專題名稱：漸進式郵件分類模式
年級與班級：資訊管理學系四年 B 班
指導教師：楊士霆老師
聯絡人姓名與 E-mail：吳朝宏、james36915@gmail.com
聯絡人電話：0921241455
專題學生：吳朝宏、丁偉哲、莊貽任、郭建良

摘要

目前處理郵件管理分類之方式，通常是採用人工閱讀內容、判斷其信件性質，再進行分類與管理，但此種做法不僅需要聘請大量人工，亦需要花費許多時間等待，才可以整理分類好所有資料，且分類正確率往往不佳。本研究所提之「漸進式郵件分類模式」乃以郵件中所包含之郵件內容為基礎，先行將此些資料進行資料層級之歸類（包含為「數字型」、「數據型」、「文字型」、「附檔案型」及「郵件聯絡人」等資料型別）；當中，「數字型」與「數據型」為第一層級資料、「文字型」與「附檔案型」為第二層級資料、「郵件聯絡人」則為第三層級資料；待資料彙整完成後，即可以本研究所建構之「郵件數據型、數字型資料解析模組」分析郵件數據型及數字型資料，先行過濾垃圾郵件，以判定適合分類之郵件群；之後，利用「郵件文字型資料解析」模組，計算目標郵件與各類別之隸屬關係；最後，當目標郵件之類別判定結果不甚理想，即可利用「郵件聯絡人資料解析模組」重新計算目標郵件與各類別之隸屬關係，進而形成一套漸進式郵件類別判定模式，期望以具效率之方式進行郵件類別判定任務。

關鍵字：*郵件分類、關鍵字擷取、知識管理*

一、研究動機與目的

隨著全球人們漸漸習慣網路之通訊方式，利用網路收發電子郵件已成為人們生活中重要部分。由於電子郵件之多樣性（如廣告信件、教育信件、會議信件等），使用者常需要花費時間去管理分類，並從中擷取、儲存其所需要之知識。目前管理郵件類別之方式，通常是採用人工閱讀內容、判斷其信件性質，再進行分類與管理，但此種做法需要花費許多時間，才可整理分類好所有資料，且分類效果往往不佳。有鑑於此，目前多數研究乃著重於解析電子郵件標頭欄位內「Date」、「From」、「Subject」以及本文內容之文字擷取，並且發展各種技術及運算法以解析郵件內容，然而當中僅解析電子郵件之郵件主旨、郵件內容等特徵值，並非分析郵件中所有內容，因此可能遺漏電子郵件中之重要分類訊息（目前之郵件分類既有模式如圖 1 所示）。

有鑑於此，本研究除解析電子郵件之寄件者、郵件主旨、郵件內容和檔案內文等特徵值外，亦解析件標頭欄位之 Message-ID、Received、聯絡人及本文內容之附加檔案之檔名及副檔名，以及本文之寫作格式等內容。基於此些郵件需解析之內容，本研究乃先行將郵件中所包含之資料進行層級歸類（包含「數字型」、「數據型」、「文字型」、「附檔案型」及「郵件聯絡人」等資料型別），之後逐一解析同層級之資料（如文字型內容、數據型內容等），進而形成一套「漸進式郵件類別判定模式」；當中，此模式乃郵件資料進行資料層級之歸類（包含為「數字型」、「數據型」、「文字型」、「附檔案型」及「郵件聯絡人」

等資料型別）；當中，「數字型」與「數據型」為第一層級資料、「文字型」與「附檔案型」為第二層級資料、「郵件聯絡人」則為第三層級資料；待資料彙整完成後，即可以本研究所建構之「郵件數據型、數字型資料解析模組」分析郵件數據型及數字型資料，先行過濾垃圾郵件，以判定適合分類之郵件群；之後，利用「郵件文字型資料解析」模組，計算目標郵件與各類別之隸屬關係；最後，當目標郵件之類別判定結果不甚理想，即可利用「郵件聯絡人資料解析模組」重新計算目標郵件與各類別之隸屬關係，進而形成一套漸進式郵件類別判定模式，期望以具效率之方式進行郵件類別判定任務。

是故，本研究所提出之漸進式郵件類別判定模式，除能考量郵件中所有包含之資料外（亦即郵件中所有資料），並能以「漸進式」之概念逐層解析並判定類別；因此，本研究之郵件類別判定模式除可具備郵件分類較佳之分類效果外，亦可兼具較佳之分類效率。本研究所提郵件分類之期望模式（TO-BE Model）如圖 2 所示。

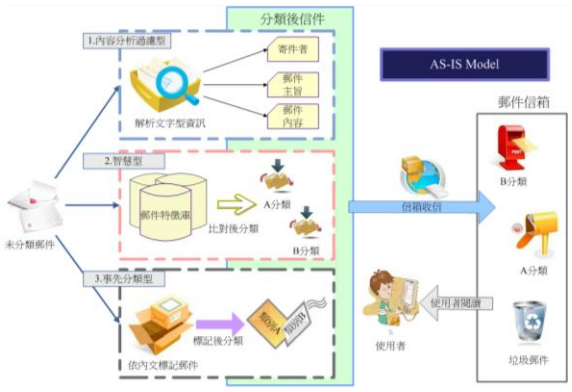

圖 1、郵件分類之既有模式（AS-IS Model）

圖 2、郵件分類管理之期望模式（TO-BE Model）

二、文獻探討

於郵件分類技術之議題中，過去相關研究乃以「郵件分類演算法」及「郵件資料解析」兩方面進行探討。

2.1 郵件分類演算法

現今郵件分類演算法中，主要包含 K-NN、決策樹、自然貝式、隨機森林、支持向量機等技術；此外，由於垃圾郵件之氾濫及郵件安全性逐漸受重視，故現有研究亦著重於垃圾郵件過濾與保密協定演算法之發展，如以下說明之。

