南　　華　　大　　學

資訊管理學系

碩士論文

個人化郵件類別推論模式

A Personalized E-mail Classification Model

研 究 生：吳朝宏

指導教授：楊士霆

中華民國 102 年 7 月 14 日

誌謝

隨著研究所兩年結束，人生中求學階段也暫時畫下句點。在這兩年內無數人的教導、支持、協助、陪伴與叮嚀，這些力量讓我順利完成碩士論文，謝謝各位的力量。

首先，謝謝教導我最多且時常叮嚀我的指導教授楊士霆博士，感謝老師再三的叮嚀與指導，讓我面對事務處裡更細心並具有規劃性的處理且更謹慎面對人與人的溝通，同時亦培養我面對問題思考的能力及解決問題的思維。此外感謝老師兩年來對學生犯錯時的指正並鼓勵學生成長，感謝老師指導讓我成長，謝謝老師給我肯定，讓我的求學階段中有個特別又充實的碩士生經歷，因此向楊士霆博士致上由衷的感謝與敬意。

於論文最後完成之際，非常感謝大葉大學的葉子明教授與本系主任洪銘建主任，謝謝老師們花費心思與時間審視學生論文，並提供寶貴建議與修正讓學生論文更加完整。此外，特別感謝系上老師們這兩年來的照顧及對學生的指導，特別感謝系助理伊汝姐的照顧與關心及學位口試申請的協助。

接著感謝已畢業的山田學長、西瓜學姐，在自己的論文繁忙之餘中關心學弟並鼓勵學弟成長；此外，感謝 C204 和 C208 實驗室中的學長姐、同學、學弟妹，大家一起吃飯、一起完成課堂作業，讓研究生在研究室生活並不孤單；還有感謝我的好友小威、貽任、阿 Mo、默默在我低落時願意傾聽我發洩的垃圾言語，讓我找回信心和情緒。感謝 H408 的學弟妹帶給我很多歡笑，忍受學長在 H408 失控的情緒，且也感謝系辦義工大學部學弟妹資沂、家熒、雅云和其他學弟妹在系上事務的協助，也感謝學校秘書室的照顧及工讀機會，也為我的碩士生涯有著更多經驗。接著，感謝最重要的 EDS Lab 成員：怕老鼠的丸子學姐、怕蜘蛛的小偉、小胖子阿昌、俞佑和晏晏，謝謝丸子學姐一起陪我談心、熬夜，幫助我面對論文難題還分給我便當一起吃，也要感謝小偉、阿昌、俞佑和晏晏，我們一起熬夜、一起被罵、一起低落、一起思考、一起笑、一起助教時間，最難忘的是參加競賽的準備前夕，大家一起扶持修改報告內容，因此參賽所獲得的殊榮，是屬於 EDS Lab 成員大家的，EDS Lab 就像是一個大家庭，讓我們在研究室一點都不孤單，謝謝 EDS Lab，謝謝丸子學姐、小偉、阿昌、俞佑和晏晏你們的陪伴與扶持，謝謝！

最後感謝在屏東鄉下等待我的家人，謝謝你們諒解我失聯時的不解釋，謝謝你們在我需要休息時的照顧，謝謝我的家人的包容，讓我專心完成碩士論文與碩士學位。再次感謝所有人，讓我在此與你們分享我畢業的喜悅，謝謝你們！

摘要

電子郵件為現今主要傳遞訊息管道之一，但電子郵件盛行導致電子郵件使用者容易大量接收各界資訊，而造成郵件過量問題，使用者亦難以檢索重要郵件資訊，故現今多數電子郵件信箱乃需使用者以人工方式區分郵件，但人工分類需使用者閱讀郵件後逐一區分，且電子郵件數量龐大導致使用者需耗費大量時間管理電子信箱；此外，現今多數研究針對郵件過量議題發展郵件分類技術，然而多數研究接針對特定領域或公共郵件發展解析技術，然而一般使用者所接收與寄發私人郵件常因使用者個人專業知識與生活經驗不同，而郵件內容所涉及領域亦不相同且複雜，進而導致多數郵件分類技術無法解析或分類錯誤。因此，本研究乃針對使用者個人之私人郵件發展解析法則，並建立一套「個人化郵件類別推論模式」，針對不同使用者推論個人專屬的郵件類別，以供使用者透過個人化郵件類別區分個人郵件，以協助使用者更有效率管理郵件。

本研究所建立之「個人化郵件類別推論模式」針對郵件內容特徵擷取與個人化類別名稱推論兩議題分別發展「語意擷取郵件關鍵字模組」與「個人化郵件類別推論模組」兩大模組。首先「語意擷取郵件關鍵字模組」乃針對一般使用者之私人郵件分析內容詞彙語言涵義，進而取得該郵件中各筆敘述主題及使用者隱含特殊涵義之語意詞彙；此外，多數私人郵件常依據使用者個人背景（專業知識與生活環境）不同，進而導致郵件內所涉及專業領域複雜，故本研究於「個人化郵件類別推論模組」中以郵件中語意詞彙解析郵件中代表性特徵，再以郵件內容特徵及「階層式分群法」（Hierarchical Clustering）為基礎推論郵件群集樹，並根據使用者個人需求擷取符合個人需求郵件群集數量，且為方便使用者透過具代表性涵義名稱區分郵件群集，本研究透過群集內代表性特徵與本研究制定類別名稱詞彙庫分析相似性類別名稱，並賦予具代表性郵件類型名稱，以作為使用者個人專屬郵件類別，即針對使用者量身建立個人化專屬郵件類別，以提高私人郵件分類準確率，進而增進使用者郵件管理之效率。本研究除發展模式與方法論外，並依此方法論建構一套「個人化郵件類別推論」系統以進行案例驗證，從而確認方法論與技術之可行性。

關鍵詞：個人化郵件、郵件分類、郵件分群、語意分析。

Abstract

For e-mail management, users need to read e-mail content to identify the e-mail properties, and then classify and manage the e-mails by themselves. The huge number of e-mails easy leads time-consuming of manual classification; therefore, previous researches develop related e-mail classification techniques. However, the e-mail users usually send private e-mails and e-mal contents majority implicit personal life experiences and expertise. The existing classification techniques based on specified filed easily decreases the accuracy of personal e-mail classification. That is, based on personal knowledge, life and other factors, the required e-mail categories of each e-mail user might be different. Therefore, this paper establishes a Personalized E-Mail Classification model to generate personalized e-mail categories to efficiently manage their own e-mails. In this proposed model, the semantic analysis technology is used to analyze and extract the critical features from e-mail content. Based these critical features, clustering technology is employed to distinguish e-mails groups and generate the corresponding titles as personalized e-mail categories. Therefore, this paper can automatically generate personalized e-mail categories to assist e-mail user in managing their private e-mails.

Keywords: *Personalized E-mail, Email Classification, Clustering, Semantic Analysis*

目錄

圖目錄

表目錄

第一章、研究背景

本章將針對本論文之研究動機、研究目的與研究步驟依序進行說明。首先乃透過研究動機與目的之描述，點出本論文研究之緣由。其次，於研究步驟中，說明本論文規劃進行之流程。各小節詳細內容說明如下。

1.1 研究動機與目的

電子郵件低成本與傳達快速特性已成為現代人們主要溝通管道，如：醫療病歷傳遞、教學溝通、商業管理等應用（Alberts 和 Forest，2012；Cornwall 等人，2008；Hassini,2006；Stuit 和 Wortmann，2012）。但隨著電子郵件廣告盛行，且電子郵件使用者於網路上隨意公開自身電子郵件地址，而造成使用者接收大量非重要郵件，形成電子郵件過量問題。故現行電子信箱（如：「Gmail」、「Hotmail」等）為協助使用者於過量郵件中搜尋與篩選出所需郵件，因此提供使用者於瀏覽郵件主旨或閱讀郵件內容後以人工方式建立郵件類別（亦即郵件資料夾），以協助使用者透過自訂郵件類別檢索郵件。然而使用者所接收郵件數量過於龐大，無法詳細瀏覽每封郵件內容後再根據郵件類型建立類別，且郵件寄件者常以個人慣用詞彙或個人知識領域名詞撰寫郵件內容，容易導致使用者誤解郵件內容而產生分類錯誤問題。

由於現今使用者乃透過人工方式瀏覽所有郵件後制定所需類別，因此，過去研究提出相關自動化郵件分類技術，如：主成份分析、倒傳遞類神經網路、天眞貝氏分類、支援向量機等技術發展郵件分類與解析法則（Gomez 和 Moens，2012；Li 和 Huang，2012；Salcedo-Campos 等人，2012；Yang 等人，2011），以協助使用者分類郵件。但 Chang 及 Poon（2009）發現一般使用者之私人郵件無法準確透過自動化郵件分類技術進行分類，此外 Yu 和 Xu（2008）亦指出一般使用者所接收郵件類型與類別過於複雜，容易導致系統無法準確分類郵件，此外該研究亦說明過去研究所發展郵件分類技術皆以特定領域或公共郵件（如：公司通知訊息）為分析對象，因此所區分類別較為單純且一致性。但電子郵件為使用者和寄件者間溝通工具，因此電子郵件所述內容乃為使用者與寄件者間之對話（Zajic，2008），其郵件中所敘述內容皆隱含著使用者專業領域知識或個人生活經歷，故郵件使用領域較為複雜，且容易隨著時間、身分等使用情境改變而不同，而產生不同郵件類別（Alberts 和 Forest，2012），因此，面臨接收郵件類性較為複雜使用者時，現有郵件分類技術容易產生失效問題。

根據上述內容，現有電子信箱多數無法提供使用者以自動化方式建立郵件類別，因此使用者須閱讀完所有郵件後以人工方式逐一分類，且過去研究所提之分類技術能有效分類特定領域或公共郵件，但無法針對使用領域較為複雜之私人郵件進行準確分類，進而降低電子郵件個人化管理效益。綜合上述，其既有之運作模式如圖 1.1 之 As-Is Model 所示。

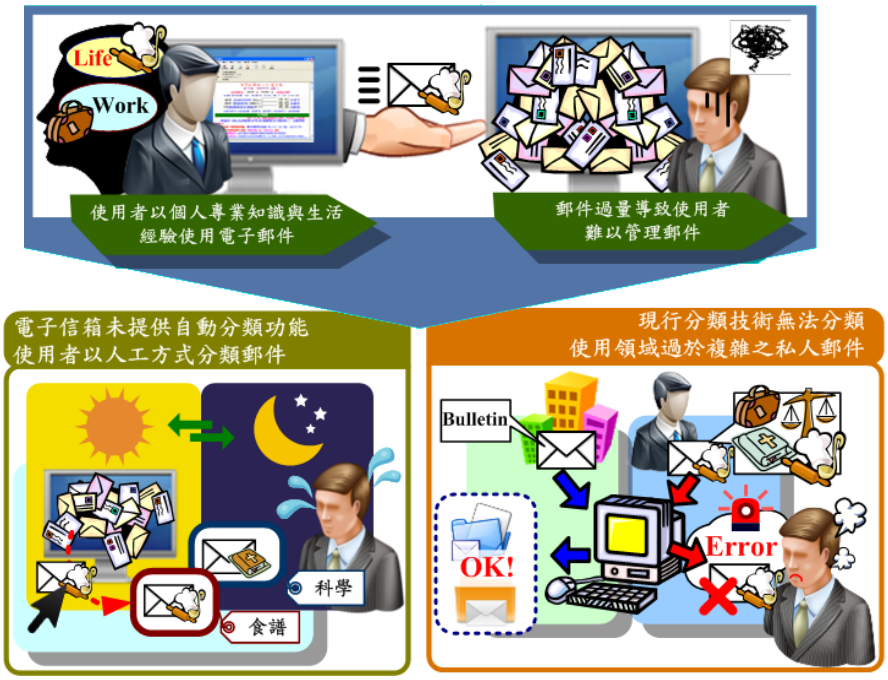

圖 1.1、電子郵件分類之既有模式 As-Is Model

如圖 1.1 所示，目前現有電子信箱與郵件分類技術無法滿足使用者管理郵件上需求，且當面臨郵件內容涉及使用者專業領域知識或個人生活經歷等相關資訊時，則無法有效分類郵件，故本研究乃將問題彙整並列點如下：

1. 電子信箱未提供自動分類功能，使用者需以人工方式分類郵件：現有電子信箱只提供使用者建立類別功能，但並未提供使用者自動分類技術或未提供個人郵件類別分類之建議，因此，使用者需透過人工瀏覽方式區分郵件類別，若使用者面臨郵件過量問題時，則龐大郵件數量往往浪費使用者過多整理時間。
2. 多數郵件技術無法隨著使用者使用情境改變，自動建立個人化郵件類別：私人郵件涉及使用者專業領域知識或個人生活經歷，因此，不同使用者根據個人專業領域不同，其所需郵件類別皆不相同，但現有郵件分類技術大多針對單一領域郵件作為分析對象，故針對不同使用者的個人專業領域或生活經歷，現有技術無法量身制定使用者個人所需郵件類別，導致分類結果往往不符使用者期望。

舉例說明，假設使用者透過電子郵件與寄件者互相分享「工作」與「生活」等議題郵件時，若一般使用者通常以人工方式閱讀完整郵件內容或主旨後，再進行人工分類。若以現有郵件分類技術進行分類時，則因私人郵件內容涉及領域過於複雜（如：生活經歷、法律、工作、信仰...等等）而降低郵件分類技術之準確率，導致郵件分類系統失效（Chang 及 Poon，2009；Yu 和 Xu，2008）。

有鑑於此，為掌握電子郵件使用者使用郵件習慣並針對不同使用情境自動變換適當郵件類別，本研究認為可在使用者接收郵件類型上分析使用者個人使用情境，加強郵件分類效果，並協助電子郵件使用者於不同使用情境下自動推論適用郵件類別，以改善郵件過載之困擾。因此，期望之運作模式如圖 1.2 所示。

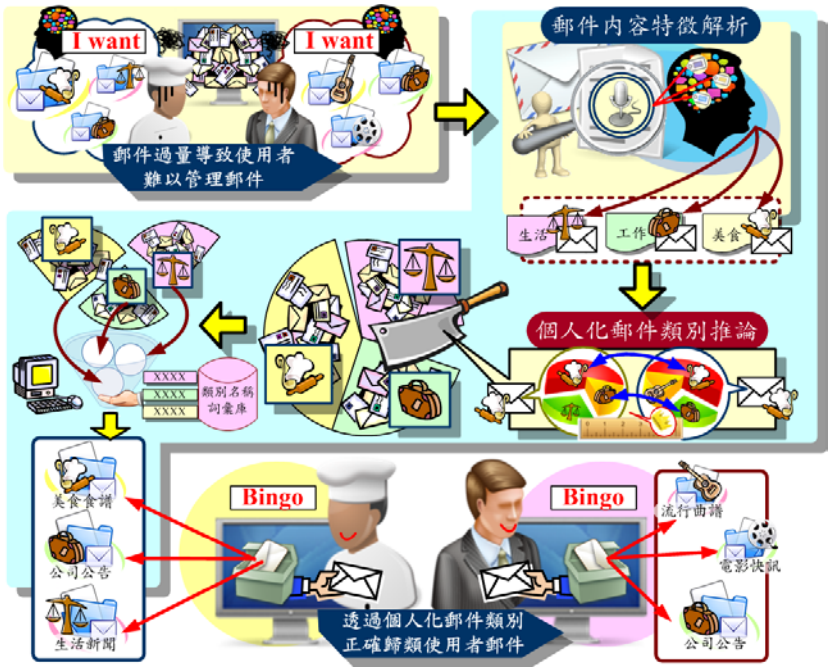

圖 1.2、電子郵件分類之期望模式 To-Be Model

本論文之研究動機與目的可歸納為以下兩點：

1. 建立郵件內容特徵解析法則，以擷取郵件中使用者個人知識之代表性特徵，進而強郵件類型區分準確性：由於電子郵件內容為使用者撰寫，且接收郵件敘述內容多與使用者個人專業知識或生活形態相關。因此，為能擷取使用者個人專業知識作為郵件區分之依據，本研究乃將使用者接收郵件進行解析並取得郵件內容之特徵字詞，並與使用者寄發郵件之特徵字詞進行語意關聯，以推論使用者接收郵件類型。
2. 發展個人化郵件類別推論法則，並自動推論使用者個人所屬郵件類別：郵件分群方式能根據郵件類型區分不同使用者所接收郵件種類，故為本研究乃以郵件中所擷取

郵件內容特徵進行郵件分群並自動推論郵件類型名稱，以加強郵件群集之解釋性並作為使用者歸類所需郵件類別，以協助使用者整理過量郵件。

整體而言，為協助電子郵件使用者快速且有效整理並管理郵件，本研究乃解析電子郵件之郵件主題、文章內容以及相關性語彙，並以此為基歸納使用者特徵資訊，進而與使用者寄發郵件之特徵字詞進行語意關聯，以推論使用者接收郵件類型。另外，本研究亦建立個人件類別推論法則，以分群法則為基礎並透過郵件內容特徵區分郵件類型並賦予郵件類型名稱，以作為個人化郵件類別，協助使用者解讀該類別所歸納之郵件類型。本研究之研究目的即根據每位使用者之專業領域知識不同自動推論適合郵件類別，以達到為使用者量身制定郵件類別效果。

1.2 研究步驟

本研究之目的乃根據電子郵件使用者不同自動推論符合使用者使用情境之郵件類別，以協助使用者有效率管理電子郵件，本研究之研究架構可分為五大步驟。藉由研究動機與目的以進行文獻回顧與探討，並於文獻回顧後確立本研究定位方向，以發展本研究之方法論，並依照方法論開發雛形系統，最後以一實例進行系統案例驗證，以確認系統之運作與成效，並評估本研究方法論之實用性，其各步驟詳細說明分述如下：

步驟一、背景資料蒐集與探討

根據本論文之研究背景、動機與目的進行相關資料蒐集，本研究乃針對電子郵件使用者管理郵件方式以及透過郵件分類方式以進行探討，並彙整所涉及主題包含「郵件資料探勘」與「郵件使用者行為探討」等議題，透過文獻蒐集與研讀，以了解電子郵件特徵解析與郵件分類之現行作法與發展方向，進而建構本研究之方法論與系統模型。

步驟二、研究方向定位

透過文獻蒐集與探討即可得知，郵件過載問題容易導致電子郵件使用者工作績效下降，故需透過郵件分類等方法整理郵件，但現行郵件分類技術多以制式化郵件類別進行分類，故透過「郵件資料探勘」之相關研究得知現有郵件分類技術多以單一特定領域郵件進行分析，此外過去文獻中發現部分研究亦提出多種針對郵件內容特徵擷取技術，以提高郵件分類準確性；「郵件使用者行為探討」之相關研究以了解個人化郵件類別推論與分析方式，並了解相關研究主要乃根據郵件詞彙語意解析等方法分類並推論個人化郵

件類別。由於既有郵件分類研究於郵件分類技術上，大多僅以單一領域郵件進行分析，無法解析內容涉及領域過於複雜之私人郵件，且無法根據使用者使用情境自動建立個人化郵件類別。因此，本研究乃針對電子郵件內容先行解析郵件使用領域特徵，並取得電子郵件之代表性特徵詞彙，進而發展一套「個人化郵件類別推論模式」，此模式將解析後之代表性特徵詞彙進行語意關聯，以作為郵件內容特徵，再透過個人化郵件類別推論並建立使用者之郵件類別，同時亦根據郵件內容特徵分析類別名稱，以取得最具代表性名詞作為類別名稱，以協助電子郵件使用者辨識郵件類型且提高郵件整理效率。

步驟三、研究模式之建立與系統開發

本研究共有四大主題需完成，分別為「個人化郵件類別推論模式規劃」、「郵件內容特徵解析法則」、「個人化郵件類別推論模式」與「系統功能開發」，以下為各項主題之細述說明：

主題一、個人化郵件類別推論模式規劃

- 蒐集並回顧郵件類別建立與郵件內容解析等議題之相關文獻
- 瞭解郵件內容解析之相關技術，以建立郵件內容特徵解析法則
- 瞭解郵件類別推論技術之差異，並建構個人郵件類別推論之運作模式

主題二、郵件內容特徵解析法則

- 解析電子郵件具代表性詞彙
- 建立電子郵件內容特徵解析法則

主題三、個人化郵件類別推論模式

- 建立郵件語意關聯與郵件類型區分法則
- 建立個人化郵件類別推論法則

主題四、系統功能開發

- 開發郵件代表性詞彙擷取功能
- 開發郵件內容特徵分析功能
- 開發郵件內容特徵語意關聯建立功能
- 開發個人郵件類別推論功能

步驟四、案例驗證

於此步驟乃將本研究所建構之「個人化郵件類別推論系統」為基，並尋求相關之實

務應用案例進行驗證，以確認本研究所發展系統模組之正確性與實用性。

步驟五、成果分析與結論

透過案例驗證之執行成效與分析，了解本研究預期成果與實際成效間之符合程度，藉，並評估本論文所發展之方法論與系統模組之實用性與準確性。最後，藉由分析評估結果規劃本研究之未來發展與應用方向。

綜合上述之研究步驟說明可知，本論文首先乃根據研究動機與目的蒐集電子郵件解析方式、郵件使用者管理習慣、郵件分類等相關文獻，以釐清研究定位並確立研究方向；其次，依據研究方向建立本研究之研究模式，並以此為基礎開發一套個人郵件動態分類系統。最後，再以案例驗證本研究所發展之方法論與系統模組，確認本研究之實用價值與發展性。本論文之研究架構如圖 1.3 所示。

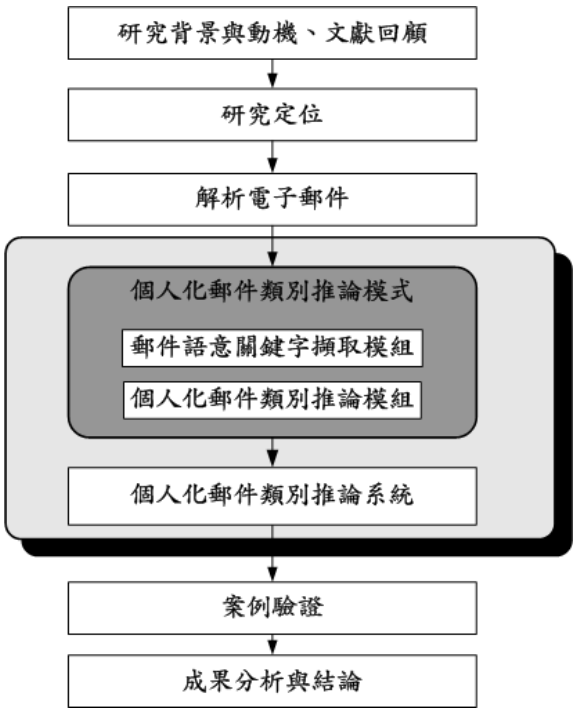

圖 1.3、研究架構

第二章、文獻回顧

本研究之目的乃協助電子郵件使用者於龐大之電子郵件中，迅速且有效地搜尋使用者欲閱讀之電子郵件，且根據使用者個人使用習慣與郵件類型偏好建立電子郵件類別，以節省使用者整理與篩選電子郵件之時間。因此，於探討相關文獻前先行釐清本研究之研究定位，以瞭解本研究與現今相關研究之差異性及本研究之研究價值。

2.1 研究定位

本研究所涉及之研究主題乃包括「郵件資料探勘」與「郵件使用者行為探討」等兩大研究方向，以下即針對此兩大主題之相關研究進行文獻回顧及探討。

於郵件資料探勘之議題中，根據過去研究可將郵件資料探勘分為「郵件分類技術」、「郵件過濾技術」以及「郵件解析方式」三方面進行探討，於郵件分類技術中，相關研究之成果可以詞彙關聯性或資料探勘方法進行分類，而郵件過濾技術課題而言，根據過去相關研究可分為「以郵件標頭檔進行過濾」、「以傳輸協定進行過濾」及「以郵件內容進行過濾」等三方面進行探討，此外，於郵件解析方式之相關研究可分為以郵件文字資料、郵件傳輸協定資料與郵件連絡人資料進行分析。

於郵件使用者行為探討之議題中，過去相關研究可歸納為「使用者郵件管理習慣分析」、「使用者慣用郵件類別分析」與「使用者郵件應用領域分析」等三方面進行探討。於使用者郵件管理習慣分析議題中，大多乃針對使用者郵件分類習慣或使用者郵件寫作風格進行探討，而使用者慣用郵件類別分析課題而言，過去相關研究之成果可分為使用者偏好類型判定以及使用者特徵資料擷取等兩議題進行探討，而使用者郵件應用領域分析之課題透過相關研究針對各產業之分析進行劃分，可分為醫療領域、教育領域及其他領域進行探討。

綜合以上所述，本論文所涉及之各項主題領域可以圖 2.1 呈現架構關係；圖中灰色部分乃代表本研究所強調之研究主題。如圖 2.1 所示，以本研究所發展之方法論為主軸，藉由過去許多相關之研究成果，本研究乃依不同研究議題搜尋相關文獻資料，並針對不同主題進行細節說明。本章文獻回顧即針對「郵件資料探勘」與「郵件使用者行為探討」兩大議題進行說明。

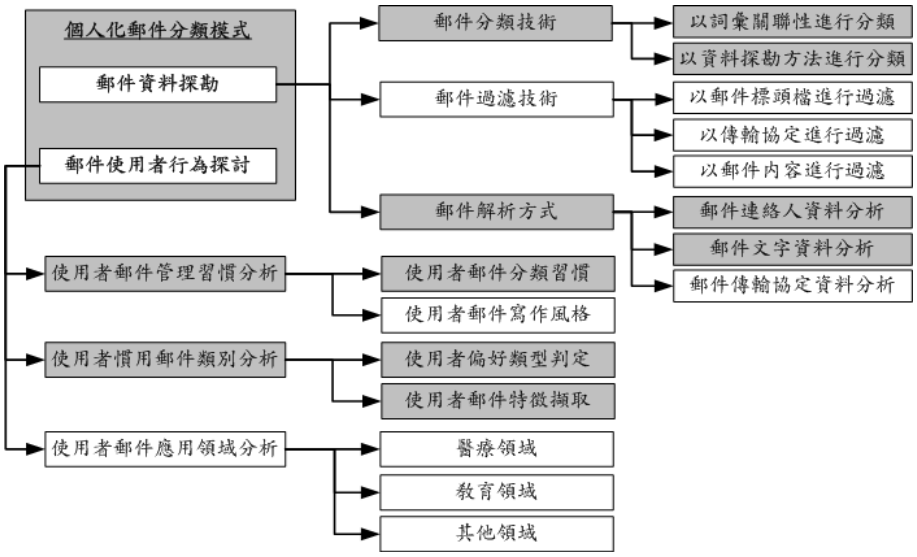

圖 2.1、研究定位圖

2.2 郵件資料探勘

對於郵件資料探勘議題而言，本研究乃針對「郵件分類技術」、「郵件過濾技術」及「郵件解析方式」進行相關文獻探討，期望於其中觀察此議題應用於不同類型之不同角度與層面解析，以更深層瞭解郵件探勘特性。

2.2.1 郵件分類技術

於郵件分類技術中，本研究乃針對「以詞彙關聯性進行分類」以及「以資料探勘方法進行分類」等兩主題進行相關文獻探討，期望從中探討郵件分類所涉及之範圍與領域。

(A) 以詞彙關聯性進行分類

針對詞彙關聯性進行分類部分，Yu 與 Zhu（2009）提出結合倒傳遞類神經網路（Back-Propagation Neural Network）和語意特徵空間（Semantic Feature Space）以進行郵件之分類。傳統上倒傳遞類神經網路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，因此該研究乃修改倒傳遞類神經網路，增加引進網路起點和使用適應性學習法以調整學習比率，即可有效改善上述之缺點。此外，於在郵件分類中，電子郵件之內容解析易產生詞意誤判之情形，故該研究乃以語意特徵空間技術，以減少相似的語意特徵，進而提高分類的精度和效率。此外，亦有研究針對郵件分類技術應用進行探討，Buffett 和 Geng（2010）發現多數使用者透過電子郵件通知工作行程，當通知行程郵件數量大增時，容易造成使用者忽略個人日誌上重要行程。因此，該研究藉由關鍵字標籤方式提醒使用者重要行程，該研究關鍵字標

籤之建立乃先擷取已分類電子郵件中所通知之重要行程，並透過行程內容獲取關鍵字，再以天眞貝氏（Naive Bayes）計算類別與關鍵字發生機率，並作為該筆關鍵字之郵件行程標籤之評估指標，且根據標籤機率判斷郵件所述行程之代表性關鍵字，以作為行程之關鍵字標籤並標示於重要行程上，提醒使用者重要行程並助於使用者整理個人日誌。

而對企業組織而言，由於客户所寄電子郵件數量龐大，因此回答每封客户所提問之電子郵件極為耗時，為幫助企業客户服務部門回答電子郵件，Scheffer（2004）先制定回答選項類別，並將客户提問之電子郵件以支援向量機（Support Vector Machine；SVM）擷取郵件內容特徵，再以天眞貝氏分析郵件特徵，以區分客户於郵件中所提問題之所屬類別，再根據郵件所屬回答類別回應客户對應回答內容，以達到客户提問自動化回覆，進而幫助企業客户服務部門回答客户於電子郵件中所提問題。此外，客户服務系統中所提供郵件回覆功能乃先行提供各類建議範本，客服人員進一步對範本設定編號，以方便檢索所需範例，進而減少搜索相關回覆資訊時所消耗時間。然而此郵件分類技術仍以單一概念為基礎，未能眞正回覆使用者之問題。是故，Sung 及 Chih（2004）乃使用多樣概念，並結合概念與分類兩者關聯，進行簡單郵件分類。該系統乃透過不同概念及分類結合成一套動態統一電子郵件機制，以建議不同內容範本，進而提升建議範本準確性，即可減輕客服人員回覆郵件往返次數，增加回覆郵件效率。

(B) 以資料探勘方法進行分類

針對資料探勘方法進行分類之課題，Bouguila 與 Amayri（2009）結合潛在狄式分配（Latent Dirichlet Allocation）與支援向量機建立一套能判別垃圾郵件與分類圖片之郵件分類模式。首先，該研究乃將郵件中詞彙頻率或圖片中之色彩空間作為分析數據，先透過潛在狄式分配分析詞彙或圖片所屬主題機率，接著將主題機率透過支援向量機建立向量特徵矩陣，以分析各郵件或圖片間向量距離並歸類，進而達成分類效果。由於該研究所建立之分類模式據有圖片與文字分析技術，透過文字與圖片兩者分類數據整合達到精確分類，且於垃圾郵件判別或文件分類等議題中皆能達到準確判別效果。此外，Poon 及 Chang（2003）利用電子郵件之詞彙相關性與 K 值鄰近演算法（K-Nearest Neighbor）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。此兩階段之分類方式即可獲得較佳之郵件分類效果。而 Islam 等人（2009）則透過統計學習機制（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，所存在的詞彙誤判（即具備灰色地帶之詞彙集）以及準確度之問題。最後，Irena 等人（2007）乃以監督學習與半監督學習兩種機器學習技術，針對隨機森林（Random Forest）、支援向量機、決策樹（Decision Tree）及天眞貝氏四種分類法分別進行郵件分類與垃圾郵件過濾，以比較四種分類法之執行績效。於監督學習模式中，該研究發現隨機森林優於其他三項，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。而在半監督學習機制中，針對四種分類法進行垃圾郵件過濾精確度之比較，發現此四種分類法於半監督模式之分類精確率並無差異，因此，該研究又以此四種分類以協同式訓練（Co-Training）方式應用於垃圾郵件過濾中，亦發現此訓練方式後之過濾郵件精確率較原模式高，因此可得知半監督模式對於垃圾郵件的範圍分類定義較具競爭力，亦可提高郵件分類的準確度。

綜合上述，針對郵件分類技術層面皆有許多文獻研究，無論根據郵件詞彙關聯性或資料探勘方法進行分類，皆使郵件分類方式更加多元、其結果更加精準，但多數分類方式皆單獨針對企業組織或未考量使用者使用習慣，而無法協助使用者進行個人化分類。本研究乃依其不同應用與技術彙整於表 2.1。

表 2.1、郵件分類技術文獻彙整表

文獻議題	過去研究	使用技術	分析對象	研究方法
以詞彙關聯性進行分類	Yu 與 Zhu（2009）	倒傳遞類神經網路、語意特徵空間	郵件內容文字	以語意特徵空間技術調整倒傳遞類神經網路學習比率。
	Buffett 和 Geng（2010）	天眞貝氏	通知郵件內容	以天眞貝氏計算郵件代表性標籤。
	Scheffer（2004）	支援向量機、天眞貝氏分類法	客户郵件內提問內容	以支援向量機擷取郵件內容特徵，並透過天眞貝氏分類法區分並產生對應回覆郵件內容。
	Sung 及 Chih（2004）	動態統一電子郵件機制	回覆郵件範本	首先進行簡單郵件分類，再根據分類建議不同郵件內容回覆範本。
以資料探勘方法進行分類	Bouguila 與 Amayri（2009）	狄式分配、支援向量機	郵件中內容文字或夾帶圖片	透過潛在狄式分配分析詞彙或圖片，並透過支援向量機建立向量特徵矩陣，並區分郵件類別。
	Poon 及 Chang（2003）	K 值鄰近演算法	郵件內容文字	事先所建立詞彙相關性並進行第一次分類，再利用 K 值鄰近演算法進行第二次郵件分類。
	Islam 等人（2009）	郵件多層分類器	郵件內容文字	建構一套郵件多層分類器，以改善現詞彙誤判（即具備灰色地帶之詞彙集）以及準確度之問題。
	Irena 等人（2007）	隨機森林、支援向量機、決策樹及天眞貝氏	郵件內容文字	分別以監督學習模式與半監督模式驗證四種分類法績效，並發現監督學習模式中隨機森林優於其他三項方法，而半監督模式中無明顯差異。

2.2.2 郵件過濾技術

於郵件過濾技術中，過去研究乃針對「以郵件標頭檔進行過濾」、「以傳輸協定進行過濾」以及「以郵件內容進行過濾」等三部分中，探討郵件過濾技術所涉及之範圍與領域。

(A) 以郵件標頭檔進行過濾

針對郵件標頭檔進行過濾之課題，Ying 等人（2010）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，Guzella 等人（2008）以人工免疫系統（Artificial Immune System；AIS）建構一套垃圾郵件篩選系統。其中，人工免疫系統乃根據生物免疫理論中，生物透過病毒入侵自我產生抵抗之觀念發展而成，因此該研究首先乃將外界接收之郵件視為生物病原體，並透過解析郵件內文中詞彙作為分析病原特徵，接著比對透過訓練資料而得之抗體資料進行篩選，當中，抗體資料乃透過匯入大量垃圾郵件作為訓練資料，且分析訓練資料之特徵，並加以重新組合成抗體資料，以過濾與抗體資料吻合之垃圾郵件，且該方法亦能自我產生機器學習機制，進一步提高垃圾郵件篩選準確率。但人工免疫系統常因錯誤訓練資料而導致篩選準確性過低，故 Mohammad 和 Zitar（2011）針對人工免疫系統並結合基因演算法（Genetic Algorithm），建構一套垃圾郵件篩選系統。其中人工免疫系統乃先透過大量垃圾郵件作為訓練資料，且該系統將訓練資料拆解並重新組合成抗體資料（即篩選時所需比對資料），而達到系統自我學習。且為能幫助系統組合有效抗體資料，故乃結合基因演算法計算訓練資料最佳解組合，進而提高垃圾郵件篩選效果，且為郵件篩選議題帶來新的技術。

