係值越高代表這兩個部門之關係越密切。因為其部門間之E-Mail聯繫密切，想當然其互動一定是很密切的。

其下列公式代表其關係值：Relation1之算法。

$$\mathrm{Re}lation1(i,j) = \sum_{k \in Group} \left|G_{k,i}\right| * \left|G_{k,J}\right|$$

$G_{k,i} \equiv$ Department i 在Group k 裡頭的人數

$G_{k,\mathrm{j}} \equiv$ Department j 在Group k 裡頭的人數

$$\sum_{K \in Group} \quad \sum_{\substack{i,j \in Dept \\ i \neq j}} Department \quad .$$

1. 藉由上面之方程式定義，計算出每個群組的每兩個部門之相關數值。
2. 針對每個月每個群組計算出之數值加總得到，每兩個部門之關連性加總數值。即得到當月份之i、j部門之間當月相關係數。數值越大代表其當月關係越密切，反之越小則代表關係越不密切。
3. 比較每個月之關係係數是否有所變化？

例如： Group k 有七個人分屬公司不同部門之收信者

表 4 部門關係之舉例

部門	人數
PC(Product Control Dept.)	4 人
PE(Product Engineering Dept.)	2 人
PR(Purchase Dept.)	1 人

關係部門	Realtion 計算方式與值
PC-PE	4*2=8
PC-PR	4*1=4
PE-PR	2*1=2

3.5.2 群組內收件者部門與寄件者供應鏈關係

每個收件者隸屬於不同部門，我們將每個收件者在每個群組中每個部門之人員總數加總，其數值代表這個部門在這個群組之值。而寄件者，依照其對公司供應鏈之區分為：上游供應商、下游客戶、其他信件分別對應其值，與部門人數之乘積，既代表群組內部門與寄件者供應鏈之間的關係值，關係值越高代表這個部門與上游供應商、下游客戶或其他信件之關係越密切。因為其間之E-Mail聯繫密切，想當然其互動一定是很密切的。

其下列公式代表其關係值：
Relation上游供應商、下游客戶或其他信件之算法。

$$Relation_{上游}(m,n) = \sum_{G \in Group} \left|G_{l,m}\right| * \left|G_{l,n1}\right|$$

$$Relation_{下游}(m,n) = \sum_{G \in Group} \left|G_{l,m}\right| * \left|G_{l,n2}\right|$$

$$Relation_{其他}(m,n) = \sum_{G \in Group} \left|G_{l,m}\right| * \left|G_{l,n3}\right|$$

$G_{l,m}$ ≡ Department m 在 Group k 裡頭的人數
$G_{l,n1}$ ≡ 上游供應商 n1 在 Group k 裡頭的人數
$G_{l,n2}$ ≡ 下游客戶 n2 在 Group k 裡頭的人數
$G_{l,n3}$ ≡ 其他信件 n3 在 Group k 裡頭的人數

$$\sum_{L \in Group} \sum_{m \in Dept.,n \in Cust.} Depa.*Cust.$$

1. 藉由上面之方程式定義，計算出每個群組與寄件者之供應鏈之相關數值。
2. 針對每個月每個群組計算出之數值加總得到，每個部門之關連性與寄件者之供應鏈之加總數值。即得到當月份之m部門，n廠商分類之間當月相關係數。數值越大代表其當月關係越密切，反之越小則代表關係越不密切。
3. 每個部門之關連性與寄件者之供應鏈之加總數值，亦可判斷每個部門對外之屬性，是否與我們認知中有差距？
4. 比較每個月之關係係數是否有所變化？

例如： Group L 有七個人分屬公司不同部門之收信者

表 5 部門與供應戀關係之舉例

部門	人數
PC(Product Control Dept.)	4 人
PE(Product Engineering Dept.)	2 人
PR(Purchase Dept.)	1 人

寄件者	人數
上游供應商	10 人
下游客戶	2 人
其他信件	5 人

關係部門	PC	PE	PR
上游供應商	4*10=40	2*10=20	1*10=20
下游客戶	4*2=8	2*2=4	1*2=2
其他信件	4*5=20	2*5=10	1*5=5

3.5.3 群組關係研究實作

3.5.3.1 基本資料分類

由公司 E-Mail Log 與前一章節中得到的分組資料，我們就部門與部門之間、部門與供應鏈之關係，來做分析。當然預先要把人員與外部之信件先歸類，人員我們先將其部門歸類，外部寄發之信件也將其歸類為上游供應商(WAFER 加工各站之廠商)、下游客戶(成品銷售之各廠商)、其他信件(包含其他支援之廠商如運送貨物之廠商、政府機關之聯繫、無法分類、垃圾郵件…等)。

我們將每個月分組資料，共有十二個檔案，分別對這十二個檔案執行其群組關係分析之程式。人員之部門資料如下圖所示(如圖 20)，外部信件之供應鏈分類如下圖所示(如圖 21)，