2.1.1 K-NN、決策樹、自然貝式等資料探勘演算法

Zhou 等人[24]結合圖片信息測量（Picture Information Measurement；PIM）及關鍵字擷取技術，建構一套智慧型電子郵件分類系統；該系統乃使用簡易貝式分類機（Native Bayes Classifier）先行過濾郵件（以郵件內文之關鍵字以作為判斷特徵），之後，以圖片信息測量分析郵件附錄中所包含之圖像訊息，將各郵件進行分類。此外，Poon 及 Chang[17]利用電子郵件之詞彙相關性與 K 值鄰近演算法以進行電子郵件分類。該研究係依照郵件事先所建立之詞彙相關性，及各關鍵詞彙與類別之關係進行第一次分類，之後再利用 K 值鄰近演算法進行郵件最後之分

類。此外，Yu 與 Zhu[23]提出結合倒傳遞類神經網路 MBPNN （Modified Back-Propagation Neural Network）和語義特徵空間 SFS（Semantic Feature Space）以進行郵件之分類。傳統上倒傳遞類神經網路 BPNN（Back-Propagation Neural Network）採用陡坡下降法的觀念，可減少錯誤或總平均誤差輸出計算，然而具備減緩學習速率與易落入區域最小值之缺點。

Segal 及 Kephart[20]利用 TF-IDF 分類器，建構一套智慧型郵件分類機制（SwiftFile），以取代目前使用者需自行分類郵件之任務。該系統係先預設三個分類資料夾，使用者可先自行分類郵件至各資料夾，系統即透過分類器分析資料夾內各郵件內容定義與資料夾類型之關係，以學習使用者分類方式，使郵件寄送至使用者時自行分類至此三個資料夾，以達到郵件自動分類效果。為了改善過去研究需要使用者自定關鍵字的缺點，王瑄榕[1]使用 MMB（Multimembership Bayesian）先擷取郵件文字內容，利用斷詞技術使文句適當切割並留下名詞，之後合併重覆名詞產生組合名詞，最後利用 MMB 運算法，獲得各郵件隸屬類別。甚者，亦有研究針對郵件內文之字詞對郵件分類之影響[4]。

此外，Islam 等人[11]乃以統計學習機制（Statistical Learning Algorithms）建構一套郵件多層分類器（Multi-classifier），能夠有效地改善現今郵件分類系統中，存在的詞彙誤判（即具備灰色地帶之詞彙集）以及準確度之問題。最後，Irena 等人[10]以監督學習之機器學習技術，針對隨機森林（Random Forest）、支持向量機（SVMs）、決策樹（Decision Tree）及天真貝式（Naïve Bayes）四種分類法之郵件分類與垃圾郵件過濾之執行績效進行比較。該研究發現隨機森林（Random Forest）優於其他三項，甚至於較為龐大之資料庫中可迅速讀取、易於調整且精確率較高。

2.1.2 垃圾郵件過濾與保密協定演算法

另於過濾垃圾郵件方面，Shih 等人[21]整合現有的郵件分類方法（貝式、貝式網路及決策樹）之優點，以建構一套檢測惡意郵件分類器，以自動掃描與過濾所接受到之郵件，並阻擋病毒式電子郵件與封鎖潛在之惡意郵件。此外，Ying 等人[22]結合決策樹、支持向量機和倒傳遞類神經網路，提出一套整合性模式以分類垃圾郵件。該方法先將郵件內容、標題、寄件等資料歸納得 14 種郵件特徵，再以綜合分類法分析此 14 種特徵，以判定是否為垃圾郵件，進而達到過濾效果。此外，為了有效改善郵件過濾之效能，Gonzaʹlez[8]整合先前研究所提之黑名單（Blacklist）與白名單（Whitelist）解決方案之特性而成，經由實驗結果可得知，該程式可提升過濾垃圾郵件之績效。此外，Herzberg[9]結合路由器和 DKIM（Domain Keys Identified Mail）以協助電子郵件過濾任務，此機制之任務主要著重在不受歡迎電子郵件訊息之過濾，以改善目前垃圾郵件過濾效果不彰之情況。更甚者，Duan 等人[6]發展一套區別轉寄郵件協定，該協定乃允許收件者可控制由不同寄件者於網路上之遞送郵件。

最後，保密協定於電子郵件系統中代表郵件使用者乃擁有密碼鑰匙，以確保郵件訊息在傳遞與接收過程中不受到外來因素的影響。以往電子郵件系統所使用的保密協定是基於用戶使用 PKI（Public Key Infrastructure）公開金鑰基礎建設之公共密鑰認證，有鑑於此提出一項基於密碼認證方式之電子郵件保密協定，以提供保密的功能，該協定可於不要求公共密鑰認證的情形之下，能充分有效的執行在資源受到限制的移動裝置上[13,16]。

2.2 郵件資料解析

就郵件資料分析之議題而論，資料挖礦與分析技術不斷發展與進步，其主要分析郵件中「郵件社群資料」、「郵件垃圾訊息」、「郵件文字型資料」及「郵件使用者資料」等範疇，如以下分別說明之。

2.2.1 郵件社群資料分析

針對社群探勘之議題中，過去研究乃試圖由通聯記錄中（如電子郵件）找出熟識的使用者社群，此等社群關係包含家人、鄰居、同事、以及同學等[6]。此外，Chundi 等人[5]提出以時間為觀點以分析嵌入時間序列之分割（Time Series Segmentation），應用於發掘時間點跨越模式及隨時間變化之電子郵件溝通模式，並計算用戶端郵件溝通模式中時間序列項目集（Item-Set），尋找目標使用者之個人化溝通模式；此外，亦計算所有使用者的電子郵件資料中時間序列項目集，用以建立以社會為中心之溝通模式。