過去研究常透過解析電子郵件內文方式篩選垃圾郵件，但此解析郵件內文之行為將涉及侵犯收件者個人隱私問題。故 Salcedo-Campos 等人（2012）為避免侵犯收件者個人信件隱私，則改以郵件標頭檔為解析對象。由於標頭檔案主要包含郵件伺服器、收件者、日期等數據化資訊，該研究為能將標頭檔資訊參數化，故乃以符號（如：「@」）作為擷取標記，並根據符號不同而賦予不同參數定義，如「<」乃用以區隔多位收件者信箱之符號，故定義為寄件者大量寄送之郵件。接著將標頭檔之參數透過隱藏式馬可夫模型（Hidden Markov Model）並以貝氏定理（Bayes Theorem）估算郵件歸屬機率，以判別垃圾郵件，由於該研究於解析過程並未分析郵件內文，因此，可保障郵件收件者個人隱

私權益。

(B) 以傳輸協定進行過濾

過去研究透過傳輸協定進行分析，以發展垃圾郵件之過濾技術，如 Herzberg（2009）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。而 Duan 等人（2007）則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，此外，該研究亦發展簡單的數學模式控制垃圾郵件過濾機制。相較於目前簡易信件傳輸協定所建立之電子郵件系統，Duan 等人（2007）所建構之電子郵件系統能使垃圾郵件於網路上緩衝較長之時間，以使垃圾郵件過濾機制可即時修正郵件黑名單。

(C) 以郵件內容進行過濾

針對郵件內容進行過濾之課題，過去垃圾郵件判斷大多只發分為垃圾郵件、非垃圾郵件兩類別，但現今多樣化發展之垃圾郵件已無法用二元制方式篩選，因此多數研究朝向分類方法為基礎發展垃圾郵件篩選機制，當中，多數分類法則皆透過擷取郵件文本特徵值進行分類，但郵件文本特徵擷取方式對分類結果影響程度皆不同，為探討各郵件特徵擷取之成效，Yang 等人（2011）針對資訊增量（Information Gain）、卡方檢定（Chi Square Test）、堅尼係數（Gini Index）、卜瓦松分配（Poisson Distribution）等郵件特徵擷取方法進行分類成效評估，當中評估方式將四個方法分別與天眞貝氏及支援向量機兩分類演算法組合進行郵件分類模擬，但為評估各擷取方法於不同類型郵件中分類成效差異，故將各分類組合分別以 6 種不同語料庫作為類別資料進行郵件分類。於模擬中發現，以天眞貝氏為分類法則時，資訊增量與堅尼係數兩方法之分類成效佳，但四種方法在以支援向量機進行分類時分類效果差距不大。

部分研究曾以倒傳遞類神經網路發展垃圾郵件篩選模式，但倒傳遞類神經網路於機器學習時太過緩慢，且容易產生分析參數最小值，進而降低判斷垃圾郵件準確性。因此，Li 和 Huang（2012）結合基因演算法將各神經元所得數據推算最適數據，以減少分析參數最小值產生，再擷取郵件內文中詞彙與同義詞，利用潛在語意分析（Latent Semantic Analysis；LSA）分析詞彙與同義詞之詞彙頻率及詞彙參數，以作為垃圾郵件判別依據，此外，為減少機器學習時間並自動產生訓練資料，該研究乃將郵件內文中詞彙與同義詞

之數據資料彙整並建立郵件詞彙庫，以作為機器學習所需訓練資料，進而減少倒傳遞類神經網路學習過程緩慢及分析參數產生最小值等問題，並提升郵件篩選效率與準確性。亦或以正向天眞貝氏分類法（Positive Naïve Bayes；PNB）及正向範例學習法（Positive Example-Based Learning；PEBL）發展垃圾郵件過濾法則，但 PNB 受權重設定正確性而影響判定結果，而 PEBL 受限於訓練資料正確，含有錯誤判斷之訓練資料將影響 PEBL 判斷結果。因此，為解決此兩項問題，Wei 等人（2008）乃結合 PNB 不受訓練資料影響判斷準確性特點及 PEBL 無需設定權重特性，以發展 E2 分析法則。該研究乃分為兩階段進行，第一階段分別以 PNB 與 PEBL 分析郵件內文，其中 PNB 乃根據郵件詞彙之詞頻計算詞彙機率，並彙整為詞彙特徵矩陣，而 PEBL 乃先拆解郵件內文之文句，再透過支援向量機分析文句特徵數據，以作為郵件分析權重。第二階段則結合上述所得之詞彙特徵矩陣及郵件分析權重，分別作為判別數據及判別權重，進而判別垃圾郵件，以避免錯誤訓練資料及設定錯誤權重等問題，並增加垃圾郵件篩選準確性。

過去部分垃圾郵件篩選模式無法針對多國語言進行郵件過濾，雖然仍有部分篩選模式支援多國語言郵件，但篩選所需執行時間過長。故 Çıltık 和 Güngör（2008）乃發展一套多國語言郵件過濾模式，其中乃將郵件內容文章以多詞彙（N-Words）拆解方式取得文章內容詞彙，並將文章詞彙透過 N-Gram 模式（N-Gram Model）分析郵件文字，但為能進行多國語言解析，因此該研究乃針對詞彙上下關聯界定詞性，且又因各語言詞性排列方式皆不同，故根據主詞、動詞、受詞排列順序計算自由詞序規則且計算排列機率，以幫助系統分析不同語言之文句，進而達到判別多國語言垃圾郵件。此外 Zorkadis 等人（2005）為減少郵件誤判與漏報問題，乃透過擷取郵件內文中詞彙並分析詞性，再以詞彙之詞性建立郵件特徵矩陣，以作為各分類法則之分析資料，並改善各分類法於郵件內容特徵擷取方式。該研究為能確實降低分類結果所產生誤判與漏報問題，因此於各分類法則中融入分類結果之評比機制，乃於分類判定後建立各分類結果誤報機率值，並根據各筆結果訊息之誤報機率判斷分類結果可信度，以改善各分類法則於郵件篩選上績效。

除了上述方法外，多數研究曾以天眞貝氏、神經網路（Neural Network；NN）、支援向量機及關聯向量機（Relevance vector machine；RVM）等演算法發展垃圾郵件篩選法則，為探討各演算法篩選效能與效果，Yu 和 Xu（2008）則針對上述四種演算法進行比較實驗，其中實驗乃分為兩階段測試四種演算法之篩選效果，其中第一階段以 6000 份電子郵件（垃圾郵件占 37.04%）作為實驗資料進行郵件篩選，於第二階段再以 5000 份電子郵件（垃圾郵件占 45.04%）進行測試，於實驗結果中發現，神經網路判定結果

常受限於訓練資料多寡與資料完整性影響，且四項分類法則中以支援向量機及關聯向量機為最佳，但關聯向量機訓練學習所耗費時間較高於支援向量機，故較適用於接收郵件種類複雜性低之使用者。

由於現今垃圾郵件已占據三分之二網路流量，且大部分垃圾郵件皆含有惡意程式或電腦病毒，造成電子郵件使用者接收郵件上困擾。Marsono 等人（2009）為幫助使用者分類且篩選垃圾郵件，乃建立一套三層式郵件內容分類法則，首先於第一層乃拆解郵件文句、字詞與符號作為郵件特徵，第二層再以郵件特徵進行天眞貝氏分類法（Naive Bayes Classifier）判定郵件是否歸類於垃圾郵件，第三層則再次檢視郵件附件，以過濾因第二層誤判導致惡意程式入侵使用者信箱。透過電子郵件收發模擬實驗，證實三層式郵件內容分類法則能準確判斷垃圾郵件，以預防使用者接收含惡意程式之信件。

綜合上述，針對郵件過濾之問題大多數文獻透過郵件標頭檔、郵件傳輸協定、郵件內容等郵件資料為基礎加以分析，使郵件過濾更加多元且精準篩選郵件（如表 2.2 所示）。

表 2.2、郵件過濾技術文獻彙整表

文獻議題	過去研究	使用技術	分析資料	郵件過濾方式
以郵件標頭檔進行過濾	Ying 等人（2010）	決策樹、支持向量機、倒傳遞類神經網路	郵件內容、標題、寄件者等資料	先將郵件內容、標題、寄件者等資料歸納得 14 種郵件特徵，再以綜合分類法分析判定垃圾郵件。
	Guzella 等人（2008）	人工免疫系統	郵件內容文字	透過解析郵件內文中詞彙作為分析病原特徵，並比對透過訓練資料而得之抗體資料篩選垃圾郵件。
	Mohammad 和 Zitar（2011）	人工免疫系統、基因演算法	郵件內容、郵件標頭檔	以基因演算法計算訓練資料最佳解組合作為抗體資料，而達到系統自我學習。
	Salcedo-Campos 等人（2012）	可夫模型、貝氏定理	郵件標頭檔	將標頭檔之參數透過隱藏式馬可夫模型並以貝氏定理估算郵件歸屬機率，以判別垃圾郵件。
以傳輸協定進行過濾	Herzberg（2009）	DKIM	郵件傳輸方式	結合路由器和 DKIM 以協助電子郵件過濾任務，以改善目前垃圾郵件過濾效果不彰之情況
	Duan 等人(2007)	區別轉寄郵件協定	郵件轉寄協定	發展一套區別轉寄郵件協定，並發展一套簡單的數學模式控制垃圾郵件過濾機制。
	Duan 等人(2007)	黑名單過濾機制	郵件傳輸方式	建構一套電子郵件系統，作為網路上垃圾郵件接收緩衝時間，以協助使垃圾郵件過濾機制即時修正郵件黑名單。
以郵件內容進行過濾	Yang 等人(2011)	資訊增量、卡方檢定、堅尼係數、卜瓦松分配	郵件內容文字	將四個方法分別進行郵件分類模擬，且發現資訊增量與堅尼係數兩方法之分類成效較佳。
	Li 和 Huang 等人（2012）	基因演算法、潛在語意分析	郵件內容文字	以基因演算法推算最適數據，再分析郵件內文中詞彙之詞頻及詞彙參數，並篩選垃圾郵件。
	Wei 等人（2008）	正向天眞貝氏分類法、正向範例學習法	郵件內容文字	第一階段透過支援向量機分析文句特徵數據，並於第二階段以文句特徵數據判別垃圾郵件。
	Çıltık 和 Güngör（2008）	多國語言郵件過濾模式、N-Gram 模式	多國語言郵件內容文字	針對各語言詞性不同界定不同詞彙上下關聯詞性，並根據詞性排列順序計算自由詞序規則與排列機率，以判別多國語言垃圾郵件。
	Zorkadis 等人（2005）	詞性分析	郵件內容文字	透過擷取郵件內文中詞彙建立郵件特徵矩陣，以改善各分類法於郵件內容特徵擷取方式。
	Yu 和 Xu（2008）	天眞貝氏、神經網路、支援向量機、關聯向量機	郵件內容文字	針對此四項法則進行驗證比較，且四項分類法則中以支援向量機及關聯向量機之篩選績效最佳。
	Marsono 等人（2009）	天眞貝氏分類法	郵件文句、字詞與符號	以郵件文句、字詞與符號作為郵件特徵並進行天眞貝氏分類法判定郵件是否歸類於垃圾郵件。

2.2.3 郵件解析方式

於郵件解析方式中，過去研究多數針對「郵件連絡人資料分析」、「郵件文字資料分析」以及「郵件傳輸協定資料分析」等議題，探討目前郵件解析方式。

(A) 郵件連絡人資料分析

於郵件連絡人資料分析議題中，Chundi 等人（2009）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，並計算用户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。最後以安然公司的電子郵件數據集（Enron Email Data Set）進行實證，亦獲得具正面效益之驗證。

企業組織為能防範員工對外洩漏機密資料或企業間諜，常以監控網路資訊傳遞方式或審核郵件寄件者與收件者方式防止資訊外流。因此，針對企業間諜防範，Okolica 等人（2007）乃透過員工傳遞之電子郵件內容進行解析，並預測員工成為內部間諜可能性。該研究乃透過作者主題模型（Author Topic Model）分析寄件者平時寄送郵件之內容主題，且主題當中亦包含機密文件主題。該研究乃跟據郵件中詞彙出現次數分析詞彙於該郵件中出現之機率，再根據詞彙於郵件中出現機率加以估計該郵件所屬主題，進而預測該郵件寄件者發送此主題類型郵件之機率，並將郵件寄件者根據發送主題機率分為多個寄件者群集。根據此方法所進行員工寄送郵件主題類型預測，若預測而得主題為機密文件，則建議管理者減少託付重要機密訊息於該員工，進而減少企業資訊外流危險。

於現今社會中，電子郵件常被用於詐騙、販毒或黑函中傷等犯罪中，因此電子郵件亦被作為犯罪工具，但犯罪相關郵件數量龐大且內容複雜，故調查人員需耗費大量時間從中尋得犯罪證明。為縮短尋得犯罪證明所耗費時間，Hadjidj 等人（2009）乃以資料探勘技術為基礎發展自動化郵件調查系統，並幫助調查人員減少搜尋犯罪郵件時間。該研究乃先解析郵件內容文字並劃分為多個郵件群集作為後續處理資料，接著該研究乃以寄件者、郵件主題等郵件特性建立多維度類別方式，以詳細區分郵件敘述犯罪性質歸屬，並以群集為單位，分別以郵件內容之詞彙頻率及寄件者，以作為特徵值進行兩種不同郵件分類，使各筆郵件分別含有寄件者類別及郵件主題類別，並透過郵件所屬群集、寄件者類別及郵件主題類別呈現多維度郵件歸屬類別。調查人員透過多維度郵件歸屬類別檢索與罪犯相關郵件，並做為犯罪證明幫助司法判決。

過去研究雖能透過傳輸協定分析協助垃圾郵件篩選，但現今詐騙郵件不斷改變標頭

結構來預防系統過濾郵件，且詐騙郵件為能提高收件者閱讀率，故乃透過修改寄件日期使郵件列於第一順位，故此類型詐騙郵件容易造成使用者閱讀信件困擾。Banday 等人（2011）探討並分析詐騙郵件修改寄件日期對使用者之影響。該研究設計一份問卷主要訪問使用者面對此類型垃圾郵件處理方法與對使用者之影響，且為提高調查可信度，該研究並針對 14 種知名電子信箱服務使用者進行調查，且經調查後歸納出七項對使用者之影響，分別為：使用者產生收件困惑、使用者容易忽略重要信件而降低工作績效、容易導致郵件過濾系統誤判重要郵件、容易造成使用者忽略重要郵件時效、詐騙郵件亦將標題改為使用者熟知親人並造成使用者對親人不信任、使用者難以追蹤郵件寄發時間及造成使用者忽略寄件日期重要含意等問題等影響。

(B) 郵件文字資料分析

郵件文字資料分析相關領域中，已有多數研究針對電子郵件分類與篩選等議題進行探討，且各研究所提之分類法則皆能達到分類或篩選惡意郵件效果。但於眾多郵件分類或篩選法則中，並非所有法則結果皆能精準分類郵件，因此為比較各電子郵件分類法則之成效，Cai 等人（2007）透過哥倫比亞大學（Columbia University）所建之入侵檢測系統網站擷取 4754 份郵件作為實驗資料，並以支援向量機、天眞貝氏、最大熵法則（Maximum Entropy）等分類法進行系統績效比較，實驗設計乃分為 5 組郵件群，當中 1 組為測試資料，其餘 4 組作為訓練資料，分別進行各法則系統測試，並記錄分類器之分類準確度、分類穩定性及系統訓練時間等三個方向，作為系統績效評估依據；於實驗中發現，相較於其他法則，支援向量機所需系統訓練時間較短，且預測分類結果較為準確。此外，部分研究亦融入圖片信息加強解析效果，如 Zhou 等人（2007）乃結合關鍵字擷取技術及圖片信息測量（Picture Information Measurement；PIM），建構一套智慧型電子郵件分類系統；該系統乃使用天眞貝氏分類先行過濾郵件（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，並將各郵件進行分類；是故，藉由關鍵字過濾與圖片資訊分析後，即可針對目標郵件進行有效地分類。

此外，過去亦有研究透過解析文字作為垃圾郵件篩選依據，增加垃圾郵件過濾效能並加強電子郵件管理與應用，因此，如 Gomez 和 Moens（2012）以分類郵件概念為基礎建立一套垃圾郵件篩選法則，以達到篩選垃圾郵件亦能分類管理電子郵件。該法則先以主成份分析（Principal Component Analysis；PCA）分析郵件內容文句，以獲取郵件之

特徵向量，接著以相同方式分析垃圾郵件之特徵向量作為郵件類別歸屬標準，以確認郵件歸屬類別，再以兩者之特徵向量計算歐式距離（Euclidean distance），並以兩向量距離作為郵件與類別相似度判斷郵件所屬類別。且為證實方法可行性，該研究乃將郵件分為：廣告郵件、詐騙郵件、一般信件等類別，並與支援向量機比較分類準確率，證實該研究所建立之郵件分類模式能準確分類且篩選郵件。而 Laorden 等人（2011）則以天眞貝氏建立垃圾郵件篩選器，以擷取郵件中詞彙表達方式篩選垃圾郵件，且為提高篩選精準度，該研究以詞義消歧（Word Sense Disambiguation；WSD）方式，根據詞性與詞彙關聯擷取詞彙並將詞彙含意單一化，以提高郵件篩選器準確度。

部分研究則針對企業之客户郵件探討郵件文字解析議題。由於客户建議信件往往蘊含各式重要資訊，過往研究雖針對客户建議信件進行分析並分類，但少有研究透過客户建議信件分析客户類型與興趣偏好。有鑑於此，Sakurai 和 Suyama（2005）乃以模糊歸納學習法（Fuzzy Inductive Learning）為基礎建立電子郵件文本分析模式。首先，該研究為能根據郵件內文詞彙進行關聯分析並取得客户興趣偏好相關詞彙，故乃先行匯入訓練文本，透過分析訓練文本中詞彙之詞彙頻率並建立關鍵概念詞庫，再以模糊歸納學習法分析郵件內容文章，獲取文章中重要資訊並根據資訊內容分類郵件，且依據文章中重要資訊與關鍵概念詞庫互相比對詞彙關聯，進而取得客户興趣偏好及客户相關資訊，以幫助企業針對不同客户群進行不同需求回應。

(C) 郵件傳輸協定資料分析

電子郵件須透過伺服器間傳輸協定審核與分析才得以傳送，然而郵件軟體於發送與接收郵件上尚存速度過於緩慢之問題。因此，Laura 及 Maria（2001）提出以 SMTP 與 POP3 協定模式建構於 SPECmail2001 軟體中，以改善郵件收發效率。該研究先於郵件伺服器中模擬大量郵件之寄送或接收作業，並測量兩者協定之傳送特徵以結合成一套協定模式；之後，該研究將此建構模式建置於 SPECmail2001 軟體中，並以此模式之協定視為郵件傳送工作量分配協定，以達有效之傳送郵件作業。實驗結果亦證實，依此模式之協定使寄發郵件速度提升，使用者亦可迅速獲得郵件訊息。

此外，Appavu 等人（2009）提出以決策樹演算法為基礎，建構一套智慧型過濾恐怖信息郵件系統。該系統乃提供儲存、編輯、搜尋和進階搜尋選項功能，當中可編輯、搜尋選項乃包含寄件者和寄件帳號等項目，且該系統亦提出一個郵件資料探勘之變更技術，並結合給予不同使用者之自行設定資料需求功能，以使系統具備易調整性與高度準

確性之績效，進而協助調查員於收集郵件線索和證據之任務。

傳輸協定過程中郵件資料隱私亦為一項重要議題，Phan（2008）為驗證聲稱能提供最完善保密機制的兩項電子郵件協定，是否具有其功效進行實驗，此兩項電子郵件協定乃使用驗證以及密鑰交換技術（Key-Exchange Techniques），並包含了完美轉送私密（Perfect Forward Secrecy）、已知金鑰安全（Known-Key Security）、金鑰及時性（Key Freshness）和未知金鑰分享安全（Unknown Key-Share Resilience）等技術，經過實驗證實，此兩項協定無法抵制反覆攻擊（Replay Attacks），並進一步指出，在第一次攻擊時便突破了未知金鑰分享安全技術，而第二次攻擊時就已突破完美轉送私密技術，此結果明顯違背了開發人員的要求，故得知，目前較新的技術並不一定意味著更安全。

綜合上述，郵件解析技術議題中多數研究著重於郵件文字解析，亦有不少研究透過郵件中文字分析使用者寫作習慣及語意，進而加強郵件過濾或郵件分類之效果，使郵件解析之議題更加多元且應用範圍更為廣泛（如表 2.3 所示）。

表 2.3、郵件解析方式文獻彙整表

文獻議題	相關研究	解析技術	解析郵件類型	解析內容
郵件連絡人資料分析	Chundi 等人（2009）	時間序列	企業組織郵件	郵件聯絡人、郵件寄發時間
	Okolica 等人（2007）	作者主題模型	企業組織郵件	郵件內容文字、寄件者與收件者關聯
	Hadjidj 等人（2009）	多維度類別	匿名犯罪郵件	郵件內容用詞、符號、寄件者寫作方式
郵件文字資料分析	Cai 等人（2007）	支援向量機、天眞貝氏、最大熵法則	校內公告郵件	郵件內容文字
	Zhou 等人（2007）	關鍵字擷取技術、圖片信息測量、天眞貝氏	一般郵件	郵件內容文字、郵件夾帶圖片
	Gomez 和 Moens（2012）	主成份分析	廣告郵件、詐騙郵件	郵件內容文字
	Laorden 等人（2011）	天眞貝氏、詞義消歧	廣告郵件、詐騙郵件	郵件內容文字
	Sakurai 和 Suyama（2005）	模糊歸納學習法	企業所接收之客戶郵件	郵件內容文字、客戶資料
郵件傳輸協定資料分析 郵件傳輸協定資料分析	Laura 及 Maria（2001）	SPECmail2001、SMTP、POP3	一般郵件	SMTP 與 POP3 等郵件傳輸協定方式
	Appavu 等人（2009）	決策樹演算法	廣告郵件、詐騙郵件	郵件傳輸協定方式、寄件者、寄件者帳號
	Phan（2008）	密鑰交換技術	一般郵件	郵件保密協定方式

2.3 郵件使用者行為探討

針對郵件使用者行為探討之課題，本研究主要乃針對「使用者郵件管理習慣分析」、「使用者慣用郵件類別分析」及「使用者郵件應用領域分析」等議題進行相關文獻探討，以更深層瞭解郵件使用者行為探討特性，並從中探討使用者之電子郵件使用習慣。

2.3.1 使用者郵件管理習慣分析

於使用者郵件管理習慣分析議題中，本研究乃針對「使用者郵件分類習慣」以及「使用者郵件寫作風格」等兩主題進行相關文獻探討，期望從中探討郵件使用者個人化特質與一般使用者管理郵件之習性。

(A) 使用者郵件分類習慣

電子郵件已成為人與人之間最主要溝通方式，然而電子郵件應用普及卻容易導致信件過量問題。Szóstek（2011）認為若提供電子郵件標籤檢索或結構化整理等功能，則可提高使用者閱讀郵件效率，並方便使用者將電子郵件進行歸類整理。為證實此項觀點，該研究乃以紙條作為信件進行電子信箱收信模擬實驗，且要求 16 位受測者以未整理、結構化整理、檢索標籤等三種方式分別進行模擬，並調查受測者對三種模擬收信之意見，多數受測者認為結構化與檢索標籤能將郵件有系統整理，幫助使用者快速取得重要信件，且檢索標籤歸類方式更為靈活，能依照使用者需求進行調整歸類方式。故調查結果證實標籤檢索與結構化整理等功能有助於提高使用者郵件閱讀效率，並建議未來電子郵件使用端，於設計時需考量信件結構化或檢索標籤建立功能，以助於使用者管理電子郵件。

為減少郵件過量對工作績效之影響，Soucek 和 Moser（2010）則認為若有效教導使用者信件文章撰寫、電子信箱內標籤與通訊錄使用方式、導入或學習電子郵件分類管理思維等相關知識，應可解決現今電子郵件所存在問題。因此，該研究設計一項實驗邀請 90 位受測者接受電子郵件相關課程培訓，並分三階段方式分別調查受測者接受培訓前後使用電子信箱狀況，以了解培訓課程是否有助於改善電子郵件問題。透過調查中發現，受測者接受培訓課程學習後皆能有效減少電子郵件過量、缺乏效率管理方式、信件內容缺乏文章結構等問題困擾，且發現若組織內全部成員皆接受此課程培訓，更能防範此些問題發生，故該研究建議企業組織應將電子郵件使用方式納入員工培訓中，以提高員工工作績效。但接受培訓課程之員工是否能改善郵件管理上耗時問題仍備受質疑。故 Huang

等人（2011）以 16 位大學生接受 3 小時培訓，並於培訓後填寫一份李克特量表（Likert Scale），該量表內容主要調查接受訓練前後差異，當中評量準則可分為整理電子信箱時間、檢索郵件時間、使用者控制管理電子郵件所耗費時間、整體使用電子郵件時間等，接著再計算量表共變異數以評估改善狀況，並發現經過培訓後對電子郵件收發整理電子信箱時間並無明顯差別，但使用者檢索或尋找郵件等行為上可大幅縮短時間。此外，Joung 及 Yang（2009）提出一套個人化電子郵件篩選機制，當中用户可依照本身所設立郵件內容與條件，允許郵件進入用户信箱，並儲存至對應之類別中。該機制利用所設立訊息使外來電子郵件進入系統後，系統隨即根據用户所設定之規則判定郵件類別。雖然現今電子郵件系統的功能已相當完善，但缺乏較為直覺且簡單之操作模式。

網際網路蓬勃發展促使電子郵件廣泛應用於工作上溝通，以提高工作上執行效率。但電子郵件除了影響工作績效外，亦可能對生活或工作產生正面或負面影響，因此 Mano 和 Mesch（2010）乃針對個人使用電子郵件特性，如應用於工作程度、檢視信箱頻率、非工作信件數量、信件接收量等特性設計李克特量表，並透過美國 Pew Internet 研究中心之美國生活樣本（The Pew and American Life Sample）中，挑選 354 位受測者接受量表測試，再由多變量分析（Multivariate Analysis）與迴歸分析（Regression Analysis）分析量表資料，以探討受測者職位、電子郵件特性及工作績效等三項因素間關聯。調查結果中發現，受測者職位越大，電子郵件中應用於工作程度、檢視頻率、接收量等特性越高，且此些特性越高相對工作績效亦越佳，但容易增加受測者工作壓力與負面情緒，且電子郵件也因受測者年齡、婚姻、性別等因素不同，產生影響程度也不相同。且亦有研究發現工作情形也會影響郵件傳遞數量，如 Sumecki 等人（2011）乃以李克特量表方式設計一份問卷，以了解企業內部使用電子郵件方式與企業文化關聯，並從中分析郵件過量主因。其中，該問卷調查內容主要包括幾項問題包含管理電子郵件策略、如何定義重要郵件、組織文化等，並透過迴歸分析方式分析問卷中資料。分析後發現，重要郵件收發量隨使用者於工作量越多及組織地位越高而增加，若員工外出公差而離開崗位時，則因無法即時處理工作轉以電子郵件方式進行導致郵件數量增加，且員工於職位上不滿時，也會透過電子郵件與親人溝通，導致私人信件過多而造成企業組織內郵件過量，是故，此些分析結果可提供企業管理者減少郵件過載問題之參考。

(B) 使用者郵件寫作風格

隨著電子郵件的使用，有許多企業組織意識到電子郵件所伴隨而來的問題，如電子郵件收發數量上的限制，以及撰寫內文的文筆方面等問題（Burgess 等人；2004）。且現行網際網路中，以隱匿寄件者方式寄送匿名郵件之犯罪者盛行，因此，過去部分研究為能辨識匿名郵件之寄件者乃透過解析匿名郵件之文字語意，以比對寄件者文章寫作方式尋找匿名者。其中，Iqbal 等人（2010）乃建立電子郵件寫作觀點分析模式，以協助司法調查局尋得具犯罪意圖之匿名郵件寄件者。該研究首先乃擷取郵件中詞彙，並保留語句中語助詞及標點符號，以作為寄件者語言風格特徵，接著乃分析該寄件者文章詞彙用法及文句串接方式，接著，該研究透過寄件者寫作風格與語言風格進行歸納，以透過寄件者用語及文句探討寄件者於文章內容中觀點，如：悲觀、積極等寫作觀點。最後透過實驗證實能成功比對出匿名郵件寄件者，因此該研究建議於未來相關研究融入寄件者寫作觀點，以增強文字分析調查上準確性。

由於電子郵件具有回覆、轉寄等功能，多數郵件使用者乃透過多封郵件來回傳遞進行對話討論，使電子郵件內容皆可串聯前後敘述內容成串聯郵件（Email Threads）。但由於串聯郵件內容大多為單一事件討論與私人對話，易導致敘述文句不完整或無法對應，更產生使用者無法理解郵件內容之問題。為能將串聯郵件所述內容以更清楚且有效率方式供使用者閱讀，Zajic 等人（2008）乃發展多方訊息彙整（Collective Message Summarization；CMS）方法，彙整郵件內文形成具結構性文章，然而此份文章內容敘述並非簡潔易懂，因此，為建立簡潔易懂之郵件總結，該研究乃透過隱藏式馬可夫模型（Hidden Markov Model）與解析樹（Parse Trees）兩法則建立個人訊息彙整（Individual Message Summarization；IMS），以擷取重要文句並構築為串聯郵件總結，方便使用者閱讀且容易理解串聯郵件內容。且現今使用者逐漸習慣透過行動裝置傳遞郵件，然而並非所有行動裝置可輔助郵件同步與整合。因此，為提供使用者一套能同步各裝置電子郵件軟體，Rao 等人（2003）乃建立 iMail 電子郵件管理系統，當中，該研究乃整合行動裝置與個人電腦之電子郵件，並以無線應用協定（Wireless Application Protocol；WAP）方式同步傳送電子信箱內郵件接收狀況，且為幫助使用者管理及閱讀郵件，並提出三項功能便於電子信件的讀取：信件過濾、二階段擷取、回覆建議等功能，其中，信件過濾使用者可選擇較為重要信件進行裝置同步，幫助使用者即時取得重要資訊，二階段擷取乃減少顯示郵件內容時傳送不必要的資料，回覆建議則針對使用者目前閱讀信件，以選單方式建議回覆內容與收件者，透過此等功能幫助使用者即時閱讀信件，且避免使用者於

不同裝置上產生整合郵件困擾。

綜合上述，針對使用者郵件分類習慣中許多研究證實不當郵件管理容易影響使用者工作績效，且使用者為能以自我慣用之分類方式管理郵件信箱將耗費更多時間，而使用者郵件寫作風格之相關研究則提出各式使用者特徵結取方式，以使郵件分類或管理等相關系統更加多元化，而依其不同應用與技術分別彙整成表 2.4。

表 2.4、使用者郵件管理習慣分析文獻彙整表

文獻議題	相關研究	研究議題	郵件類型	研究方法	研究過程與結果
使用者郵件分類習慣	Szóstek（2011）	以標籤檢索或結構化整理郵件之可行性。	企業組織郵件、私人郵件	模擬實驗、訪談調查	使用者期望電子信箱能提供標籤檢或郵件分類功能。
	Soucek 和 Moser（2010）	培訓課程對郵件管理效率影響。	企業組織郵件	模擬實驗、訪談調查	受測者接受培訓後皆能有效提高郵件管理效率。
	Huang 等人（2011）	調查郵件管理培訓課程之效益。	企業組織郵件	模擬實驗、訪談調查	培訓前後無明顯差別，但可大幅縮短使用者檢索或尋找郵件時間。
	Mano 和 Mesch（2010）	電子郵件對工作與生活影響。	企業組織郵件、私人郵件	量表統計、數據分析	受測者職位越大，使用郵件頻率越高且工作績效越佳，但容易增加受測者工作壓力與負面情緒。
	Sumecki 等人（2011）	電子郵件使用頻率與工作影響。	企業組織郵件、私人郵件	量表統計、數據分析	郵件收發量隨著使用者職位增加，且員工於職場不滿易透過郵件與親人表達不滿。
使用者郵件寫作風格	Sumecki 等人（2011）	電子郵件使用頻率與工作影響。	企業組織郵件、私人郵件	量表統計、數據分析	郵件收發量隨著使用者職位增加，且員工於職場不滿易透過郵件與親人表達不滿。
	Iqbal 等人（2010）	分析匿名郵件之寄件者。	匿名犯罪郵件	技術發展	首先乃擷取並分析郵件中詞彙用法及文句串接方式，以尋得郵件寄件者。
	Zajic 等人（2008）	串聯郵件敘述內容供使用者閱讀。	私人郵件	技術發展	透過隱藏式馬可夫模型與解析樹分析郵件文章結構，並串聯郵件內容。

2.3.2 使用者慣用郵件類別分析

於使用者郵件管理習慣分析議題中，本研究乃針對「使用者偏好類型判定」以及「使用者特徵資料擷取」等兩主題進行文獻探討，期望從中瞭解使用者郵件管理習慣分析議題所涉及之範圍與領域。

(A) 使用者偏好類型判定

現今多數人以電子郵件進行郵件溝通，致使郵件數量過於龐大，而超乎個人之處理能力，產生郵件超載問題。為解決此問題，Schuff 等人（2006）建構一套自動類聚管理

系統（Automatic Clustering E-Mail Management System），該系統乃運用多屬性（Multi-Atribute）、多重權重分配（Multi-Weight）類聚之方法，於電子郵件進入此系統後進行郵件解析任務：(1)電子郵件檢索：每封郵件另存為獨立之文字檔案，並且置於伺服器磁碟機之含有唯一 Session 之識別字之目錄中；(2)權重分配：用户可根據郵件標題、內容、發件人之屬性以作權重分配；(3)群集分析：收集郵件標題、內容、發件人之屬性，並且消除常見文字與片語，以取得關鍵片語；(4)分層群集：將群集郵件依據關鍵片語再度細分，讓使用者可以看清郵件之關鍵片語與郵件分類，藉由上述之機制後，即可協助郵件使用者進行郵件管理之任務。此外，針對使用者偏好分類，Sun 和 Dong（2010）乃建立一套動態層級化郵件分類系統，幫助使用者整理過量郵件。該研究乃以潛在語意分析（Latent Semantic Analysis；LSA）與非負矩陣分解法（Nonnegative Matrix Factorization；NMF）兩方法解析並歸類電子郵件內容，再以歸類郵件群建立層級化郵件類別，且為預防推論結果不符合使用者需求，故另行發展動態分層重建法則（Dynamic Category Hierarchy Reconstruction；DCHR）以供使用者重新推論層級化郵件類別，以強化電子郵件整理效果，並解決郵件過量而難以閱讀郵件之問題。

過往研究曾以分群方式整理郵件，進而解決郵件過量問題。但部分郵件分群方法需額外建立系統學習參數提高分群效果，導致需有效設定系統參數才得以準確分群。有鑑於此，Yang（2009）乃發展無參數文本分群演算法（Nonparametric Text Clustering Algorithm），將相似郵件分群歸類達到整理效果。首先，該研究乃拆解郵件內文詞彙並計算各詞彙之詞彙頻率進行分群，並為能使系統自動產生分群參數並驅使系統具學習能力，故乃將關鍵字之詞彙頻率透過 Hubert's Γ 統計（Hubert's Γ Statistics）分析郵件關鍵字之分群參數，再利用向量空間（Vector Space）以分群參數為依據計算郵件相似距離，並根據郵件相似距離判斷郵件歸屬群集，以有效將郵件分群並幫助電子郵件使用者檢索郵件。