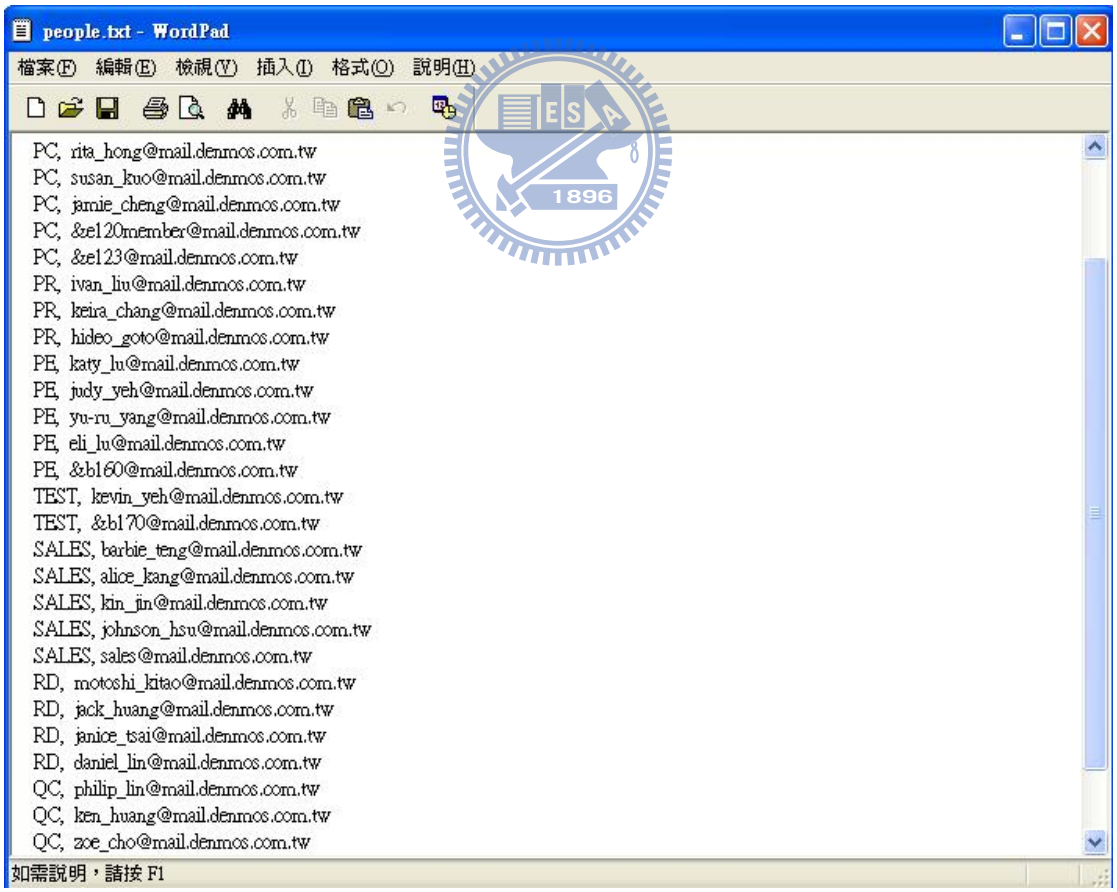


圖 20 公司人員隸屬之部門分類

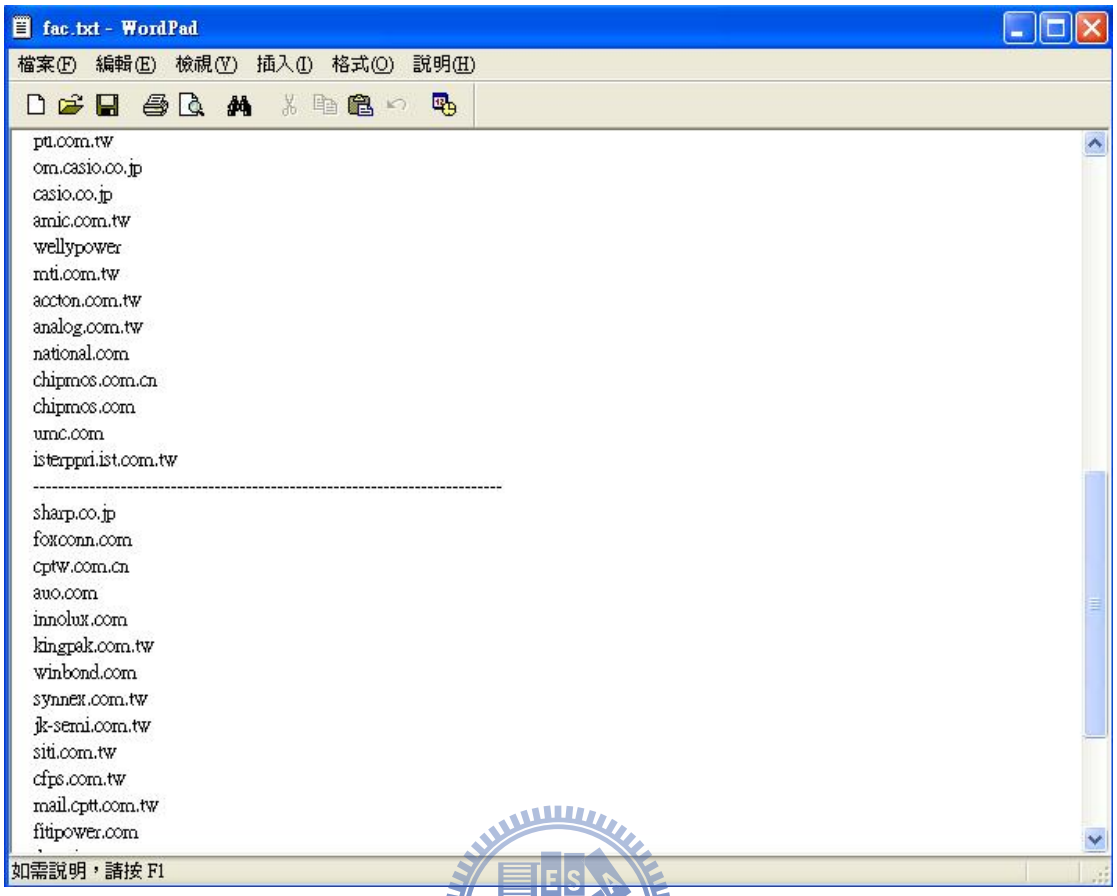


圖 21 外部信件之供應鏈分類

3.5.3.2 執行程式與獲取結果

由2007年每個月的資料，分別由以下指令執行，產生出我們需要的分析資料。其執行將產生其他信件之資訊與未分類之人員名單，讓我們可以調整基本資料是否建檔正確。如下圖所示(如圖22)，

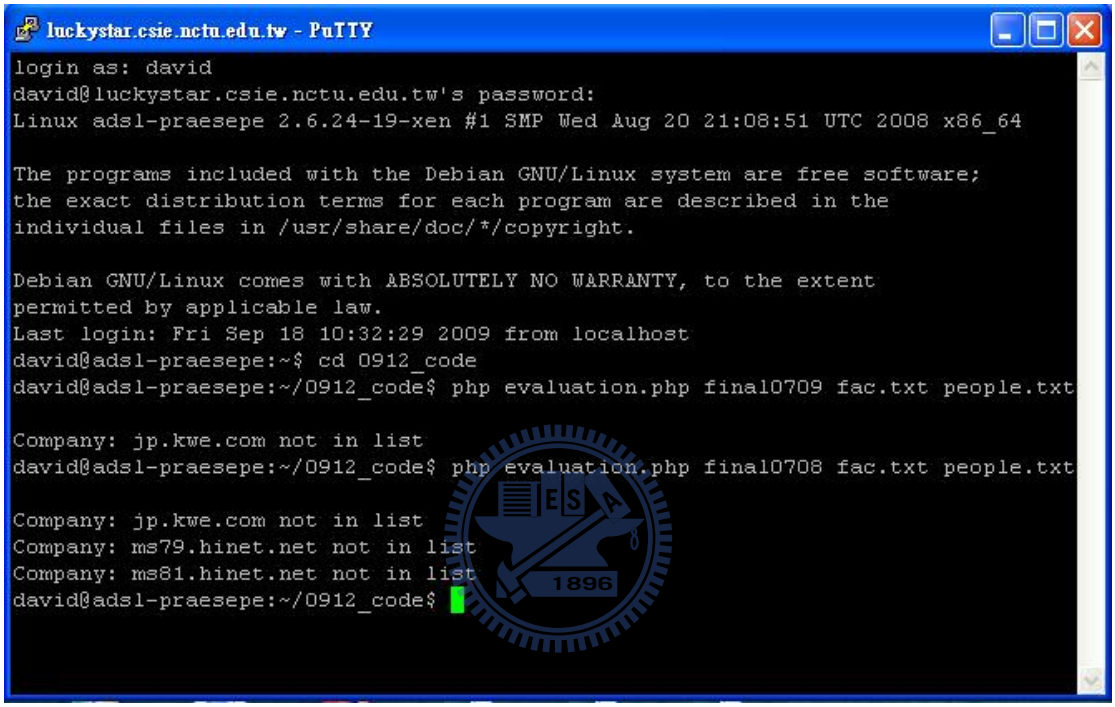


圖 22 執行八、九月完畢後之圖

3.5.3.3 群集純度與亂度

針對文中使用LCM-freq演算法，產生出來的分組，判斷其純度(Purity)與亂度(Entropy)，並分別套用下面之公式產生每個月之純度與亂度值。純度數值越大越好、反之越不好；亂度數值越小、反之越不好。

純度(Purity)：

我們全部之人數Dataset D，使用LCM-freq分組後產生K個群組，第j個群組Cj個數用|Cj|來表示，|Cj|class=i 表示在群組j中class i 有多少個數。找出最大之class，算出之純度。

$$purity(C_j) = \frac{1}{|C_j|} \max_i \left(|C_j|_{class=i} \right)$$

算出一個月(D)，每個月k個群組，總純度。

$$purity = \sum_{j=1}^{k} \frac{|C_j|}{|D|} purity(C_j)$$

亂度(Entropy):

使用LCM-freq分組後產生K個群組，群組全部之人數|D|，第j個群組Cj個數用|Cj|來表示，Pi 表示在群組中class i 有多少個數之機率。套用下面之公式算出每個月之亂度Hs。

$$P_i = \frac{|C_j|}{|D|}$$

$$\mathrm{H}_s = \sum_{i=1}^{n} p_i I_e = -\sum_{i=1}^{n} p_i \log_2 p_i$$

第四章 實驗與討論

4.1 資料處理與實驗結果

4.1.1 資料來源與預處理

由於，於業界工作任職資訊部門，E-Mail 的備份資料極其重要，每個月進出的 Log 檔案的資料對公司是有意義的。本研究的資料就是由任職公司的實際進出資料取得後加以分析研究，由於公司規模在 2007 年時人員大概為 100 人左右，其資料大概每個月為數萬筆。

下表即為 2007 年一整年的進出紀錄筆數，其不包含公司內部郵件之流通，經過預先處理(PreProcess)後，條件如下:

1. 失敗投遞之郵件
2. 離職員工之郵件
3. 不在公司收件者名單之郵件

表 6 2007 年每個月 E-Mail 進出數量與預處理後之數量

2007 年 月 / 筆數	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
未處理	65205	44316	58642	53559	44187	43942	49980	64305	50301	52461	45697	45282
預處理後	61384	41526	54615	50469	42198	40253	45452	53778	44167	47217	41562	42006

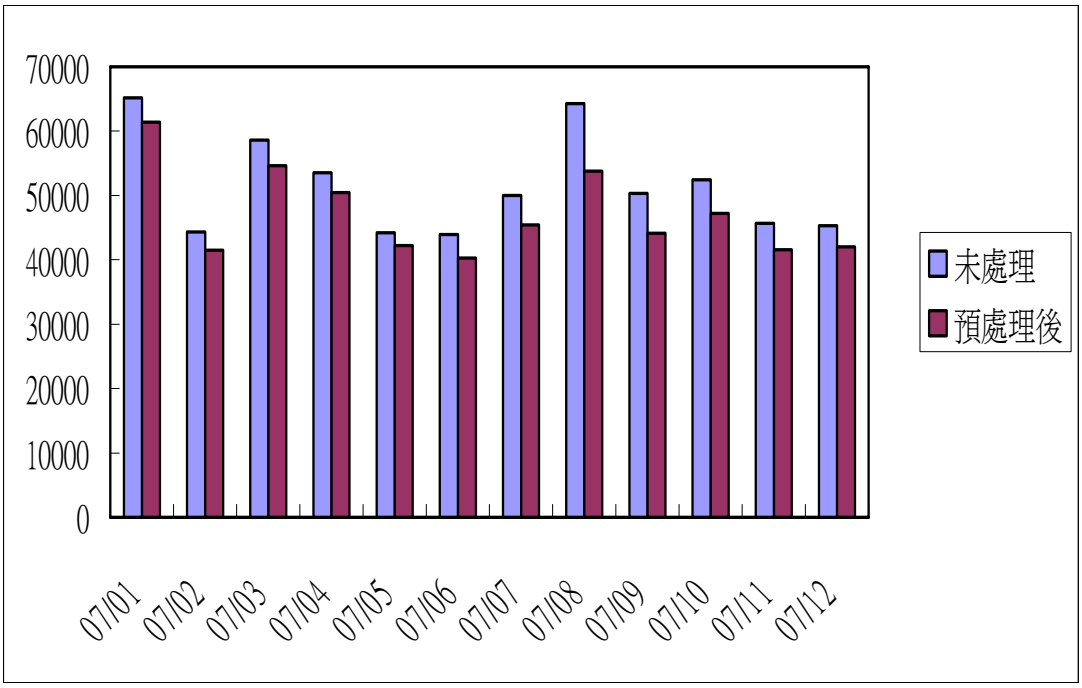


圖 23 2007 年每個月 E-Mail 進出數量與預處理後之數量

4.1.2 資料參數

E-Mail 資料做分析，要判斷哪些收信者與某些發信者之關係，其收信者群組與發信者群組，參數的決定是非常重要的，我們下了幾個參數做實驗，其結果直來判斷最適合之參數。太小的參數範圍將產生許多群組，對於資料的判斷將非常困難；太大的參數範圍將產生極少群組，對於資料的判斷將非常不準確。