2.2.2 郵件垃圾訊息分析

由於電子郵件濫用的增長，調查員需要高效率的電子郵件自動化分析工具。Appavu 等人[3]提出以決策樹（Decision Tree）演算法為基礎，建構一套智慧型過濾恐怖信息郵件系統。此外，因電子郵件具備高度之便利性，利用電子郵件以進行消費者詐騙之情況日趨嚴重。因此，Neese 等人[15]分析電子郵件內容之產品推銷、產品、價錢、行銷通路方式之關鍵詞彙以判別該郵件是否為詐欺郵件，以降低消費者受騙機率。而隨著網路技術發展，電子郵件已經成為重要通訊工具，許多職業依靠者電子郵件傳達訊息（如：商業貿易或者教育機構），當中為避免錯過重要之郵件訊息，Kadoya 等人[12]提出確認郵件回覆時間，並依照多重屬性規則檢測時間語意表達，判斷使用者郵件重要訊息以及分類，此系統過濾郵件可讓使用者直接讀取重要郵件，忽略不必要之郵件。

2.2.3 郵件使用者資料、文字型資料解析

Sallis 和 Kassabova[19]為瞭解以電腦為媒介溝通中電子郵件所扮演的角色，該研究使用乃針對電子郵件進行定量和定性的分析；當中實驗資料係取自於新聞群組大型數據組，並透過分析電子郵件中文本屬性（如文字的數目、句子的長度與其他文體之特點），並以分數加權機制評估該電子郵件之易讀性。而 Nagabhushan 等人[14]則提出應用軟性演算模型（Soft Computing Model），使郵件地址可準確映射至確定之目的地，甚者若存不完整或相似之郵件地址，則運用可讀性之通訊地址，以進行象徵性分析，並依據內容相似以判別送達地點，進而提高郵件寄出正確目的之準確率。此外，吳文峰[2]考慮郵件中不同特徵（如字詞等特徵），並結合適用之分類器以預測郵件類別，之後使應用軟體能依單一標記直接將郵件分派到各分類目錄中，以減少使用者進行人工分類或訂定複雜分類規則。最後，Sakurai 及 Suyam[18]使用文字探勘技術分析顧客之郵件資料，該研究之分析資料包含郵件主題、郵件內容、及郵件主體以決定郵件類型，並從電子郵件中取得關鍵字和統計資料（如關鍵字數等資訊），藉由上述之郵件類型與關鍵資訊加以判定此顧客之類型，進而協助顧客中心操作人員主動地將各類資訊予顧客得知。

三、系統特色

本研究所提之「漸進式郵件分類模式」乃以郵件中所包含之郵件內容為基礎，先行將此些資料進行資料層級之歸類（包含為「數字型」、「數據型」、「文字型」、「附檔案型」及「郵件聯絡人」等資料型別）；當中，「數字型」與「數據型」為第一層級資料、「文字型」與「附檔案型」為第二層級資料、「郵件聯絡人」則為第三層級資料；待資料歸類完成後，即可以本研究所建構之「郵件數據型、數字型資料解析模組」分析郵件數據型及數字型資料，先行過濾垃圾郵件，以判定適合分類之郵件群；之後，利用「郵件文字型資料解析」模組，計算目標郵件與各類別之隸屬關係；最後，當目標郵件之類別判定結果不甚理想，即可利用「郵件聯絡人資料解析模組」重新計算目標郵件與各類別之隸屬關係，進而形成一套漸進式郵件類別判定模式，期望以具效率之方式進行郵件類別判定任務。因此本研究之主要流程可分為三大部份，分別為「郵件數據型、數字型資料解析模組」、「郵件文字型資料解析」及「郵件聯絡人資料解析模組」。

四、系統架構規劃

針對前一章節所發展之方法論與模式，本研究乃開發一漸進式郵件分類系統，以確認方法論與模式之可行性。此系統之功能重點係使用者可透過上傳郵件並擷取郵件資料後，系統即進行漸進式郵件類別分析與判定，以達具效率郵件自動分類之目的；另外，系統管理者除可使用一般使用者之功能外，亦可進行系統參數調整，使郵件分類效果更加完善。本章即針對本研究所提之「漸進式郵件分類系統」，分別以系統核心架構、系統功能架構、資料模式定義及開發工具等主題進行深入說明。

4.1 郵件類別自動判定模式之流程架構

本研究論文所提出之「漸進式郵件自動判定系統」，依其進行流程可分為「郵件資料上傳」及「漸進式郵件類別判定」等二大模組層次，此系統之運作流程架構如圖4.1所示，各功能層次之詳細流程說明如下。

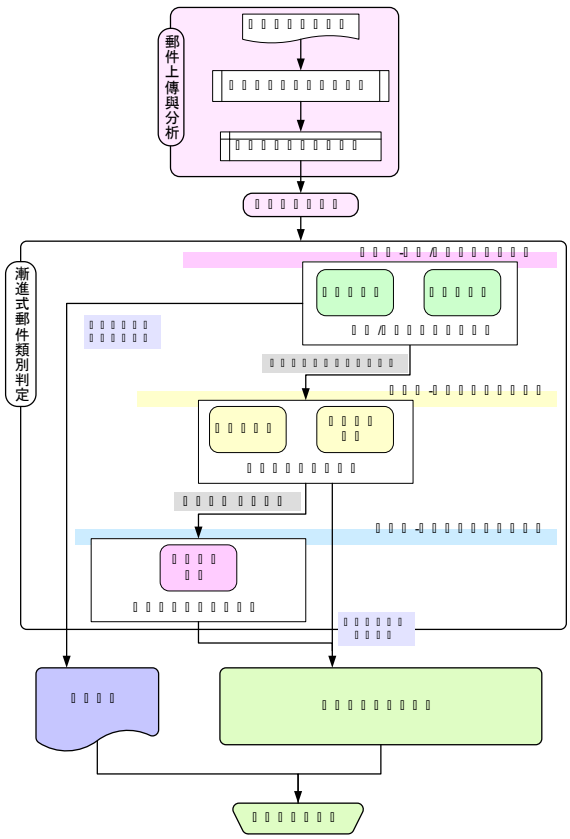

圖4.1、郵件類別自動判定系統核心架構

➢ 郵件上傳與解析

系統使用者可將未分類郵件上傳至系統，系統即擷取此目標郵件之分析資訊（包含文字型郵件資料，如內容、郵件主旨等，以及數據郵件型資料郵件等），並由後續「漸進式郵件類別判定模組」判定該目標郵件所屬類別。