電子郵件過量使組織內員工容易遺漏重要訊息，但現今郵件分類法則無法針對組織性質不同、不同員工業務作適合郵件分類，故 Alberts 和 Forest（2012）乃建立適合不同組織且能符合需求之分類法則，當中分為兩階段，第一階段乃針對組織員工調查電子郵件使用習慣，如：電子郵件傳遞期限、主題、寄件對象及收件者地位關係等特性，並將此些特性彙整為非詞彙資料，以作為郵件分類時附加條件；第二階段擷取郵件中詞彙語意與文法透過 K 值鄰近演算法分析郵件與類別相似程度，且為符合各組織不同分類需求，於郵件與類別相似度分析時結合非詞彙資料區分郵件歸屬，再透過郵件與類別相似

度將郵件歸類於相似類別，以協助組織員工以個人需求方式分類郵件。

此外，短文分類技術已廣泛應用於各種文件檔案議題中，如郵件過濾或文件檢索等議題中，但過去分類技術於詞彙擷取時並未考量專業術語與各詞彙間關聯特徵，進而降低分類技術精準度。因此，Meng 和 Yu（2011）透過特徵貢獻度模型（Feature Contribution Degree）與潛在語意索引（Latent Semantic Indexing）進行兩階段短文特徵詞彙擷取與分類。該研究於第一階段乃將短文拆解出若干詞彙，並透過特徵貢獻度模型將各詞彙以特徵向量方式呈現各詞彙於短文中關聯程度，接著以奇異值分解（Singular Value Decomposition）將詞彙特徵向量維度縮小並建立各短文間關聯，再以第二階段以潛在語意索引分析詞彙特徵向量，並根據各短文間詞彙關聯程度進行分類。且該研究以一般郵件與垃圾郵件匯入系統中進行垃圾郵件判別測試，證實該研究所提方法能有效解析短文中文章，並以詞彙之分析數據將短文準確分類至所屬類別。但短文分類過程中容易受到額外所擷取特徵干擾，而降低分類績效。因此，為能有效且精準擷取短文特徵，Ugˇuz（2011）透過基因演算法結合主成分分析提高短文特徵擷取精準度與分類準確性。其中，該研究共劃分為兩階段進行分類，首先於第一階段擷取短文中詞彙並計算詞彙頻率後，再將詞彙頻率透過資訊獲利計算增強詞彙特徵性並排序；第二階段則透過主成分分析將增強後詞彙特徵性分析短文中代表性詞彙，以避免短文分類受到額外非重要詞彙干擾，接著再以基因演算法將各筆詞彙特徵性加以組合並計算，以推論最適特徵組合降低短文分類誤判性。該研究藉由資訊獲利方式增強詞彙特徵性，並透過主成分分析方法過濾額外擷取之詞彙或特徵性較低之詞彙，進而可協助短文分類系統擷取具有效益之詞彙並提升分類系統績效。

(B) 使用者特徵資料擷取

於使用者特徵資料擷取議題中，過去少有研究針對電子郵件寄件者之郵件傳送意圖進行擷取分析，Mao 等人（2011）提出一套自動學習之電子郵件意圖分析系統。首先乃解析郵件中內容文字之詞彙頻率擷取關鍵詞，並根據郵件所屬寄件者歸類關鍵詞，接著為達到系統自動學習且能判斷郵件傳送意圖，該研究乃結合社群網路，再根據寄件者於社群網路中郵件寄件行為彙整為數據庫，並根據寄件者之社群網路行為數據及郵件中內容文字進行垃圾郵件判斷，由於該系統具自動化學習機制，且能根據郵件傳送意圖加以分析，故能以少量訓練資料且準確篩選垃圾郵件。

過去研究曾以傳統分類方式分析郵件內文並歸類出相同寄件者之郵件，但無法針對

匿名電子郵件寄件者之眞實寫作風格進行分析。因此，Iqbal 等人（2008）針對罪犯寫作方式建立一套郵件文本分析方法，該研究首先收集犯罪嫌疑人過往發送郵件，再以此些郵件分析文章中詞彙詞性、用語及符號用法，並歸類出嫌疑人之寫作方式與寫作風格，並整理成罪犯寫作特徵庫，接著乃分析匿名電子郵件之內文並擷取郵件內文寫作特徵，再與罪犯寫作特徵庫比對寫作特徵相似性，進而歸屬匿名電子郵件寄件者，並幫助調查者進行犯罪調查。

於郵件分類方面，為能探討分類器與特徵擷取關聯，Crawford 等人（2004）利用分類器中之向量空間、K 值鄰近演算法、決策樹、天眞貝氏與最大概似法（Maximum-Likelihood）及詞語分析中詞彙庫（Bag of Words）與特徵選擇進行交叉組合，以比較當中各組合對於電子郵件之分類具備有最佳效益，實驗結果乃顯示最大概似法與特徵選擇之組合具備電子郵件分類之最佳績效。此外，為瞭解郵件內文之詞彙對郵件分類之影響，Chang 及 Poon（2009）於電子郵件分類器中，利用郵件字詞作為分類特徵進行分類效能之評估，該研究係透過天眞貝氏分類器、K 值鄰近演算法與詞彙頻率（TF-IDF）結合第 K 鄰近之演算法等分類器，並以郵件集合以分別探討「片語長度」、「涵蓋區域大小」及「最鄰近點長度」對分類之影響，並計算各方法可改善分類準確率之效益；然而實驗結果發現，僅使用字詞為分類特徵之分類效果上，對於各分類器皆無明顯之顯著差異，且於此實驗中亦發現公共郵件較私人郵件易於分類。

使用者特徵資料擷取亦應用於企業客户投訴信件管理中，其因客户投訴信件管理信箱內所接收信件並非皆為客户投訴信件。為能區分客户投訴信件，Coussement 和 Poel（2008）乃建立一套客户信件分類法則，以區分電子信箱內所接收信件。由於客訴信件與一般信件寫作用詞截然不同，因此該研究乃以信件寫作格式作為區分基準，先以向量空間（Vector Space）拆解郵件內文詞彙，並根據詞彙之詞彙頻率建立郵件寫作風格特徵作為郵件區分依據，再以 Adaboost 分類演算法分析郵件寫作風格特徵，判斷郵件寫作用詞所趨向類別，達到區分客户投訴信件與一般信件。

綜合上述，針對使用者慣用郵件類別分析層面皆有許多文獻研究，無論根據「使用者偏好類型判定」及「使用者特徵資料擷取」等以提之相關研究，皆使資訊解析方式更加多元且精準分類郵件，並依不同應用與技術彙整成表 2.5。

表 2.5、使用者偏好類型判定文獻彙整表

文獻議題	相關研究	分析技術	分析郵件類型	擷取資料
使用者偏好類型判定	Schuff 等人（2006）	自動類聚管理系統	私人郵件	郵件標題、內容、寄件人之屬性
	Sun 和 Dong（2010）	潛在語意分析、非負矩陣分解法	私人郵件	郵件內容文字
	Yang（2009）	無參數文本分群演算法、Hubert's Γ 統計、向量空間	一般郵件	郵件內容文字
	Alberts 和 Forest（2012）	K-NN 演算法	企業組織郵件	員工使用習慣、郵件內容文字
	Meng 和 Yu（2011）	特徵貢獻度模型、潛在語意索引、奇異值分解	一般郵件	郵件內容文字
	Ugˇuz（2011）	基因演算法、主成分分析、資訊獲利	一般郵件	郵件內容文字
使用者特徵資料擷取	Mao 等人（2011）	自動學習之電子郵件意圖分析系統	私人郵件	郵件內容文字、寄件者之社群網路行為
	Iqbal 等人（2008）	寫作風格	匿名電子郵件	文章中詞彙詞性、用語、符號用法
	Crawford 等人（2004）	向量空間、K 值鄰近演算法、決策樹、天真貝氏、最大概似法	一般郵件	郵件內容文字
	Chang 及 Poon（2009）	天真貝氏分類器、K 值鄰近演算法、詞彙頻率	企業組織郵件、私人郵件	郵件內容文字
	Coussement 和 Poel（2008）	向量空間、Adaboost 分類演算法	客戶投訴信件	郵件內容文字

2.3.3 使用者郵件應用領域分析

於使用者郵件應用領域部分，本研究乃針對「醫療領域」、「教育領域」以及「其他領域」等三領域進行探討，期望從中此議題所涉及之範圍與領域。

(A) 醫療領域

醫學領域對「電子郵件幫助病人與醫療人員溝通」此項觀點給予保留態度，多數醫師依然無法接受電子郵件方式與病人溝通，故 Bhor 和 Mason（2006）為探討電子郵件無法在醫療領域盛行原因，以擴散理論研究分析使用電子郵件之優勢、電子郵件與現行醫療體系相容性、電子郵件使用複雜性及醫療人員實際使用成效等因素，再以問卷方式調查 1500 位醫療人員對電子郵件既有印象與態度。該研究於調查中發現電子無法盛行乃因醫療程序具有嚴謹程序規範，但並未明確定立透過電子郵件傳遞病歷資料之程序，導致醫療人員為避免錯誤治療而拒絕使用。

此外，藥品供應商對郵件關注程度、病人使用郵件方便以及郵件使用效率等因素容易導致電子郵件在醫學上難以推廣，是故，White 等人（2004）為克服上述問題，從 3007 封病人與醫師之間的電子郵件隨機抽取樣本，並另隨機抽取 10%的電子郵件作為對照組，經過分析，最常見的郵件內容包括醫師的資訊更新（41.4%）、處方更新（24.2%）、健康問題（13.2%）、對測試結果的問題（10.9%）、推薦（8.8%）、“其他”（包括致感謝信，道歉）（8.8%）、任用（5.4%）、請求非健康有關的信息（4.8%）和計費問題（0.3%），在這之中，只有 43%的郵件需要醫師的回覆，故得知，要推廣電子郵件系統以作出適當的初級醫療保健，患者們必須遵守以下幾點準則：(1)內容重點、(2)限制每封郵件請求的數量、以及(3)避免緊急要求或高度敏感的內容以達到電子郵件的有效利用。而 Cornwall 等人（2008）乃探討電子郵件在專科護士與肺癌病人及其家屬之間的溝通效果，對使用電子郵件為溝通方式的兩位專科護士以及 16 位病人與家屬進行問卷調查，資料的收集來自下列三種來源：(1)電子郵件在專科護士與病人/家屬成員之間的接觸、(2)病人/家屬成員問卷、(3)專科護士們在小組/檢討會議上的焦點，待收集結束後，該研究乃將資料作量化之處理，整理出具敘述性的彙整統計；而實驗結果顯現，對病人及其家屬而言，此種通訊方式較以往來的更快速、簡易，並對其反應速度感到滿意，另外專科護士們也對此種通訊方式給予正面的肯定。

最後為能探討電子郵件於醫療領域之接受程度，Hobbs 等人（2003）乃評估電子郵件於整合傳遞系統中，醫生與病人間之使用情形，以及促進此通訊形式的發展方向。該研究使用問卷調查 94 位初級醫療照顧之醫師，問卷內容乃著重於醫師們在平日中，電子郵件所扮演的角色，以及與病人通訊之間所遇到的障礙；根據評估的結果，該研究認為透過充足的事前篩選、分流以及償還機制的話，將會有效的提升醫師與病人之間以電子郵件做為通訊形式的發展。

(B) 教育領域

由於電子郵件上盛行，電子郵件已用來學生與教師之間教育與溝通上之應用。Hassini（2006）認為電子郵件清單可提供學生與教師之間一個良好的溝通平台，亦即可作為教學內容作之補救措施；如一個介紹操作的研究課程中，當中包括一個電子郵件訊息交換的研究案例，該案例說明如何藉由電子郵件的策略性應用來為溝通提供一個額外媒介，引導出更豐富之學習經驗，以及提供一個可用來改善未來課程編排上的反饋數據庫。最後根據實際在班級上應用的結果，明顯增加了學生與教師之間的互動性，學生們

也利用此平台對教師們的教學方式給予了大量的評價。

此外，Hu 等人（2009）藉由調查 2998 位具備 24 至 48 個月教學經驗之教師，試圖解析學校教師們對於電子郵件的使用模式。該研究之調查結果顯示以下兩種情形：(1)當教師們任教於較高等級的學校之時，教師們主要將電子郵件使用在與同事的溝通上面、(2)當教師們任教於較低等級的學校之時（特別是小學教師），則主要用於與家長們的溝通中；此結果可應用於討論教師們的職前教育以及專業發展中。有鑑於學生們在課堂上，經常因為準備不足以至於無法參與課堂上的活動與討論，使得學生們無法順利完成讀書或其他的工作，對學生們主動參與學習和討論的動力有著負面的影響；是故，Couzenza（2009）提出以電子郵件為媒介，將相關資料層次性地提供給學生，以協助學生完成作業；該研究結果顯示，學生們參與課堂上的活動與討論情形有大幅度的改善。

(C) 其他領域

使用者應用領域方面，除了醫療與教育領域外亦有其他各式領域，如客户行銷、犯罪調查等領域，如 Merisavo 和 Raulas（2004）考核使用行銷式電子郵件以維持與加強品牌忠誠度之成效，以及消費者族群對於郵件內容之評價。該研究之資料來自 890 位國際品牌化妝品，以及定時接收到行銷部門寄發之電子郵件訊息消費者。研究結果顯示，定期發送行銷式電子郵件對於維持與加強品牌的忠誠度具正面之影響，且消費者也樂意的將這些電子郵件內容推薦給朋友們，藉由此研究結果可有效激勵行銷部門透過電子郵件來與顧客維持頻繁接觸，以加強其對於品牌的忠誠性。

企業常以電子郵件傳遞方式取代業務流程，但部分業務流程皆含有隱藏或不明確執行流程，導致業務處理花費更多時間，因此為能訂立明確業務流程，Stuit 和 Wortmann（2012）建立互動式電子郵件探勘模式（E-mail Interaction Mining Method；EIM），並透過擷取電子郵件中寄件者、時間、郵件類型（如：轉寄、回覆、副本等）等資訊，以建立業務資料傳遞流程，接著乃根據寄件者傳遞電子郵件動向，以確立寄件者於業務流程中角色（如：決策者、執行者等），最後彙整業務角色與資料傳遞流程，則可歸納出從決策到執行之線性業務流程。該研究再以荷蘭天然氣運輸公司之電子郵件進行推論業務流程，且證實推論流程與既有流程相符，故建議企業可透過互動式電子郵件探勘模式作為業務流程分析輔助工具，以助於企業有效修正過於冗長業務流程。

於最後，現今社會因網路普及而出現「網路霸凌」等相關問題，其中「網路霸凌」乃指施暴者利用電腦網路上傳影片、手機簡訊、網路分享或電子郵件黑函等方式，將對

受暴者不利訊息快速蔓延，使受暴者遭受歧視、恥笑等遭遇，Baruch（2005）調查在大型國際公司中，使用電子郵件進行的網路霸凌對工作所造成的影響。該研究結果顯示受到網路霸凌影響的員工，在工作的滿意度與效率上有著負面的影響，並帶有強烈的焦慮感、意圖離開組織之行為，是故根據這些結果可得知，一個在心理上受到威脅與衝擊以及資訊的誤用，將會對個人或組織帶來難以想像的傷害。

綜合上述，針對使用者郵件應用領域中，凡舉「醫療領域」、「教育領域」或「其他領域」中皆有相關研究針對使用者應用方式進行探討（如表 2.6 所示）。

表 2.6、使用者郵件應用領域分析文獻彙整表

郵件應用領域	相關研究	研究議題	各領域應用方式與研究結果
醫療領域	Bhor 和 Mason（2006）	分析電子郵件無法盛行於醫療體系之因素。	以擴散理論分析並發現醫療領域無法盛行電子郵件，因未明確定立透過電子郵件傳遞病歷資料之程序，導致醫療人員為避免錯誤治療而拒絕使用。
	White 等人（2004）	推廣電子郵件系統之成功要素。	參加調查醫師期望病患以電子郵件系統進行醫療行為需制定使用規範，且必須遵守手用規範。
	Cornwall 等人（2008）	醫療體系對電子郵件應用滿意度分析。	針對已使用電子郵件為溝通方式的護士及病人與家屬進行問卷調查發現，此種通訊方式快速、簡易，且受測者皆給予正面的肯定。
	Hobbs 等人（2003）	探討電子郵件於醫療領域之接受程度。	受測醫師認為透過充足的事前篩選、分流以及償還機制，將有效的提升電子郵件接受程度。
教育領域	Hassini（2006）	電子郵件對教師教學影響。	認為電子郵件清單可提供學生與教師之間一個良好的溝通平台，亦即可作為教學內容作之補救措施。
	Hu 等人（2009）	教師應用電子郵件方式。	高年級教師主要應用於與學校同事間溝通，低年級教師主要使用於與學生家長溝通。
	Couzenza（2009）	電子郵件對學生學習影響。	以電子郵件為課堂資料傳遞媒介時，學生們參與課堂上的活動與討論情形有大幅度的改善。
其他領域	Merisavo 和 Raulas（2004）	行銷式電子郵件以維持與加強品牌忠誠度之成效。	定期發送行銷式電子郵件對於維持與加強品牌的忠誠度具正面之影響，且消費者也樂意的將這些電子郵件內容推薦給朋友們。
	Stuit 和 Wortmann（2012）	以電子郵件探勘簡化業務流程	根據寄件者傳遞電子郵件動向，以確立寄件者於業務流程中角色，並可歸納線性業務流程。
	Baruch（2005）	電子郵件的網路霸凌對工作所造成的影響	調查發現，受到網路霸凌影響的員工，在工作的滿意度與效率上有著負面的影響，並帶有強烈的焦慮感，意圖離開組織之行為。

2.4 小結

本研究所建立「個人化郵件類別推論模式」，主要目的乃分析使用者私人郵件，並為使用者自動建立專屬郵件類別。本研究乃針對「郵件探勘技術」與「郵件使用者行為探討」等兩大議題進行比較，並歸納本研究與過往研究之差異，如表2.7、表2.8所示。

於2.2節「郵件探勘技術」議題中得知，過去研究主要著重於郵件特徵分析技術，其主要解析特徵乃包含使用者郵件內容文字、郵件聯絡人關聯等郵件資料，進而提高郵件

分類或過濾之技術之績效。但過去多數研究之郵件解析技術發展乃著重於公共郵件解析（如：企業組織郵件、客户投訴信件等），故本研究乃針對私人郵件發展郵件內容解析法則，並擷取私人郵件中語意詞彙建立郵件內容特徵。

另外，於2.3節「郵件使用者行為探討」之相關文獻中，乃針對使用者個人管理郵件方式進行探討，且多數研究透過訪談使用者發現，現今多數電子郵件系統無法針對使用者個人自動產生專屬郵件類別，亦部分研究以郵件標籤方式呈現郵件類別。本研究乃著重於使用者個人化郵件類別建立，透過使用者個人專業知識或生活經驗量身制定所屬個人化郵件類別，達到郵件類別個人化與自動化建立。

綜上所述，本研究考量私人郵件內容，因個人知識領不同導致郵件內容涉及領域廣泛等因素。因此，本研究所建立「個人化郵件類別推論模式」乃透過語意方式解析郵件內容擷取郵件內容所含特徵，並透過郵件內容特徵區分郵件所屬類型，此外，為能提供使用者明確且具解釋性郵件類別，本研究乃建立郵件類別名稱詞彙庫，並以郵件內容特徵進行語意關聯建立，以取得各郵件類別所屬代表性名稱，形成郵件類別自動化產生效果，並達到為使用者量身制定專屬郵件類別之效果。

表 2.7、個人化郵件類別推論模式與過去文獻差異彙整表(1)

比較研究	解析郵件類型	自動建立郵件類別	個人化郵件類別	郵件分群	解析郵件詞彙語意	郵件敘述主題分析	郵件標籤	郵件過濾	技術開發	使用者觀點調查	優勢
本研究	私人郵件	●	●	●	●	●			●		推論使用者專屬類別
Yu與Zhu（2009）	企業組織郵件				●				●		提高分類準確率
Buffett和Geng（2010）	企業組織郵件					●	●		●		透過郵件分析日誌標籤
Scheffer（2004）、Sung及Chih（2004）、Coussement和Poel（2008）	客户郵件				●				●		以語意解析提高準確性
Bouguila與Amayri（2009）、Zhou等人（2007）	一般郵件				●				●		結合圖片解析技術
Poon及Chang（2003）、Islam等人（2009）	企業組織郵件				●				●		多階段式分類，以減少失誤發生
Hadjidj 等人（2009）、Iqbal 等人（2008）、Iqbal 等人（2010）	匿名犯罪郵件				●	●			●		寄件者寫作方式
Sakurai 和 Suyama（2005）	客户郵件				●	●			●		以模糊理論預測郵件類型
Zajic 等人（2008）	私人郵件				●	●			●		串聯使用者郵件內容
Schuff 等人（2006）、Sun 和 Dong（2010）	私人郵件			●	●		●		●		建立多種郵件標籤供使用者選擇
Alberts 和 Forest（2012）	企業組織郵件			●	●		●		●		分析員工使用習慣
Mao等人（2011）	私人郵件			●			●		●		分析寄件者社群行為
Ugˇuz（2011）	一般郵件				●				●		以語意解析提高準確性
Yang（2009）	一般郵件			●					●		以寄件者分群
Chundi等人（2009）	企業組織郵件							●	●		過濾垃圾郵件
Okolica等人（2007）	企業組織郵件			●					●		預測內部間諜
Gomez 和 Moens（2012）、Laorden等人（2011）、	廣告郵件、詐騙郵件				●			●	●		以語意解析提高準確性
Duan 等人（2007）、Salcedo-Campos 等人（2012）、Mohammad 和 Zitar（2011）、Herzberg（2009）	廣告郵件、詐騙郵件							●	●		過濾垃圾郵件技術之提升

表 2.8、個人化郵件類別推論模式與過去文獻差異彙整表(2)

比較研究	解析郵件類型	自動建立郵件類別	個人化郵件類別	郵件分群	解析郵件詞彙語意	郵件敘述主題分析	郵件標籤	郵件過濾	技術開發	使用者觀點調查	優勢
本研究	私人郵件	●	●	●	●	●			●		推論使用者專屬類別
Zorkadis等人（2005）	廣告郵件、詐騙郵件				●				●		以語意解析提升郵件解析技術之準確性
Ying等人（2010）、Yang等人（2011）、Yu和Xu（2008）	廣告郵件、詐騙郵件							●			比較過濾垃圾郵件技術之差異
Li和Huang等人（2012）、Wei等人（2008）、Marsono等人（2009）、Çıltık和Güngör（2008）	廣告郵件、詐騙郵件				●			●	●		以語意解析提升過濾垃圾郵件技術
Irena等人（2007）	一般郵件							●			比較過濾垃圾郵件技術之差異
Crawford等人（2004）	一般郵件				●						比較郵件分類技術之差異
Laura及Maria（2001）、Phan（2008）	一般郵件							●	●		保有使用者信件隱私
Guzella等人（2008）、Appavu等人（2009）	廣告郵件、詐騙郵件							●	●		過濾垃圾郵件技術之提升
Cai等人（2007）	校內公告郵件				●						比較郵件分類技術之差異
Szóstek（2011）、Mano和Mesch（2010）、Sumecki等人（2011）	企業組織郵件、私人郵件									●	探討使用者管理郵件需求
Soucek和Moser（2010）、Huang等人（2011）	企業組織郵件									●	探討郵件對使用者影響

第三章、個人化郵件類別推論模式

本研究所提之「個人化郵件類別推論模式」乃以郵件中所包含之郵件文字內容為基礎，先行解析郵件中詞彙語意，並歸納出郵件語意詞彙集合，接著再以主成份分析擷取語意詞彙集合中郵件具代表性之郵件特徵詞彙，並根據郵件特徵詞彙分析郵件之相似性，並根據郵件相似性以「階層式分群法」（Hierarchical Clustering）推論郵件群集樹，再以此樹狀結構之郵件群集進行類別名稱推論。接著，先行以網頁新聞文件作為基礎資料，並擷取並彙整網頁新聞文件詞彙為類別名稱詞彙庫，再以類別名稱詞彙庫為名稱挑選依據，並與各郵件群集內郵件特徵詞彙建立詞彙關聯，再根據詞彙關聯最高關聯性之名稱詞彙作為類別名稱，即達到郵件群集命名效果（如圖 3.1 所示）。因此，本研究之主要流程可分為兩大部份，分別為「語意擷取郵件關鍵字模組」與「個人化郵件類別推論模組」，如下說明。

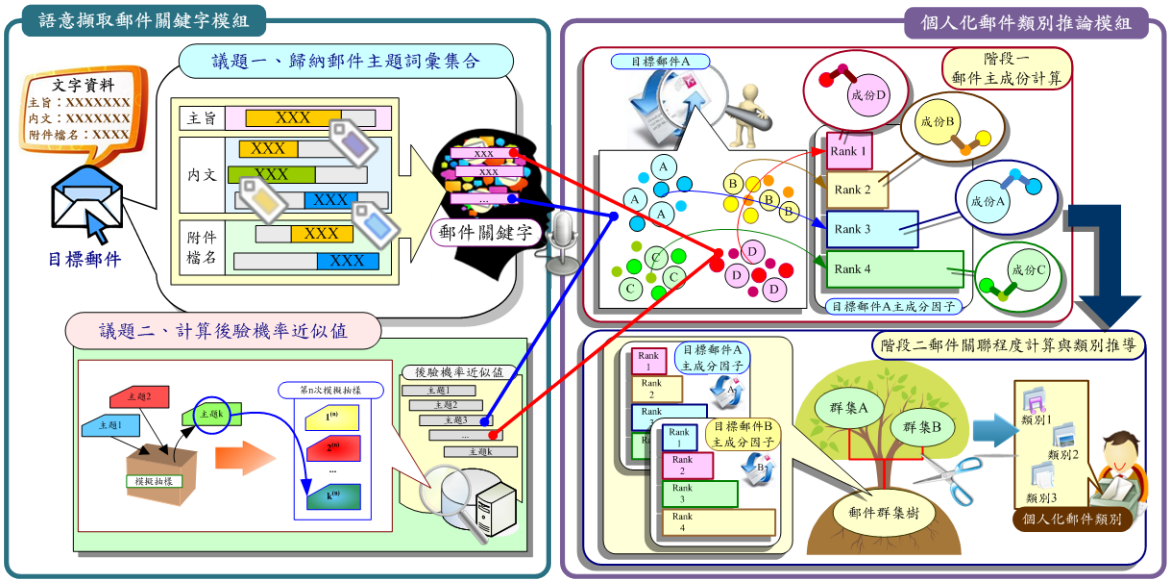

圖 3.1、個人化郵件類別推論模式架構圖

3.1 語意擷取郵件關鍵字

個人郵件意指郵件內容並非制式化之正式郵件，因此郵件所敘述內容較屬於私人對談，且內容、主題較為發散，其敘述涵蓋較多類型，但過去研究於擷取郵件關鍵字過程中，僅考量字詞單一含意，並沒有考量字詞與字詞串連後所具有另一種涵義，而無法正確保留郵件所傳達眞正內容。因此本研究欲以語意技術方式擷取郵件關鍵字，其中以 Asuncion 等人（2011）所提之 LDA 模式（Latent Dirichlet Allocation）進行擷取關鍵字任務。該模式主要著重於文件摘要建立，當中乃以語意方式擷取字詞，以達文章摘要建

立之目的，本研究乃以個人郵件分類為主要目的，並非郵件摘要之建立，因此本研究並不考量 LDA 模式中推論文摘法則部份，僅以模式中主題詞彙建立方技術為基礎，發展本研究之關鍵字擷取法則。由於 LDA 模式所形成之主題詞彙可串聯成文件之摘要，而摘要具有保留文章中最主要含意與重點詞彙，故相較於以往擷取方式，透過此法所取得之關鍵字，因考量字詞與字詞間含意性，較能保留郵件中文句涵義，故本研究乃以 LDA 模式為基礎，針對個人郵件內容進行語意關鍵字擷取，方可獲得較具郵件中具含意之關鍵字詞，以進行後續個人郵件之分類。

於 LDA 模式中，摘要是以主題詞彙所組成，因此為取得能建構出具郵件涵義之詞彙，需將各筆詞彙進行重要程度計算，而模式中則以計算詞彙機率作為判斷依據劃分出主題詞彙，當中詞彙機率計算時需取得詞彙之後驗機率（即為透過證實之發生機率）輔以計算，但本研究所分析之郵件內容皆為未知情況下，使後驗機率無法直接取得；有鑑於此，LDA 模式則分別透過各式演算法進行計算取得後驗機率之近似值，以計算主題詞彙之機率；由此得知除進行主題詞彙機率取得之法則外，亦須進行詞彙後驗機率近似值之取得，故本研究將分為兩項議題分別進行推論，兩項議題分別為：

- 議題一、歸納郵件主題詞彙集合
- 議題二、計算後驗機率近似值

透過「議題一、歸納郵件主題詞彙集合」計算詞彙之重要性，再以「議題二、計算後驗機率近似值」法則取得詞彙後驗機率近似值，並給予議題一計算，以獲取主題詞彙，如圖 3.2 所示，當中，亦可透過議題二中所獲取之近似值觀察推論正確關聯性，以了解模式推論績效與參數調整方式。

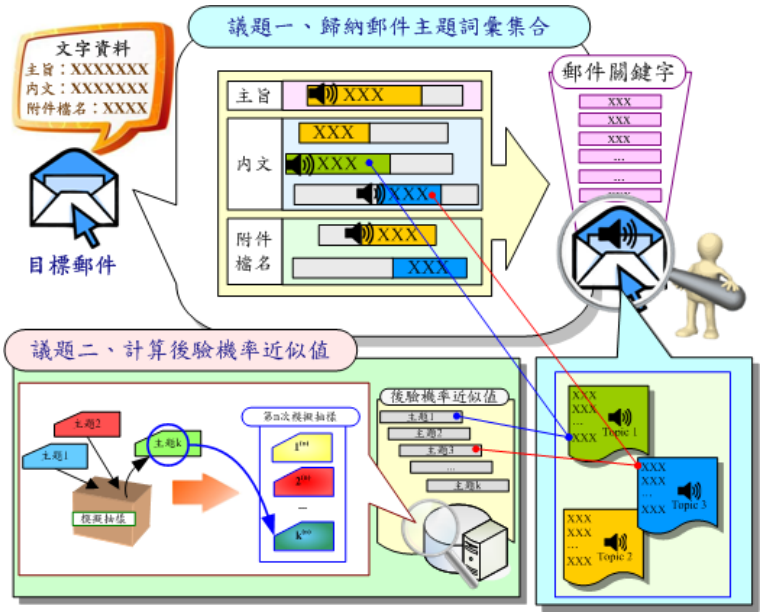

圖 3.2、語意擷取郵件關鍵字示意圖

3.1.1 議題一、歸納郵件主題詞彙集合

本研究於「語意擷取郵件關鍵字」中乃以 LDA 模式為基礎，進行個人郵件關鍵字擷取，以取得個人郵件敘述中較具代表性關鍵字。其中，LDA 模式乃以假設文件包含 k 個主題，當中主題與各類別皆無關聯，接著計算主題機率與字詞機率關係，產生主題詞彙集，並串聯成文件摘要。於此模式中，分別需 α、β 兩參數作為主要計算，其中 α 參數代表文件中 k 個主題機率於狄氏分配（Dirichlet Distribution）中所需先驗機率（即未考量任何情況下主題所發生機率），而 β 參數則代表各字詞可能屬於各主題之機率（矩陣參數），當兩項參數生成後得以計算出文件邊際分佈，取得各字詞所對應主題之分佈。但其計算過程中 α、β 兩項參數之計算無法直接取得，因此，本研究乃整合 Asuncion 等人（2011）所提之 LDA 模式乃以吉布斯抽樣（Gibbs Sampling）方式取得 α、β 兩項參數之近似值，以計算文件之邊際機率分佈。

是故，於議題一法則中（如圖 3.3 所示），需先行去除郵件內容中非關鍵字集，以成為重要詞彙（即為非關鍵字集外，並未被歸類之字集，本研究乃將此類字集定義為「重要詞彙字集」），接著，決定 k 個主題機率 θ，並以主題先驗機率 α（即未考量任何情況下主題所發生機率）作為狄氏分配之機率參數，於本法則中 k 代表郵件類別數量，接著，選取重要詞彙，計算該詞彙屬於各主題之機率，反覆計算重要詞彙之機率，形成 β 矩陣參數，接著計算 α、β 兩項參數間關聯性；然而，由於 α、β 兩項參數兩參數無法以直接計算方式取得關聯性，於計算中須透過詞彙後驗機率（經證實後所獲得事件發生機率）輔以計算，但因郵件中詞彙與主題之後驗機率皆為未知情況，因此利用吉布斯抽樣方式

取得詞彙後驗機率之近似值，並計算後續兩項參數間關聯性，進而取得郵件中較具代表性詞彙，而吉布斯抽樣取得近似值之法則於議題二中進行探討。於說明議題一之模式推導前，先行定義議題一所使用之符號。

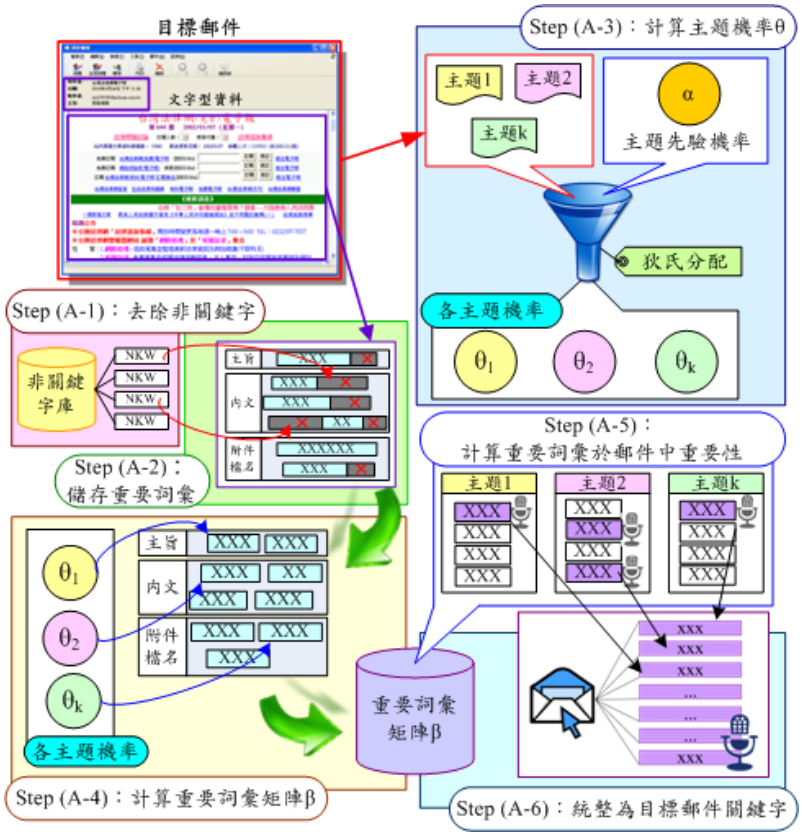

圖 3.3、歸納郵件主題詞彙集合示意圖

符號定義

UI	電子郵件所包含之文字型資料集合
UI_i	電子郵件中第 *i* 項文字型資料，i=1,2,3；UI_1 為「郵件主旨」、UI_2 為「郵件內文」及 UI_3 為「附加檔名」
NKW	非關鍵字集
FUI_i	第 i 項重要文字型資料（即經 NKW 篩選過之第 i 項文字型資料）
$IW_{i,j}$	第 i 項重要文字型資料中第 j 筆重要詞彙
$N(IW_i)$	第 i 項重要文字型資料中重要詞彙之數量
$N(IW_{i,j})$	第 i 項重要文字型資料中第 j 筆重要詞彙於郵件中出現次數
$P(IW_{i,j})$	第 i 項重要文字型資料中第 j 筆重要詞彙於郵件中出現機率
$P(IW_{i,j}\|Z_k)$	於第 k 個主題發生條件下第 i 項重要文字型資料中第 j 筆重要詞彙於郵件中發生機率