我們用了三組參數去產生實驗組數，分別為

1. 3 個收信者，10 個發送者。
2. 4 個收信者，12 個發送者。
3. 5 個收信者，15 個發送者。

下表即為其實驗結果之組數

表 7 三組參數去產生實驗組數

2007 年 月 / 參數	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
3 收信者 10 發信者	102	49	56	41	63	44	27	17	20	24	15	17
4 收信者 12 發信者	31	20	27	17	12	11	13	13	16	17	4	9
5 收信者 15 發信者	14	15	15	5	6	8	8	6	4	6	8	7

上表 3 個收信者，10 個發送者，其結果組數過於龐大，經分析其代表意義不夠詳細與精確，大部份並沒有跨部門的組數，故將其捨去；而 5 個收信者，15 個發送者，其結果組數甚少，經分析其代表意義不能包含大部分的資訊，大部份有代表性的組數被捨去，也不適合分析；4 個收信者，12 個發送者，其結果組數適中，經分析其中已經幾乎包含全部的資訊，組數也適合分析，故選擇此參數。

4.1.3 E-Mail 數量與營業額之關係

本實驗在探討 E-mail 進出的數量與公司營業額多寡的關係。針對這兩項數據我們來加以評估比較，就兩種資料的特性來預估，應該是有正相關的關係才對，以下是公司 E-Mail 進出數量與營業額之統計數量表。

將此統計圖表用折線圖表示為圖 24。

表 8 E-Mail 與公司營業額之統計表

2007 年 月 / 項目	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
E-Mail	65205	44316	58642	53559	44187	43942	49980	64305	50301	52461	45697	45282
營業額(K)	158073	155001	155194	103997	105309	55848	76621	90687	49905	68691	38658	24936

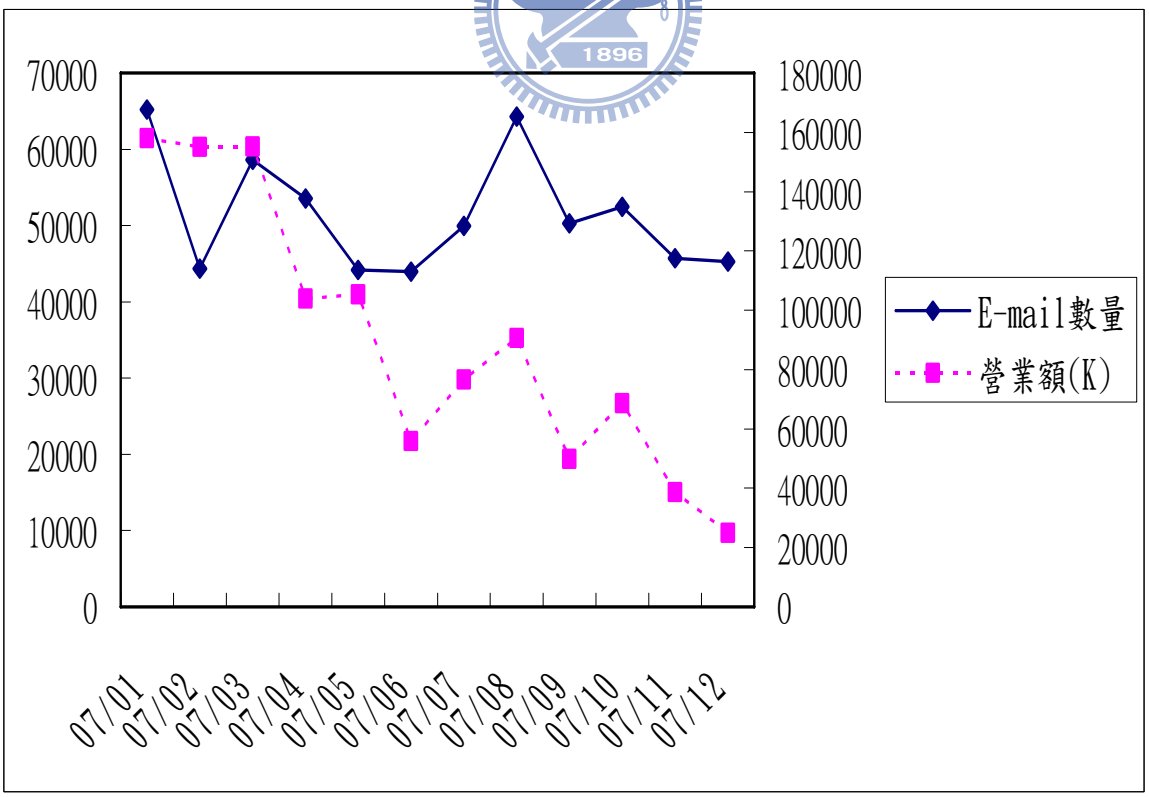


圖 24 E-Mail 與公司營業額之相對折線圖

就上圖分析 E-Mail 進出數量，與公司營業額之相關關係，其兩個資料之曲線圖，公司營業額之每個月增減也影響到 E-Mail 進出的數量。二月在過年上班日減少，造成 E-Mail 數量相對較少並未與營業額數量同步，其結果也相當合理。

就我們在本節前面所敘述與預估，E-Mail 進出數量與公司營業額之多寡成正相關。也就是當營業額增加時，當月之 E-Mail 進出數量也將增加；而當營業額減少時，當月之 E-Mail 進出數量也隨之減少。

4.1.4 E-Mail 群組與部門之關係

分析 E-Mail 進出數量，與公司各個部門之相關關係，將上節實驗中產生出來的組數，進一步去把各組數之收發者分析其所屬之部門，整理後得到下表 9，我們也知道這些部門之間與公司外之聯繫，互相都有密切的合作關係。 表 10 表示 9 中各部門代表之全名。

表 9 群組 E-Mail 收發者所屬之部門

2007 年 月 / 部門	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
PC-PR	9	5	7	4	7	6	5	5	3	4	2	1
PC-PE	9	5	1	4	3	1	1	1	0	0	0	3
PC-PR-PE	2	3	9	1	0	0	2	2	3	3	2	3
PC-SALES	2	1	4	6	1	0	4	3	9	8	0	1
PC	2	2	2	1	1	3	0	2	1	0	0	0
QC	0	0	0	1	0	1	1	0	0	1	0	1
PR-PE-RD	3	0	0	0	0	0	0	0	0	0	0	0
PC-TEST	1	0	0	0	0	0	0	0	0	0	0	0
PR-PE	1	0	0	0	0	0	0	0	0	0	0	0
PE-TEST	1	0	2	0	0	0	0	0	0	0	0	0
PC-RD	1	0	0	0	0	0	0	0	0	0	0	0
PC-QC	0	2	0	0	0	0	0	0	0	0	0	0
SALES	0	1	0	0	0	0	0	0	0	1	0	0
PE-TEST-RD	0	1	0	0	0	0	0	0	0	0	0	0
PC-TEST-RD	0	0	2	0	0	0	0	0	0	0	0	0
Total	31	20	27	17	12	11	13	13	16	17	4	9

表 10 表示 9 中各部門代表之全名

代號	部門
PC	Production Control Dept.
PE	Product Engineering Dept.
PR	Purchase Dept.
SALES	Sales Dept.
QC	Quality Control Dept.
RD	Design Dept.
TEST	Testing Dept.

把上表進一步分析，將公司各個部門在每個組數之相對關係分開，統計每個部門曾經出現在幾個組數上，得到下面的表 11。

圖 25-圖 31 中各部門代表之全名。

表 11 統計每個部門曾經出現在幾個組數

2007 年 月 / 部門組數	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
PC	26	18	25	16	12	10	12	13	16	15	4	8
PR	15	8	16	5	7	6	7	7	6	7	4	4
PE	6	9	12	5	3	1	3	3	3	3	2	6
SALES	2	2	4	6	1	0	4	3	9	8	0	1
QC	0	2	0	1	0	1	1	0	0	1	0	1
TEST	2	1	4	0	0	0	0	0	0	0	0	0
RD	4	1	2	0	0	0	0	0	0	0	0	0