➢ 漸進式郵件分類判定

本本研究所發展之漸進式郵件分類系統其中最重要的部分為此，此部分可提供

郵件進行分類的功能，並與以往郵件分類系統不同的地方為本系統分為漸進式的多層次分類，其中包含「數字/數據型資料分析層」、「文字類型資料分析層」及「聯絡人類型資料分析層」等三層級，透過此三層以進行郵件分析。

4.2 系統功能架構

本研究所建置之漸進式郵件分類系統乃架構於網際網路環境下。使用者可藉由瀏覽器（如IE）透過網際網路登入本系統，方能使用本系統所提供之各項功能。當使用者登入系統後，系統即根據使用者帳號判斷該使用者於系統中之功能權限。

在本系統平台之權限管理架構下乃將系統使用者區分為一般使用者與系統管理者，以下即分別針對此兩種不同身份使用者所能使用之功能加以說明：

一般使用者

1. 可上傳未分類郵件，以執行郵件分類功能
2. 可瀏覽符合使用者權限之所有郵件
3. 可查詢使用者上傳之未分類郵件所隸屬類別

系統管理者

1. 可瀏覽/編輯系統資料庫內之所有郵件
2. 可上傳訓練具分類代表性關鍵字庫之已知類別郵件
3. 可上傳未分類郵件，以執行郵件分類功能
4. 可查詢、新增、修改或刪除關鍵字庫之相關資訊
5. 可查詢、新增、修改或刪除非關鍵字庫之相關內容
6. 可查詢、新增、修改或刪除具分類代表性關鍵字庫之相關資訊內容
7. 可維護郵件類別（包含分類類別之查詢、新增、修改或刪除）
8. 可修改/查詢系統參數與門檻值
9. 可將郵件判定結果維護至系統中

本系統所開發之重點模組共有「郵件資料維護模組」、「漸進式郵件類別判定模組」、「關鍵字/非關鍵字維護模組」、「郵件類別維護模組」、「訓練郵件模組」及「系統參數設定」等六大模組；圖4.2即表示此郵件類別自動判定系統之核心模組架構。

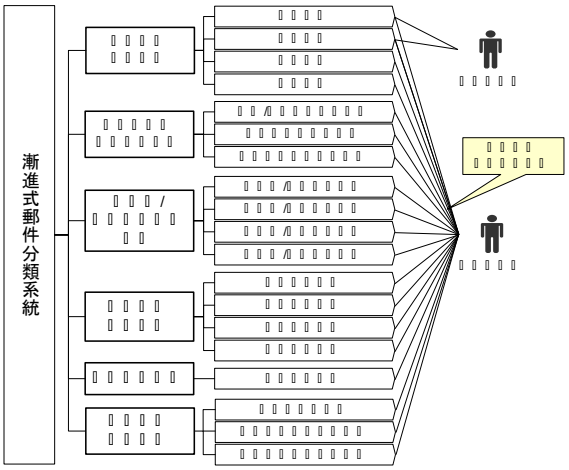

圖 4.2、漸進式郵件分類系統之模組與功能架構

針對上述系統架構所包含之基本功能模組說明如下：

(一) **郵件資料維護模組**

- 郵件新增：一般使用者將郵件上傳後，系統將擷取完畢之資料，傳送至資料庫，以待後續模組使用。
- 郵件查詢：此部分提供一般使用者郵件基本資料查詢功能。
- 郵件修改：此部分提供系統管理者郵件基本資料刪除功能。
- 郵件刪除：此部分提供系統管理者郵件基本資料修改功能。

(二) 漸進式郵件分類判定模組

- 數字/數據型資料解析模組：此部份提供系統管理者垃圾郵件判定之功能，使用者必須先輸入查詢字串、選取查詢條件、選擇上傳時間範圍及點選邏輯運算子，待查詢完畢後，即可選擇需要判別電子郵件，即可完成垃圾郵件判定。
- 文字型資料解析模組：此部份提供系統管理者郵件類別判定之功能，使用者必須先輸入查詢字串、選取查詢條件、選擇上傳時間範圍及點選邏輯運算子，待查詢完畢後，即可選擇需要判別文字型資料解析之電子郵件，即可完成郵件類別判定。
- 聯絡人資料解析模組：此部份提供系統管理者郵件類別判定之功能，使用者必須先輸入查詢字串、選取查詢條件、選擇上傳時間範圍及點選邏輯運算子，待查詢完畢後，即可選擇具高度關聯之聯絡人資料，以完成郵件類別判定結果之修正。

(三) 關鍵字/非關鍵字維護模組

關鍵字維護功能如下：

- 關鍵字新增：提供系統管理者將關鍵字資料匯入並維護於系統資料表中。
- 關鍵字查詢：提供系統管理者查詢已維護之關鍵字。
- 關鍵字修改：提供系統管理者修改錯誤之關鍵字。
- 關鍵字刪除：提供系統管理者刪除錯誤之關鍵字。

非關鍵字維護功能如下：

- 非關鍵字新增：提供系統管理者將非關鍵字資料匯入並維護於系統資料表中。
- 非關鍵字查詢：提供系統管理者查詢已上傳之非關鍵字。
- 非關鍵字修改：提供系統管理者修改錯誤之非關鍵字。
- 非關鍵字刪除：提供系統管理者刪除錯誤之非關鍵字。

(四) 郵件類別維護模組

- 郵件類別新增：提供系統管理者將郵件類別匯入系統資料庫內。
- 郵件類別查詢：提供系統管理者查詢所有郵件類別。
- 郵件類別修改：提供系統管理者進行修改錯誤之郵件類別。
- 郵件類別刪除：提供系統管理者進行刪除不必要之郵件類別。