$P(Z_k \mid IW_{i,j})$	在第 i 項重要文字型資料中第 j 筆重要詞彙屬發生條件下第 k 個主題之後驗機率發生機率，亦作為計算主題詞彙重要性所需之後驗機率（即為經證實後所獲得事件發生機率）
$I(IW_{i,j} \mid Z_k)$	第 i 項重要文字型資料中第 j 筆重要詞彙於第 k 個主題對郵件重要性，即為重要詞彙於主題集合中重要性，亦表示詞彙於郵件中重要含意程度
N(ST)	郵件類別數量
Z_k	郵件中第 k 個主題
$ZUW_{i,k}$	第 i 項重要文字型資料中第 k 個主題詞彙集合
$ZIW_{i,k,n}$	第 i 項重要文字型資料中第 k 個主題詞彙集合中第 n 筆主題詞彙
θ	主題出現機率之集合
θ_k	第 k 個主題於郵件中出現機率，為狄氏分配之隨機變數
α	主題先驗機率（即未考量任何情況下主題所發生機率）之集合
α_k	郵件中第 k 個主題之先驗機率
β_i	第 i 項重要文字型資料之重要詞彙機率矩陣
$I(\beta_i)$	第 i 項重要文字型資料之重要詞彙於郵件中重要性矩陣
$P(\theta \mid \alpha)$	郵件中所有主題於狄氏分配下發生機率

Step (A-1)：篩選非關鍵字集

為能於電子郵件文字型資料中獲得重要詞彙，以推論關鍵字。本模式於蒐集郵件內容中所包含之各項郵件文字型資料（UI_i，當中 i=1,2,3；UI_1 為「郵件主旨」、UI_2 為「郵件內文」及 UI_3 為「附檔名」）後，先以各式標點符號為斷據點進行斷句，並經由非關鍵字集（NKW）篩選，去除各項郵件文字型資料之非關鍵字，即成為各項重要郵件文字型資料集合（FUI_i；即經 NKW 篩選過之郵件內文集合），如公式(3.1)所示。

$$FUI_i=UI_i\text{-}NKW \quad \text{當中 } i=1,2,3 \tag{3.1}$$

Step (A-2)：儲存重要詞彙

本步驟乃將各項重要郵件文字型資料（FUI_i），利用二至六字詞之解析方式將其拆解成若干字詞（即 $FUI_i=\{IW_{i,1}, IW_{i,2},\ldots, IW_{i,j},\ldots\}$），以視為各項重要郵件文字型資料之重要詞彙（$IW_{i,j}$），彙整如表 3.1。

表 3.1、重要詞彙統整表

	字詞 1	字詞 2	…	字詞 j	…
FUI_1	$IW_{1,1}$	$IW_{1,2}$	…	$IW_{1,j}$	…
FUI_2	$IW_{2,1}$	…	…	…	…
…	…	…	…	…	…
FUI_i	$IW_{i,1}$	…	…	$IW_{i,j}$	…

Step (A-3)：計算主題機率

各詞彙與詞彙間具有一定關聯性，於 LDA 模式中，將互為關聯詞彙劃分為一個群集，而此群集則成為主題詞彙群集，當中，分群方式乃以隨機假設各主題出現機率，並符合狄氏分配之隨機變數以作為後續詞彙重要性計算之依據，接著，以此主題出現機率進行詞彙分群，形成主題詞彙群集。因此，本步驟乃先行隨機設定第 k 個主題於郵件中出現機率 θ_k（其中主題數量即為郵件類別數量 N(ST)），且符合狄氏分配之隨機變數，並以郵件中各主題之先驗機率 α_k（即未考量任何情況下主題所發生機率）作為狄氏分配之機率參數，以獲得主題於郵件中出現機率 θ_k，如公式(3.1)所示。

$$\begin{aligned}&\theta=\{\theta_1,\theta_2,\theta_3,...,\theta_k,...\}\\&\alpha=\{\alpha_1,\alpha_2,\alpha_3,...,\alpha_k,...\}\\&\theta\sim \text{Dirichlet}(\alpha)\quad \text{and}\,\theta_k>0\quad \text{and}\sum_{\text{all k}}\theta_k=1\end{aligned}\tag{3.1}$$

為計算後續詞彙於分群後，主題詞彙群集於郵件中重要性，故本步驟計算郵件中主題機率集合 θ 於郵件中發生機率 $P(\theta|\alpha)$，如公式(3.2)所示，作為修正主題詞彙於郵件中重要性。

$$P(\theta|\alpha)=\frac{\Gamma(\sum_{all\ \text{k}}\alpha_k)\times\prod_{all\ \text{k}}\theta_k^{\alpha_{k-1}}}{\prod_{all\ \text{k}}\Gamma(\alpha_k)}\tag{3.2}$$

Step (A-4)：計算重要詞彙與各主題間條件機率

於 LDA 模式中，各詞彙皆與各主題間互為關聯性，當中，乃以詞彙於文件中出現機率，並輔以主題出現機率計算，以獲得詞彙與各主題間條件機率（即為文件中包含此主題情況下，詞彙所出現機率），成為主題詞彙機率以計算各詞彙於文件中重要性。因此，本步驟乃以取得重要詞彙於郵件中發生機率，首先，針對每筆重要詞彙 $IW_{i,j}$ 由 k 個主題中選定主題 Z_k，且為多項式分配（Multinomial Distribution）之隨機變數，並以上一步驟中所得郵件中主題機率集合 θ 作為多項式分配之機率參數，進而將重要詞彙選定主題 Z_k 賦予主題機率 θ_k，如公式(3.3)所示，接著，計算重要詞彙 $IW_{i,j}$ 於主題 Z_k 為條件下之條件機率 $P(IW_{i,j}|Z_k)$，其中重要詞彙之機率 $P(IW_{i,j})$，乃計算重要詞彙於郵件中出現次數 $N(IW_{i,j})$與第 i 項重要文字型資料中重要詞彙之數量 $N(IW_i)$之比例，如公式(3.4)所示。當中，詞彙之後驗機率 $P(Z_k|IW_{i,j})$ 由於無法直接取得進行計算，因此於議題二中探討其近似值作為計算依據。

$$Z_k \sim \text{Multinomial}(\theta) \tag{3.3}$$

$$P(IW_{i,j}) = \frac{N(IW_{i,j})}{N(IW_i)}$$
$$P(IW_{i,j}|Z_k) = \frac{P(Z_k|IW_{i,j}) \times P(IW_{i,j})}{\theta_k} \tag{3.4}$$

如上述取得重要詞彙與各主題間條件機率 $P(IW_{i,j}|Z_k)$ 後，彙整為 k×j 之重要詞彙之機率矩陣 β_i，則 $P(IW_{i,j}|Z_k)$ 表示重要詞彙於第 K 個主題下所發生機率，如公式(3.5)所示，作為後續計算各詞彙於郵件中重要性之依據。

$$\beta_i = \begin{bmatrix} P(IW_{i,1}|Z_1) & P(IW_{i,1}|Z_2) & \cdots & P(IW_{i,1}|Z_k)\cdots \\ P(IW_{i,2}|Z_1) & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ P(IW_{i,j}|Z_1) & \cdots & \cdots & P(IW_{i,j}|Z_k)\cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{3.5}$$

Step (A-5)：計算重要詞彙於郵件中重要性

經由上述步驟所取得重要詞彙之矩陣 β_i，但僅表達重要詞彙與主題關聯性，未能明確表達於郵件中重要。因此本步驟為取得重要詞彙之重要程度，以郵件中主題機率集合 θ 於郵件中發生機率 $P(\theta|\alpha)$ 為依據，修正重要詞彙與各主題間條件機率 $P(IW_{i,j}|Z_k)$，以

取得重要詞彙於郵件中重要性 $I(IW_{i,j}|Z_k)$，彙整為重要詞彙重要性矩陣 $I(\beta_i)$，如公式(3.6)所示。

$$I(IW_{i,j,k}) = P(\theta|\alpha) \times P(IW_{i,j}|Z_k)$$

$$I(\beta_i) = \begin{bmatrix} I(IW_{i,1,1}) & I(IW_{i,1,2}) & \cdots & I(IW_{i,1,k}) \cdots \\ I(IW_{i,2,1}) & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ I(IW_{i,j,1}) & \cdots & \cdots & I(IW_{i,j,k}) \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \quad (3.6)$$

Step (A-6)：統整主題詞彙集合

經由上述步驟取得重要詞彙於郵件中重要性，而此重要程度則代表重要詞彙於郵件中所含語意重要性。因此，為取得重要程度較高詞彙，首先，將重要詞彙以重要性程度 $I(IW_{i,j}|Z_k)$ 為依據，並以詞彙重要性門檻值 W 進行篩選，去除重要性較低之詞彙，獲得主題詞彙集合 $ZUW_{i,k}$（如公式(3.7)所示），並成為郵件關鍵字（如表 3.2 所示）。

$$IF\ I(IW_{i,j,k}) > W\ THEN\ IW_{i,j} = ZIW_{i,k,n}\ Where\ 0 < W < 1$$
$$ZUW_{i,k} = \{ZIW_{i,k,1}, ZIW_{i,k,2}, ZIW_{i,k,3}, ..., ZIW_{i,k,n}\} \quad (3.7)$$

表 3.2、主題詞彙統整表

主題	字詞 1	字詞 2	…	字詞 n
Z_1	$ZIW_{i,1,1}$	$ZIW_{i,1,2}$	…	$ZIW_{i,1,n}$
Z_2	$ZIW_{i,2,1}$	$ZIW_{i,2,2}$	…	$ZIW_{i,2,n}$
…	…	…	…	…
Z_k	$ZIW_{i,k,1}$	$ZIW_{i,k,2}$	…	$ZIW_{i,k,n}$

由上述步驟中所得之主題詞彙則為目標郵件中較具代表性詞彙，但於計算各重要詞彙與各主題間機率無法直接計算其中後驗機率（即為透過證實之發生機率），因在於目標郵件各筆重要詞彙與主題是不可預測（由於一封未分類郵件資料被視為一封未閱讀信件）；有鑑於此，本研究係以 Asuncion 等人（2011）所提之法則，以吉布斯抽樣方式間接取得近似值，於「議題二、計算後驗機率近似值」中進行步驟詳述。

3.1.2 議題二、計算後驗機率近似值

由語意擷取郵件關鍵字法則中包含兩項重要參數，分別為：主題先驗機率 α 及重要詞彙之機率 β，透過貝氏定理方式計算以取得主題詞彙重要程度，並彙整成郵件關鍵字。但於貝氏定理計算中須計算主題與重要詞彙間之後驗機率（即為透過證實之發生機率），以進行後續主題詞彙重要程度統整。而目標郵件乃為一封未閱讀信件，因此無法從中得知重要詞彙、主題等此些資料於信件中比例，而導致無法計算後驗機率；因此於 LDA 模式中皆導入各式方法，進行計算後驗機率之近似值，而本研究係以 Asuncion 等人（2011）所提中所提之吉布斯抽樣方式間接取得近似值，以進行後續主題詞彙統整。

承上所述，於語意擷取郵件關鍵字法則中亦可察覺，LDA 模式主要乃計算主題先驗機率 α 及重要詞彙之機率 β 兩項參數，並透過兩參數相互計算後，需輔以兩參數所得之後驗機率作為主要計算依據。因此，於模式中主要影響整體模式績效之參數則為後驗機率近似值之計算，則為能了解後驗機率近似值分布狀況與郵件關鍵字擷取績效之關聯性，故須建立後驗機率近似值計算之法則。

後驗機率近似值計算乃透過吉布斯抽樣進行推論，如圖 3.4 所示，首先賦予主題於模擬抽樣前之初始值，接著進行模擬抽樣，反覆假設抽出各主題情況下機率，當各主題皆有一次抽樣機率後則完成第一次抽樣，反覆執行多次模擬抽樣則形成多組模擬抽樣機率組合，而當中模擬抽樣次數則以重要詞彙數量作為模擬次數；完成模擬抽樣後，由於前幾次之抽樣受於初始值影響，故其抽樣結果並不準確，因此篩選並去除前 m 個模擬抽樣，並將保留之結果進行計算，則取得後驗機率之近似值，並於議題一中作為郵件關鍵字計算之依據。

舉例而言，當郵件中共 n 個詞彙 k 個主題，因此進行 n 次模擬抽樣，於第一次抽樣時設定各詞彙所屬主題一個數字且介於 1 至 k 之間，其數字代詞彙所屬主題共 n 個數字成為一次抽樣集合，接著輪流計算第 n 個詞彙所屬主題發生機率，當抽樣集合中每筆詞彙所屬主題皆獲得一次機率後，則以獲得數據作為計算值計算下一階段模擬抽樣，反覆進行 n 次模擬抽樣後，去除前 M 個模擬抽樣集合，當中 m 個集合可由使用者自行設定，並將每個詞彙所得之各次模擬結果計算其平均值，則成為後驗機率近似值。

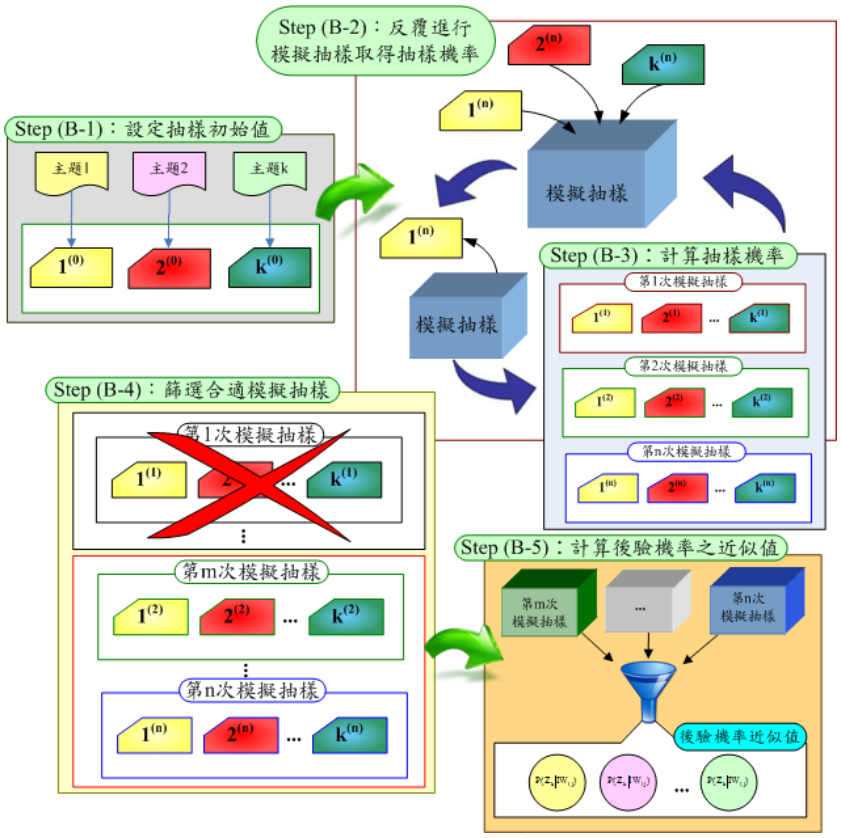

圖 3.4、後驗機率近似值計算步驟示意圖

符號定義

$IW(Z)_{i,j}$	第 i 項重要文字型資料中第 j 筆重要詞彙所屬之主題
$N(Z)$	郵件中全部主題數量
$PIW(Z)^{(n)}$	第 n 次模擬抽樣後所得機率集合
$PIW(Z)^{(0)}_{i,j}$	第 i 項重要文字型資料中第 j 筆重要詞彙所屬之主題於模擬抽樣中系統給定初始值
$PIW(Z)^{(n)}_{i,j}$	第 i 項重要文字型資料中第 j 筆重要詞彙所屬之主題於第 n 次模擬抽樣中所得機率
$N(IW_{i,j})^{Z_k}$	第 i 項重要文字型資料中第 j 筆重要詞彙被分配於第 k 主題之頻率
$N(IW_i)^{Z_k}$	第 i 項重要文字型資料中屬於第 k 主題之詞彙數量
$\gamma_{i,j}$	第 i 項重要文字型資料中第 j 筆重要詞彙之後驗機率近似值

Step (B-1)：設定主題模擬抽樣之初始值

後驗機率為一種透過證實之發生機率，而吉布斯抽樣乃以多次之抽樣模擬作為依據，並將抽樣模擬後之結果綜合計算而獲得後驗機率，因此可發現，吉布斯抽樣所得之

機率亦同於透過證實所得之機率，因此透過吉布斯法則進行後驗機率之推論。

首先，進行模擬抽樣前給予詞彙所屬之主題於模擬抽樣時所需抽樣出現次數 $PIW(Z)_{i,j}^{(0)}$，當中出現次數介於 1 至 N(Z)（郵件中全部主題數量）之整數，如公式(3.8)所示。

$$PIW(Z)_{i,j}^{(0)} \in \{|a| : a \in Z\} \quad \text{and} \quad 1 \le PIW(Z)_{i,j}^{(0)} \le N(Z) \tag{3.8}$$

Step (B-2)：反覆進行模擬抽樣取得抽樣機率

透過主題初始值之設定後，即可進行模擬抽樣。抽樣過程當中，首先針對每筆詞彙之主題 $IW(Z)_{i,j}$ 進行抽樣，計算其被抽出情況下所獲得機率，即為主題於模擬抽樣中所得機率 $PIW(Z)_{i,j}^{(n)}$（如公式(3.9)所示），完成第一次模擬後，以第一次模擬中取得之數據進行第二次模擬，如此反覆進行 n 次抽樣（當中 n 以重要詞彙數量 $N(IW_i)$為模擬次數），則獲得結果如公式(3.10)所示，取得各筆詞彙主題於各次模擬後所得之抽樣機率 $PIW(Z)_{i,j}^{(n)}$。

$$\begin{aligned}
PIW(Z)_{i,1}^{(1)} &= p_1(IW(Z)_{i,1} \,|\, PIW(Z)_{i,2}^{(0)}, PIW(Z)_{i,3}^{(0)}, ..., PIW(Z)_{i,j}^{(0)}) \\
PIW(Z)_{i,2}^{(1)} &= p_2(IW(Z)_{i,2} \,|\, PIW(Z)_{i,1}^{(1)}, PIW(Z)_{i,3}^{(0)}, ..., PIW(Z)_{i,j}^{(0)}) \\
&\vdots \\
PIW(Z)_{i,j}^{(1)} &= p_j(IW(Z)_{i,j} \,|\, PIW(Z)_{i,1}^{(1)}, PIW(Z)_{i,2}^{(1)}, ..., PIW(Z)_{i,j-1}^{(1)})
\end{aligned} \tag{3.9}$$

$$\begin{aligned}
PIW(Z)^{(1)} &= \{PIW(Z)_{i,1}^{(1)}, PIW(Z)_{i,2}^{(1)}, ..., PIW(Z)_{i,j}^{(1)}\} \\
&\vdots \\
PIW(Z)^{(n)} &= \{PIW(Z)_{i,1}^{(n)}, PIW(Z)_{i,2}^{(n)}, ..., PIW(Z)_{i,j}^{(n)}\}
\end{aligned} \tag{3.10}$$

Step (B-3)：計算模擬抽樣之抽樣機率

承上述步驟，透過假設性之模擬抽樣取得須透過證實之機率，其中，模擬抽樣所得之抽樣機率 $PIW(Z)_{i,j}^{(n)}$ 計算方式，乃透過模擬抽樣集合 $PIW(Z)^{(n)}$ 中，與第 j 筆重要詞彙相同且分配於同一主題之詞彙數量 $N(IW_{i,j})^{Z_k}$，以及第 i 項文字資料中所含之主題數量 $N(IW_i)^{Z_k}$ 進行計算，如公式(3.11)所示。

$$PIW(Z)_{i,j}^{(n)} = \frac{N(IW_{i,j})^{Z_k} + \beta}{N(IW_i)^{Z_k} + N(IW_i) \times \beta} \times \frac{N(IW_i)^{Z_k} + \alpha}{N(IW_i) + N(Z) \times \alpha} \tag{3.11}$$

Step (B-4)：篩選合適模擬抽樣

由於模擬抽樣中初始值為自行設定，因此於後續抽樣中，前次抽樣接受初始值之影響，於主題詞彙中將影響所得結果。有鑑於此，為取得不受初始值影響之機率數值，由 n 次抽樣中，以後驗機率篩選條件捨去前 m 組抽樣，保留 m 至 n 組（即為 $N(PIW_i)$組模擬抽樣）進行估計。

Step (B-5)：計算後驗機率之近似值

由上述步驟中取得 m 至 n 組之抽樣機率 $PIW(Z)_{i,j}^{(n)}$ 後，為使其成為議題一中計算主題與詞彙之條件機率所需後驗機率 $P(Z_k|IW_{i,j})$，因此本研究以吉布斯抽樣模擬，取得經證實之發生機率，並根據吉布斯抽樣之法則計算邊際計機率，使其成為後驗機率 $P(Z_k|IW_{i,j})$ 之近似值 $\gamma_{i,j}$，如公式(3.12)所示，其中 $N(PIW_i)$代表捨去前 m 組抽樣後數量。

$$\gamma_{i,j} \approx P(Z_k|IW_{i,j})$$
$$\gamma_{i,j} = \frac{1}{N(PIW_i)} \times \sum_{n=m+1}^{N(PIW_i)} PIW(Z)_{i,j}^{(n)} \quad (3.12)$$

所得近似值彙整如表 3.3，並將近似值 $\gamma_{i,j}$ 帶回議題一中「Step (A-4)：計算重要詞彙與各主題間條件機率」，成為該步驟所需之後驗機率 $P(Z_k|IW_{i,j})$，進而完成語意擷取郵件關鍵字之法則。

表 3.3、後驗機率近似值彙統整表

	$IW_{i,1}$	$IW_{i,2}$	…	$IW_{i,j}$
$P(Z_1\|IW_{i,j})$	$\gamma_{i,1}$	$\gamma_{i,2}$	…	$\gamma_{i,j}$
$P(Z_2\|IW_{i,j})$	$\gamma_{i,1}$	$\gamma_{i,2}$	…	$\gamma_{i,j}$
…	…	…	…	…
$P(Z_k\|IW_{i,j})$	$\gamma_{i,1}$	$\gamma_{i,2}$	…	$\gamma_{i,j}$

後驗機率之定義乃為事件發生後所證實之機率，但此後驗機率於未閱讀之電子郵件中無法直接取得，透過吉布斯抽樣之抽樣法則進行模擬抽樣，進而取得經與後驗機率相近之機率值，但此抽樣於模擬前需給予假設，因此為避免假設值之影響，造成近似值誤差，於抽樣模擬結束後將前 m 次之抽樣去除不予以計算，並將保留之模擬數據計算平均值，則此數據即為一項經證實之發生機率（因此機率乃透過模擬抽樣所獲得），並於「議題一、歸納郵件主題詞彙集合」中進行計算取得具含義之郵件關鍵字。

3.2 個人化郵件類別推論

個人化郵件意指私人通訊信件，因此個人化郵件具有幾項特性，包含無明確或顯著性之類別，且每位使用者所欲使用之郵件類別、接收郵件性質皆不相同，當中，Jonathan（1999）發現商業信件（即以企業名義傳遞之信件）之信件寫作格式較為制式，且多數商業信件皆為企業對員工通知與公告，於個人化郵件中，Zajic（2008）指出電子郵件不同於新聞報社所轉寫文章，需具有既定寫作格式與文章結構，且文章用詞並非以一般大眾所認知用語進行撰寫，無須考量文章結構與文章用詞，亦可能含有錯誤用字。因此，本研究乃彙整此兩類信件之比較，如表 3.9 所示。

表 3.4、個人化郵件與商業信件差異

	商業信件	個人化郵件
寄件者	以企業、組織為名義	寄件人私人名義
撰寫格式	制定化	無固定格式
對象	企業全體員工	朋友、工作同事、家人等
信件類型	企業內部資訊通知、組織資訊傳遞等	私人對話
信件內容	統一制式化內容	因收件者為不同人，故每封信件類型皆不相同
收件者單一性	通常單封信件傳遞給多位員工	單封信件通常傳遞給單一特定收件人
信件回覆	無回覆信件	有回覆信件，且多為互動式對話內容

由表 3.9 中發現，若以個人名義所傳遞之電子郵件，因郵件內容為使用者與寄件者間私人對話，故不同使用者其郵件內容類型亦不相同，因此郵件類別劃分方式也因使用者而不盡相同，導致無法明確定立郵件類別，有鑑於此，個人化郵件類別因使用者身分、收件對象等個人因素而改變分類類別數量或類型，是故本研究乃發展「個人化郵件類別推論」模組以解決個人化郵件類別變動性之特質。

承上所述，本研究乃以「語意擷取郵件關鍵字」所擷取之郵件關鍵字乃將使用者原有信件進行分群，並將各郵件群集視為不同屬性之郵件類別，而此類別推論結果乃根據使用者之信件進行分析，故根據使用者個人習慣推論類別數量與郵件類型。故本法則欲將「語意擷取郵件關鍵字」模組所擷取之郵件關鍵字，以主成分分析方式將關鍵字重要程度精確化，則可明顯表達郵件中每筆關鍵字之關聯性，並將這些關鍵字成為郵件分群之分群依據。因此，此模組有兩項重要議題，「階段一、郵件主成份計算」、「階段二、

郵件關聯程度計算與類別推論」(如圖 3.5 所示);其中,透過「階段一、郵件主成份計算」取得郵件與郵件間關聯程度,且擷取郵件中具明確關聯性之關鍵字,並將此關鍵字進行「階段二、郵件關聯程度計算與類別推論」推論出各郵件類別。

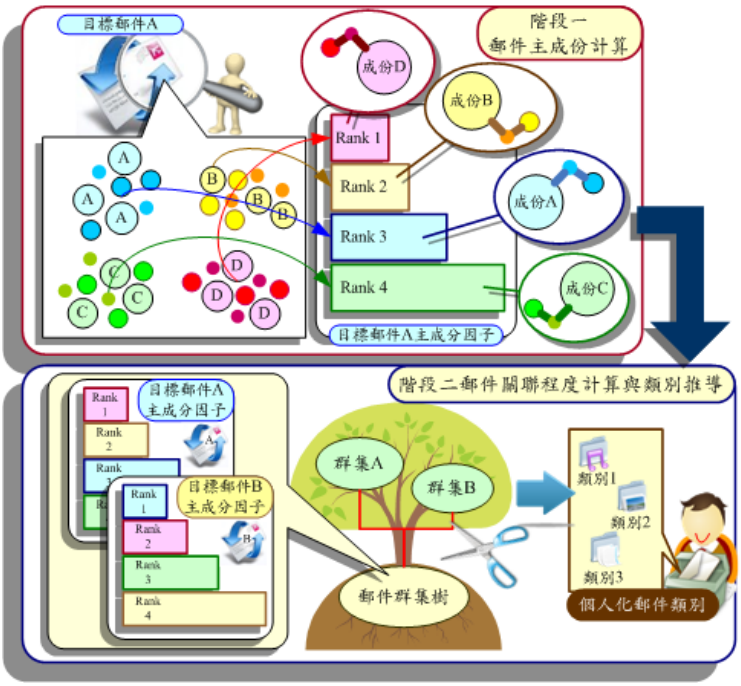

圖 3.5、個人化郵件類別推論模組示意圖

3.2.1 階段一、郵件主成份計算

個人化郵件類別推論之郵件主成份計算部分乃以 Šmídl 與 Quinn (2007) 所提之「貝氏主成份分析」為基礎制定法則,在此法則目的乃以「貝氏主成份分析」尋找影響郵件分群之關鍵字,並以此關鍵字作為「階段二、郵件關聯程度計算與類別推論」之郵件群集推論依據。因此,本研究乃以「語意擷取郵件關鍵字」模組中所獲得之主題詞彙與詞彙重要性作為「貝氏主成份分析」依據,由目標郵件之主題詞彙中尋找出具郵件所屬群集代表性關鍵字,以作為目標郵件之郵件分群依據。

根據上述,以下乃說明郵件主成份分析之過程(如圖 3.6 所示),此階段乃以主題詞彙所含詞彙重要性進行分析,以分析主題共變異數與郵件特徵值。當中,主題共變異數即目標郵件中各兩主題所含詞彙互相比較下,兩主題所含詞彙影響目標郵件分群之影響程度是否相同,假設目標郵件中含有「電影」與「文藝」此兩主題,若此兩主題共變異數較低,則代表此兩主題對目標郵件之分群影響力趨於相等,接著,將此些主題共變異數彙整成郵件共變異數矩陣,以進行郵件特徵值分析;之後,郵件特徵值乃透過郵件共變異數矩陣計算其矩陣特徵值取得,郵件特徵值乃分別代表目標郵件與各項主題關聯

性，假設目標郵件中「電影」主題所對應特徵值較高，則表示目標郵件內容所含詞彙與「電影」主題具較大關聯，故將此郵件特徵值與主題詞彙之詞彙重要性，計算其線性組合以強化各詞彙與目標郵件關係，即可獲得各詞彙與目標郵件關聯程度，並挑選關聯程度較強之詞彙作為郵件分群之關鍵字，以作為「階段二、郵件關聯程度計算與類別推論」目標郵件分群之依據。

圖 3.6、郵件主成份計算示意圖

符號定義

$ZIW_{i,k}$	第 i 項重要文字型資料中第 k 個主題詞彙集合
$ZIW_{i,k,n}$	第 i 項重要文字型資料中第 k 個主題詞彙所含第 n 筆主題詞彙
$I(ZIW_{i,k,n})$	第 i 項重要文字型資料中第 n 筆重要詞彙於第 k 個主題對郵件重要性係數

QW	主題詞彙集合之詞彙總數
$\overline{X}_{i,k}$	第 i 項重要文字型資料之第 k 個主題詞彙集合中詞彙重要性程度平均值
$Cov(ZIW_{i,k}, ZIW_{i,j})$	第 i 項重要文字型資料中第 k 個主題詞彙集合與第 j 個主題詞彙集合之詞彙重要性程度共變異數
Cov_i	第 i 項重要文字型資料之主題共變異矩陣
$\lambda_{i,k}$	第 i 項重要文字型資料中第 k 筆郵件特徵值
$N(\lambda_i)$	第 i 項重要文字型資料中所含郵件特徵值總數
I	k×k 形式之單位矩陣
$Det(Cov_i)$	第 i 項重要文字型資料中主題共變異矩陣之行列式
$EV_{i,k}$	第 i 項重要文字型資料中第 k 筆郵件特徵值向量矩陣
$\eta_{i,k}$	第 i 項重要文字型資料中第 k 個主題詞彙重要性係數之向量矩陣
$R_{i,k}$	第 i 項重要文字型資料中第 k 筆重要詞彙與郵件關聯性
$\omega_{i,k}$	第 i 項重要文字型資料中第 k 筆郵件特徵值，相對於所含郵件特徵值組合之解釋比例
Q	第 i 項重要文字型資料中解釋比例組合之第三四分位數指標
W_i	第 i 項重要文字型資料中郵件特徵值組合之篩選門檻值
TR_i	第 i 項重要文字型資料中，經篩選後所保留重要詞彙與郵件關聯性之郵件關聯集合
$MKW_{i,k}$	第 i 項重要文字型資料中第 k 個郵件特徵詞彙集合

Step (C-1)：計算主題詞彙重要性之平均值

本模組之法則分為兩階段，於第一階段中乃以「語意擷取郵件關鍵字」中所取得關鍵字之重要程度為分析數據，當中，為保留每筆關鍵字之主題特性，乃以關鍵字之主題將關鍵字劃分作為既有成分。因此，於本法則乃以「語意擷取郵件關鍵字」模組所得之主題詞彙 $ZIW_{i,k,n}$ 與主題詞彙重要性係數 $I(ZIW_{i,k,n})$ 進行分析。而主成份分析中，特徵值乃透過各主題詞彙集合 $ZIW_{i,k}$ 之共變異數進行共變異數矩陣計算後取得。因此，為計算各主題詞彙集合 $ZIW_{i,k}$ 之共變異數，因此，本步驟乃計算共變異數所需之平均值 $\overline{X}_{i,k}$，首先，以詞彙重要性 $I(ZIW_{i,k,n})$ 進行加總，並與主題詞彙集合之詞彙總數 QW 進行計算，進而取得各項詞彙重要性程度平均值 $\overline{X}_{i,k}$，如公式(3.13)所示。

$$\overline{X}_{i,k} = \frac{1}{QW} \times \sum_{\text{all n}} I(ZIW_{i,k,n}) \tag{3.13}$$

Step (C-2)：計算主題詞彙集合之共變異數

步驟(C-2)乃計算各主題與主題間之主題共變異數，以計算郵件特徵值，進而分析出影響郵件分群之關鍵字。目標郵件中，郵件內文乃由主題詞彙所構成，故藉由分析各主題內詞彙及詞彙重要性不同（如：「明星」此詞彙於電影與文藝兩主題中重要性差異）進而分析出影響郵件分群之關鍵字所屬主題。因此，本步驟乃透過主題與主題間各自所含主題詞彙重要性係數 $I(ZIW_{i,k,n})$ 計算主題共變異數（即為主題詞彙重要性共變異數 $Cov(ZIW_{i,k}, ZIW_{i,j})$），以透過主題內詞彙及詞彙重要性進而分析出具影響郵件分群之關鍵字。其中，主題共變異數 $Cov(ZIW_{i,k}, ZIW_{i,j})$ 乃以各主題詞彙集合 $ZIW_{i,k}$ 之平均值 $\overline{X}_{i,k}$ 為計算依據，計算第 k 個主題詞彙集合 $ZIW_{i,k}$ 與第 j 個主題詞彙集合 $ZIW_{i,j}$ 中兩集合之主題詞彙重要性共變異數 $Cov(ZIW_{i,k}, ZIW_{i,j})$，如公式(3.14)所示，並以此主題共變異數 $Cov(ZIW_{i,k}, ZIW_{i,j})$ 進行後續步驟之分析。

$$Cov(ZIW_{i,k}, ZIW_{i,j}) = \frac{\sum_{n=1}^{QW} (I(ZIW_{i,k,n}) - \overline{X}_{i,k}) \times (I(ZIW_{i,j,n}) - \overline{X}_{i,j})}{QW} \quad \text{and} \quad k \le j \tag{3.14}$$