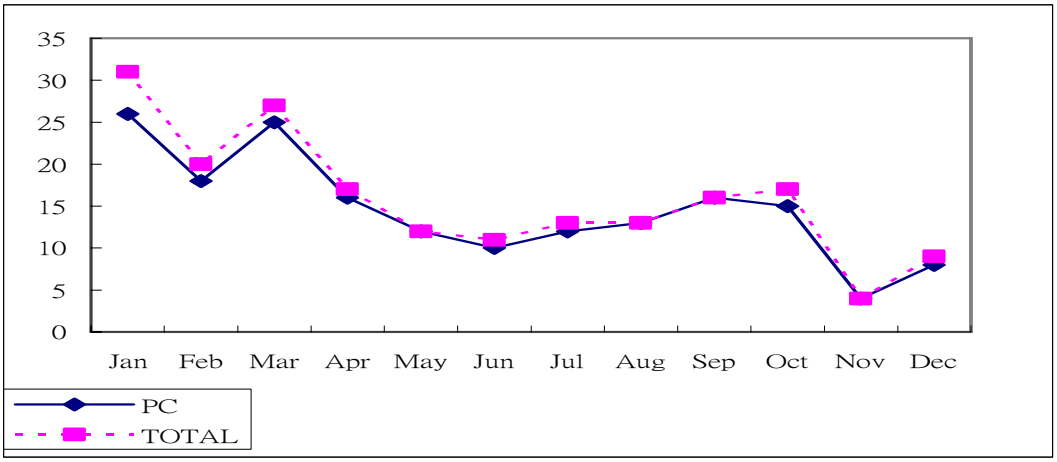


圖 25 PC 部門與所有群組之比較

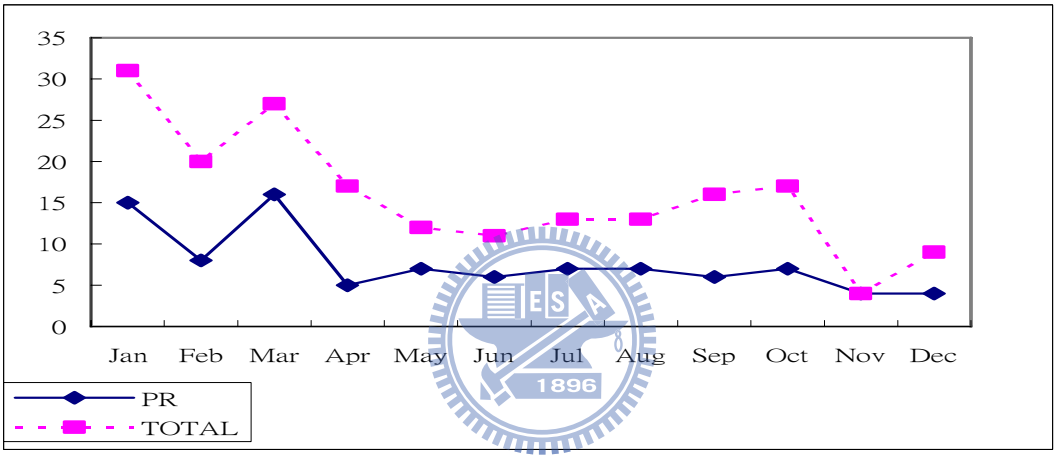


圖 26 PR 部門與所有群組之比較

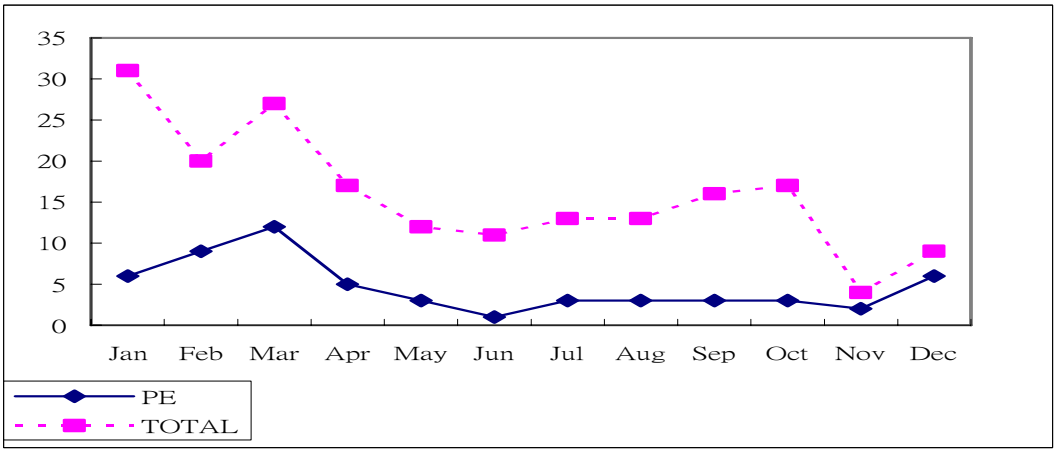


圖 27 PE 部門與所有群組之比較

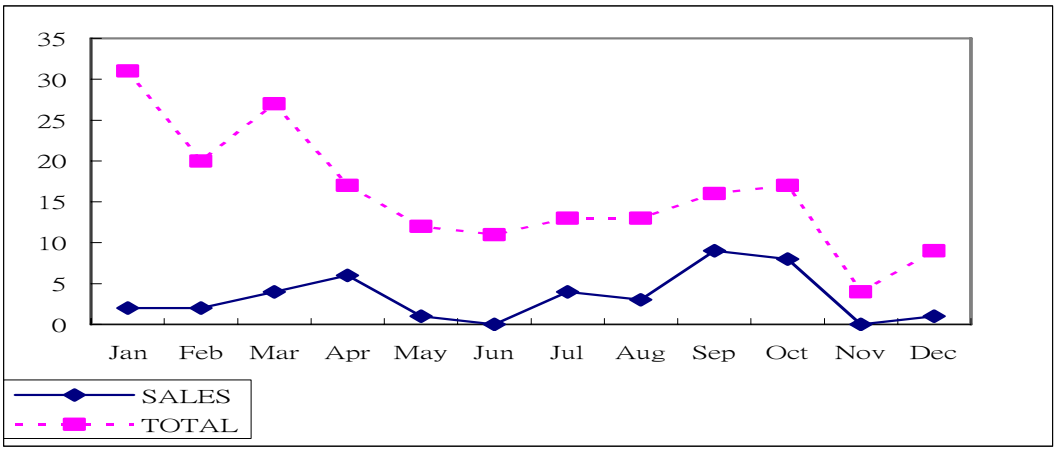


圖 28 SALES 部門與所有群組之比較

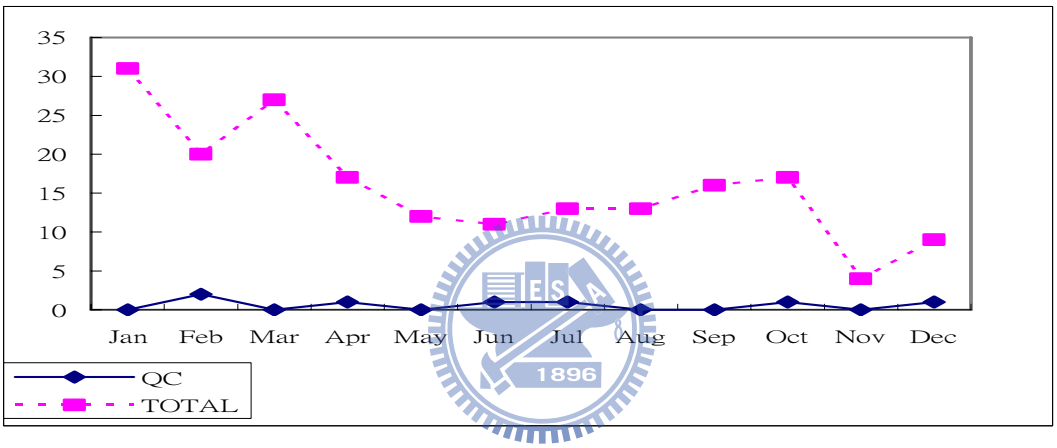


圖 29 QC 部門與所有群組之比較

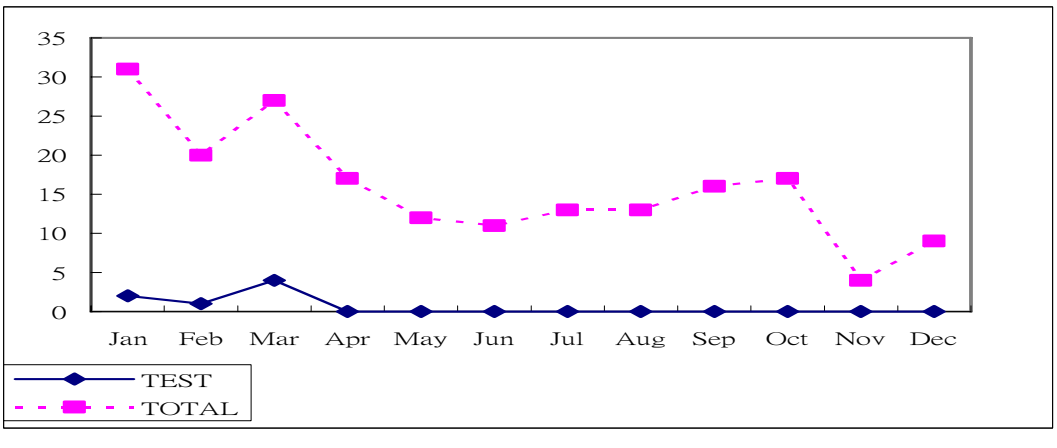


圖 30 TEST 部門與所有群組之比較

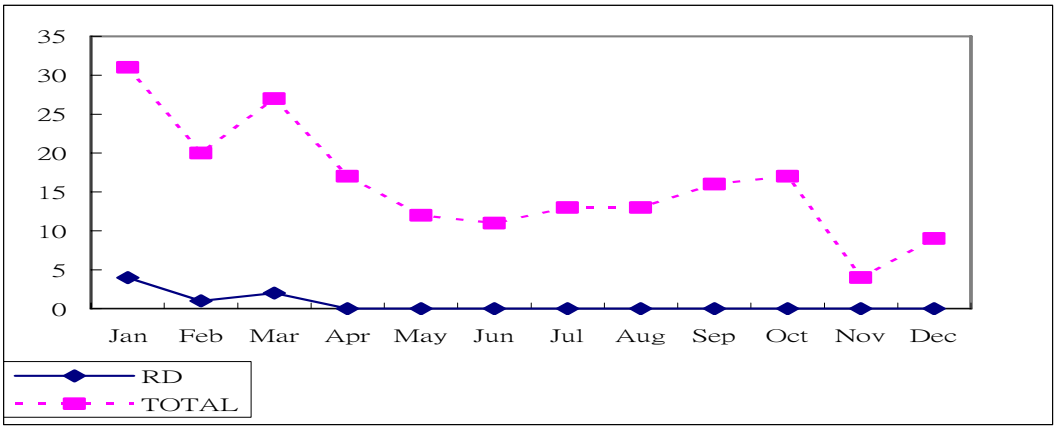


圖 31 RD 部門與所有群組之比較

由圖 25 至 31 分析 E-Mail 群組與各部門之相關程度，了解各部門對外的聯絡程度，推論出以下的排序

表 12 部門對外連絡程度

排序	代號	部門
1	PC	Production Control Dept.
2	PR	Purchase Dept.
3	PE	Product Engineering Dept.
4	SALES	Sales Dept.
5	QC	Quality Control Dept.
6	TEST	Testing Dept.
7	RD	Design Dept.