(五) 訓練郵件模組

- 「訓練郵件上傳」功能：提供系統管理者透過上傳功能，將已知分類之訓練郵件樣本上傳至系統中，並結合系統判定結果與本身之領域知識，進而修訂「關鍵字/類別隸屬係數」與「郵件/類別隸屬係數」，而系統管理者透過此功能之修訂以提高各解析模組於運算時之準確率。

(六) 系統參數設定模組

- 系統門檻值設定：提供系統管理者進行修改與維護錯誤之系統門檻值，進而保持系統門檻值之正確性。
- 數字/數據型資料權重值設定：提供系統管理者進行借定篩選區間之參數並修改錯誤之參數值，進而保持判定之正確性。
- 文字型資料權重值設定：提供系統管理者進行修改錯誤之權重值，進而保持系統權重值之正確性。
- 聯絡人資料權重值設定：提供系統管理者進行修改錯誤之權重值與預選等級級數，進而保持系統權重值與預選等級級數之正確性。

本系統之運作架構如圖4.3所示，由於此系統乃架構於網際網路環境下，故可允許多個使用者經由網際網路登入方式進入系統，使用者分為一般使用者及系統管理者，依權限不同分別可執行：一般使用者可執行郵件新增、查詢功能及郵件類別查詢功能（如圖4.3之一般使用者）；系統管理者則可執行訓練郵件上傳、系統參數設定、郵件類別判定、關鍵字/非關鍵字維護、郵件類別維護（如圖4.3 之系統管理者）；一般使用者上傳未分類郵件後，系統管理者即執行漸進式郵件類別判定模組乃判定郵件所屬類別，待判定完畢後，乃將判定結果輸出予系統管理者以及儲存至資料庫中，而後一般使用者即可透過郵件類別查詢，以得知郵件類別判定結果為何。

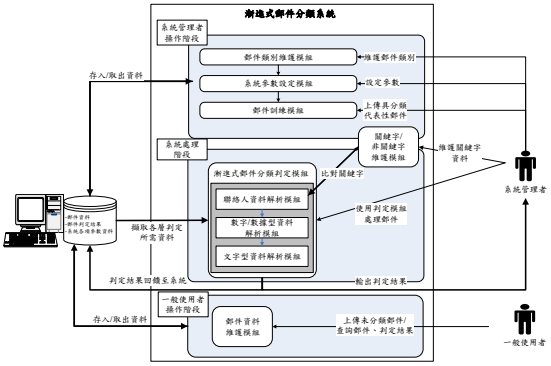

圖4.3、漸進式郵件分類系統運作架構

4.3 資料模式定義

本研究所發展之漸進式郵件分類模式乃以網際網路環境為基礎，並配合資料庫技術以開發系統之各項功能，期能使郵件管理、郵件關鍵資訊擷取、分類依據關鍵字分析及目標郵件類別判定等任務可即時完成。依據系統運作之需要，將漸進式郵件分類模式之資料分為「郵件內容與擷取資料」、「專業領域資料」、「類別判定資料」與「系統參數資料」等四大部分，以下即就各部分所包含之資料表說明其資料定義。

(一) 郵件內容與擷取資料

此資料之目的乃記錄郵件相關資料內容與解析時所需關鍵資訊，以有效進行郵件管理與判定目標郵件之類別；其所屬之資料表及其相關定義說明如下：

- 郵件基本資料表：紀錄郵件之基本資料，如郵件編號、郵件檔案名稱、郵件上傳時間、郵件語言與郵件簡述等資訊。
- 數字型資料表：紀錄目標郵件所擷取數字資料，以作為數字/數據型資料解析模組應用之基礎。
- 數據型資料表：紀錄目標郵件所擷取之數據資料，以作為數字/數據型資料解析模組應用之基礎。

- 文字型資料表：紀錄目標郵件所擷取之所有文字型資料，以作為文字型資料解析模組應用之基礎。
- 附件檔資料表：紀錄目標郵件所擷取之附件檔案名稱及副檔名，以作為文字型資料資料解析模組應用之基礎。
- 聯絡人資料表：紀錄目標郵件所擷取之聯絡人資料，以作為聯絡人型資料解析模組應用之基礎。
- 聯絡人類型資料表：紀錄所有聯絡人之類型，提供解析模組讀取聯絡人資料時之參考依據。

(二) 專業領域資料

此資料之目的乃記錄郵件內容之關鍵資訊，如關鍵字/非關鍵字、關鍵字所屬類別之隸屬關係等資料，以有效進行郵件之類別判定；其包含之資料表及與相關定義如下：

- 關鍵字基本資料表：定義不同字數之參照用關鍵字集合，作為關鍵字比對之基礎。
- 關鍵字/類別隸屬係數資料表：定義關鍵字與郵件類別之相關性，作為解析模組判定參考數據。
- 非關鍵字基本資料表：定義不同字數之參照用非關鍵字集合，以作為非關鍵字比對之基礎。
- 標點符號資料表：定義郵件中可能使用之各種符號。

(三) 類別判定資料

此部分乃記錄郵件類別之所有相關資料，包含目標郵件經系統判定後所隸屬之類別與系統內各郵件類別之相關資料。其包含之資料表與相關定義說明如下：

- 郵件類別資料表：乃維護各項郵件分類所屬分類之類別資料，如：類別名稱、類別修改時間等資料。
- 郵件/類別隸屬係數資料表：乃維護所有郵件解析模組中所判定結果之數據，並供全縣內使用者查詢郵件判定狀況與結果。

(四) 黑名單郵件資料

此部分乃記錄黑名單垃圾郵件類別之相關資料，包含目標黑名單垃圾郵件IP、黑名單垃圾郵件Message-ID與黑名單垃圾郵件Received，以提供數字/數據型資料解析模組判斷，完成垃圾郵件判定。其包含之資料表與相關定義說明如下：

- 郵件IP黑名單資料表：乃維護垃圾郵件IP與新增日期，以供數字/數據型資料解析模組判斷垃圾郵件與往後資料新增。
- 郵件Message-ID黑名單資料表：乃維護垃圾郵件Message-ID與新增日期，透過資料表維護以維持數字/數據型資料解析模組判斷垃圾郵件準確性。
- 郵件Received黑名單資料表：乃維護垃圾郵件Received與新增日期，數字/數據型資料解析模組透過資料表資料比對是否為垃圾郵件。