Step (C-3)：彙整主題共變異數得郵件共變異矩陣

步驟(C-3)乃將主題共變異數進行彙整為郵件共變異矩陣，以作為進行郵件特徵值分析時之分析依據。藉由步驟(C-2)取得各主題與主題間之主題共變異數後，為能針對目標郵件整體所含蓋主題進行分析出各主題對目標郵件分群之影響程度，並將此作為郵件特徵值，因此，本步驟乃進行共變異數矩陣之彙整，以進行目標郵件之郵件特徵值分析。首先，由上述步驟中獲得主題詞彙集合之詞彙重要性程度共變異數 $Cov(ZIW_{i,k}, ZIW_{i,j})$ 後，本步驟乃將共變異數 $Cov(ZIW_{i,k}, ZIW_{i,j})$ 彙整為主成份分析進行所需之郵件主題共變異數矩陣 Cov_i，如公式(3.15)所示。

$$Cov_i = \begin{bmatrix} Cov(ZIW_{i,1}, ZIW_{i,1}) & Cov(ZIW_{i,1}, ZIW_{i,2}) & \cdots & Cov(ZIW_{i,1}, ZIW_{i,j}) \\ Cov(ZIW_{i,2}, ZIW_{i,1}) & Cov(ZIW_{i,2}, ZIW_{i,2}) & \cdots & Cov(ZIW_{i,2}, ZIW_{i,j}) \\ \vdots & \vdots & \cdots & \vdots \\ Cov(ZIW_{i,k}, ZIW_{i,1}) & Cov(ZIW_{i,k}, ZIW_{i,2}) & \cdots & Cov(ZIW_{i,k}, ZIW_{i,j}) \end{bmatrix} \tag{3.15}$$

where $Cov(ZIW_{i,k}, ZIW_{i,j}) = Cov(ZIW_{i,j}, ZIW_{i,k})$

Step (C-4)：計算郵件特徵值

透過步驟(C-3)取得郵件共變異數矩陣 Cov_i 後，於此步驟乃進行郵件特徵值分析，並以郵件特徵值針對各筆詞彙強化詞彙與目郵件關聯，進以尋得具影響郵件分群之關鍵字。郵件特徵值乃表示主題與目標郵件之間兩者關聯，例如：郵件主旨分別為「電影分享」與「花東旅行」兩封郵件，此兩封郵件因所含詞彙所屬主題不同，因此「花東旅行」之郵件則偏向「旅遊」相關主題，則此郵件與主題間關係即為郵件特徵值。是故，透過此郵件特徵值與各筆詞彙之詞彙重要性進行計算則可強化詞彙與郵件關聯，進而凸顯出較能影響郵件分群之關鍵字。是故，本步驟乃以郵件共變異數矩陣 Cov_i 進行分析郵件特徵值 λ_i，首先，將郵件特徵值 λ_i 與 k×k 之單位矩陣 I 相乘，並以步驟(C-3)取得主題共變異數矩陣 Cov_i 與郵件特徵值 λ_i（即與單位矩陣 I 相乘後之郵件特徵值 λ_i）相減，且計算行列式 $Det(Cov_i)$，當中，行列式 $Det(Cov_i)$計算結果必等於 0，如公式(3.16)所示，由於行列式 $Det(Cov_i)$計算需等於 0 則可求得郵件特徵值 λ_i，並成為行列式 $Det(Cov_i)$最佳解。

$$\begin{aligned} &Det(Cov_i) = \left|Cov_i - \lambda_i \times I\right| \\ &\text{where} \quad Det(Cov_i) = 0 \quad \text{and} \quad I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \end{aligned} \tag{3.16}$$

當中所求得郵件特徵值 λ_i 具有 k 個最佳解，且郵件特徵值 $\lambda_{i,k}$ 總和等於郵件共變異數矩陣 Cov_i 中各主題詞彙集合 $ZIW_{i,k}$ 之變異數加總（即為 k=j 時所計算之共變異數 $Cov(ZIW_{i,k}, ZIW_{i,j})$ 之加總），且郵件特徵值 $\lambda_{i,k}$ 之乘積與互變異矩陣之行列式相等，如公式(3.17)所示。

$$\begin{aligned} &\lambda_i = \{\lambda_{i,1}, \lambda_{i,2}, \lambda_{i,3}, ..., \lambda_{i,k}\} \\ &\sum_{\text{all k}} \lambda_{i,k} = \sum_{\text{all k}} Cov(ZIW_{i,k}, ZIW_{i,k}) \\ &\prod_{k=1}^{m} \lambda_{i,k} = \left|Cov_i\right| \end{aligned} \tag{3.17}$$

Step (C-5)：計算郵件特徵向量

步驟(C-5)乃透過郵件特徵值 $\lambda_{i,k}$ 計算郵件特徵向量，作為詞彙與目標郵件關聯分析之權重，以減少詞彙與目標郵件關聯受非詞彙所屬主題干擾。於郵件內文中各筆詞彙並非單屬於單一主題，故因主題不同其詞彙重要性亦不相同，如：「明星」此詞彙於電影

與文藝兩主題中皆含有詞彙重要性，但乃以電影主題所對應之詞彙重要性較為凸顯；為使詞彙計算詞彙與目標郵件之關聯時，不受詞彙重要性較低之主題干擾，於計算詞彙與目標郵件之關聯時以郵件特徵向量作為比重，降低其詞彙重要性較低主題之影響。因此本步驟乃以主題共變異數矩陣 Cov_i 與郵件特徵值 $\lambda_{i,k}$ 與郵件特徵向量 $EV_{i,k}$ 進行計算且等式為零（如公式(3.18)所示），當中，郵件特徵向量 $EV_{i,k}$ 為非零向量且有 p 個最佳解並彙整為矩陣，於後續步驟以此郵件特徵向量矩陣 $EV_{i,k}$ 計算，則可獲得詞彙與目標郵件之關聯。

$$\left(Cov_i - \lambda_{i,k} \times I\right) \times EV_{i,k} = 0$$
$$\text{where} \quad EV_{i,k} = \begin{bmatrix} EV_{i,k,1} \\ EV_{i,k,2} \\ \vdots \\ EV_{i,k,p} \end{bmatrix} \tag{3.18}$$

Step (C-6)：計算詞彙與目標郵件關聯性

步驟(C-6)乃以郵件特徵值向量與詞彙重要性進行計算，以獲得詞彙與目標郵件關聯性，並作為挑選郵件特徵詞彙之指標。目標郵件中各詞彙與各主題之間皆具有互相對應之詞彙重要性，而此詞彙重要性卻無法凸顯詞彙與目標郵件之關係，故本步驟乃以詞彙與主題所對應之詞彙重要性為分析依據，並考量郵件特徵值向量減少非該詞彙所屬主題干擾，以分析得各筆詞彙與目標郵件之關聯性。本步驟乃以詞彙與各主題所對應之詞彙重要性 $I(ZIW_{i,k,n})$ 轉化為詞彙重要性向量矩陣 $\eta_{i,k}$，並進行線性組合計算，當中乃以郵件特徵值向量矩陣 $EV_{i,k}$ 作為比重加入於計算線性組合中，其線性組合即為詞彙與郵件關聯性 $R_{i,k}$（如公式(3.19)所示），其結果可整理如表 3.5 所示。

$$R_{i,k} = \eta_{i,1} \times EV_{i,k,1} + \eta_{i,2} \times EV_{i,k,2} + \eta_{i,3} \times EV_{i,k,3} + \cdots + \eta_{i,p} \times EV_{i,k,p}$$
$$\text{where} \quad \eta_{i,k} = \left[I(ZIW_{i,k,1}) \quad I(ZIW_{i,k,2}) \quad \cdots \quad I(ZIW_{i,k,n})\right] \tag{3.19}$$

表 3.5、詞彙與郵件關聯性之彙整表

郵件特徵值	郵件特徵向量	詞彙與郵件關聯性
$\lambda_{i,1}$	$EV_{i,1}$	$R_{i,1}=\eta_{i,1}\times EV_{i,1,1}+\eta_{i,2}\times EV_{i,1,2}+\eta_{i,3}\times EV_{i,1,3}+\cdots+\eta_{i,p}\times EV_{i,1,p}$
$\lambda_{i,2}$	$EV_{i,2}$	$R_{i,2}=\eta_{i,1}\times EV_{i,2,1}+\eta_{i,2}\times EV_{i,2,2}+\eta_{i,3}\times EV_{i,2,3}+\cdots+\eta_{i,p}\times EV_{i,2,p}$
…	…	…
$\lambda_{i,k}$	$EV_{i,k}$	$R_{i,k}=\eta_{i,1}\times EV_{i,k,1}+\eta_{i,2}\times EV_{i,k,2}+\eta_{i,3}\times EV_{i,k,3}+\cdots+\eta_{i,p}\times EV_{i,k,p}$

Step (C-7)：計算郵件特徵值之特徵比例

本步驟乃計算各筆郵件特徵之解釋比例以進行篩選詞彙與郵件關聯性 $R_{i,k}$。郵件特徵值乃表示目標郵件之詞彙所偏向之主題，故為去除關聯性較低之主題（即詞彙與郵件關聯性 $R_{i,k}$ 較低之線性組合），以尋得與目標郵件關係較高之主題，進而尋得影響郵件分群之關鍵字。故本步驟乃以郵件特徵值 $\lambda_{i,k}$ 與郵件特徵值加總後數據相除計算解釋比例 $\omega_{i,k}$（如公式(3.20)所示），使以郵件特徵之解釋比例 $\omega_{i,k}$ 作為挑選詞彙與郵件關聯性 $R_{i,k}$ 之依據。

$$\omega_{i,k}=\frac{\lambda_{i,k}}{\sum_{\text{all a}}\lambda_{i,a}} \tag{3.20}$$

Step (C-8)：挑選高度關聯之詞彙線性組合

步驟(C-8)乃透過郵件特徵之解釋比例 $\omega_{i,k}$ 進行篩選詞彙與郵件關聯性 $R_{i,k}$，以保留具高度關聯性之詞彙線性組合（即詞彙與郵件關聯性 $R_{i,k}$）。透過步驟(C-7)所得之郵件特徵之解釋比例 $\omega_{i,k}$ 乃代表該詞彙與郵件關聯性 $R_{i,k}$ 於目標郵件中關聯程度比例，故以此郵件特徵之解釋比例 $\omega_{i,k}$ 進行挑選高度關聯之詞彙線性組合，當中本步驟乃制定下列多種篩選方式，並依使用者需求選定其篩選方式，篩選方式如下：

(a)解釋比例平均值

解釋比例平均值乃表達目標郵件整體解釋比例之解釋強度，故郵件特徵之解釋比例 $\omega_{i,k}$ 若低於整體目標郵件之解釋比例平均值，則代表其郵件特徵所對應詞彙線性組合之關聯度未能影響郵件分群。因此，乃計算解釋比例平均值 W_i 挑選高關聯性之詞彙線性組合（如公式(3.21)所示），當中，若該郵件特徵之解釋比例 $\omega_{i,k}$ 高於（或相等）解釋比例平均值 W_i 則表示該詞彙與郵件關聯性 $R_{i,k}$ 屬於高度郵件關聯性之郵件關聯集合 TR_i。

$$W_i = \frac{\sum_{\text{all } k} \omega_{i,k}}{N(\lambda_i)} \quad , \quad \text{IF} \quad \omega_{i,k} \geq W_i \quad \text{Then} \quad R_{i,k} \in TR_i \tag{3.21}$$

(b)解釋比例中位數

為避免解釋比例中極大值與極小值差距過大而影響解釋比例平均值，因此本法則亦可透過解釋比例中位數作為目標郵件整體解釋比例之解釋強度，並以解釋比例中位數 W_i 挑選高度關聯之詞彙線性組合（如公式(3.22)所示），若該郵件特徵之解釋比例 $\omega_{i,k}$ 高於（或相等）解釋比例中位數 W_i 則表示該詞彙與郵件關聯性 $R_{i,k}$ 屬於高度郵件關聯性之郵件關聯集合 TR_i。

$$W_i = \begin{cases} \omega_{i,\frac{k+1}{2}} & \text{IF } N(\lambda_i) \text{ is odd} \\ \frac{1}{2} \times \left(\omega_{i,\frac{k+1}{2}} + \omega_{i,\frac{k+1}{2}+1} \right), & \text{IF } N(\lambda_i) \text{ is even} \end{cases} \tag{3.22}$$

$$\text{and} \quad \text{IF} \quad \omega_{i,k} \geq W_i \quad \text{Then} \quad R_{i,k} \in TR_i$$

(c)解釋比例四分位數

為能更精確挑選挑選高度關聯之詞彙線性組合，且不受極端值影響挑選結果，於此乃制定解釋比例四分位數 W_i（於此乃第三四分位數進行篩選）。首先計算解釋比例之第三四分位數指標 Q，以界定第三位之解釋比例四分位數 W_i，進已挑選詞彙線性組合（如公式(3.23)所示），當中，若該郵件特徵之解釋比例 $\omega_{i,k}$ 高於（或相等）解釋比例四分位數 W_i 則表示該詞彙與郵件關聯性 $R_{i,k}$ 屬於高度郵件關聯性之郵件關聯集合 TR_i。

$$Q = N(\lambda_i) \times 75\%$$

$$W_i = \begin{cases} \omega_{i,k\times 75\%} & \text{IF } Q \notin \{X : |X| \in N\} \\ \frac{1}{2} \times \left(\omega_{i,k\times 75\%} + \omega_{i,k\times 75\%+1} \right), & \text{IF } Q \in \{X : |X| \in N\} \end{cases} \tag{3.23}$$

$$\text{and} \quad \text{IF} \quad \omega_{i,k} \geq W_i \quad \text{Then} \quad R_{i,k} \in TR_i$$

(d)解釋比例門檻值

使用者亦可自行制訂解釋比例門檻值，以進行挑選詞彙與郵件關聯性 $R_{i,k}$（如公式(3.24)所示），若該郵件特徵之解釋比例 $\omega_{i,k}$ 高於（或相等）解釋比例門檻值 W_i 則表示該詞彙與郵件關聯性 $R_{i,k}$ 屬於高度郵件關聯性之郵件關聯集合 TR_i，並於後續步驟中以此郵件關聯集合 TR_i 尋找出郵件特徵詞彙。

$$\text{IF} \quad \omega_{i,k} \geq W_i \quad \text{Then} \quad R_{i,k} \in TR_i \quad \text{where } 0 \leq W_i < 1 \tag{3.24}$$

Step (C-9)：挑選郵件特徵詞彙

透過前述步驟尋得高度關聯之詞彙線性組合（即詞彙與郵件關聯性 $R_{i,k}$）後，步驟(C-9)乃以高度關聯之詞彙線性組合彙整為郵件特徵詞彙合。由於各組詞彙線性組合中所含詞彙可能重複，為避免郵件分群關鍵字重疊而影響郵件分群之準確性，因此，乃彙整郵件關聯集合 TR_i 中各高度關聯之詞彙線性組合所含詞彙重要性向量矩陣 $\eta_{i,k}$ 去除重複詞彙，以彙整為郵件特徵詞彙集合 MKW_i（如公式(3.25)所示）。

$$MKW_i = \sum_{\text{all } k} \eta_{i,k} - \bigcap_{\text{all } k} \eta_{i,k} \tag{3.25}$$

當中，郵件特徵詞彙集合 MKW_i 亦可彙整為表 3.6。

表 3.6、郵件特徵詞彙彙整表

郵件主旨	MKW_1	$ZIW_{1,k,1}$	$ZIW_{1,k,2}$	…	$ZIW_{1,k,n}$
郵件內文	MKW_2	…	$ZIW_{2,k,2}$	…	…
附件檔案名稱	MKW_3	$ZIW_{3,k,1}$	$ZIW_{3,k,2}$	…	$ZIW_{3,k,n}$

綜上所述，本階段乃以詞彙對各主題重要性分析詞彙主題與目標郵件關聯，以了解目標郵件內含詞彙所偏向主題（如：郵件含有「電影」、「旅遊」兩主題詞彙，但因「旅遊」詞彙較多，故目標郵件則偏向「旅遊」主題），並透過目標郵件主要主題中分析其詞彙與郵件關聯性程度（即詞彙影響郵件分群之程度），從中篩選出郵件特徵詞彙，由於目標郵件關鍵字（即郵件特徵詞彙）乃透過郵件所偏向主題中分析而得，故透過階段一法則分析已知類別郵件所含詞彙作為類別關鍵字，以作為個人化郵件分類中各個類別之類別關鍵字，亦可透過此法則推論各郵件所偏向主題後，更進一步推論使用者常用類別，滿足各使用者所需不同需求之郵件類別。

3.2.2 階段二、郵件關聯程度計算與類別推論

目標郵件乃透過「階段一、郵件主成份計算」尋得郵件分群之關鍵字後，於此階段乃進行郵件類別推論。由於個人化郵件並非以結構性文章進行撰寫，亦含有不完整之文句或隱藏含意性文句，而無法以制式化類別進行分類，且 Zajic（2008）亦指出電子郵件所述內容乃為使用者與寄件者間之對話，故多數郵件皆具互相關聯性，但卻無明顯關聯結構，需使用者自行彙整串聯得以瞭解郵件所述主要內容。故為能推論符合使用者個人之郵件類別，且凸顯郵件與郵件間關聯，本研究乃以既有郵件進行郵件分群，並以「階層式分群法」（Hierarchical Clustering）之樹狀架構分群特性，以凸顯個人化郵件關聯結構並建立關聯架構（如圖 3.7 所示）。

圖 3.7、個人化郵件與階層式分群法關聯示意圖

「階段二、郵件關聯程度計算與類別推論」乃以階段一所得之郵件分群關鍵字彙整發送郵件關鍵字並分析各郵件內容相似性，形成階層式群集以推論為個人化郵件類別。因此，本法則乃先行彙整使用者之發送郵件關鍵字作為階層式群集推論之使用者分群規則，乃因發送郵件關鍵字（即使用者所寄出信件透過階段一所分析而得郵件分群之關鍵字）皆為使用者根據自身使用經驗所書寫，故能表達使用者於使用上所需類別之類型，因此，本研究乃先彙整使用者所寄發信件之關鍵字作為發送郵件關鍵字，並將發送郵件關鍵字與非寄發信件（即非使用者所撰寫之信件）進行內容相似性分析，以獲得第一階層之郵件群集，再將第一階層之郵件群集以平均連結聚合法分析群集距離，並整併為第

二階層郵件群集，如此反覆分析各階層郵件群集內容相似性則形成層級群集樹（即階層式郵件群集），再以此樹狀結構之郵件群集進行類別名稱推論，首先乃以領域文件作為基礎資料，參考 Atkinson 等人（2009）以自然語言處裡（Natural Language Processing；NLP）方式擷取並彙整領域文件詞彙為類別名稱詞彙庫，再以典型相關分析法（Kernel Canonical Correlation Analysis；KCCA）並參考 Li 和 Shawe-Taylor（2007）之方法分析郵件群集與名稱詞彙集，尋得與郵件群集高關聯性之名稱詞彙集，並以名稱詞彙集中最具代表性詞彙作為類別名稱，達到郵件群集命名效果，並推論符合使用者之個人化郵件類別。郵件類別名稱推論過程，如圖 3.8 所示。

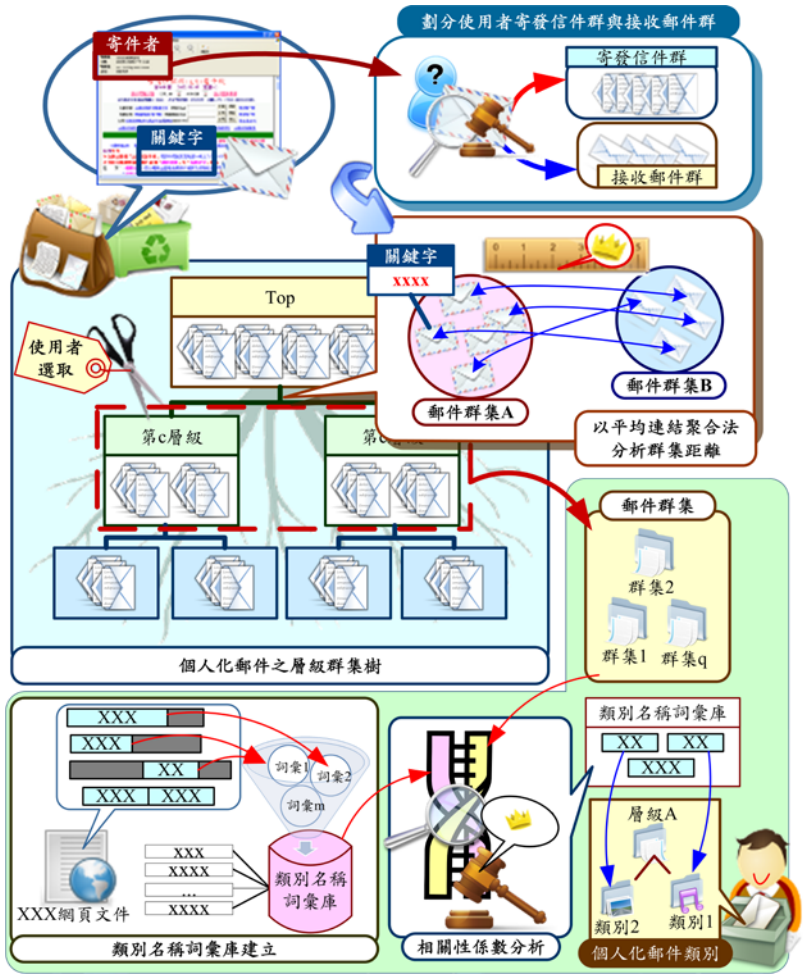

圖 3.8、個人化郵件類別推論示意圖

符號定義

AP_T	目標郵件
UI_i	電子郵件中第 i 項文字型資料，i=1,2,3；UI_1 為「郵件主旨」、UI_2 為「郵件內文」及 UI_3 為「附加檔名」

RM	使用者接收郵件群
SM	使用者寄發郵件群
Send(AP_T)	目標郵件所含之寄件人信箱地址
UseSend	使用者之電子郵件信箱地址
Fw_Set	轉寄信件集合
Rw_Set	回覆信件集合
FW	轉寄信件標籤
RW	回覆信件標籤
MKW_i	電子郵件中第 i 項文字型資料所含郵件特徵詞彙
CW_Set(RM)	接收郵件分群字集
CW_Set(RM_t)	第 t 封接收郵件之分群字集
CW_Set(SM)	寄發郵件分群字集
CW_Set(SM_v)	第 v 封寄發郵件之分群字集
RM_t	第 t 封接收郵件
SM_v	第 v 封寄發郵件
$M(RM_t, SM_v)$	第 t 封接收郵件與第 v 封寄發郵件間郵件內容相似度
$M[SM_v]$	第 v 個郵件內容相似度矩陣
$Group_{c,q}$	於層級群集樹中第 c 層級之第 q 個郵件群集
$Group_{c,q,n}$	第 c 層級之第 q 個郵件群集所含第 n 封郵件
$M(Group_{c,a,f}, Group_{c,b,e})$	第 c 層級中第 a 個郵件群集之第 f 封郵件與第 b 個郵件群集之第 e 封郵件間郵件內容相似度
$D(Group_{c,a}, Group_{c,b})$	第 c 層級之第 a 個郵件群集與第 b 個郵件群集間群集距離（即兩群集所含郵件之平均距離）
$D[Group_c]$	第 c 層級之郵件群集距離矩陣
HCTree	層級群集樹矩陣
HC(Y)	層級群集樹矩陣分割座標點
NTop	層級群集樹之層級數量
Category	個人化郵件類別矩陣
$CW_{HC(Y),q,z}$	第 HC(Y)層級之第 q 個郵件群集所含第 z 筆分群關鍵字
$ParW_{w,e}$	第 w 份領域文件中第 e 段文章段落之詞彙集合

$PW_{w,e,m}$	第 w 份領域文件中第 e 段文章段落之第 m 筆詞彙
$P(PW_{w,e,m} \mid PW_{w,e,m-1})$	第 w 份領域文件之第 e 段文章段落中第 m-1 筆詞彙出現時第 m 筆詞彙亦出現之機率
$N(PW_{w,e,m})$	第 w 份領域文件中第 e 段文章段落之第 m 筆詞彙出現次數
$N(PW_{w,e,m-1}, PW_{w,e,m})$	第 w 份領域文件中第 e 段文章段落之第 m-1 筆詞彙與第 m 筆詞彙同時出現次數
$P(PW_{w,e})$	第 w 份領域文件中第 e 段文章段落之關聯機率矩陣
$Sim(PW_{w,a}, PW_{w,b})$	第 w 份領域文件中第 a 段文章段落與第 b 段文章段落相似度
$\|PW_{w,e}\|$	第 w 份領域文件中第 e 段文章段落之矩陣範數（即該段落關聯機率矩陣之向量大小）
$NamW_{w,e}$	第 w 份領域文件中第 e 個名稱詞彙集
$EV_{q,n}$	第 q 個郵件群集所含第 n 封郵件特徵值向量矩陣
$F(Group_{q,n})$	第 q 個郵件群集所含第 n 封郵件之線性組合
$F(NamW_{w,e})$	第 w 份領域文件中第 e 個名稱詞彙集線性組合
$Cov(Group_{q,(a\|b)})$	第 q 個郵件群集中第 a 封郵件與第 b 封郵件之郵件群集共變異數
$Cov[G_{q,\bullet}]$	第 q 個郵件群集之共變異矩陣
$Cov(NamW_{w,(a\|b)})$	第 w 份領域文件中第 a 個名稱詞彙集與第 b 個名稱詞彙集之名稱詞彙共變異數
$Cov[N_{w,\bullet}]$	第 w 份領域文件之名稱詞彙共變異矩陣
$\alpha_{q,n,p}$	第 q 個郵件群集所含第 n 封郵件之第 p 個郵件群集加權係數
$\beta_{w,e,f}$	第 w 份領域文件中第 e 個名稱詞彙集所含第 f 個名稱詞彙集加權係數
$Cov(G_q^F, N_w^F)$	第 q 個郵件群集線性組合與第 w 份領域文件線性組合之共變異數
$\eta(G_q, N_w)$	第 q 個郵件群集與第 w 份領域文件之相關性係數
$[\alpha]_{q,n}$	第 q 個郵件群集所含第 n 個郵件群集加權係數矩陣
$[\beta]_{w,e}$	第 w 份領域文件中第 e 個名稱詞彙集加權係數矩陣

Step (D-1)：劃分使用者寄發信件群與接收郵件群

第二階段之法則乃以使用者之慣用詞彙作為分群準則，以推論符合使用者個人情境之郵件群集，進而尋得符合使用者之個人化類別。於此步驟，乃以目標郵件之寄件人信箱地址 Send(AP_T)與使用者之電子郵件信箱地址 UseSend 比對，若為相同則目標郵件 AP_T 屬於使用者接收郵件群 RM，反之則屬於使用者寄發郵件群 SM，如公式(3.26)所示。

$$AP_T \in \begin{cases} RM & IF\ Send(AP_T) \neq UseSend \\ SM' & IF\ Send(AP_T) = UseSend \end{cases} \tag{3.26}$$

接著以使用者寄發郵件群 SM 之郵件分析主旨 UI_1 所含寄發型別標記，以推論郵件所屬寄發型別集合，若目標郵件主旨 UI_1 含有轉寄信件標籤 FW 則屬於轉寄信件集合 Fw_Set，若含有回覆信件標籤 RW 則屬於回覆信件集合 Rw_Set，如公式(3.27)所示。

$$\begin{aligned} &UI_T = \{UI_1, UI_2, UI_3\} \\ &IF\ RW\ exist\ in\ UI_1 \quad Then \quad AP_T \in Rw\_Set \\ &IF\ FW\ exist\ in\ UI_1 \quad Then \quad AP_T \in Fw\_Set \end{aligned} \tag{3.27}$$

Step (D-2)：彙整郵件分群關鍵字

使用者之慣用詞彙乃以使用者寄發郵件群 SM 之郵件分群關鍵字彙整而得，因此 Step (D-2)乃彙整各郵件由「階段一、郵件主成份計算」所得之郵件特徵詞彙 MKW_i，當中，若目標郵件 AP_T 歸類為接收郵件群 RM，則郵件特徵詞彙 MKW_i 屬於接收郵件之分群字集 CW_Set(RM)；反之，若歸類為寄發郵件群 SM，則郵件特徵詞彙 MKW_i 屬於寄發郵件之分群字集 CW_Set(SM)，如公式(3.28)所示。

$$MKW_i \in \begin{cases} CW\_Set(RM) & IF\ AP_T \in RM \\ CW\_Set(SM)' & IF\ AP_T \in SM \end{cases} \quad and\ i = 1,2,3 \tag{3.28}$$

Step (D-3)：第一層級之郵件群集劃分

於階層式分群中，各層級之群集乃透過比對各郵件之分群字集相似度進行分群，故本步驟乃以寄發郵件 SM_v 與接收郵件 RM_t 所含分群字集中，將各郵件所含相同詞彙數作為內容相似性比對依據，判斷該接收郵件與寄發郵件所述內容之相似度，並以郵件內容相似度 $M(RM_t,SM_v)$作為接收郵件 RM_t 群集歸屬依據，如公式(3.29)所示。

$$
\begin{aligned}
&CW\_Set(RM)=\{CW\_Set(RM_1),CW\_Set(RM_2),...,CW\_Set(RM_t)\}\\
&CW\_Set(SM)=\{CW\_Set(SM_1),CW\_Set(SM_2),...,CW\_Set(SM_v)\}\\
&M(RM_t,SM_v)=\frac{N(CW\_Set(RM_t)\cap CW\_Set(SM_v))}{N(CW\_Set(SM_v))}
\end{aligned}
\tag{3.29}
$$

且本步驟乃將第 v 封寄發郵件 SM_v 所對應之郵件內容相似度 $M(RM_t,SM_v)$彙整為郵件內容相似度矩陣 $M[SM_v]$，於郵件內容相似度矩陣 $M[SM_v]$中，並計算郵件內容相似度 $M(RM_t,SM_v)$之最大值後，將第 t 封接收郵件 RM_t 歸類於層級群集樹中第一層級之郵件群集 $Group_{1,v}$ 中，如公式(3.30)所示。

$$
\begin{aligned}
&M[SM_v]=\begin{bmatrix} M(RM_1,SM_v)\\ M(RM_2,SM_v)\\ \vdots\\ M(RM_t,SM_v)\\ \vdots \end{bmatrix}\\
&Group_{1,v}=\{RM_t \mid Max(M[SM_v])=M(RM_t,SM_v),\forall t\}
\end{aligned}
\tag{3.30}
$$

Step (D-4)：郵件群集距離計算

為能取得所含郵件相似之群集並合併，本步驟乃計算郵件群集距離 $D(Group_{c,a},Group_{c,b})$ 作為群集合併之依據。其中，郵件群集距離 $D(Group_{c,a},Group_{c,b})$ 乃透過兩群集內所有郵件互相計算郵件內容相似度 $M(Group_{c,a,f},Group_{c,b,e})$計算而得，因此，本步驟乃先以郵件群集內各郵件所含分群字集 $CW_Set(Group_{c,q,n})$計算相同詞彙數比例，並以此作為郵件內容相似度 $M(Group_{c,a,f},Group_{c,b,e})$，如公式(3.31)所示。

$$
M(Group_{c,a,f},Group_{c,b,e})=\frac{N(CW\_Set(Group_{c,a,f})\cap CW\_Set(Group_{c,b,e}))}{N(CW\_Set(Group_{c,b,e}))}
\tag{3.31}
$$

之後本步驟乃以平均連結聚合法計算兩群集間距離，當中，平均連結聚合法乃以郵件內容相似度 $M(Group_{c,a,f},Group_{c,b,e})$作為各郵件間距離，並透過計算郵件內容相似度 $M(Group_{c,a,f},Group_{c,b,e})$之平均值，以獲得兩郵件群集間平均距離（即郵件群集距離）$D(Group_{c,a},Group_{c,b})$，如公式(3.32)所示，當中，郵件群集 $Group_{c,q}$ 與群集本身所對應之距離為零。

$$
D(Group_{c,a},Group_{c,b})=\frac{\sum_{all\ f,e} M(Group_{c,a,f},Group_{c,b,e})}{N(Group_{c,a})\times N(Group_{c,b})}\ \text{where}\ \ a>b\ \text{for all}\ a,b
\tag{3.32}
$$

其中，各郵件群集所對應之郵件群集距離 $D(Group_{c,a}, Group_{c,b})$ 乃彙整為郵件群集距離矩陣 $D[Group_c]$，如公式(3.33)所示。

$$D[Group_c] = \begin{bmatrix} 0 & D(Group_{c,2}, Group_{c,1}) & \cdots & D(Group_{c,a}, Group_{c,1}) & \cdots \\ D(Group_{c,1}, Group_{c,2}) & 0 & \cdots & D(Group_{c,a}, Group_{c,2}) & \cdots \\ \cdots & \cdots & 0 & \cdots & \cdots \\ D(Group_{c,1}, Group_{c,a}) & D(Group_{c,2}, Group_{c,a}) & \cdots & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & 0 \end{bmatrix} \quad (3.33)$$

and $D(Group_{c,a}, Group_{c,b}) = D(Group_{c,b}, Group_{c,a})$

Step (D-5)：合併相似郵件群集

由於郵件群集距離乃透過各群集內郵件所含相同詞彙數計算而得，因此若郵件群集距離數值越大，則表示兩群集所含郵件之郵件敘述內容相似，故兩群集距離相近。因此，本步驟乃判斷郵件群集距離矩陣 $D[Group_c]$ 中郵件群集距離 $D(Group_{c,a}, Group_{c,b})$ 最大值，則所對應之第 c 層級中兩郵件群集形成聯集，並歸屬於第 c+1 層級之郵件群集 $Group_{c+1,q}$，如公式(3.34)所示。

$$Group_{c+1,a} = \left\{ Group_{c,a} \cup Group_{c,b} \middle| D(Group_{c,a}, Group_{c,b}) \text{ is Max}, \forall b \right\}$$
$$\text{and } Group_{c+1,\bullet} = Group_{c+1,1} \cup Group_{c+1,2} \cup \cdots \cup Group_{c+1,a} \cup \cdots, \text{ for all a} \quad (3.34)$$

Step (D-6)：建立完整層級群集樹與群集關聯

本法則乃以層級群集樹將使用者之郵件建立關聯結構，並於層級群集樹中擷取某一層級之郵件群集作為個人化郵件類別。故本法則乃反覆進行上述之步驟(D-4)與步驟(D-5)至所有郵件群集合併為同一群集，即完成層級群集樹，並彙整為層級群集樹矩陣 HCTree，使用者透過層級群集樹矩陣 HCTree 擷取所需層級，以作為個人化郵件類別 Category，以下乃詳細說明各步驟：

1. **郵件群集距離計算**：乃以步驟(D-4)之平均連結聚合法群集距離計算法則，以取得第 c 層級中各群集間郵件群集距離 $D(Group_{c,a}, Group_{c,b})$。
2. **合併相似郵件群集**：乃以步驟(D-5)之法則判斷兩群集間郵件群集距離 $D(Group_{c,a}, Group_{c,b})$，並合併距離相近之群集為第 c+1 層級之郵件群集 $Group_{c+1,q}$。
3. **建立完整層級群集樹矩陣**：反覆進行步驟(D-4)與步驟(D-5)至所有郵件群集合併為同一群集後，即完成層級群集樹，並彙整為層級群集樹矩陣 HCTree，如公式(3.35)