4.2 部門與部門關係實驗結果與細節

4.2.1 群組內之部門與部門之關係

由前一章節中我們提出之方法，計算群組內部門與部門之間的關係數值，數值越高代表這兩個部門關係越密切；反之，數值越低代表這兩個部門關係並不是很密切。

我們針對2007年每個月去做計算，得到的結果如下表13：

表 13 2007 年部門與部門之關係值

一月	PC	PR	PE	TS	SL	RD	QC
PC		62	57	3	12	3	0
PR	62		12	0	0	6	0
PE	57	12		3	0	12	0
TEST	3	0	3		0	0	0
SALES	12	0	0	0		0	0
RD	3	6	12	0	0		0
QC	0	0	0	0	0	0	

二月	PC	PR	PE	TS	SL	RD	QC
PC		54	39	0	6	0	6
PR	54		3	0	0	0	0
PE	39	3		2	0	2	0
TEST	0	0	2		0	1	0
SALES	6	0	0	0		0	0
RD	0	0	2	1	0		0
QC	6	0	0	0	0	0	

三月	PC	PR	PE	TS	SL	RD	QC
PC		103	38	0	15	0	0
PR	103		15	0	0	0	0
PE	38	15		11	0	4	0
TEST	0	0	11		0	2	0
SALES	15	0	0	0		0	0
RD	0	0	4	2	0		0
QC	0	0	0	0	0	0	

四月	PC	PR	PE	TS	SL	RD	QC
PC		49	20	0	22	0	0
PR	49		4	0	0	0	0
PE	20	4		0	0	0	0
TEST	0	0	0		0	0	0
SALES	22	0	0	0		0	0
RD	0	0	0	0	0		0
QC	0	0	0	0	0	0	

五月	PC	PR	PE	TS	SL	RD	QC
PC		45	11	0	6	0	0
PR	45		2	0	0	0	0
PE	11	2		0	0	0	0
TEST	0	0	0		0	0	0
SALES	6	0	0	0		0	0
RD	0	0	0	0	0		0
QC	0	0	0	0	0	0	

六月	PC	PR	PE	TS	SL	RD	QC
PC		49	7	0	0	0	0
PR	49		2	0	0	0	0
PE	7	2		0	0	0	0
TEST	0	0	0		0	0	0
SALES	0	0	0	0		0	0
RD	0	0	0	0	0		0
QC	0	0	0	0	0	0	

七月	PC	PR	PE	TS	SL	RD	QC
PC		58	14	0	14	0	0
PR	58		4	0	1	0	0
PE	14	4		0	0	0	0
TEST	0	0	0		0	0	0
SALES	14	1	0	0		1	0
RD	0	0	0	0	1		0
QC	0	0	0	0	0	0	

八月	PC	PR	PE	TS	SL	RD	QC
PC		46	9	0	10	0	0
PR	46		3	0	0	0	0
PE	9	3		0	0	0	0
TEST	0	0	0		0	0	0
SALES	10	0	0	0		0	0
RD	0	0	0	0	0		0
QC	0	0	0	0	0	0	

九月	PC	PR	PE	TS	SL	RD	QC
PC		45	9	0	33	0	0
PR	45		5	0	4	0	0
PE	9	5		0	0	0	0
TEST	0	0	0		0	0	0
SALES	33	4	0	0		0	0
RD	0	0	0	0	0		0
QC	0	0	0	0	0	0	

十月	PC	PR	PE	TS	SL	RD	QC
PC		46	12	0	22	0	0
PR	46		4	0	4	0	0
PE	12	4		0	0	0	0
TEST	0	0	0		0	0	0
SALES	22	4	0	0		0	0
RD	0	0	0	0	0		0
QC	0	0	0	0	0	0	

十一月	PC	PR	PE	TS	SL	RD	QC
PC		35	11	0	0	0	0
PR	35		3	0	0	0	0
PE	11	3		0	0	0	0
TEST	0	0	0		0	0	0
SALES	0	0	0	0		0	0
RD	0	0	0	0	0		0
QC	0	0	0	0	0	0	

十二月	PC	PR	PE	TS	SL	RD	QC
PC		35	21	0	4	0	0
PR	35		5	0	0	0	0
PE	21	5		0	0	0	0
TEST	0	0	0		0	0	0
SALES	4	0	0	0		0	0
RD	0	0	0	0	0		0
QC	0	0	0	0	0	0	

由上面圖表觀察整年的部門關係，其中某些部門相關性非常密切(如：PC-PR)，也有一些部門相關性微乎其微(如：TEST-SALES)，我們就幾個相關性較高的部門試著來解釋，這些部門的相關性。

1. PC(Production Control Dept.)-PR(Purchase Dept.)：

PC生產部門負責生產流程的掌控，對於每個生產流程需要隨時注意是否缺料，PR採購部門需要隨時支援生產部門之缺補料。這兩個部門之關係想當然耳一定是非常密切的。這也在我們實驗結果中證實了。

2. PC(Production Control Dept.)-PE(Product Engineering Dept.)：

PC生產部門負責生產流程的掌控，對於每個生產流程之良率也是要相當注意的，PE產品工程部門對於良率的分析，與產品各加工流程產生的問題，亦需要去分析解決。這兩個部門之關係應該也是有一定程度的密切。其在實驗中也證明其相關性。

3. PC(Production Control Dept.)-SALES(Sales Dept.)：

PC生產部門負責生產流程的掌控，對於每個成品之掌控是非常重要的，對於產出量與交期一點都不能馬虎，SALES銷售部門對於成品的掌控是一定要非常清楚，對於客戶的訂單與反應才能有效表達，才能對公司的銷售能掌控。這兩個部門之關係當然有一定程度的密切。其在實驗中也證明其相關性。

4. PR(Purchase Dept.)-PE(Product Engineering Dept.)：

PR採購部門需要隨時透過良率的分析，對於不良率偏高的半成品對廠商求償，PE產品工程部門對於良率的分析，與產品各加工流程產生的問題，要隨時反映給採購部門，也許採購部門將變更外包商。這兩個部門之關係似乎不事項以上之部門那麼密切。我們針對部門屬性似乎也可以了解。

當然公司還有許多的部門，在我們的實驗結果來看，除了上述的幾個，其他部門的相關性，並沒有這麼的突顯。QC(Quality Control Dept.)品質控制部門，針對產品品質檢測、甚至於 ISO 文件的處理；TEST(Testing Dept.)測試部門，對於設計部門設計的產品，開發將如何測試其可用性之部門；RD(Design Dept.)，設計部門然是設計新產品、或依客戶需求改版現有產品。這幾個部門與其他部門在 E-Mail 的聯繫與相關性，並不特別明顯。當然，就我們的了解，這幾個部門，幾乎都是比較有獨立性的作業，在聯繫上當然也是相對減少。以實驗結果來看似乎也是蠻符合的。

4.2.2 部門關係與營業額之關係

由前一章節中我們計算部門與部門的關係之值，我們在前段提到的關係較密切之部門與公司營業額來做比較，計算群組內部門與部門之間的關係數值，數值越高代表這兩個部門關係越密切；反之，數值越低代表這兩個部門關係並不是很密切。

表 14 相關部門與公司營業額之統計表

Monthly	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
PC-PR	62	54	103	49	45	49	58	46	45	46	35	35
PC-PE	57	39	38	20	11	7	14	9	9	12	11	21
PC-SALES	12	6	15	22	6	0	14	10	33	22	0	4
PR-PE	12	3	15	4	2	2	4	3	5	4	3	5
營業額(百萬)	158	155	155	104	105	55.8	76.6	90.7	49.9	68.7	38.7	24.9