(五) 系統參數資料

此資料之目的乃紀錄系統參數之資料，如系統門檻值、文字型資料權重值等資料，透過設定以有效提高郵件類別判定之準確率；其所屬資料表及其相關定義如下：

- 系統參數名稱資料表：記錄系統參數名稱與參數敘述等資料。

- 系統參數值：紀錄系統參數之數值，以影響判定模組之數據結果準確率。

上述各資料乃為系統中各功能模組所需使用或產生之各項資訊，並依其所規劃之資料表形式記錄於資料庫中，用以支援系統各功能模組執行其任務。此外，透過各項資料表間之關聯性（Entity Relationship Model；ER Model）設計，使本研究所發展之漸進式郵件分類系統可方便地進行郵件與資料控管，並有效提升系統之彈性、效率性與正確性。各資料表間之關聯性如圖4.4所示。

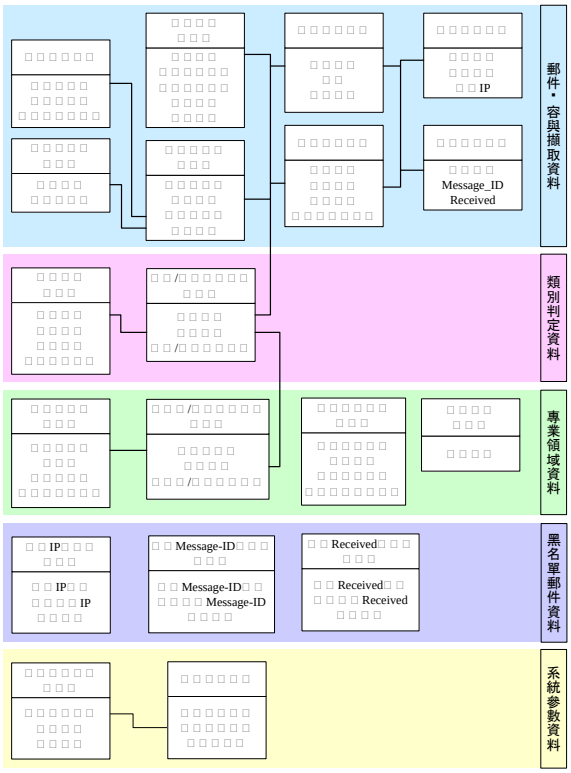

圖4.4、系統之資料模式關聯

4.4 系統流程

本節乃針對「系統功能流程」與「系統資料流程」兩部分進行說明；其中，系統功能流程將介紹使用者於各功能模組之功能流程規劃，而系統資料流程則介紹系統內各項資料傳遞之流程關係。

4.4.1 系統功能流程

如 4.2 節所述，本系統實際運作乃依不同功能進行區分，包括「郵件資料維護模組」、「漸進式郵件類別判定模組」、「關鍵字/非關鍵字維護模組」、「郵件類別維護模組」、「訓練郵件維護模組」及「系統參數設定模組」等六大模組，以下即說明各系統功能之流程規劃。

郵件資料維護模組

此模組可供系統人員上傳欲分析之郵件，並作解析與擷取，作為郵件類別判定模組之分析資料。此外，系統人員亦可根據系統中所維護之郵件內容，執行郵件之查詢、新增、修改與刪除功能。

漸進式郵件類別判定模組

系統管理者可於「數字/數據型資料解析模組」、「文字型資料解析模組」、與「郵件聯絡人資料解析模組」等三大類別判定模組功能中，擇其優先順序來進行解析，除可加速分類之效率外，並可獲得較為正確之分類效果。

關鍵字/非關鍵字維護模組

此模組提供系統管理者可執行新增、查詢、修改與刪除等功能維護各類關鍵字集（其中包括關鍵字集以及非關鍵字集）。

郵件類別維護模組

此模組乃提供系統管理者新增、查詢、刪除及修改各郵件類別資料。郵件

類別維護模組包含「郵件類別新增」、「郵件類別查詢」、「郵件類別刪除」與「郵件類別修改」等四大功能；其中，「郵件類別新增」功能乃提供系統管理者將郵件類別匯入系統資料庫內。「郵件類別查詢」功能乃提供系統管理者查詢所有郵件類別，以方便使用者瞭解系統內各項郵件類別之維護結果。此外，「郵件類別修改」與「郵件類別刪除」功能乃提供系統管理者進行修改與維護錯誤之郵件類別，進而保持郵件類別之正確性。

訓練郵件維護模組

系統管理者可透過「訓練郵件上傳」功能將訓練郵件資料上傳匯入系統資料庫中，並建立關鍵字與類別之隸屬係數，以及維護領域郵件與領域類別之隸屬關係。

系統參數設定模組

為使系統管理者方便維護各系統相關資料，此模組乃提供系統管理者於線上修改各系統參數資料。郵件資料維護模組包含「系統門檻值設定」、「數字/數據型資料權重值設定」、「文字型資料權重值設定」及「聯絡人資料權重值設定」等四大功能；其中，「系統門檻值設定」功能乃提供系統管理者進行修改與維護錯誤之系統參數，進而保持系統門檻值之正確性；「數字/數據型資料權重值設定」、「文字型資料權重值設定」及「聯絡人資料權重值設定」功能乃提供系統管理者進行修改錯誤之權重值與預選等級級數，進而保持系統權重值與預選等級級數之正確性。