所示。

4. **個人化郵件類別擷取：**完成層級群集樹矩陣 HCTree 後，使用者乃設定矩陣分割座標點 HC(Y)，以擷取使用者所需之 HC(Y)層級之郵件群集，以作為個人化郵件類別矩陣 Category ，如公式(3.36)所示，其中，各矩陣分割座標點 HC(Y)擷取層級如表 3.7 所示。

$$\text{HCTree} = \begin{bmatrix} \text{Group}_{1,1} & \text{Group}_{1,2} & \text{Group}_{1,3} & \cdots & \text{Group}_{1,v} \\ \text{Group}_{2,1} & \text{Group}_{2,2} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \text{Group}_{c,1} & \text{Group}_{c,2} & \cdots & \cdots & 0 \\ \text{Group}_{\text{Top},1} & 0 & 0 & \cdots & 0 \end{bmatrix} \tag{3.35}$$

$$\text{Category} = \left[\text{Group}_{\text{HC(Y)},1} \quad \text{Group}_{\text{HC(Y)},2} \quad \cdots \quad \text{Group}_{\text{HC(Y)},q} \cdots \quad 0 \right] \tag{3.36}$$
$$\text{and } 1 \le \text{HC(Y)} \le \textit{NTop} \text{ and } \text{HC(Y)} \in \{ |\text{X}| : \text{X} \in \text{Z} \}$$

表 3.7、分割座標點與各層級之郵件群彙整表

分割座標點	擷取層級	郵件群集數	群集說明
HC(Y) = 1	層級 1	N(SM)	使用者寄發信件為分群依據之群集
HC(Y) = c	層級 c	$N(Group_c)$	相似內容之郵件所形成郵件群集
HC(Y) = Top	NTop	1	使用者所有信件（即使用者收件匣）

透過上述步驟計算後，即可完成層級群集樹之建立，並擷取使用者所需之層級，當中，該層級所含郵件群集即為使用者所需個人化郵件類別。

Step (D-7)：類別名稱詞彙庫建立

類別名稱推論乃透過建立類別名稱詞彙庫，並透過上述之郵件群集知識本體分析兩者關聯，再由類別名稱詞彙庫中提取關連較高之詞彙作為類別名稱，是故本研究乃參考 Atkinson 等人（2009）以自然語言處理（Natural Language Processing；NLP）方式擷取領域文件詞彙，並建立類別名稱詞彙庫。其中，Atkinson 等人（2009）乃標記文件詞彙屬性與段落，再以潛在語意分析（Latent Semantic Analysis；LSA）分析詞彙關聯與重要

性，故本步驟乃先行拆解領域文件成二至六字詞，並根據詞彙所在段落進行屬性標記，再以該段落詞彙進行潛在語意分析，再以標記屬性與重要性彙整其詞彙關聯組合，此詞彙關聯組合乃表達各筆詞彙之上下層級含意關聯，如：「運動」詞彙包含「籃球」、「羽球」此種關聯，取得詞彙關聯組合後則整理成類別名稱詞彙庫，作為擷取類別名稱所需詞彙集。以下乃詳細說明各步驟：

1. **領域文件內容詞彙拆解**：乃先行歸類各領域文件之類別，並由領域文件中文章段落分別拆解成二至六字詞，以取得文章段落詞彙集合 $ParW_{w,e}$。
2. **潛在語意分析詞彙出現機率**：取得文章段落詞彙集合 $ParW_{w,e}$ 後，乃以段落詞彙 $PW_{w,e,m}$ 與文句中銜接之前一筆詞彙 $PW_{w,e,m-1}$ 分析條件機率（即第 m-1 筆詞彙出現於文句時，第 m 筆詞彙出現之機率），當中，段落詞彙集合 $ParW_{w,e}$ 中排列順序乃以詞彙在文句中出現位置所擷取，故乃以第 m-1 筆詞彙於段落中出現次數 $N(PW_{w,e,m-1})$，與兩詞彙（即第 m-1 筆與第 m 筆詞彙）一同出現次數 $N(PW_{w,e,m-1},PW_{w,e,m})$ 進行計算，以獲得兩詞彙之出現機率 $P(PW_{w,e,m}|PW_{w,e,m-1})$，如公式(3.37)所示。
3. **段落關聯機率矩陣建立**：乃以詞彙之出現機率 $P(PW_{w,e,m}|PW_{w,e,m-1})$ 彙整為該段落之關聯機率矩陣 $P(PW_{w,e})$，當中，第一筆詞彙因前無連接詞彙，故無法計算條件機率，因此乃以第一筆詞彙於段落中出現機率代替，並以此段落關聯機率矩陣 $P(PW_{w,e})$，於後續進行相似度分析，如公式(3.38)所示。
4. **段落關聯機率矩陣計算段落相似度**：本步驟乃以段落關聯機率矩陣 $P(PW_{w,e})$ 計算矩陣範數 $\| PW_{w,e} \|$（如公式(3.39)所示），並以此矩陣範數為依據計算各段落之餘弦相似度，作為各段落相似度 $Sim(PW_{w,a},PW_{w,b})$（如公式(3.40)所示），當中，所得結果越大則兩矩陣越相似。
5. **彙整詞彙關聯為名稱詞彙集**：為整併領域文件內所含詞彙相似之段落，於此將段落相似度 $Sim(PW_{w,a},PW_{w,b})$ 較高兩段落分析相同詞彙，接著合併相同詞彙，並保留詞彙前後關聯，以成為名稱詞彙集 $NamW_{w,e}$（所得結果如表 3.8 所示），以作為類別名稱命名時所需類別名稱詞彙庫。

$$ParW_{w,e} = \{PW_{w,e,1}, ParW_{w,e,2}, PW_{w,e,3}, ..., PW_{w,e,m}\}$$
$$P(PW_{w,e,m}|PW_{w,e,m-1}) = \frac{N(PW_{w,e,m-1}, PW_{w,e,m})}{N(PW_{w,e,m-1})} \quad (3.37)$$

$$P(PW_{w,e}) = \left[P(PW_{w,e,1} \middle| PW_{w,e,0}) \quad P(PW_{w,e,2} \middle| PW_{w,e,1}) \quad \ldots \quad P(PW_{w,n,m} \middle| PW_{w,n,m-1}) \quad \ldots \right] \quad (3.38)$$

$$and \quad P(PW_{w,e,1} \middle| PW_{w,e,0}) = \frac{N(PW_{w,e,m})}{N(ParW_{w,e})}$$

$$\left\| PW_{w,e} \right\| = \sqrt{\sum_{all\ m} \left| P(PW_{w,e,m} \middle| PW_{w,e,m-1}) \right|^2} \quad (3.39)$$

$$Sim(PW_{w,a}, PW_{w,b}) = \frac{P(PW_{w,a}) \times P(PW_{w,b})}{\left\| PW_{w,a} \right\| \times \left\| PW_{w,b} \right\|} \text{ where } a > b \text{ for all a, b} \quad (3.40)$$

表 3.8、名稱詞彙集與整併條件彙整表

名稱詞彙集	段落	層級 1	層級 2	層級 3	…	層級 x	整併條件式
$NamW_{w,e}$	$ParW_{w,a}$	$PW_{w,a,1}$	$PW_{w,a,2}$		…	$PW_{w,a,x}$	$Max(Sim(PW_{w,a}, PW_{w,b}))$
	$ParW_{w,b}$		$PW_{w,b,2}$	$PW_{w,b,3}$	…		

Step (D-8)：郵件群集與名稱詞彙集之線性組合建立

完成類別名稱詞彙庫建立後，為能從中挑選適當詞彙作為類別名稱，故須分析郵件群集 $Group_{HC(Y),q}$ 與名稱詞彙集 $NamW_{w,e}$ 相關性。本研究參考 Li 和 Shawe-Taylor（2007）以典型相關分析法（Kernel Canonical Correlation Analysis；KCCA）分析兩種不同語言文件之方法，解析郵件群集 $Group_{HC(Y),q}$ 與名稱詞彙集 $NamW_{w,e}$ 相關性，當中，本步驟乃將郵件群集 $Group_{HC(Y),q}$ 與名稱詞彙集 $NamW_{w,e}$ 視為兩種不同文件進行分析。郵件群集 $Group_{HC(Y),q}$ 乃以「階段一、郵件主成份計算」所得各郵件特徵向量矩陣 $EV_{q,n}$ 作為分析數據（如公式(3.41)所示），名稱詞彙集 $NamW_{w,e}$ 則以段落關聯機率矩陣 $P(PW_{w,n})$ 作為分析數據（如公式(3.38)所示），接著以此分析數據建立兩組資料之線性組合，分別為郵件群集線性組合 $F(Group_{q,n})$ 與名稱詞彙集線性組合 $F(NamW_{w,e})$，如公式(3.42)所示，且各線性組合之加權係數 $\alpha_{q,n,p}$、$\beta_{w,e,f}$ 加總皆為 1。

$$EV_{q,n} = \begin{bmatrix} EV_{q,n,1} \\ EV_{q,n,2} \\ \vdots \\ EV_{q,n,p} \end{bmatrix} \quad (3.41)$$

$$
\begin{aligned}
&F(Group_{q,n}) = \alpha_{q,n,1} \times EV_{q,n,1} + \alpha_{q,n,2} \times EV_{q,2} + \cdots + \alpha_{q,n,p} \times EV_{q,n,p} \\
&F(NamW_{w,e}) = \beta_{w,e,1} \times P(PW_{w,e,1} \mid PW_{w,e,0}) + \beta_{w,e,2} \times P(PW_{w,e,2} \mid PW_{w,e,1}) + \cdots \\
&\qquad\qquad + \beta_{w,n,f} \times P(PW_{w,e,m} \mid PW_{w,e,m-1}) \\
&\text{and} \sum_{\text{all p}} \alpha_{q,n,p} = 1 \text{ and } \sum_{\text{all f}} \beta_{w,e,f} = 1
\end{aligned}
\tag{3.42}
$$

Step (D-9)：郵件群集與名稱詞彙集之共變異矩陣建立

本步驟乃先計算郵件群集線性組合 $F(Group_{q,n})$與名稱詞彙集線性組合 $F(NamW_{w,e})$之共變異數，以獲得郵件群集共變異數 $Cov(Group_{q,(a|b)})$（如公式(3.43)所示）與名稱詞彙共變異數 $Cov(NamW_{w,(a|b)})$（如公式(3.44)所示），接著，乃彙整為郵件群集共變異矩陣 $Cov[G_{q,\bullet}]$（如公式(3.45)所示）及名稱詞彙共變異矩陣 $Cov[N_{w,\bullet}]$（如公式(3.46)所示）。

$$
Cov(Group_{q,(a|b)}) = \frac{\sum_{all\ n} (EV_{q,a,p} - \overline{GX}_{q,a}) \times (EV_{q,b,p} - \overline{GX}_{q,b})}{N(EV_{q,n})} \quad \text{and} \quad a \le b
\tag{3.43}
$$

$$
\begin{aligned}
&Cov(NamW_{w,a|b}) \\
&= \frac{\sum_{\text{all n}} (P(PW_{w,a,m} \mid PW_{w,a,m-1}) - \overline{NWX}_{w,a}) \times (P(PW_{w,b,m} \mid PW_{w,b,m-1}) - \overline{NWX}_{w,b})}{N(P(PW_{w,e}))} \\
&\text{and} \quad a \le b
\end{aligned}
\tag{3.44}
$$

$$
Cov[G_{q,\bullet}] = \begin{bmatrix}
Cov(Group_{q,(1|1)}) & Cov(Group_{q,(2|1)}) & \cdots & Cov(Group_{q,(a|1)}) & \cdots \\
Cov(Group_{q,(1|2)}) & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
Cov(Group_{q,(1|b)}) & \cdots & \cdots & Cov(Group_{q,(a|b)}) & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots
\end{bmatrix}
\tag{3.45}
$$

$$\text{and } Cov(Group_{q,(a|b)}) = Cov(Group_{q,(b|a)})$$

$$
Cov[N_{w,\bullet}] = \begin{bmatrix}
Cov(NamW_{w,1|1}) & Cov(NamW_{w,2|1}) & \cdots & Cov(NamW_{w,a|1}) & \cdots \\
Cov(NamW_{w,1|2}) & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
Cov(NamW_{w,1|b}) & \cdots & \cdots & Cov(NamW_{w,(a|b)}) & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots
\end{bmatrix}
\tag{3.46}
$$

$$\text{and } Cov(NamW_{w,(a|b)}) = Cov(NamW_{w,(b|a)})$$

Step (D-10)：郵件群集與名稱詞彙集之相關性係數計算

典型相關分析法（Kernel Canonical Correlation Analysis；KCCA）乃以相關性係數作為兩分析資料之相似判斷依據，其中，相關性係數 $\eta(G_q,N_w)$乃以郵件群集線性組合 $F(Group_{q,n})$與名稱詞彙集線性組合 $F(NamW_{w,e})$之變異數（如公式(3.48)所示），與兩線性組合之共變異數 $Cov(G_q^F,N_w^F)$計算而得（如公式(3.49)所示），並以計算所得最大值作為郵件群集與名稱詞彙集相關性係數 $\eta(G_q,N_w)$（如公式(3.50)所示），此相關係數可作為郵件群集與名稱詞彙集之關聯判斷依據。

$$\begin{aligned} Var\left(F(Group_{q,n})\right) &= [\alpha]_{q,n} \times Cov\left[G_q\right] \\ Var\left(F(NamW_{w,e})\right) &= [\beta]_{w,e} \times Cov\left[N_w\right] \end{aligned} \tag{3.48}$$

$$Cov(G_q^F,N_w^F) = Var\left(F(Group_{q,n})\right) \times Var\left(F(NamW_{w,e})\right) \tag{3.49}$$

$$\eta(G_q,N_w) = Max\left(\frac{Cov(G_q^F,N_w^F)}{\sqrt{Var\left(F(Group_{q,n})\right)^2 \times Var\left(F(NamW_{w,e})\right)^2}}\right) \tag{3.50}$$

Step (D-11)：郵件類別名稱擷取

郵件群集 $Group_{HC(Y),q}$ 之類別名稱乃由相關係數 $\eta(G_q,N_w)$ 較高之名稱詞彙集 $NamW_{w,e}$ 擷取詞彙 $PW_{w,e,m}$ 而得。因此，為能使名稱詞彙集 $NamW_{w,e}$ 中所擷取詞彙 $PW_{w,e,m}$ 具高度郵件內容解釋性（即能完整表達群集內郵件所述內容之名稱），故本步驟乃擷取名稱詞彙集 $NamW_{w,e}$ 中最高層級之詞彙 $PW_{w,e,m}$ 作為類別名稱，並彙整如表 3.9，其中，名稱詞彙集 $NamW_{w,e}$ 所含各詞彙 $PW_{w,e,m}$ 層級乃以詞彙與詞彙間關聯進行劃分而得，因此最高層級之詞彙 $PW_{w,e,m}$ 亦具有較高詞彙解釋性，故能作為群集內郵件所述內容之代表性名稱，以作為個人化郵件類別名稱。

表 3.9、郵件類別名稱與擷取條件彙整表

郵件群集	名稱詞彙集	類別名稱	整併條件式
$Group_{HC(Y),1}$	$NamW_c$	$PW_{c,a,1}$	$Max\left(\eta(G_1,N_c)\right)$
$Group_{HC(Y),2}$	$NamW_h$	$PW_{h,a,1}$	$Max\left(\eta(G_2,N_h)\right)$
…	…	…	…
$Group_{HC(Y),q}$	$NamW_w$	$PW_{w,a,1}$	$Max\left(\eta(G_q,N_w)\right)$

綜上所述，本階段乃以「階段一、郵件主成份計算」所得之郵件特徵詞彙作為分群依據，再以平均連結聚合法計算郵件群集合併條件並彙整為層級群集樹，達到階層式分

群效果。為能賦予郵件群集具具高度郵件內容解釋性之名稱，以作為符合使用者所需之郵件類別，因此，本階段乃先以領域文件建立名稱詞彙集，並彙整為類別名稱詞彙庫，以作為郵件類別名稱挑選依據，再以典型相關分析法（Kernel Canonical Correlation Analysis；KCCA）挑選與郵件群集具高度相似性之名稱詞彙集，從中尋得符合群集內郵件所述內容之名稱，以作為類別名稱，達到郵件類別自動推論與命名之效果，並滿足各使用者所需不同需求之個人化郵件類別。

3.3 結論

由於私人郵件內容隱含使用者個人生活經驗與專業知識，多數分類技術乃以特定領域郵件為分析對象，而無法針對使用者個人量身制定所需郵件類別。因此本研究針對私人郵件進行解析並建立一套個人化郵件類別推論模式，並參考 Asuncion 等人（2011）所提之 LDA 模式（Latent Dirichlet Allocation）中語意詞彙擷取技術，以制定語意擷取郵件關鍵字模組，達到郵件語意詞彙擷取效果。此外，LDA 模式（Latent Dirichlet Allocation）中需假設郵件主題進行語意詞彙計算，故本研究亦參考 Asuncion 等人（2011）以吉布斯抽樣方式間接取得主題近似值方法，並發展為本研究所需郵件主題抽樣模擬，進而完成郵件語意關鍵字擷取。

於個人化郵件類別推論模組中，為能取得郵件中據使用者特徵之詞彙，本研究乃參考 Šmídl 與 Quinn（2007）所提之主成份分析為基礎制定特徵之詞彙擷取法則，且再根據郵件中特徵詞彙與本研究所發展之分群法則進行階層式郵件分群，以形成郵件群集樹並根據使用者需求擷取郵件群集。此外為能賦予個群集專屬名稱作為使用者個人化郵件類別，故本研究乃參考 Atkinson 等人（2009）以自然語言處理（Natural Language Processing；NLP）方式擷取領域文件中據含意之詞彙作為類別名稱詞彙，並彙整為類別名稱詞彙庫，再參考 Li 和 Shawe-Taylor（2007）所提典型相關分析法（Kernel Canonical Correlation Analysis；KCCA）分析郵件群集與類別名稱詞彙庫兩者間關係，進而從類別名稱詞彙庫中挑選出郵件群集中代表性名稱作為使用者個人化郵件類別。

根據上述，本研究除了參考眾多研究延伸並發展語意詞彙、郵件特徵與類別名稱分析等方法論外亦自行發展分群法則進行階層式郵件分群，本研究彙整如表 3.10。

表 3.10、參考文獻延伸與本研究發展彙整表

探索主題	發展方法	演算法	參考文獻	本研究發展與延伸
使用者特質探勘主題	郵件語意詞彙擷取模組	Latent Dirichlet Allocation 法則（應用既有研究方法）	Asuncion 等人（2011）	● 轉化文件解析為郵件解析 ● 擷取語意詞彙解析法則
	郵件主成份計算法則	主成份分析（應用既有研究方法）	Šmídl 與 Quinn（2007）	● 加入主成分篩選法則 ● 針對語意詞彙解析語意關聯
郵件分群主題	郵件收發類型區分法則	啓發式演算法（創新研究方法）	--	● 以使用者寄發信件為基礎發展後續分群法則
	郵件相似計算法則	平均連結聚合法（應用既有研究方法）	張云濤與龔玲（2007）	● 結合詞彙語意延伸相似度計算
	郵件群集樹建立法則	階層式分群（應用既有研究方法）	張云濤與龔玲（2007）	● 建立使用者群集擷取法則 ● 以使用者寄發信件為基礎建立郵件群集樹
類別名稱推導主題	類別名稱詞彙庫建立法則	自然語言處理（應用既有研究方法）	Atkinson 等人（2009）	● 將既有法則加入詞彙集合彙整法則形成類別名稱詞彙庫
	代表性類別名稱推論法則	典型相關分析法（應用既有研究方法）	Li 和 Shawe-Taylor（2007）	● 擷取關聯分析法則部分 ● 加入詞彙關聯解析，並擷取代表性名稱

第四章、系統架構規劃

針對前一章節所發展之方法論，本研究乃開發一套個人化郵件類別推論系統，以確認方法論與模式之可行性。此系統之功能重點乃透過使用者上傳私人郵件，系統管理者先行設定系統參數與建立類別名稱詞彙庫，進而執行郵件語意關鍵字擷取、個人化郵件類別推論等主要模組。本章即針對本研究所提之「個人化郵件類別推論系統」，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個人化郵件類別推論系統之流程架構

本研究所開發之「個人化郵件類別推論系統」依其運作流程可分為「私人郵件上傳」、「領域文件上傳」、「郵件語意關鍵字擷取」、「類別名稱詞彙庫建立」及「個人化郵件類別推論」等五大階段，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。

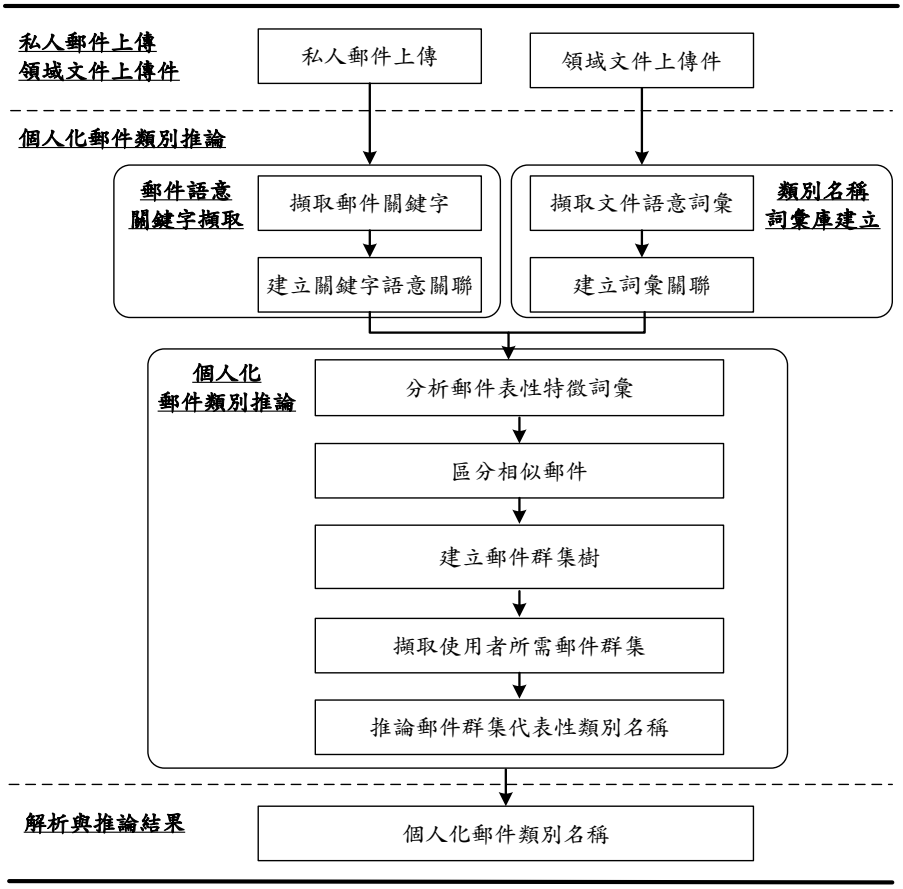

圖 4.1、個人化郵件類別推論系統之流程架構

- 私人郵件上傳

使用者可將尚未進行郵件語意關鍵字擷取與類別區分之私人郵件新增至系統，系統即擷取郵件內容，並由後續「郵件語意關鍵字擷取」及「個人化郵件類別推論」等模組分析並判定該目標郵件之代表性特徵詞彙與郵件群集並推論適當郵件類別名稱。

- 領域文件上傳

使用者將尚未進行類別名稱詞彙擷取之領域文件新增至系統，並由後續「名稱詞彙庫建立」功能分析並擷取該領域文件之代表性詞彙，並將領域文件之詞彙進行語意串聯，以作為類別名稱詞彙庫。

- 郵件語意關鍵字擷取

此部份乃以使用者所上傳之私人郵件擷取關鍵字，並進行關鍵字語意主題分析以確實擷取具有語言含意性之詞彙。因此於此部份可在細分為「郵件主題詞彙集合建立」、「詞彙後驗機率近似值計算」等功能。

- 類別名稱詞彙庫建立

此部分主要功能為解析領域文件中詞彙並進行潛在語意分析，以標記屬性與重要性彙整其詞彙關聯組合，此詞彙關聯組合乃表達各筆詞彙之上下層級關聯。於取得詞彙關聯組合後則整理成類別名稱詞彙庫，作為擷取類別名稱所需詞彙集。

- 個人化郵件類別推論

此部分乃根據「郵件語意關鍵字擷取模組」之語意詞彙分析郵件代表性特徵，並透過分析郵件特徵方式將郵件以階層化方式分群並推論各群集之名稱，以作為使用者個人化郵件類別。此部份可分為「郵件主成份計算」、「郵件關聯程度與類別推論」等功能。

4.2 系統功能架構

本研究所建置個人化郵件類別推論系統乃架構於網際網路環境下。使用者可透過網際網路登入本系統，並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. 可新增上傳未進行特徵擷取與類別區分之私人郵件至系統
2. 可瀏覽符合使用者上傳之所有私人郵件
3. 可瀏覽符合使用者所上傳郵件經分析後所得個人化郵件類別推論結果

系統管理者

1. 可上傳未擷取文件詞彙之領域文件至系統
2. 可瀏覽/編輯系統資料庫內之所有電子郵件
3. 可查詢、新增、修改或刪除領域文件內容
4. 可修改/查詢系統參數與門檻值
5. 可執行郵件語意關鍵字擷取
6. 可執行個人化郵件類別推論
7. 可執行類別名稱詞彙庫建立

本系統所開發之重點模組共有「郵件資料維護模組」、「郵件語意關鍵字擷取模組」、「個人化郵件類別推論模組」、「類別名稱詞彙庫維護模組」及「系統參數設定」等五大模組；圖4.2即表示個人化郵件類別推論系統之核心模組架構。

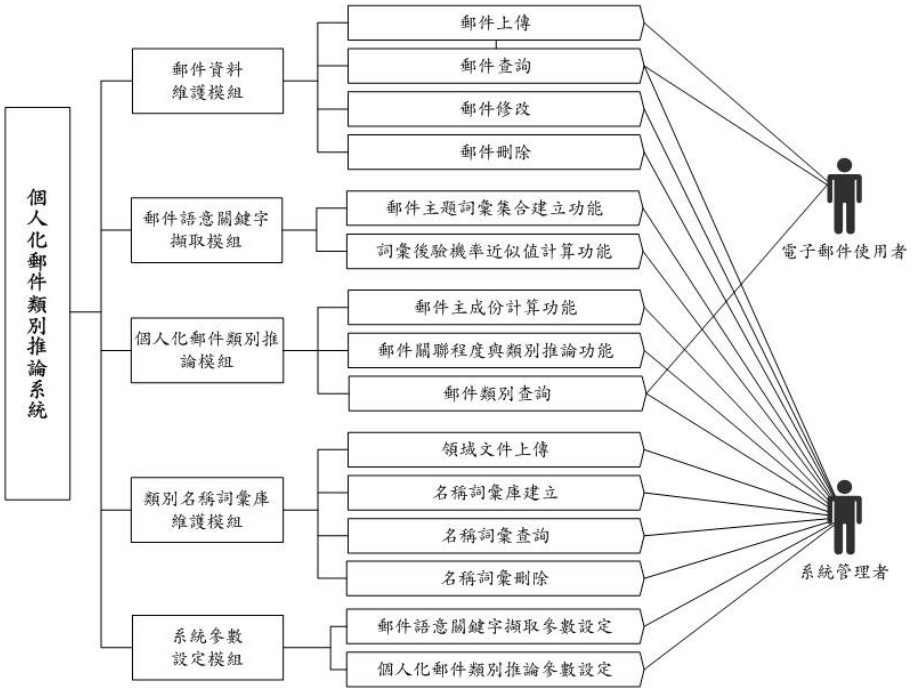

圖 4.2、個人化郵件類別推論系統之功能架構

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

- 郵件上傳：提供系統管理者與一般使用者將未分類之郵件資料匯入並維護於系統資料表中。
- 郵件查詢：提供系統管理者與一般使用者進行查詢已維護之電子郵件資料。
- 郵件修改：提供系統管理者修改錯誤之電子郵件資料。
- 郵件刪除：提供系統管理者刪除錯誤之電子郵件資料。

(二) 郵件語意關鍵字擷取模組

- 郵件主題詞彙集合建立功能：系統管理者可輸入查詢字串、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解析電子郵件，即可透過郵件中所含主題分析特定詞彙，並作為該封郵件類別推論所需詞彙。
- 詞彙後驗機率近似值計算功能：系統管理者可輸入查詢字串、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解析電子郵件，即可分析郵件中內容主題發生機率，並透過此主題比例強化詞彙主題關聯性。

(三) 個人化郵件類別推論模組

- 郵件主成份計算功能：系統管理者可輸入查詢字串、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解析電子郵件，即可根據詞彙與郵件關聯篩選出該郵件最具代表性詞彙作為郵件分群所需之郵件特徵。
- 郵件關聯程度與類別推論功能：主要劃分為五項步驟，分別為：「彙整郵件特徵詞彙」、「區分相似郵件」、「建立郵件群集樹」、「分析群集代表詞彙」、「推論郵件類別名稱」等步驟，系統管理者根據系統提示依序執行五項步驟即可根據郵件相似度進行郵件階層式分群，並取得群集中最具代表性與解釋性詞彙作為群集名稱，即完成個人化郵件類別推論。
- 郵件類別查詢：提供系統管理者與一般使用者進行查詢系統推論之個人化郵件類別名稱與郵件群集資料。

(四) 類別名稱詞彙庫維護模組

- 領域文件上傳：提供系統管理者將領域文件資料匯入並維護於系統資料表中。
- 名稱詞彙庫建立：系統管理者可輸入查詢字串、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解析領域文件，即可透過系統提示步驟建立名稱詞彙庫
- 名稱詞彙查詢：提供系統管理者進行查詢已維護之名稱詞彙資料。
- 名稱詞彙刪除：提供系統管理者刪除錯誤之名稱詞彙資料。

(六) 系統參數設定

- 郵件語意關鍵字擷取參數設定：提供權限內使用者進行修改系統分析參數，以進而提升「郵件主題詞彙集合建立」與「詞彙後驗機率近似值計算」兩功能郵件語意詞彙分析準確性。
- 個人化郵件類別推論參數設定：提供權限內使用者進行修改系統分析參數，以進而提升「郵件主成份計算」與「郵件關聯程度與類別推論」功能兩者郵件特徵與個人化郵件等分析準確性。

本系統之使用者可分為一般使用者與系統管理者，並依權限而有不同執行權力。針對一般使用者可執行郵件上傳、郵件查詢及郵件類別查詢等功能（如圖4.3之一般使用者所示），而系統管理者則可進行郵件資料維護、系統參數設定、郵件語意關鍵字擷取、個人化郵件類別推論與類別名稱詞彙庫維護等（如圖4.3之系統管理者所示）。當使用者上傳未解析之電子郵件後，系統管理者乃需完成參數設定以及領域文件上傳，待設定與上傳完畢，並建立類別名稱詞彙庫後，系統管理者即可執行郵件語意關鍵字擷取模組以取得該郵件中具有語言含意性之詞彙。系統管理者再藉由個人化郵件類別推論模組中郵件主成份計算功能分析郵件中之語意詞彙，並取得郵件中代表性郵件特徵詞彙，以作為郵件分群之分群依據，接著，透過郵件關聯程度與類別推論功能將系統中已完成主成份計算功能之郵件進行階層式分群，且根據分群結果中各郵件與名稱詞彙庫建立詞彙語意關聯，進而取得各群集中代表性名稱詞彙，即可完成個人化郵件類別推論，最後使用者即由個人化郵件類別推論模組之郵件類別查詢功能檢視系統之個人化郵件類別名稱推論與分群結果。

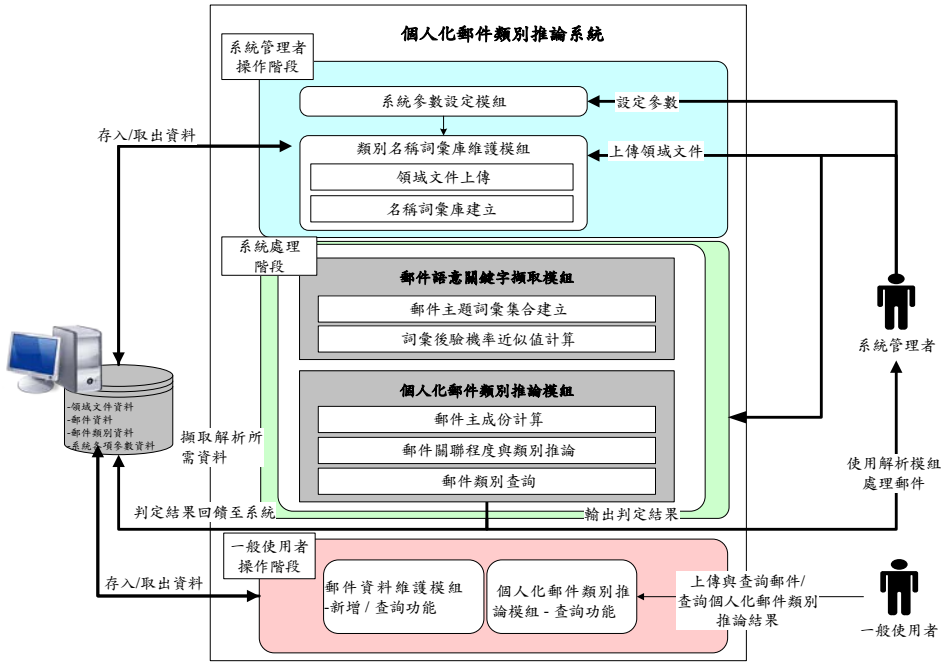

圖 4.3、個人化郵件類別推論系統運作架構

4.3 資料模式定義

本研究所發展之個人化郵件類別推論系統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，期使電子郵件管理、郵件語意關鍵字擷取與個人化郵件類別推論等任務可即時完成。依據系統運作之需要，將個人化郵件類別推論系統之資料分為「郵件內容資料」、「郵件聯絡人資料」、「語意詞彙資料」、「郵件類別資料」、「類別名稱詞彙資料」與「系統參數資料」等六大部分，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、郵件名稱、郵件上傳時間、郵件語言等資訊。
- 郵件文字內容資料表：紀錄目標郵件所擷取之所有文字資料，以作為郵件語意關鍵字擷取模組應用之基礎。
- 郵件附件檔案資料表：紀錄目標郵件所擷取之附件檔案名稱及副檔名，以作為郵件語意關鍵字擷取模組應用之基礎。

(二) 郵件聯絡人資料

此資料之目的乃記錄郵件中收件者與寄件者關聯資訊，以協助系統篩選出使用者所接收郵件進行分群；其所屬之資料表及其相關定義說明如下：

- 郵件聯絡關係資料表：紀錄目標郵件所擷取之聯絡人於該信件中所屬身份類型，如：寄件者、收件者、副本收件者...等類型，以作為個人化郵件類別推論模組應用之基礎。
- 聯絡人基本資料表：紀錄目標郵件所擷取之聯絡人資料，以作為個人化郵件類別推論模組應用之基礎。
- 聯絡人信箱資料表：紀錄目標郵件所擷取之電子信箱地址，以管理郵件寄件者多組電子信箱地址並區分郵件寄件者。

(三) 語意詞彙資料

此資料之目的乃記錄電子郵件透過郵件語意關鍵字擷取模組分析而得語意詞彙資料與郵件語意關鍵字擷取時所需相關資料，以有效進行郵件特徵詞彙擷取與個人化郵件類別名稱推論；其所屬之資料表及其相關定義說明如下：