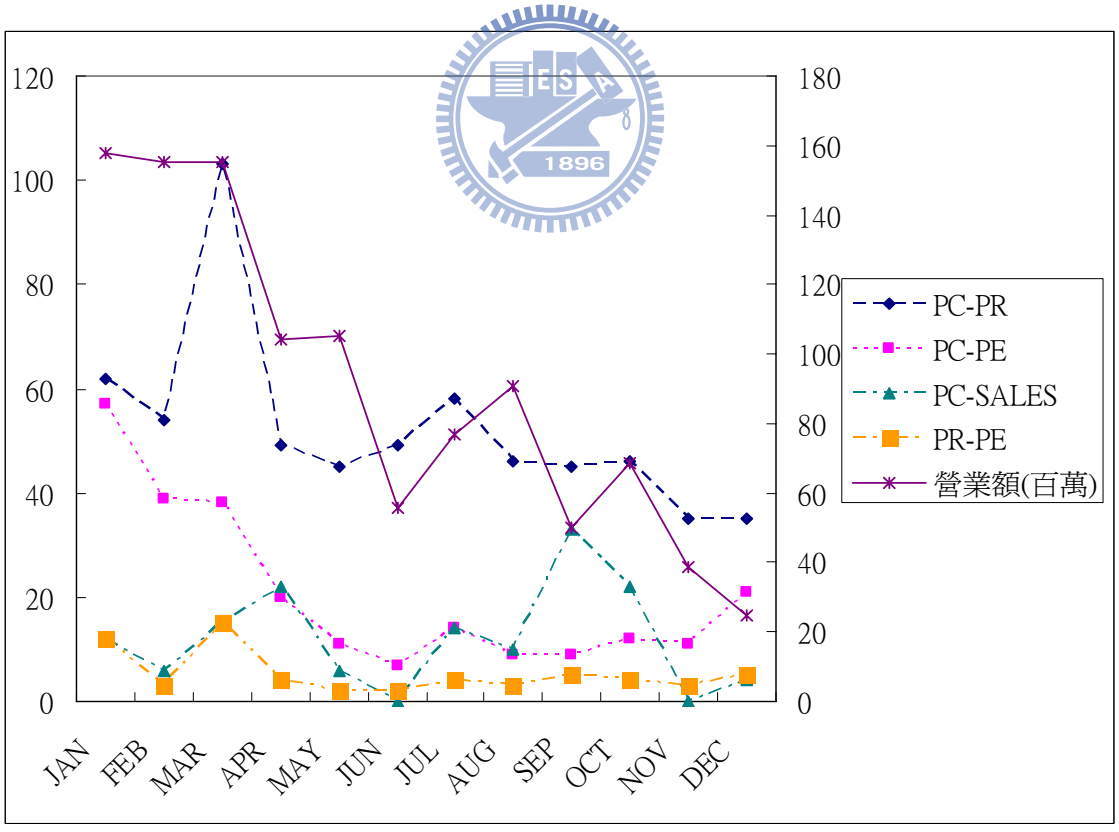


圖 32 相關部門與公司營業額之統計表

就上圖分析相關部門之係數，與公司營業額之相關關係。其每個相關係數之曲線圖與公司營業額之每個月增減看起來還蠻有關係的，相關部門當月關係係數越大，其公司營業額亦有增加；反之，相對就減少。二月在過年上班日減少，造成 E-Mail 數量相對較少並未與營業額數量同步，其結果也相當合理。

就我們在本節前面所提出的方法去所作之實驗結果，兩個部門的相關係數與公司營業額存在著相當程度的相似度。也就是當營業額增加時，當月之相關部門關係細數增加，也就是這兩個部門當月 E-Mail 進出數量增加；而當營業額減少時，當月之相關部門關係細數減少，當月之 E-Mail 進出數量也隨之減少。

4.3 部門與供應鏈實驗結果與細節

4.3.1 群組內之部門與供應鏈之關係

由前一章節中我們提出之方法，計算群組內部門與供應鏈之間的關係數值，數值越高代表這個部門與此供應鏈關係越密切；反之，數值越低代表這個部門與此供應鏈關係並不是很密切。

由於公司屬於IC Design產業，我們把供應鏈分為上游供應商、下游客戶與其他信件。

1. 上游供應商：包含Wafer廠商(如：VIS，TSMC)，製程外包商(如：ChipMos，ChipBond，ist)。
2. 下游客戶：公司賣出的成品，對公司有盈收貢獻的客戶（如：AUO，Innoulux）
3. 其他信件：不屬於前面之信件皆屬於此類別。包含運送貨物之廠商、政府機關之郵件、垃圾郵件等。

我們針對2007年每個月去分類做計算，得到的結果如下表：

表 15 上游供應商與部門之關係

月份 / 部門	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
PC	315	372	406	379	124	294	180	327	311	249	61	83
PR	54	54	118	45	28	108	46	54	85	75	15	19
PE	110	59	76	23	10	17	12	10	16	14	4	12
SALES	12	6	21	42	8	0	8	17	53	52	0	4
TEST	7	4	8	0	0	0	0	0	0	0	0	0
RD	31	4	3	0	0	0	2	0	0	0	0	0
QC	0	13	0	16	9	21	0	20	0	20	0	8

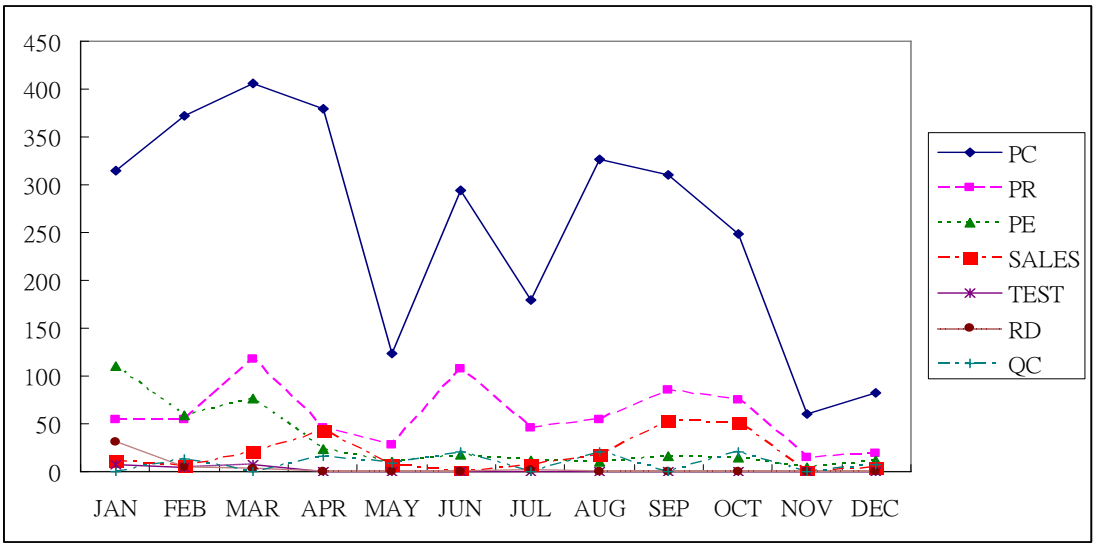


圖 33 上游供應商與部門之關係

表 16 下游廠商與部門之關係

月份 / 部門	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
PC	0	9	64	12	23	16	8	16	3	32	37	16
PR	0	3	8	0	0	12	0	0	0	0	8	7
PE	0	1	10	0	0	2	0	0	0	0	3	3
SALES	0	14	15	8	2	0	4	8	2	44	0	2
TEST	0	0	5	0	0	0	0	0	0	0	0	0
RD	0	0	2	0	0	0	0	0	0	0	0	0
QC	0	0	0	0	0	0	0	0	0	0	0	4

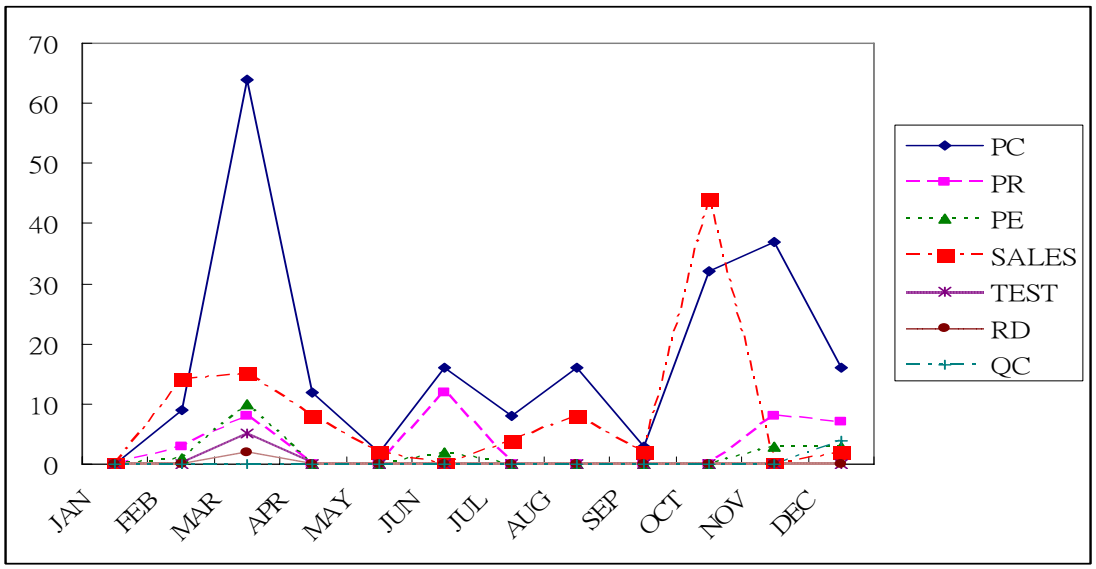


圖 34 下游廠商與部門之關係

表 17 其他郵件與部門之關係

月份 / 部門	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
PC	928	23	618	7	471	6	469	2	4	9	185	253
PR	168	0	177	0	104	0	100	0	0	3	47	56
PE	201	10	165	0	24	0	23	0	0	0	16	37
SALES	38	6	50	5	26	17	72	2	1	2	0	16
TEST	14	1	47	0	0	0	0	0	0	0	0	0
RD	59	1	2	0	0	0	18	0	0	0	0	0
QC	0	0	0	0	24	6	40	8	0	0	0	44