4.4.2 系統資料流程

本系統運作之初，系統管理者首先需將已知類別郵件匯入系統，系統管理者透過本身之領域知識選取關鍵字並修定關鍵字/類別隸屬係數；此時，系統即根據系統管理者所指定之關鍵字/類別隸屬係數與系統參數進行分類依據關鍵字推論，以取得郵件/類別隸屬係數，並存入系統資料庫內。上述步驟完成後，使用者即可將未分類郵件上傳至系統，系統乃先擷取類別判定所需之郵件關鍵資料於資料庫中，之後漸進式郵件分類模組以此郵件關鍵資料為基礎，並根據使用者所指定之系統門檻值與參數進行郵件類別判定推論，以取得目標郵件之隸屬類別。系統相關資料之存取與傳遞情形如圖4.14所示。

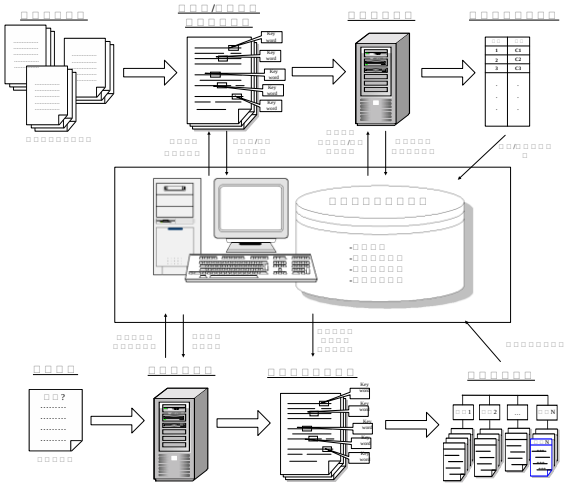

圖 4.11、系統資料流程

五、系統實作與案例分析

根據第四章所提出之雛形架構與規劃，本研究乃發展一套以網際網路為基礎之電子郵件類別自動判定系統，並針對系統中各身份別使用者可執行之功能模組詳細介紹，各功能模組之操作說明乃彙整於附錄。本章以訂閱Pchome個人電子報之文章報導，利用網路信箱工具「outlook」收信，即可獲得「*.eml」電子郵件格式之電子郵件樣本，並以其為案例，分析本研究所提出之方法論與雛形系統的可行性；於第

5.1 節中檢驗與評估本電子郵件管理系統之類別推論績效，且將本研究之類別推論績效與其他分類方法之績效進行比較，並於第5.2 節中以Pchome個人電子報為案例，闡述系統之應用情境與電子郵件之類別推論實效。

5.1 系統案例之應用流程

為驗證漸進式電子郵件類別判定系統於實務應用之可行性，本研究乃以一封電子郵件為案例驗證樣本，並以漸進式電子郵件類別判定系統之兩大核心功能模組（包含「郵件資料維護」及「漸進式郵件類別判定」等推論模組），進行以電子郵件之「郵件新增與郵件內文擷取」及「漸進式郵件類別判定處理與判定結果建議」等決策推論，以評估本研究所發展之方法論與開發之系統是否具備可行性。首先，系統管理者必須將本系統各項參數設定完畢。接著，一般使用者乃上傳一封未分類之電子郵件至系統中。而後，系統管理者乃執行主功能以判斷郵件之類別，待分析完畢，系統即將此次郵件類別之判斷結果輸出予管理者，但如判斷之門檻值過低，系統則會建議管理者繼續往下分析，直到分析至最後一層，以取得該份郵件之正確郵件類別，待判定完畢後，管理者再將此次判定結果回饋至系統中。最後使用者即可查詢其所上傳之郵件與郵件類別判定結果。其完整運作架構如圖5.1所示，以下即進行系統應用情境之詳細說明。

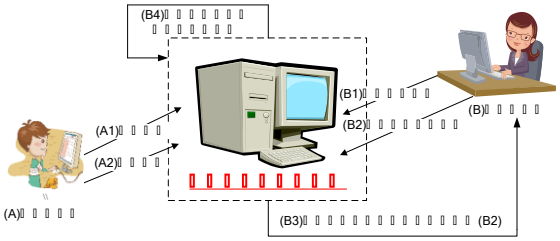

圖 5.1、漸進式電子郵件類別判定系統之應用流程

5.1.1 電子郵件類別推論之驗證與評估

為驗證本系統之電子郵件類別自動推論績效，本研究乃以Pchome個人電子報中所提供之各類電子報透過網路訂閱，並利用郵件管理功能將所訂閱之電子報（如圖5.26）以「*.eml」格式收信並彙整（如圖5.27），並以其為探討案例。而驗證過程可分為訓練與測試兩大階段，以下即針對驗證資料取得、系統驗證方式說明、評估指標定義與驗證結果分析而依序進行說明。