- 郵件語意詞彙資料表：記錄電子郵件中所擷取之語意詞彙、詞彙編號、郵件特徵代表性等資訊，以作為個人化郵件類別模組應用之基礎。
- 詞彙主題關聯資料表：記錄語意詞彙與郵件主題兩者間關聯及關聯係數，以作為個人化郵件類別推論模組應用之基礎。
- 郵件主題資料表：記錄電子郵件中各主題之主題編號、主題名稱，以作為郵件語意關鍵字擷取模組應用之基礎。
- 郵件主題機率資料表：記錄電子郵件中各主題後驗機率，以作為郵件語意關鍵字擷取模組應用之基礎。
- 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，以作為郵件語意關鍵字擷取模組應用之基礎。

(四) 郵件類別資料

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：

- 郵件類別關聯資料表：乃維護所有個人化郵件類別推論模組中所判定結果之關聯，並供權限內使用者查詢郵件判定狀況與結果。
- 郵件類別基本資料表：乃維護各項郵件分類所屬分類之類別資料，如：類別名稱、類別層級等資料。

(五) 類別名稱詞彙資料

此資料之目的乃記錄知識文件資料內容與解析時所需關鍵資訊，以有效進行知識文件管理、問答解析與結構化摘要推論；其所屬之資料表及其相關定義說明如下：

- 領域文件基本資料表：記錄領域文件之基本資料，如領域文件編號、領域文件名稱、文件內容、領域文件所屬類別與領域文件上傳日期等資訊。
- 領域文件與類別名稱詞彙關聯資料表：記錄領域文件與類別名稱詞彙關聯，以作為名稱詞彙庫建立功能之應用基礎。
- 領域文件類別基本資料表：乃維護各領域文件所屬分類之類別資料，如：領域文件類別編號、領域文件類別名稱等資料。
- 類別名稱詞彙資料表：記錄透過領域文件擷取而得之類別名稱詞彙，以作為名稱詞彙庫建立功能之應用基礎。
- 領域文件段落資料表：記錄類別名稱詞彙於領域文件中所屬段落關聯。
- 類別名稱詞彙與段落關聯資料表：記錄名稱詞彙庫建立功能分析所得類別名稱詞彙間上下層級關聯，以作為個人化郵件類別推論模組應用之基礎。

(六) 系統參數資料

此資料之目的乃記錄系統參數之資料，如系統門檻值、權重值等資料，透過設定以有效提高郵件語意關鍵字擷取、個人化郵件類別推論之準確率；其所屬資料表及其相關定義如下：

- 系統參數資料表：記錄系統參數之數值，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，使本研究所發展之個人化郵件類別推論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。

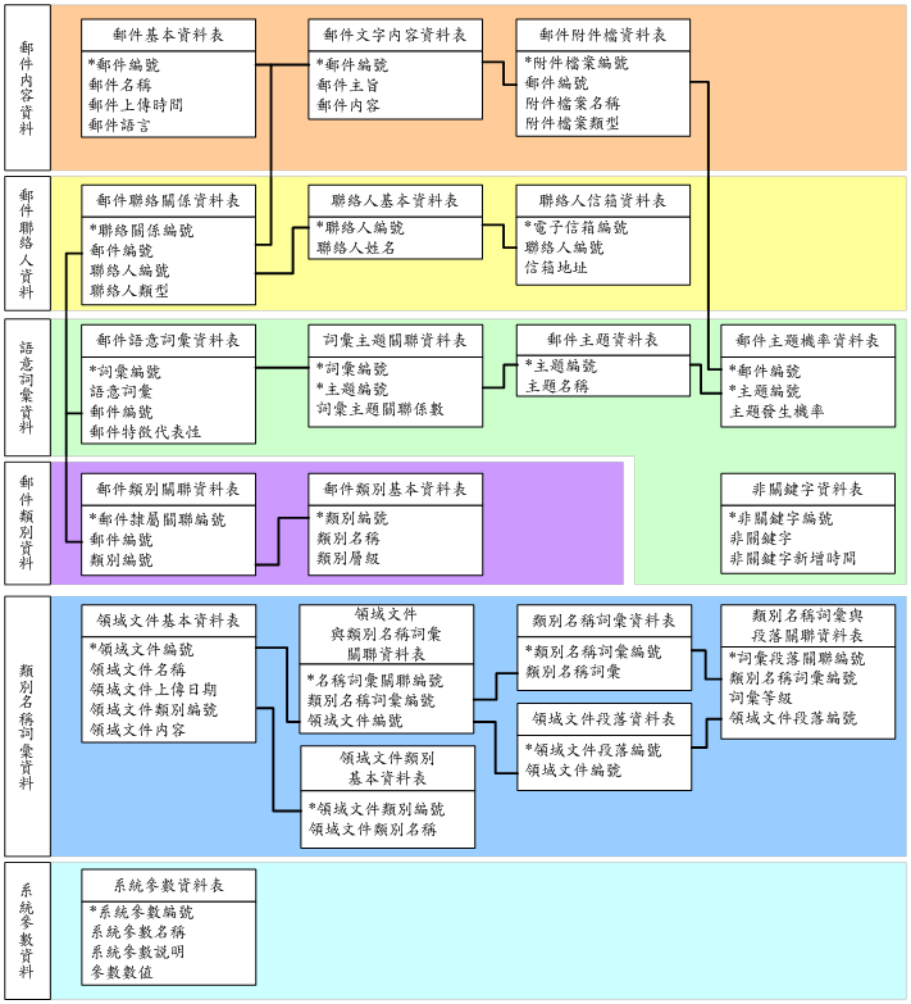

圖 4.4、個人化郵件類別推論系統之資料關聯

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」、「郵件語意關鍵字擷取模組」、「個人化郵件類別推論模組」、「類別名稱詞彙庫維護模組」及「系統參數設定」等五大模組，以下即說明各系統功能之流程規劃。

郵件資料維護模組

此模組可供權限內使用者上傳欲分析之郵件，並作解析與擷取，作為郵件類別判定模組之分析資料。此外，權限內使用者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；其流程設計概念如圖 4.5 所示。

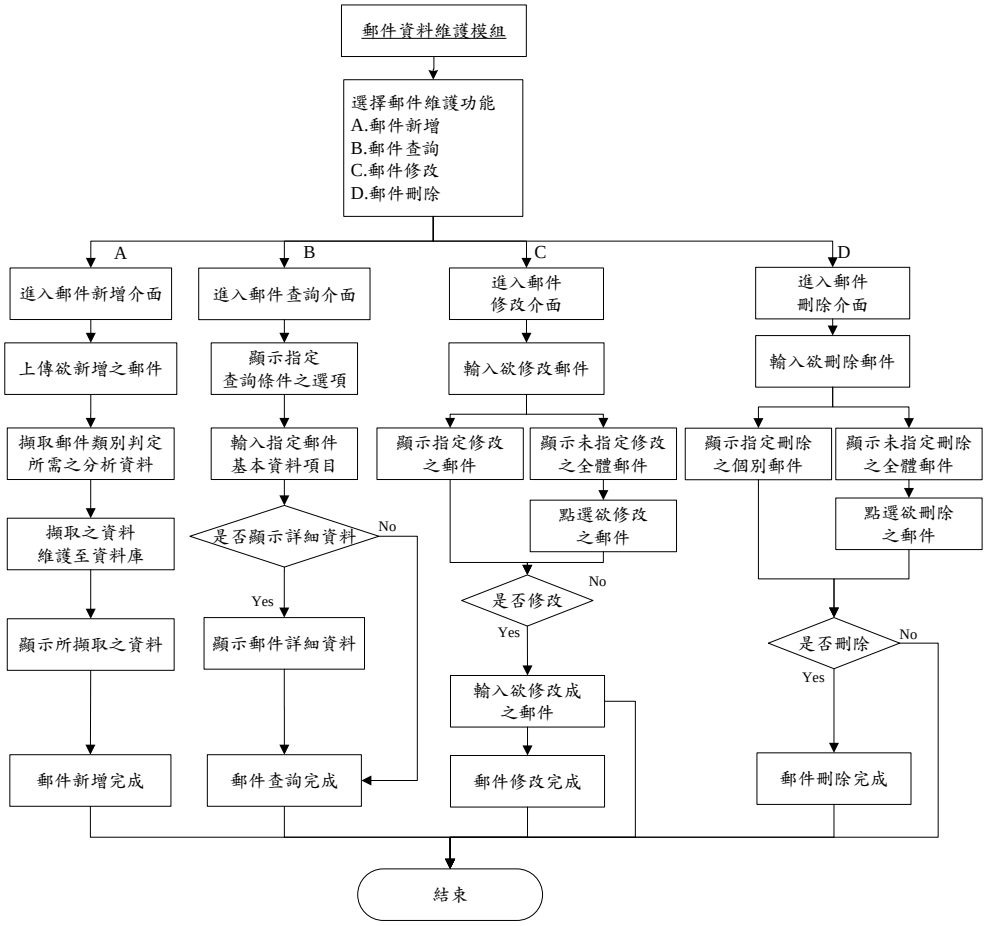

圖 4.5、「郵件資料維護模組」功能流程

郵件語意關鍵字擷取模組

本系統乃提供系統管理者進行「郵件主題詞彙集合建立」、「詞彙後驗機率近似值計算」等功能。於「郵件主題詞彙集合建立」中，首先系統管理者以搜尋字串尋得系統中欲分析之已上傳郵件並決定推論，系統則先取得該郵件資料進行斷詞並計算該詞彙之詞彙頻率與詞彙出現機率，取得詞彙之詞彙出現機率後與郵件中各主題之主題發生機率計算詞彙與主題關聯係數，接著，透過詞彙與主題關聯係數即可篩選出該郵件中所含語意詞彙。於「詞彙後驗機率近似值計算」中，系統管理者以搜尋字串尋得系統中欲分析之已上傳郵件並決定推論，系統則先取得該郵件之主題，並以隨機方式給定第一次抽樣之主題發生次數，並計算發生機率，再依此發生機率再次進行若干次抽樣後，根據系統參

數設定擷取若干組抽樣計算主題後驗機率近似值，並與各詞彙重新計算詞彙主題關聯係數，以獲得強化後詞彙主題關聯係數；其流程設計概念如圖 4.6 所示。

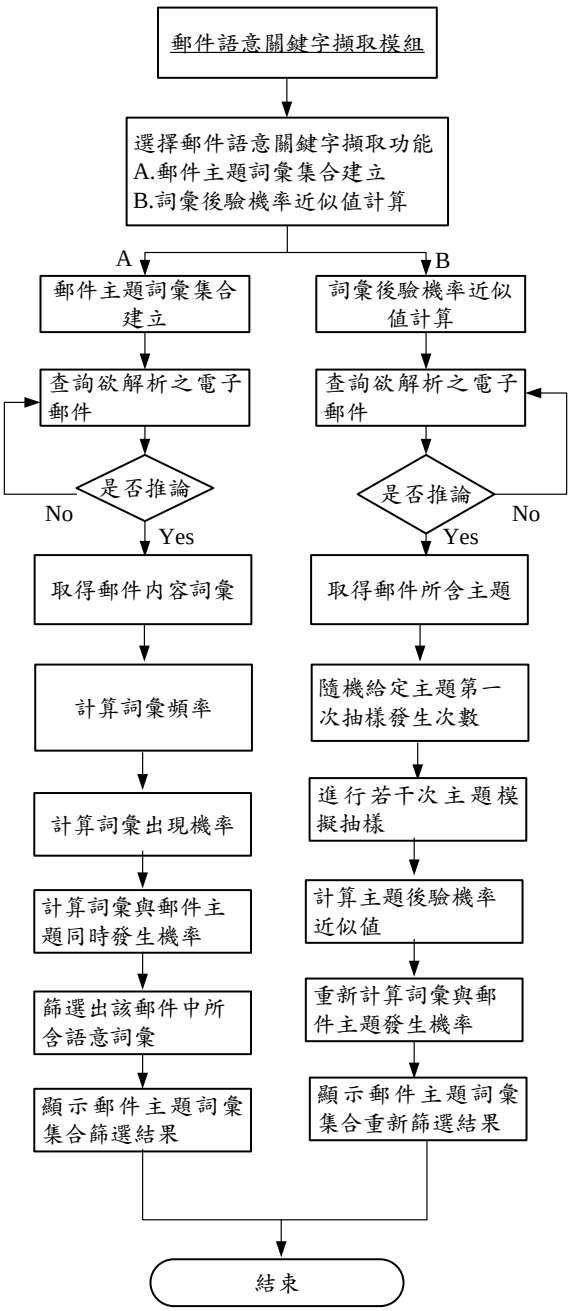

圖 4.6、「郵件語意關鍵字擷取模組」功能流程

個人化郵件類別推論模組

本系統乃提供系統管理者進行「郵件主成份計算」、「郵件關聯程度與類別推論」等功能，而一般使用者可使用「郵件類別查詢功能」功能。系統管理者進入「郵件主成份計算」，先尋得欲分析之郵件後，系統則根據該郵件之詞彙主題關聯係數計算詞彙共變異數矩陣，接著詞彙再根據詞彙共變異數矩陣中詞彙間關聯差異大小組合成特徵向量，

於取得郵件中各組特徵向量後，根據系統所制定之特徵解釋比例門檻值篩選出郵件中最具代表性特徵向量組合並彙整，進而取得郵件特徵詞彙。於「郵件關聯程度與類別推論」中，首先彙整使用者上傳之所有郵件與郵件中所含特徵詞彙，接著根據郵件中所含所有郵件特徵詞彙計算郵件相似性，並依據相似程度歸類郵件群集。系統再以郵件特徵詞彙反覆計算郵件相似距離並建立郵件群集樹，完成郵件群集樹建立後，系統於此步驟乃分析群集內中各郵件之郵件特徵詞彙，並根據特徵詞彙之詞彙關聯取得詞彙等級較高詞彙，最後，系統根據郵件群集中特徵詞彙關聯與郵件名稱詞庫進行關聯比對，以尋找郵件群集中代表性名稱。「郵件類別查詢」功能乃提供權限內使用者查詢系統已推論之個人化郵件類別與類別所含郵件；其流程設計概念如圖 4.7 所示。

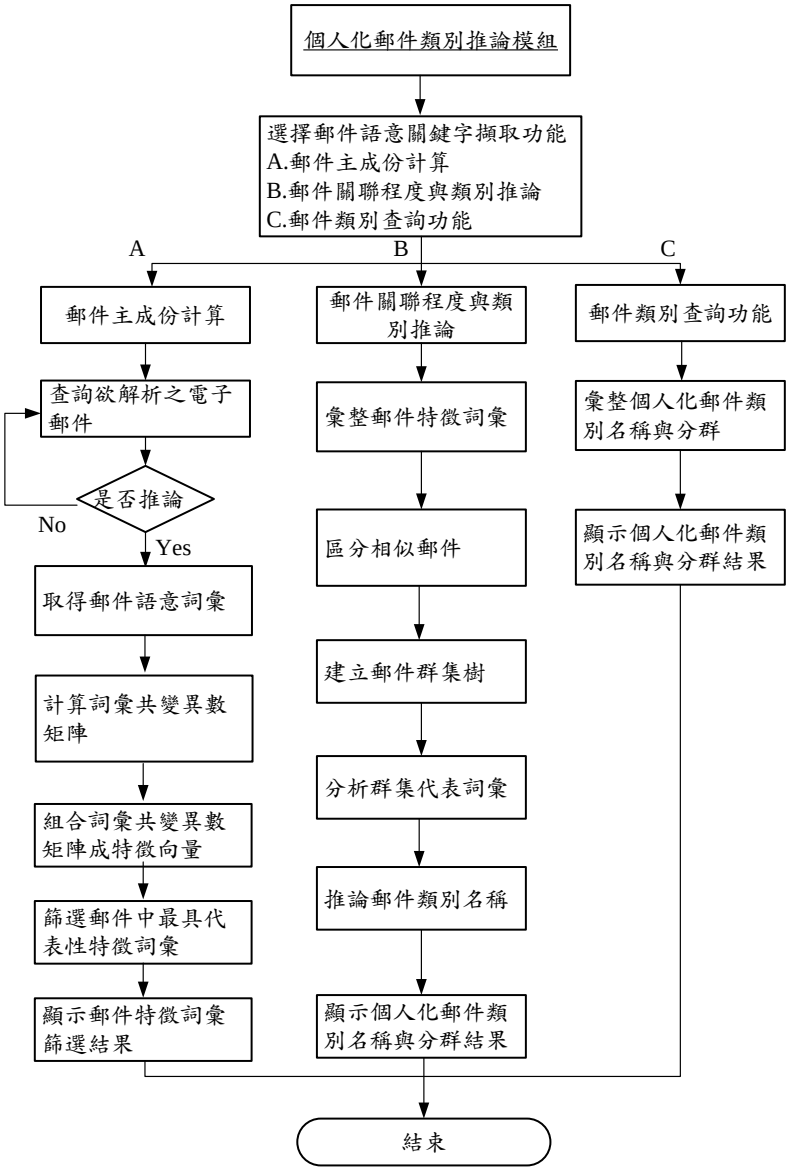

圖 4.7、「個人化郵件類別推論模組」功能流程

類別名稱詞彙庫維護模組

此模組乃提供系統管理者進行領域文件上傳、名稱詞彙庫建立、名稱詞彙查詢與名稱詞彙刪除等功能。「領域文件上傳」功能乃提供權限內使用者上傳領域文件。「名稱詞彙庫建立」功能乃解析領域文件中詞彙並進行潛在語意分析，以建立類別名稱詞彙庫。此外權限內使用者可透過「名稱詞彙查詢」、「名稱詞彙刪除」功能查詢或刪除系統內已彙整之名稱詞彙；其流程設計概念如圖 4.8 所示。

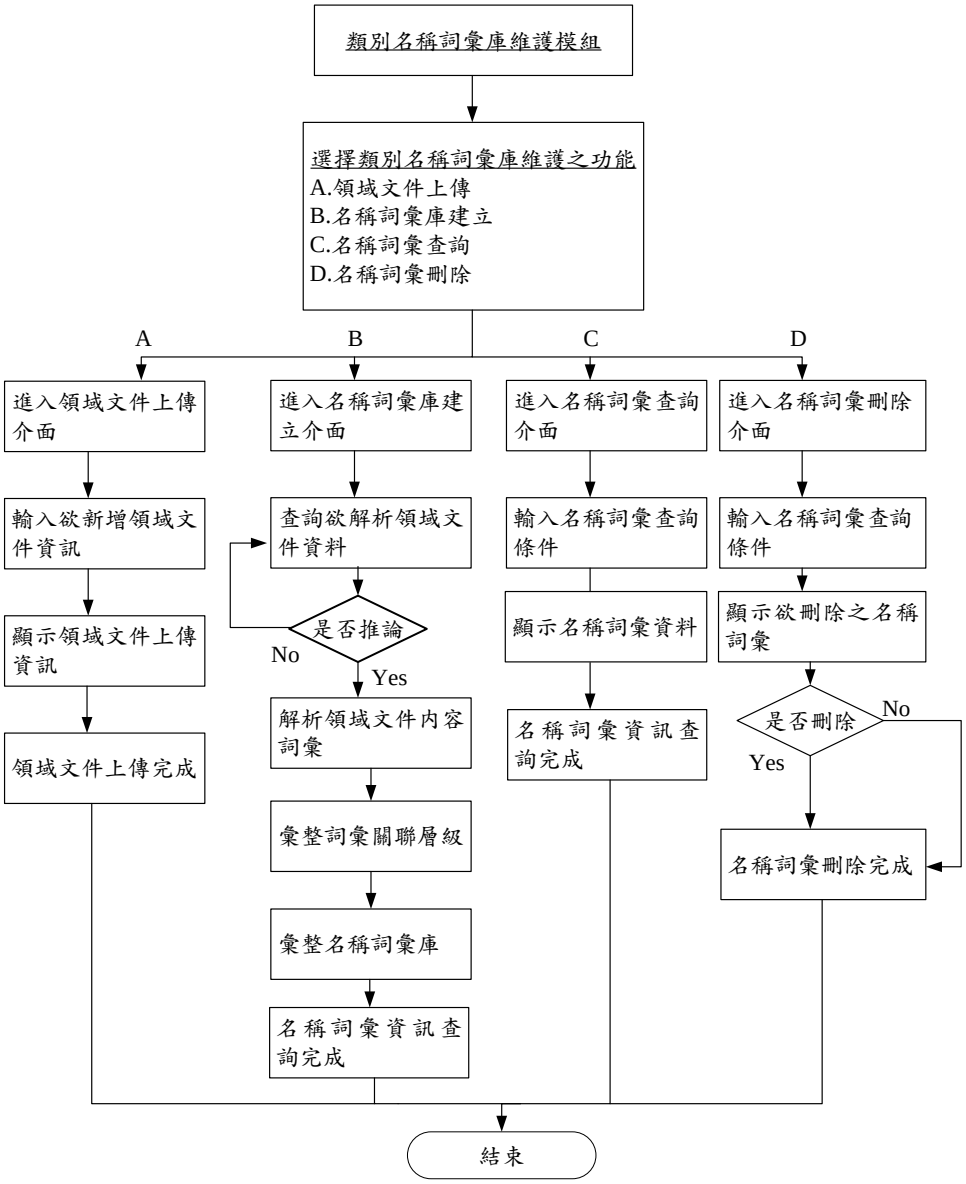

圖 4.8、「類別名稱詞彙庫維護模組」功能流程

系統參數設定

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，包含「郵件語意關鍵字擷取參數設定」與「個人化郵件類別推論參數設定」等二大功能；其中，「郵件語意關鍵字擷取參數設定」功能乃提供權限內使用者針對郵件語意關鍵字擷取模組之門檻值與權重值進行修改與維護，進而保持門檻值與權重值之正確性；「個人化郵件類別推論參數設定」功能乃提供系統管理者進行修改個人化郵件類別推論模組之門檻值與權重值，進而保持系統門檻值與權重值之正確性，其流程設計概念如圖 4.9 所示。

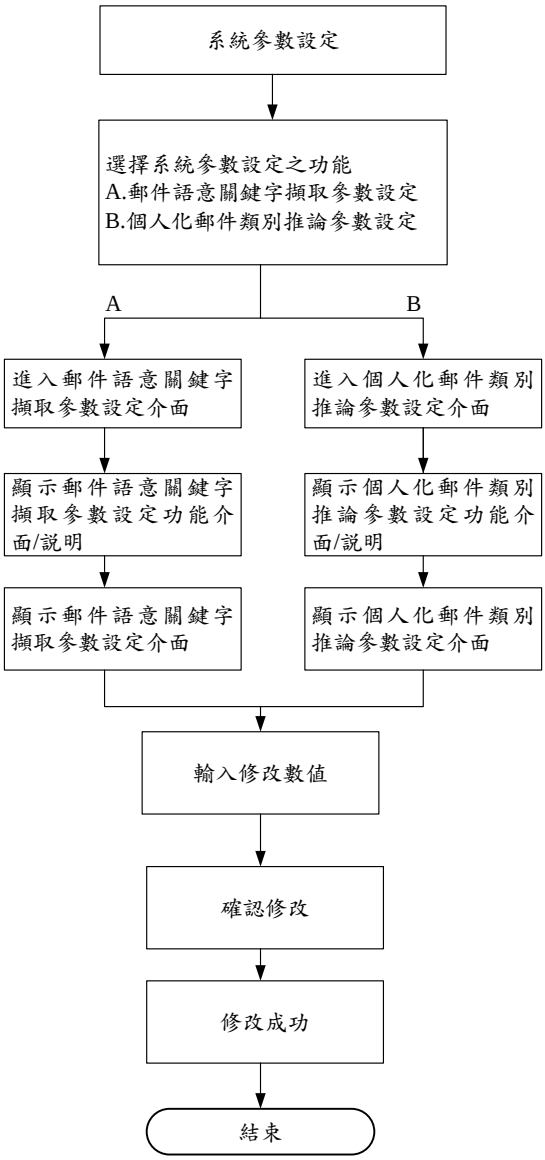

圖 4.9、「類別名稱詞彙庫維護模組」功能流程

4.4.2 系統資料流程

本系統運作之初，系統管理者將未解析之領域文件匯入系統，並進行類別名稱詞彙庫建立，代系統管理者完成類別名稱詞彙庫建置後，一般使用者即可將未解析之電子郵件匯入系統，而系統管理者則設定郵件語意關鍵字擷取參數與個人化郵件類別推論參數等各參數並設定完成，即可進行郵件主題詞彙集合建立與詞彙後驗機率近似值計算，以取得郵件具語意之主題詞彙，接著，系統管理者依序進行郵件主成份計算與郵件關聯程度與類別推論，進而取得郵件中代表性特徵詞彙，再透過根據特徵詞彙推論個人化郵件類別名稱與郵件分群，並存入系統資料庫內。上述步驟完成後，一般使用者即以郵件類別查詢功能進行查詢，即可取得系統推論之個人化郵件類別名稱與階層式郵件群集，其系統相關資料之存取與傳遞情形如圖4.10所示。

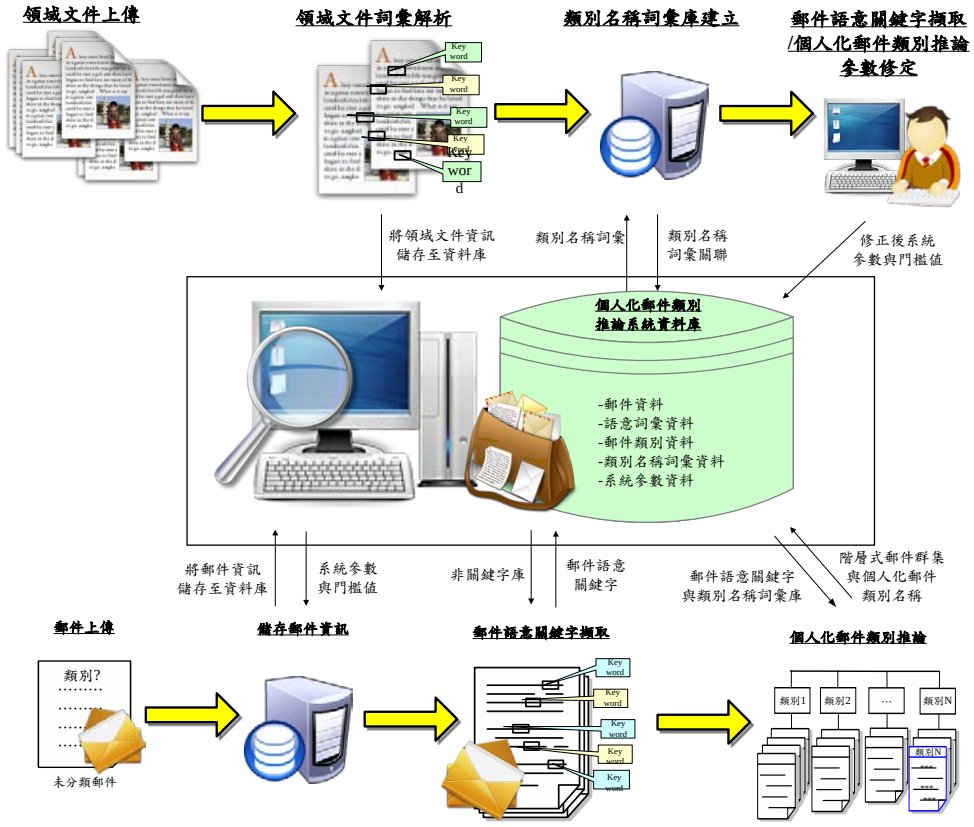

圖 4.10、系統資料流程

4.5 系統開發工具

本系統乃建置於 Microsoft Windows XP作業系統上，並以Microsoft SQL Server 2008資料庫系統儲存系統運作過程相關資料。系統之操作介面與內部演算法則採用JSP（Java Server Pages）語法進行開發，並利用SQL Server 2008來存取資料庫，輔助系統進行知識

文件分析。以下即分別介紹系統開發時所使用之工具。

- JSP（Java Server Pages）

JSP是由Sun Microsystem公司所倡導之網站伺服器描述語言程式，其乃以Java程式語言為基礎，並集結其他公司所共同建立之動態網頁技術標準，故具有Java支援跨平台與跨網站伺服器之優點，而使網頁設計更具彈性。

當使用者透過瀏覽器向伺服器端（Server）要求開啓JSP網頁時，架設於伺服器端之JSP引擎乃先將JSP網頁轉譯為Servlet程式，其次再將JSP執行後所產生之之是文件資料傳送至用户端（Client），並同時顯示執行結果於瀏覽器上。此外，JSP還具有下列特性（詹權恩，2004）：

- 瀏覽者端環境：各種網頁瀏覽器均可，如IE、Google Chrome、Fire Fox、KK Man。
- 模組程式的可重用性：JSP元件（Enterprise JavaBeans）可跨平台重複使用於任何地方。Enterprise JavaBeans 元件可存取傳統資料庫，並能以分散式系統模式於Unix 和 Windows 平台工作，減少程式開發之時間並可增加程式之彈性。
- 保護原始程式碼：延伸名為*.jsp的JSP程式碼並未顯示於Client端之瀏覽器上。
- 跨平台性：JSP可執行於任何具有Web伺服器之環境，並支援多數作業系統。
- 標籤可擴充性：由於JSP技術兼容XML標籤技術，故JSP開發者可擴展JSP標籤或制訂標籤庫，以減少對Scripting語言之依賴，並降低網頁製作者製作網頁和擴充網頁功能之複雜程度。
- 伺服器端環境：Windows XP，並加上「J2SDK」Java程式編譯工具與Tomcat等JSP伺服器。與HTML緊密整合：將JSP融入HTML標籤中，不僅提高便利性亦減少I/O問題，兼具可取代傳統CGI等直譯式語言。
- 平台和伺服器的獨立性：JSP技術一次寫入之後，可以在任何具有符合JavaTM語法結構的環境下執行。
- 伺服器端搭配資料庫：SQL Server資料庫系統。

- 關聯式資料庫-Microsoft SQL Server 2008

Microsoft SQL Server 2008為一種關聯式資料庫（Relational Database Management Systems，RDBMS）管理系統，其擁有高彈性與多元化之架構，可安裝於主從式架構之作業系統平台上或獨立伺服器主機。關聯式資料庫乃將資料分類儲存於多個二維表格

中，此些表格通稱為資料表。之後再利用兩資料表間之關聯以查詢相關資料。其優點在於各個資料表均可獨立運作，當進行資料之新增、修改或刪除時，亦不會互相影響。系統管理員可透過應用程式進入伺服器，更改資料型態，並管理及處理伺服器資源。此種資料庫常使用SQL（Structured Query Language）語法進行資料查詢，SQL語法可用以查詢資料庫、建立新表格、更新與刪除資料，並設定資料庫權限。

綜上所述，本研究乃利用上述工具進行系統開發工作，並將系統架構於 Web 環境下，以開發 4.2 節所述之各項系統功能。

第五章、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，本研究乃發展一套個人化郵件類別推論系統，並針對系統中各權限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。根據個人資料保護法第二條與第六條之規定（法務部全國法規資料庫：http://law.moj.gov.tw/)，私人郵件屬於使用者個人資料之一且不得私自隨意蒐集他人個人資料，故本研究難以收集與取得。為能完成系統驗證，本研究以「李榮陸文本分析語料庫」中郵件文本為案例驗證樣本，並以「台灣光華雜誌」之新聞內容作為領域文件樣本，再邀請試驗者進行私人郵件使用模擬實驗並以問卷統計試驗者感受，以分析本研究所提出之方法論與雛形系統可行性。

5.1 系統案例之應用流程

為驗證個人化郵件類別推論系統於實務應用之可行性，本研究乃分別針對個人化郵件類別推論與類別名稱詞彙庫建立兩方面透過眞實案例進行驗證，以證實本研究方法論之管理實務性。於個人化郵件類別推論之案例樣本，本研究乃以中國復旦大學李榮陸博士所建「李榮陸文本分析語料庫」為基礎，並從中隨機挑選多封郵件文本為案例驗證樣本，並以個人化郵件類別推論系統之兩大核心功能模組（包含「郵件語意關鍵字擷取模組」及「個人化郵件類別推論模組」等推論模組），進行以電子郵件之「郵件新增與郵件內文語意關鍵字擷取」及「個人化郵件類別推論與推論結果建議」等決策分析。於類別名稱詞彙庫建立之案例樣本，本研究乃以行政院發行「台灣光華雜誌」之新聞內容作為領域文件樣本，再以系統中「名稱詞彙庫建立」功能進行解析並建立類別名稱詞彙庫，並根據兩者之推論結果進行分析系統效能，以評估本研究所發展之方法論與開發之系統可行性。

首先，系統管理者必須蒐集類別名稱詞彙庫所領域文件，接著上傳領域文件至系統中並建立類別名稱詞彙庫，最後設定系統各項參數。接著，一般使用者乃上傳多封未解析之電子郵件至系統中。而後，系統管理者乃執行主功能以推論個人化郵件類別，待分析完畢，系統即將此次郵件類別之推論結果輸出予管理者，管理者再將此次推論結果回饋至系統中。最後使用者即可查詢系統為使用者個人推薦之個人化郵件類別。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。

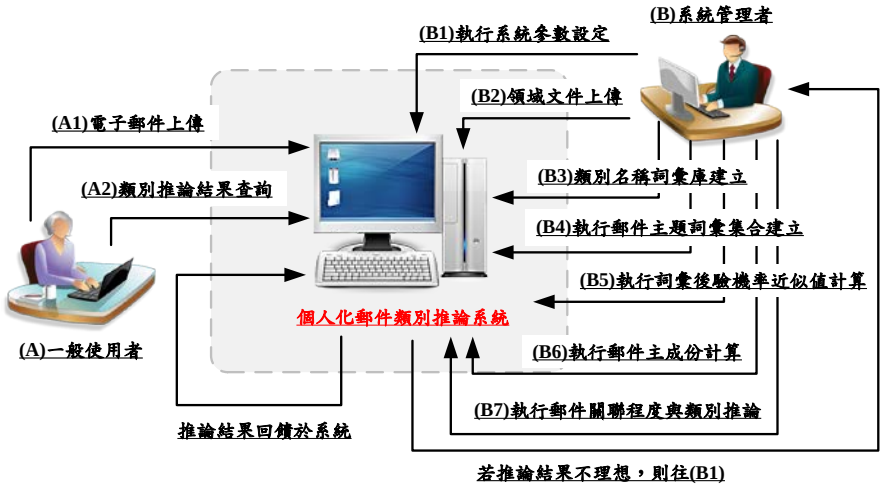

圖 5.1、個人化郵件類別推論系統之應用流程

➢ **系統管理者蒐集領域文件**

系統管理者於開放使用本系統前，先行蒐集未解析之領域文件，之後乃利用本系統幫助解析領域文件之內容詞彙並建立類名稱詞彙庫。系統管理者於使用本系統之前，蒐集了從行政院發行「台灣光華雜誌」之新聞內容（如圖 5.2 所示）並儲存文件與彙整，如表 5.1 乃列出部分系統管理者所蒐集之新聞內容。

圖 5.2、行政院「台灣光華雜誌」之新聞內容

表 5.1、系統管理者所蒐集之新聞內容

文件編號	領域文件名稱
1	尬出文化新生命——馮凱與《陣頭》
2	婚外有藍天，熟年出走潮
3	台灣珍珠奶茶攻佔全球
4	日式老宅，台北文創新基地

- **系統管理者設定系統參數及類別名稱詞彙庫建立**

系統管理者必須先設定系統內之各項參數與門檻值，以保持系統主功能判斷之正確性。首先，管理者必須設定系統參數之門檻值，之後透過類別名稱詞彙庫維護模組之領域文件上傳功能，上傳領域文件至系統中，接著進行名稱詞彙庫建立功能，以分析領域文件中詞彙並建立類別名稱詞彙庫，即完成系統參數設定及類別名稱詞彙庫建立。