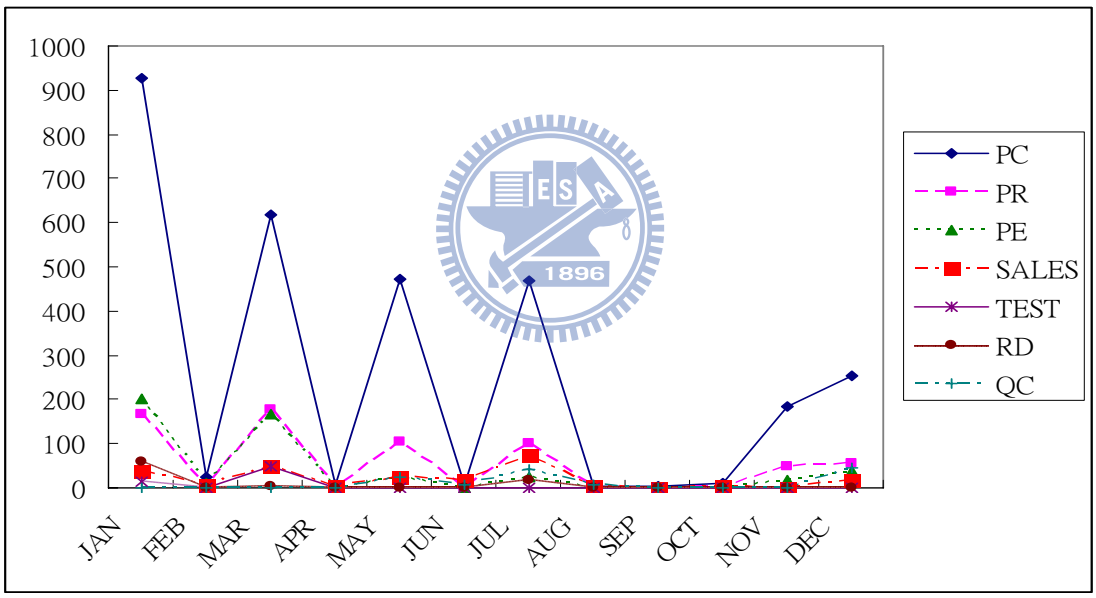


圖 35 其他郵件與部門之關係

由上面圖表觀察整年的部門與上下游廠商關係，其中某些部門與上游相關性非常密切(如：PC，PR，PE，SALES)，也有一些部門相關性微乎其微(如：TEST，RD，QC)；就下游的關係比較密切的是(如：PC，SALES)，其他的部門就似乎沒什麼關係。下面我們就幾個相關性較高的部門與上下游關係來解釋，這些相關性。

1. 上游供應商：

由於IC Design油原料製成品須時二--三個月的時間，對於生產的掌控是非常重要的。PC生產部門負責生產流程的掌控，對於每個生產流程需要隨時注意是否缺料，PR採購部門需要隨時支援生產部門之缺補

料。SALES必須了解生產進度，適時回報至客戶何時出貨，了解何時有貨可以賣。PE部門必須在生產過程中，隨時注意良率，與產品各加工流程產生的問題，亦需要去分析解決，以節省公司的生產經費。這四個部門之關係想當然耳一定是非常密切的。這也在我們實驗結果中證實了。

2. 下游客戶：

SLAES部門絕對要與客戶保持聯繫，開發新客戶與維繫既有客戶都是他們的主要工作。PC生產部門負責成品的出貨流程，隨時追蹤產品是否已平安到客戶手中。這兩個部門與下游客戶之關係一定是相當密切。其在實驗中也證明其相關性。

3. 其他信件：

由於這個供應鏈分類，較無法與各部門之關係，相關性進行分析，所以我們也較無從判斷。

4.3.2 SALES 部門與下游客戶關係值與營業額之關係

由前一章節中我們計算SALES部門與下游客戶的關係之值，每家公司SALES部門一定與客戶關係密切，我們利用此關係值與公司營業額來做比較，看看其是否成正相關，數值越高是否代表營業額越好；反之，數值越低是否代表營業額相對較不好。

表 18 SALES 部門與公司營業額之統計表

月份 / 項目	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
SALES	0	14	15	8	2	0	4	8	2	44	0	2
營業額(百萬)	158	155	155	104	105	55.8	76.6	90.7	49.9	68.7	38.7	24.9

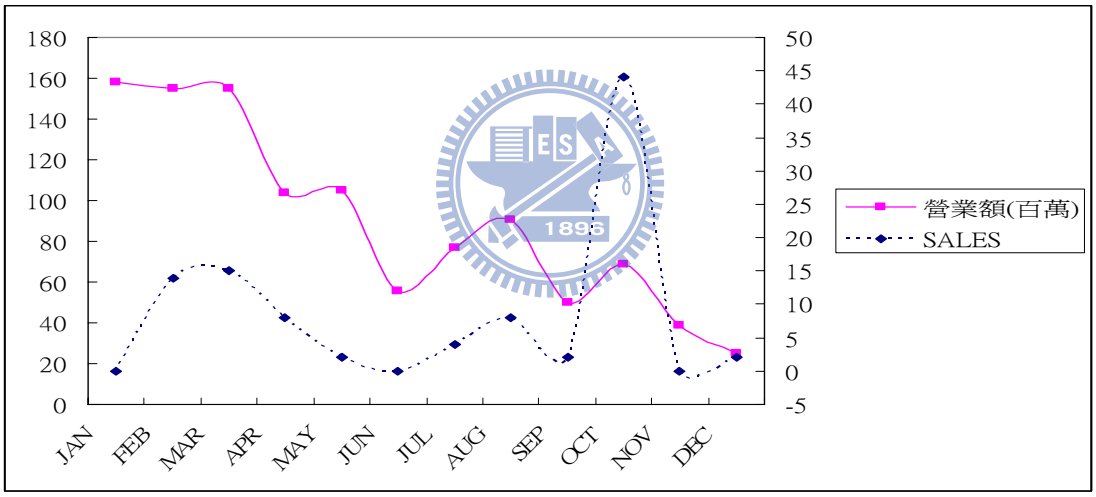


圖 36 SALES 部門與公司營業額之統計表

就上圖分析 SALES 部門與下游客戶之係數，與公司營業額之相關關係。其每個相關係數之曲線圖與公司營業額之每個月曲線增減看起來還蠻有關係的，SALES 部門與下游客戶當月關係係數越大，其公司營業額亦有增加；反之，相對就減少。十月 SALES 部門與下游客戶之係數特別高，因為在十一月公司營業額極速下降，當月 SALES 部門積極在尋找新的客戶，造成 E-Mail 數量增加許多與下游客戶之聯繫，並未與營業額數量同步減少，其結果也相當合理。

4.4 其他實驗結果與細節

4.4.1 純度(Purity)與亂度(Entropy)

由前一章節中我們提到的公式，計算出每個月之純度與亂度。純度數值越大越好(越接近1越好)、反之越不好；亂度數值越小、反之越不好。

我們每個月的純度都在0.6以上，表示我們分出來的群組相當不錯，亂度除幾個月稍差，也在控制在不錯的範圍。當然純度與亂度是相反的指標，純度越好亂度數值就越小、反之純度越差亂度就越大，從我們的實驗結果也可看出這樣的結果。

表 19 2007 年各月純度與亂度表

Item / Month	Cluster Purity	Average Entropy
JAN	0.72022161	0.81612569
FEB	0.88679245	0.40882225
MAR	0.60778443	1.10244306
APR	0.91836735	0.28113897
MAY	0.77222222	0.77122419
JUN	0.91558442	0.31889814
JUL	0.76923077	0.76745925
AUG	0.8902439	0.33225546
SEP	0.97826087	0.08162799
OCT	0.85046729	0.40995515
NOV	0.65957447	1.23300294
DEC	0.72641509	0.97729739

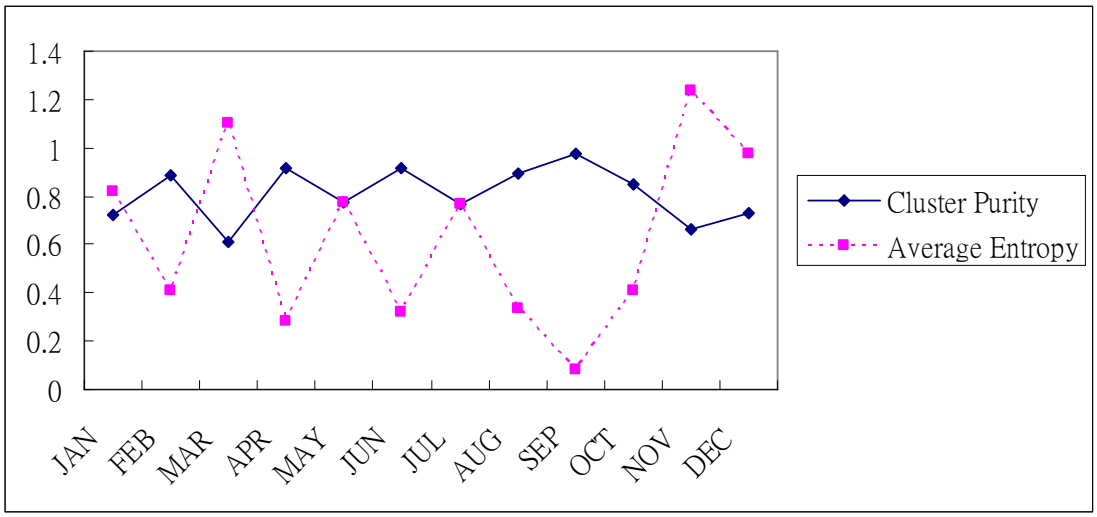

1.4
1.2
1
0.8
0.6
0.4
0.2
0
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC
Cluster Purity
Average Entropy

圖 37 純度與亂度

4.4.2 部門關係使用社會網路方式顯現

由4.2.1節中群組內之部門與部門之關係，我們使用2007年一月之部門關係值，畫出部門與部門之關係網路，其關係細數在節點跟節點之間呈現。由下圖(圖38)我們可以看出每個部門之關係程度：