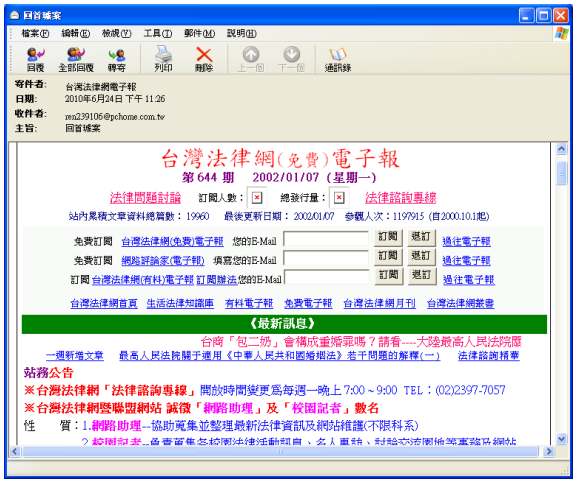

圖 5.26、以郵件系統顯示電子報內容

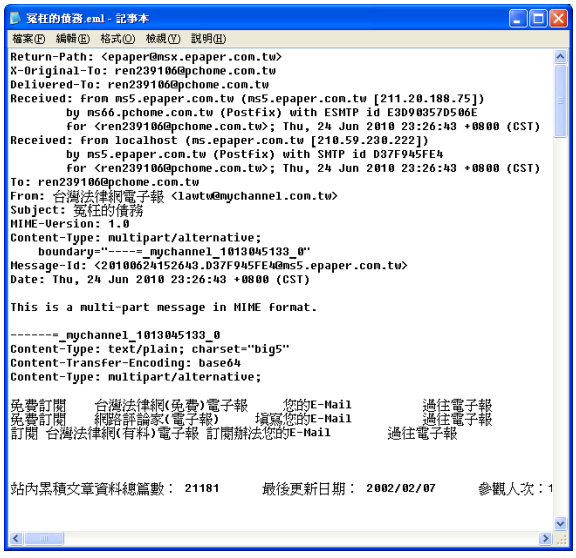

圖5.27、以記事本顯示「*.eml」之內容

5.1.1.1 驗證資料取得

於系統訓練階段中，本研究乃以 Pchome 個人電子報中所提供之各類電子報為基礎，蒐集與影視娛樂、資訊類、工商金融經濟、生活休閒人文、知識教育、政治法律及醫療保健等 7 個類別之電子郵件，總計 700 篇電子郵件。

5.1.1.2 系統驗證方式說明

為驗證本論文所提出之方法論與系統績效，首先於系統驗證第一階段（即系統訓練階段）乃自700份與「影視娛樂」、「資訊類」、「工商金融經濟」、「生活休閒人文」、「知識教育」、「政治法律」及「醫療保健」內容相關之電子郵件中，共挑選210份電子郵件作為訓練資料（各類別30份訓練資料），並逐一匯入系統中，以建構關鍵字與領域類別之關係。之後，於此所選定之七大類別中，各類隨機挑選3份文件（共計21份，整理如表5.3），以作為系統測試資料，並利用訓練階段所取得之關鍵字/類別隸屬關係，推論此21份測試電子郵件之隸屬類別，並藉由觀察系統所推論之電子郵件類別是否符合該電子郵件之實際類別，以確認本研究所提方法論之正確性。待完成上述之第一階段系統績效驗證後，於第二階段（即系統測試階段）乃分為7個週期持續匯入訓練用之電子郵件資料，每週期皆再匯入70份電子郵件（各類別10份電子郵件），總計再匯入490份電子郵件（各類別70份電子郵件）；於各週期中乃利用前述21份測試電子郵件重新進行電子郵件類別推論，以分析系統於不同訓練電子郵件數量下之長期學期趨勢。

5.2.1 第一階段驗證結果分析

於第一階段系統驗證中，本研究先由700 封與「影視娛樂」、「資訊類」、「工商金融經濟」、「生活休閒人文」、「知識教育」、「政治法律」及「醫療保健」內容相關之電子郵件中，共挑選210 封電子郵件作為訓練資料（各類別30 封訓練資料），並逐一匯入系統中，以作為第一階段驗證之基礎訓練資料。以下即針對各項指標說明系統驗證過程，並分析系統驗證之結果。

5.3 驗證結果整體分析

綜合兩階段之驗證成效後，可將各項驗證指標之相關結果整理如表 5.8。由表 5.8 可知，各項驗證指標之收斂前，文字型召回率及正確率初始只有91%，但至第四週期便以達到 100%，呈現收斂狀態，而文字型之類別隸屬係數雖然初始只有 41%，但至第四週期呈現收斂狀態，且至第七週期仍然有小幅度成長並以成長至 80%，以整體平均值而言，文字型召回率與正確率為 97%，而文字型類別隸屬係數雖然只有60%，但最後達到 80%以上之水準。

整體而言，雖然文字型判定功能判斷郵件類別正確性已達到 80%以上，但是本模式又另外建構一個聯絡人判定功能來輔助文字型，其令召回率及準確率初始便提高至 95%（原本 91%），且第四週期便已呈現收斂狀態，令類別隸屬係數初始便提高至 84%（原本41%），並且第四週期便已呈現收斂狀態（原本第五週期才呈現），以整體平均值而言，聯絡人召回率與準確率提高至 98%（原為 97%），而聯絡人類別隸屬係數則提高至 89%（原為 60%），並且最後達到 97%之水準（原本只有80%）。

雖然單獨使用文字型判定功能便能有 80%以上的正確性，但如果再用聯絡人判定功能輔助文字型，便能將正確性提高至 96%以上的高水準，不過本研究提供了可以單獨使用文字型判定功能，或者兩者互相搭配使用之選擇，並能依使用者需求來選擇使用，且如果單獨使用文字型判定功能如果判定結果低於門檻值，系統依然會建議使用者繼續往下使用聯絡人判定功能來取得更高的判定正確性，故使用者不用擔心正確性不夠高之問題。

且各項驗證指標皆於六個週期內呈現收斂狀態，因此，以本研究所選驗證個案（即 Pchome 個人電子報）為例，當本系統使用約 280 封至 350 封訓練用電子郵件時，可讓系統之各項推論績效提升至 80%以上之水準，故本研究所建置之漸進式類別判定模式從一開始數字/數據型判定功能便可以有效刪除垃圾信，並將有效信件往下繼續透過文字型判定功能來判斷類別，且如果判斷結果不理想，更可以繼續往下透過聯絡人判定功能來輔助文字型之判定結果，令判斷之正確性更進一步從 80%提高至 96%以上，因此得知，本系統可有效應用於電子郵件分類判定，並準確地依據判定結果給予使用者分類建議。

驗證指標	整體平均值	收斂週期	收斂前每週期平均成長率	整體每週期平均成長率
漸進式郵件類別判定－文字型召回率	97.28%	第四週期	2.38%	1.59%
漸進式郵件類別判定－文字型正確率	97.28%	第四週期	2.38%	1.59%
漸進式郵件類別判定－文字型類別隸屬係數	59.69%	第五週期	8.00%	6.51%
漸進式郵件類別判定－聯絡人型召回率	97.96%	第四週期	1.59%	0.79%
漸進式郵件類別判定－聯絡人型正確率	97.96%	第四週期	1.59%	0.79%
漸進式郵件類別判定－聯絡人型類別隸屬係數	89.25%	第四週期	1.75%	2.16%

表 5.8、各項驗證指標成長率之彙整表