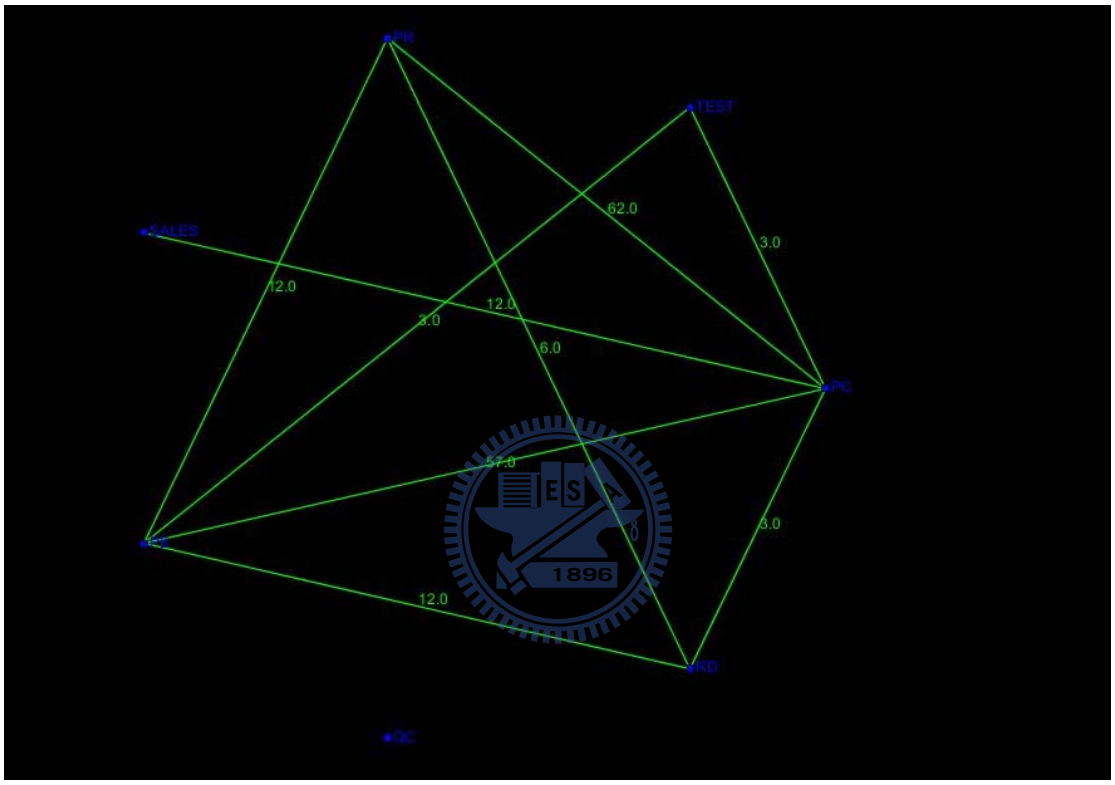


圖 38 2007 年一月部門與部門之關係

第五章 結論與未來方向

E-Mail 已經是人生活中不可或缺的工具，對企業更是不能沒有的溝通管道。我們嘗試從實際的進出紀錄中，找到個人與群組的關係、部門與部門的關係、部門與供應鏈的關係，再就對這些關係與公司營業額是否有相關的關係。就既有的主觀意識來說，本篇論文提出的方法已經證實了我們對這些關係的認知與判斷大部分都是對的。

對於身為企業老闆，希望可以每一年度的 E-Mail 進出資訊，找到各個部門的關係(相依度)與公司營業狀況的關係，找出一套思維，加強部門之間的溝通管道提升公司的業績。人資部門主管，也許希望透過這個 E-Mail 透露的資訊，一旦有一個人離職的話, 那個群組的其他的人離職的機會也會提升，這群組屬於高度不穩定的一群人，需要加強心理建設輔導，安撫剩下的員工，可以降低員工的離職率。IT 部門主管，也能透過這樣的資訊，防止公司機密資訊的外洩，甚至可以了解是否有人在對公司做不利的舉動，提早知會老闆，做有效的防範，防止公司的損害。Mining E-Mail 技術為這些問題提供了解決方案，並取得較好的效果。

我們於文中使用 LCM-freq 的資料探勘方法，加上我們提出的方法，可以分析出公司中部門與部門的關係程度、部門與供應鏈的關係。進而尋找出每個部門對營業額的影響程度、SALES 部門對供應鏈與營業額的關係。由這些關係，用於公司中，小至人事的轉變、大至公司的營運，都可提供不錯的參考資訊，應用於公司的管理與發展策略。

郵件智慧 (Mail Intelligent) 似乎是可以未來表達 Mining E-Mail 這個概念的名詞，將發掘出的郵件知識，加以分析 (Analysis)、理解 (Insight)、行動 (Action)、量化 (Measurement)，郵件內含可觀的企業資源，郵件的歷史資料與內容可以匯聚成為企業知識庫，透過 Mail Mining 技術，可以協助企業統計、挖掘與分析隱含的郵件知識。各類郵件附加檔案的再利用更可以避免資源浪費，從郵件延伸的行為，如收送時間，地點，單位，日流量等等，透過行為與資料的比對分析得到的情報，更可以提供企業重要決策的參考。是否了解企業中每日溝通的大量郵件中隱含多少有價值的資訊呢？同時又代表怎樣的組織與個人關係，又該如何進行管理？或許我們可以思考這個問題。

參考文獻

[1]. Takeaki Uno, Taisuya Asai, Yuzo Uchida, Hiroki Arimura “LCM：AN Efficient Algorithm for Enumerating Frequent Closed Item Sets,” In Proc. IEEE ICDM99 Workshop FIMI’03,2003.

[2]. Christian Bird, Alex Gourley,Prem Devanbu, Michael Gertz Dept. of Computer Science, Kemper Hall, Anand Swaminathan Graduate School of Management,University of California, Davis, Davis, California “Mning Email Social Networks,” MSR’06, May 22－23, 2006, Shanghai, China.

[3]. Cheng Yung Hsiung, Dr. Tasi Hsien Leing , A Thesis Submitted to Department of Information Management, I-Shou University, “A study of association rule mining algorithms,” Jule 2007.

[4]. Jimeng Sun, Philip S. Yu , Carnegie Mellon University, Spiros Papadimitriou,, Christos Faloutsos, IBM TJ Watson lab, “GraphScope: Parameter-free Mining of Large Time-evolving Graphs,” International Conference on Knowledge Discovery and Data Mining (KDD) 2007.

[5]. Tang Chang-jie, Liu Wei, Wen Fen-lian and Qiao Shao-jie,School of Computer Science, Sichuan University, Chengdu , ” Three probes into the social network and Consortium Information Mining--Mining the structure, Core and communication behavior of virtual Consortium,” 2006.

[6]. Olivier de Vel.,”Mining Email Authorship,” International Conference on Knowledge Discovery and Data Mining (KDD)-2000 Workshop on Text Mining, August 20, 2000, Boston.

[7]. Olivier de Vel, A.Anderson, M..Corney and G.Mohay, ” Mining E-Mail Content for Author Identification Forensics,” SIGMOD Record, 2001, Vol. 30, No. 4, 55-64 Olivier de Vel. Mining Email Authorship. KDD-2000 Workshop on Text Mining, August 20, 2000, Boston.

[8]. Marshall van Alstyne and Jun Zhang. EmailNet, ” A System for

Automatically Mining Social Networks from Organizational Email Communication," In NAACSOS2003, 2003.

[9]. R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu," Mining newsgroups using networks arising from social behavior," In WWW '03: Proceedings of the 12th international conference on World Wide Web, 2003.

[10]. Fayyad, U.M., "Data Mining and knowledge Discovery: Making Sense Out of data," IEEE Expert, Volume 11, Issue 5, pp. 20-25, 1996.

[11]. Freeman, L., "Centrality in Social Networks: I. Conceptual Clarification," Social Networks, 1979.

[12]. Wellman, B., "For a Social Network Analysis of Computer Networks: A Sociological Perspective on Collaborative Work and Virtual Community," Proceedings of ACM, 1996.

[13]. Garton, L. and Haythornthwaite, C. and Wellman, B., "Studying Online Social Networks," Journal of Computer-Medicated Communication, June 1997, http://jcmc.huji.ac.il/vol3/issue1/garton.html.

[14]. Hanneman, R.A., "Introduction to Social Network Methods," http://faculty.ucr.edu/hanneman/SOC157/NETTEXT.pdf

[15]. J.Han, J.Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation, " Proceedings of the ACM SIGMOD International Conference on Management of Data, Dallas, USA, pp. 241-250, 2000.

[16]. M.J. Zaki and C.J. Hsiao, "CHARM: An Efficient Algorithm for Closed Itemset Mining," Proc. 2002 SIAM Int' l Conf. Data Mining (SDM' 02), pp. 457-473, 2002.

[17]. H. Li, Z. Nie, W-C Lee, C. Lee Giles, J-R Wen, "Scalable Community Discovery on Textual Data with Relations," Intl. Conf on Information and Knowledge Management (CIKM) 2008.

[18]. Lars Backstrom, Dan Huttenlocher, Joh Kleinberg, Xiangyang

Lan, “Group Formation in Large Social Networks: Membership, Growth, and Evolution,” International Conference on Knowledge Discovery and Data Mining (KDD) 2006.

[19].Michael E.Houle,”The Relevant-set Correlation Model for Data Clustering,” SIAM International Conference on Data Mining (SDM) 2008.

[20].Chayant Tantipathananandh,Tanya Berger-Wolf,David Kempe , ”A Framework For Community Identification in Dynamic Social Networks,” Association for Computing Machinery’s Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) 2007.

[21].Neil Chang, “郵件探勘技術基本概念探討”, “從郵件探勘淺談關係理,” Mail Intelligence.

[22].謝德平, 南京大學 計算機科學與技術系, 江蘇 南京, ”A Study Report for Mining Email,” 2004 Journal of Software.

[23].Jiawei Han, Micheline Kamber,”DATA MINING Concepts and Techniques,”北京: 高等教育出版社, 2001.5.