- **一般使用者蒐集電子郵件**

一般使用者於使用本系統前，先行蒐集未解析郵件內容特徵之私人郵件，之後乃利用本系統幫助其建立使用者個人專屬之個人化郵件類別。於案例驗證中，本研究之測試方式乃蒐集「李榮陸文本分析語料庫」中郵件文本（如圖 5.3 所示），作為樣本資料，並儲存郵件與彙整如表 5.2 乃列出部分本研究所蒐集之郵件文本主旨。

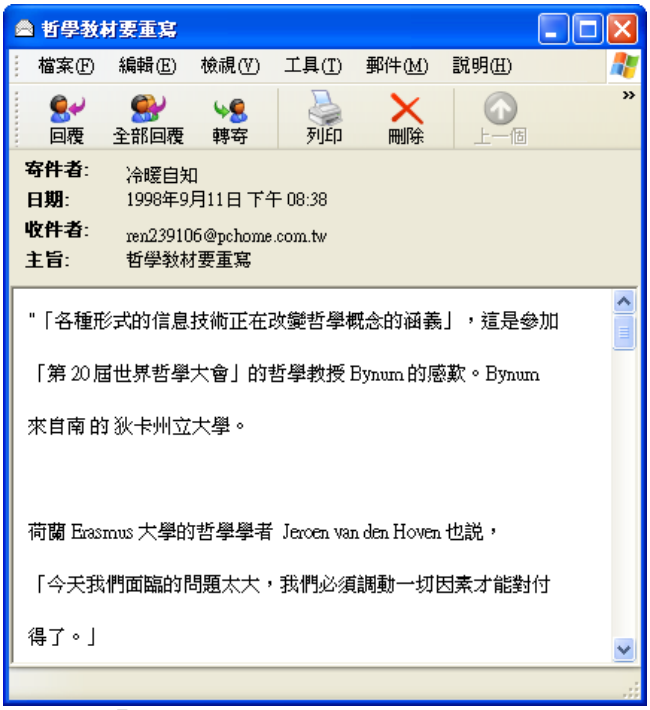

圖 5.3、「李榮陸文本分析語料庫」之郵件文本

表 5.2、李榮陸文本分析語料庫中郵件文本主旨

郵件編號	郵件文本主旨
1	審美意識的兩個層面
2	兩岸天文教育研討會舉行
3	呼喚藝術教育
4	哲學教材要重寫

- **(A1)一般使用者上傳未分類之電子郵件**

當一般使用者進入「郵件上傳」功能時，系統乃提供郵件資料新增介面予使用者（如圖 5.4 所示），此一般使用者依序輸入此郵件名稱為「審美意識的兩個層面」、郵件上傳時間為「2012 年 12 月 15 日」、郵件語言為「中文」，並點擊瀏覽選擇上傳郵件之所在位置，如「e:\Users\Leo\Desktop\審美意識的兩個層面.eml」；最後，使用者按下「上傳」鍵後，並擷取郵件中收件者為「ren239106@pchome.com.tw」、郵件主旨為「審美意識的兩個層面」等郵件資料，即完成郵件資料之上傳作業（如圖 5.5 所示）。

圖 5.4、電子郵件上傳畫面(1)

圖 5.5、電子郵件上傳畫面(2)

- **(B1)系統管理者設定系統參數**

待使用者新增郵件後，系統管理者則需設定系統內之各項參數與門檻值，以保持系統主功能判斷之正確性。管理者必須點選系統參數設定模組以設定系統參數之門檻值與權重值，即完成系統參數設定步驟（如圖 5.6 與圖 5.7 所示）。

圖 5.6、系統參數設定模組-系統參數檢視

圖 5.7、系統參數設定模組-系統參數設定

➢ **(B2)系統管理者上傳領域文件**

當系統管理者執行「領域文件上傳」功能時系統則顯示畫面，系統管理者欲上傳一份名稱領域文件時，如圖 5.8 所示，於領域文件名稱輸入「尬出文化新生命──馮凱與《陣頭》」，且於領域文件類別輸入「大眾傳播」等相關資訊後，並點擊瀏覽選擇欲上傳檔案路徑「E:\Users\Leo\Downloads…」後，點擊「上傳」按鈕系統則顯示領域文件新增成功之系統畫面，如圖 5.9 所示，系統完成資料擷取並回饋系統後乃顯示領域文件之資料，如系統管理者上傳名稱為「尬出文化新生命──馮凱與《陣頭》」領域文件後，系統則顯示領域文件名稱與內容供系統管理者確認。

圖 5.8、領域文件上傳(1)

圖 5.9、領域文件上傳(2)

➢ **(B3)系統管理者建立類別名稱詞彙庫**

系統管理者進入「名稱詞彙庫建立」功能後，權限內使用者「陣頭」進行搜尋，則系統篩選出領域文件名稱為「尬出文化新生命——馮凱與《陣頭》」之領域文件（如圖 5.10 所示）。接著系統將斷詞結果顯示於系統介面中，如圖 5.10 所示，段落「有人說，電影的情節張力與高票房...」包含「電影」、「導演」、「馮凱」、「電視劇」等詞彙，接著根據詞彙機率與詞彙關聯彙整為名稱詞彙集，如圖 5.11 所示，且系統於畫面中乃根據詞彙之關聯等級進行樹狀排列。最後於系統解析完成時，如圖 5.12 所示，本系統乃將取得詞彙與系統內既有名稱詞彙庫整併相同詞彙，當使用者點擊系統畫面中名稱詞彙「影視娛樂」系統則顯示該筆詞彙來源之領域文件明「影像魔法師 ——王小棣」，即完成名稱詞彙庫建立功能。

圖 5.10、名稱詞彙庫建立功能(1)

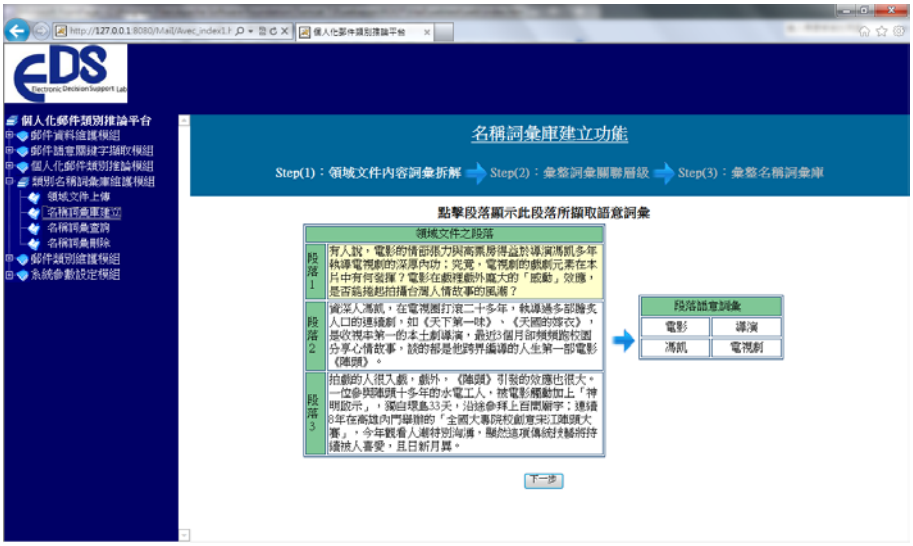

圖 5.11、名稱詞彙庫建立功能(2)

圖 5.12、名稱詞彙庫建立功能(3)

圖 5.13、名稱詞彙庫建立功能(4)

(B4)系統管理者執行郵件主題詞彙集合建立

當使用者完成郵件上傳動作，且系統管理者亦完成系統參數設定，系統管理者即可執行「郵件主題詞彙集合建立功能」擷取郵件中語意詞彙。首先，使用者搜尋郵件主旨「審美意識的兩個層面」之電子郵件並進行解析（如圖 5.14 所示），待系統計算完成後乃呈現郵件所含主題如「意識」、「哲學」等（如圖 5.15 所示），如「意識」主題包含「審美」、「層面」等詞彙，其中如詞彙「審美」與「哲學」主題之關聯係數為「0.043」，並相較於其他主題如「意識」之關聯係數還高，則代表詞彙「審美」於該封郵件中所代表語言含意與「哲學」主題較具關聯。

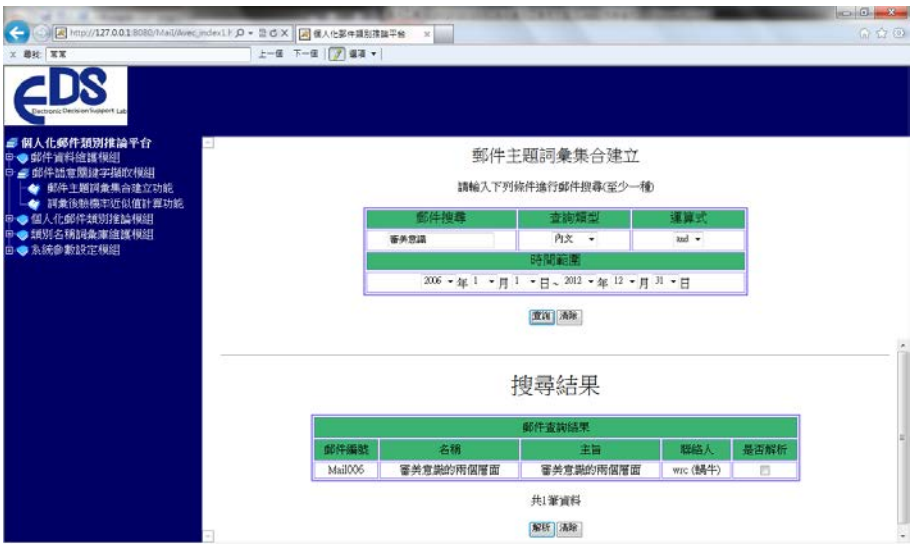

圖 5.14、郵件主題詞彙集合建立(1)

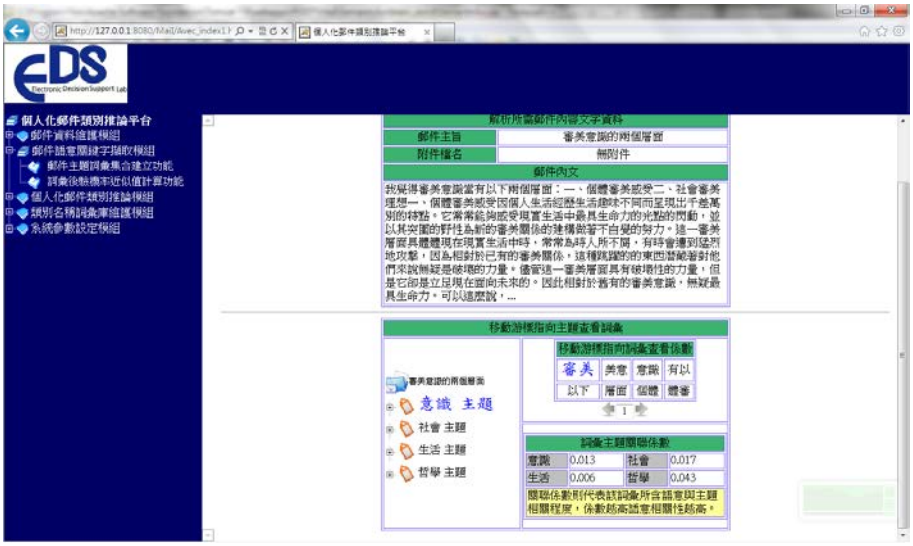

圖 5.15、郵件主題詞彙集合建立(2)

(B5)系統管理者執行詞彙後驗機率近似值計算

系統管理者進入「詞彙後驗機率近似值計算」功能後以「審美意識的兩個層面」之郵件進行解析（如圖 5.16 所示），並根據參數預設值「15」擷取適用模擬抽樣組合進行平均值計算（如圖 5.17 所示），即可取得後驗機率近似值，接著點擊詞彙主題關聯修正系統即呈現詞彙重新計算後「詞彙主題間關聯係數」供使用者檢閱，如圖 5.18 所示。

圖 5.16、詞彙後驗機率近似值計算(1)

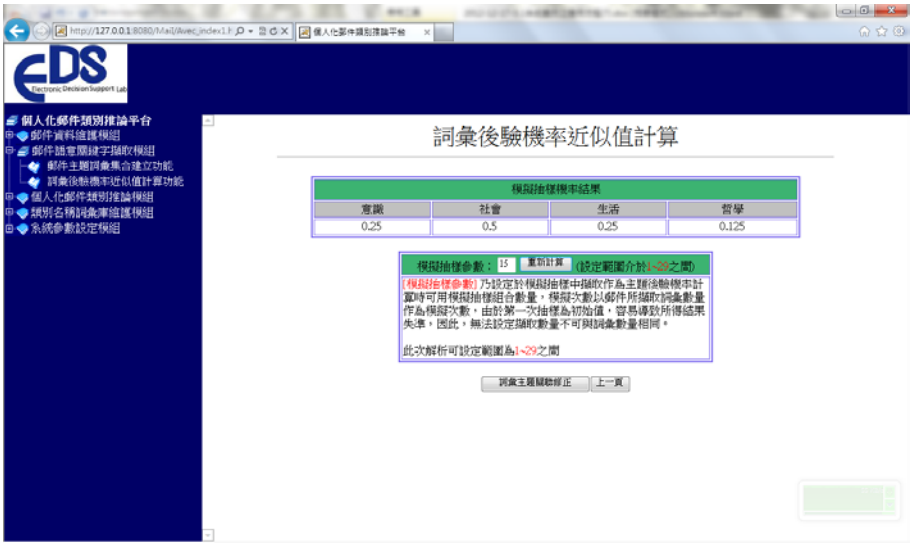

圖 5.17、詞彙後驗機率近似值計算(2)

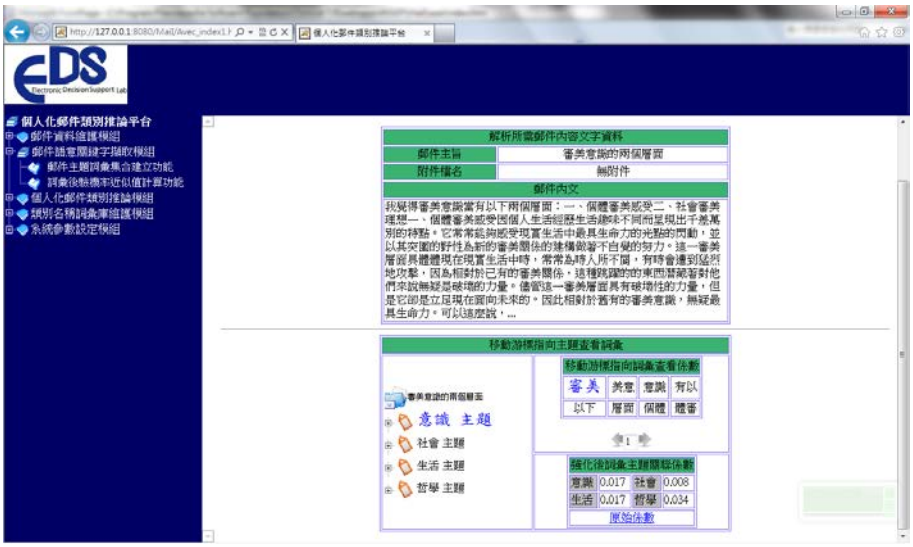

圖 5.18、詞彙後驗機率近似值計算(3)

➢ **(B6)系統管理者執行郵件主成份計算**

系統管理者進入「郵件主成份計算」功能後，如圖 5.19 所示，以「審美意識的兩個層面」之語意詞彙與詞彙主題關聯係數進行分析，即可獲得郵件特徵詞彙，如圖 5.20 所示，目標郵件「審美意識的兩個層面」之郵件特徵詞彙包含「審美」、「意識」、「力量」…等詞彙。

圖 5.19、郵件主成份計算(1)

圖 5.20、郵件主成份計算(2)

➢ (B7)系統管理者執行郵件關聯程度與類別推論

「郵件關聯程度與類別推論」功能主要目的乃將一般使用者所有上傳之郵件分析郵件群集與一般使用者適用之個人化郵件類別。以下乃分別以步驟方式說明本功能中個分析步驟使用方式。

Step(1) 彙整郵件特徵詞彙

於系統進行前，系統管理者需先行完成「郵件語意關鍵字擷取模組」與「郵件主成份計算」功能後進入本功能。系統先行彙整並分析系統內所有郵件之郵件特徵詞彙，當系統管理者點擊郵件「審美意識的兩個層面」時，系統則顯示郵件特徵詞彙（如圖 5.21 所示），以供系統管理者檢閱。

圖 5.21、郵件關聯程度與類別推論(1)

Step(2) 區分相似郵件

待系統完成「彙整郵件特徵詞彙」後，系統根據郵件中所含所有郵件特徵詞彙計算郵件相似性，並依據相似程度歸類郵件群集，如圖 5.22 所示，系統畫面中「相似郵件群 A」所含相似郵件如「審美意識的兩個層面」、「《藝術哲學》研究補白」等郵件。完成相似郵件區分後，系統管理者點擊「建立郵件群集樹」按鈕進入下一步驟。

圖 5.22、郵件關聯程度與類別推論(2)

Step(3) 建立郵件群集樹

於此步驟系統乃彙整相似郵件群集為郵件群集樹，如圖 5.23 所示，系統畫面中「層級 2」之群集「相似郵件群 C」所含相似郵件如「國際運輸與物流博覽會」、「航運安全不可忽視」等郵件。系統根據系統管理者所設定之層級擷取參數擷取所需群集。

圖 5.23、郵件關聯程度與類別推論(3)

Step(4) 分析群集代表詞彙

完成郵件群集樹建立後，系統於此步驟乃分析群集內中群集代表詞彙。使用者可點擊其他群集名稱，系統則展開該群集之郵件特徵詞彙關聯，如圖 5.24 所示，「相似郵件群 A」中根據「審美意識的兩個層面」、「《藝術哲學》研究補白」等郵件分析後，其郵件特徵詞彙關聯中「等級一」詞彙為「藝術」。

圖 5.24、郵件關聯程度與類別推論(4)

Step(5) 推論郵件類別名稱

於最後，系統根據郵件群集中特徵詞彙關聯與郵件名稱詞庫進行關聯比對，以尋找郵件群集中代表性名稱，如圖 5.25 所示，「相似郵件群 A」之群集名稱經系統與名稱詞彙庫分析詞彙關聯後，該群集之名稱命名為「藝術文化」。

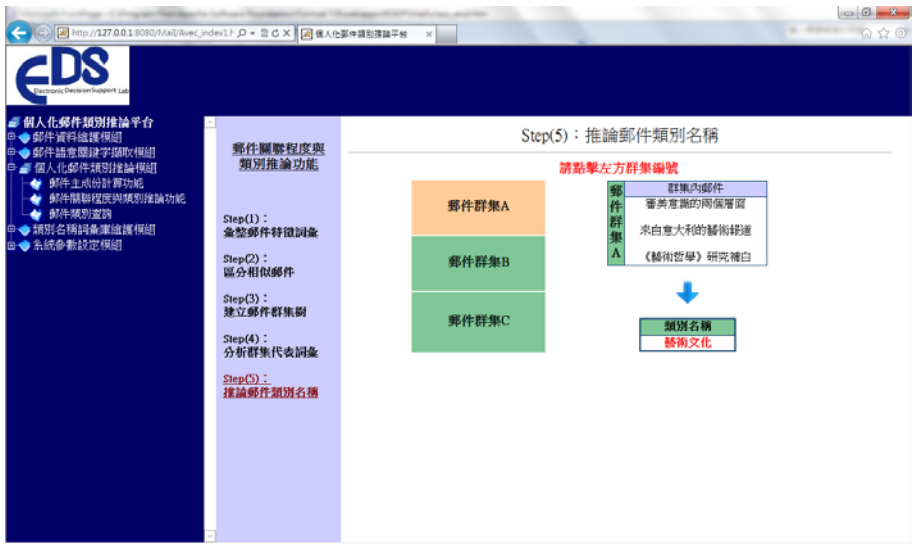

圖 5.25、郵件關聯程度與類別推論(5)

➢ **(A2)一般使用者查詢個人化郵件類別推論結果**

系統管理者完成上述推論後，一般使用者乃透過「郵件類別查詢」功能顯示系統所推論之郵件類別層級，如圖 5.26 所示，系統則顯示「藝術文化」、「人生哲學」、「交通」等郵件類別並以層級樹方式展示，當一般使用者滑鼠指標指向「人生哲學」類別時，系統則顯示「人生哲學」類別所包含郵件為「哲學的威力」、「哲學教材要重寫」等郵件資訊。

圖 5.26、郵件類別查詢

5.2 系統驗證與評估

本研究乃以「個人化郵件類別推論」之方法論為基，開發一套「個人化郵件類別推論系統」，並以中國復旦大學李榮陸博士所建「李榮陸文本分析語料庫」內蒐集郵件資料作為應用情境，並將語料庫中郵件文本作為使用者之私人郵件；此外，本研究所提出之系統乃分析使用者之私人郵件取得郵件中主要特徵詞彙，並以郵件特徵與類別名稱詞彙庫進行代表行名稱解析，是故，於驗證之前，本研究乃蒐集行政院發行之「台灣光華雜誌」新聞內容作為領域文件，以進行名稱詞彙庫建立。於系統驗證中乃以「李榮陸文本分析語料庫」郵件文本及「台灣光華雜誌」新聞內容進行解析，並以「個人化類別名稱推論」主題進行系統績效驗證。以下乃針對本系統各主題之驗證方式（即「驗證資料說明」、「驗證方式說明」與「驗證指標定義」）依序說明，以驗證「個人化郵件類別推論系統」績效與準確性。

測試資料與訓練資料取得

針對「個人化類別名稱推論」主題之驗證，本研究乃先彙整中國復旦大學李榮陸博士所建「李榮陸文本分析語料庫」中郵件文本，以作為個人化郵件類別推論之測試資料，同時亦至行政院「台灣光華雜誌」取得新聞內容並建立類別名稱詞彙庫，以作為系統驗證之訓練資料。其各驗證資料乃說明如下：

針對私人郵件之測試資料部分，由於私人郵件屬於使用者個人資料之一，然而2012年實行「個人資料保護法」第二條與第六條之規範（法務部全國法規資料庫：http://law.moj.gov.tw/），為保護使用者個人資料洩漏導致不法人士竊取並使用於違法行為，故法令規定不得公開個人資料，導致本研究難以收集與取得。為能完成系統驗證，本研究以「李榮陸文本分析語料庫」中郵件文本為測試資料，再邀請試驗者將「李榮陸文本分析語料庫」中郵件文本作為私人郵件使用並模擬實驗，再以問卷統計試驗者感受，以進行系統驗證。於「李榮陸文本分析語料庫」中所蒐集之郵件文本資料乃包含：郵件中主旨、寄件日期、寄件人與郵件內容等文字資料，且各郵件資料乃根據內容類型進行類別區分。因此，本研究乃針對「李榮陸文本分析語料庫」取得郵件文本資料，如圖5.27所示，並彙整成「郵件文本資料表」（如表5.3所示）。

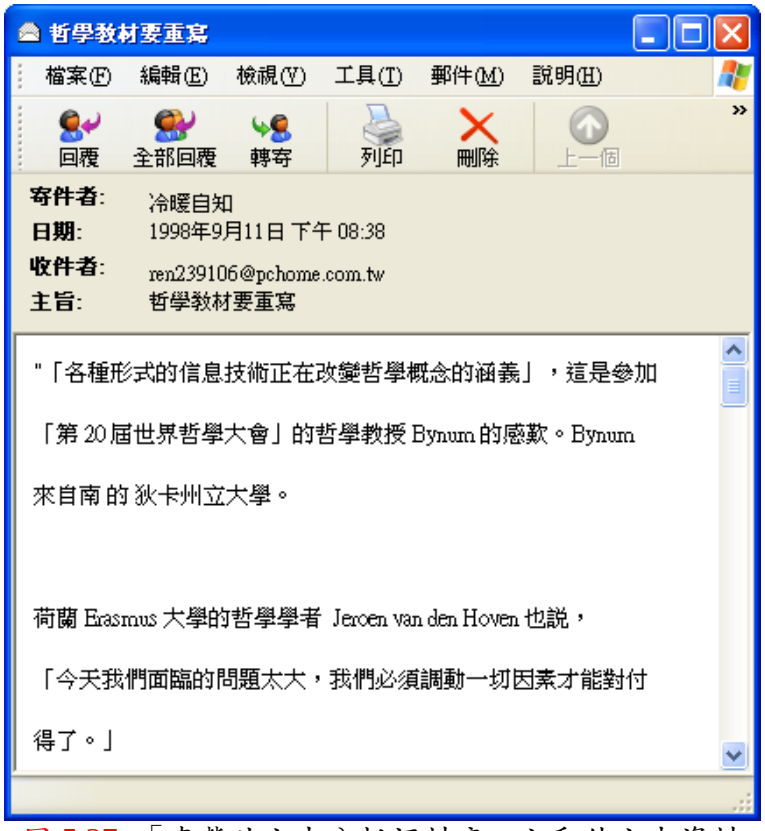

圖 5.27、「李榮陸文本分析語料庫」之郵件文本資料

表 5.3、郵件文本資料表（部分資料）

題號	郵件主旨	寄件人	寄件日期
1	哲學的威力	跳仔	1997/10/18
2	排球規則的重要修改	陳培基	1999/09/28
3	中 國 陸 軍 航 空 兵	幻影	1998/07/30
4	我們怎樣建設自己的信息網絡	張樹新	1996/07/03
5	數據庫系統	dibo	1996/05/14
6	哲學教材要重寫	冷暖自知	1998/09/11

針對類別名稱詞彙庫所需領域文件（即訓練文件），本研究乃先至行政院「台灣光華雜誌」取得新聞內容，以進行類別名稱詞彙解析並維護於系統資料庫中以完成類別名稱詞彙庫建立建置，以作系統驗證前資料。於「台灣光華雜誌」之新聞內容乃針對一般生活時事進行敘述，而大多一般使用者之私人郵件內容相較於公共郵件內容較為生活與多樣化，因此，本研究乃以「台灣光華雜誌」之新聞內容解析詞彙，並彙整為詞彙關聯作為名稱詞彙庫。本研究乃將部分結取「台灣光華雜誌」之新聞文章內容作為系統驗證之訓練資料（如圖5.28所示）。最後乃將上述資料表作為「個人化郵件類別推論系統」之驗證資料，其驗證資料分析流程如圖5.29。

圖 5.28、行政院「台灣光華雜誌」之新聞網頁畫面

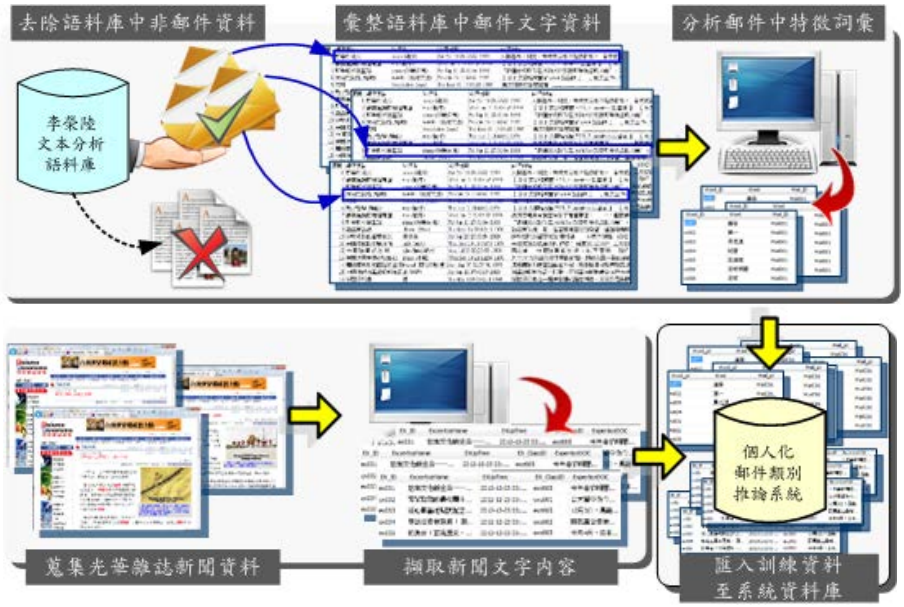

圖 5.29、個人化郵件類別推論驗證資料分析流程

5.2.1個人化郵件類別推論系統整體驗證方式

為針對「個人化郵件類別推論系統」之類別名稱推論正確性進行探討，本研究乃提出「個人化類別名稱推論」課題以進行系統績效驗證。目的乃透過郵件分群與使用者個人特徵以取得使用者個人之個人化郵件類別名稱。於此課題本研究乃以「李榮陸文本分析語料庫」中隨機挑選100筆郵件資料作為測試資料並進行分群，再依郵件群集推論類別名稱。為能驗證系統推論類別名稱與郵件內容合適度，本研究乃以25位一般郵件使用者作為試驗者，針對郵件類別名稱與群集內所含郵件類型之一致性進行評估。針對試驗者部分，本研究乃邀請25位大專院校學生（男女不限）作為一般郵件使用者以進行測試，當中，本研究針對取5所不同系所進行邀請，而各系所皆邀請5位試驗者，總計25名進行測試。首先，本研究先行提供系統推論郵件群集名稱與郵件群集所含郵件之主旨供試驗者檢閱。此外，本研究乃針對郵件管理對一般使用者效益議題之文獻作為參考並設計問卷之問項，以作為試驗者對「個人化郵件類別推論系統」評估之問卷，試驗者根據群集內郵件主旨判讀郵件類型並對照系統推論郵件群集名稱，於對照完畢後根據試驗者之主觀感受判斷群集內郵件與系統推論名稱符合性。

對此，本研究乃規劃兩階段之系統驗證，且以800份「台灣光華雜誌」新聞文章（即領域文件）作為訓練資料並建立類別名稱詞彙庫。於系統驗證第一階段乃於800份新聞文章中，隨機選取100筆作為訓練資料，建立第一階段驗證所需類別名稱詞彙庫，再以「李榮陸文本分析語料庫」中隨機挑選之100筆郵件資料進行類別名稱推論，以作為系統測試資料並提供給試驗者進行符合性評估，藉以觀察系統類別名稱推論結果之適合性

與類別名稱推論方式，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，即進行系統測試第二階段（即系統測試階段），於此階段乃分6個週期（每週期皆匯入100筆不重複之新聞文章，共計600筆資料），藉由第一二階段共7個週期之驗證持續匯入以分析系統於不同訓練測試資料下之解析效果，並從中了解個人化郵件類別推論系統之長期學習趨勢。

- **系統評估問卷設計**

本研究之系統驗證設計乃邀請試驗者進行檢索系統推論結果，並依據檢索後觀感進行問卷回答，以證實本研究所提方法論可行性與實用性。於試驗者問卷中乃分為推論結果問項評級及正確類別名稱統計兩部分。於推論結果問項評級部分，本研究乃參考 Szóstek（2011）定義「項目一、電子郵件組織結構」與「項目二、電子郵件檢索功能」之提問項目，且參考 Soucek 和 Moser（2010）、Mano 和 Mesch（2010）、Lewis 等人（2004）、Jerejian 等人（2013）與 Wang 等人（2009）等研究制定「項目三、郵件分群對郵件管理幫助」、「項目四、郵件類別名稱對郵件內容閱讀影響」與「項目五、整體推論結果對郵件應用影響」等提問項目作為郵件類別名稱推論結果評級之提問項目。此外亦根據 Soucek 和 Moser（2010）將評級分為 3 等級，分別為：1 分（不認同）、4 分（可以）與 7 分（非常認同），以凸顯評級間差距分別給定不同分數。針對本研究設計五項推論結果問項評級之項目提問內容之細述如下：

項目一、電子郵件組織結構：

- **分類結構可靠度**：針對試驗者檢索郵件整體歸類方式與層級結構後，試驗者可針對各郵件群集之階層關聯進行觀感判定。
- **分類結構變化性**：針對系統所推論可區分郵件類型之變化進行評估。
- **類別名稱解釋性**：針對試驗者檢閱後對類別名稱與群集內郵件符合行觀感分析。

項目二、電子郵件檢索功能：

- **分群準確性**：針對試驗者檢索郵件群集後，針對郵件群集歸類正確性與適當性進行觀感分析，透過群集內郵件之郵件主旨檢索郵件歸類一致性。
- **分類名稱靈活性**：試驗者針對分類名稱可劃分郵件類型數量進行評估，類別名稱含意結代表同類型含意則代表分類名稱具靈活性。

- **郵件檢索方便性：**試驗者針對郵件類別名稱之樹狀結構進行檢索，並依自我觀感根據整體郵件資料檢索時方便性進行評分。

項目三、郵件分群對郵件管理幫助：

- **過量郵件整理助益性：**試驗者針對系統推論結果之分群與類別名稱進行感觀判斷，以判斷此歸類方式對郵件過載整理幫助。
- **減少電子郵件檢查頻率：**試驗者依自我感觀判斷此次推論結果對試驗者減少郵件信箱檢查頻率之可能性。

項目四、郵件類別名稱對郵件內容閱讀影響：

- **減少郵件錯誤通訊：**透過郵件整理與歸納後，試驗者根據此次推論結果判斷減少錯誤郵件閱讀與回覆之可能性。
- **協助使用者判別郵件內容重要性：**私人郵件涉及領域廣泛，而非各領域所中信件皆含有重要訊息，故協助使用者快速尋找重要訊息信件為重要項目之一，試驗者則針對郵件分群與類別名稱進行檢索與判斷對重要訊息信件標記幫助效益。

項目五、整體推論結果對郵件應用影響：

- **協助使用者快速瀏覽收件夾中所有信件：**系統根據郵件內容進行分群與類別名稱制定，故試驗者於檢閱系統推論結果時，依個人習慣判斷樹狀類別名稱與群集對使用者快速理解與郵件搜尋之幫助效益。
- **減少使用者繁複信箱管理作業：**透過系統推論郵件類別與郵件區分群集進行郵件管理，試驗者依據此項假設判斷本研究推論郵件管理方法可否減少使用者需繁複進行郵件標籤加註、郵件刪除、移動...等工作。
- **協助使用者回覆多筆郵件內容：**試驗者假設系統所分析郵件皆需進行郵件回覆，並依據系統將郵件整理歸納方式進行分析，並判斷系統將郵件歸類方式對使用者回覆大量串聯郵件時方便性。

此外正確類別名稱統計部分，本研究乃將推論結果彙整並排列於問卷中，試驗者根據推論結果勾選正確類別名稱與郵件分群，以統計本研究推論結果正確數量，且為統計分群數量正確性，試驗者於正確類別名稱與郵件群集判斷結束後填寫此些郵件實際所需

郵件類別數量，以統計試驗者觀感認知中郵件所需實際類別與群及數量。正確類別名稱統計部分提問如下：

項目六、正確類別名稱統計

- **正確郵件類別名稱與群集勾選：**試驗者根據推論結果進行檢索並判斷該群集所歸類郵件類型一致性與郵件類別名稱適合性。該群集名稱與歸類正確則勾選正確，反之則勾選不正確。
- **使用者期望郵件類別實際數量：**試驗者根據推論結果之類別數量與郵件整理所需分類數量進行判斷，試驗者根據判斷結果填寫試驗者認為此些郵件樣本實際所需郵件類別與群集數量。

➢ **試驗方法設計**

根據上述系統整體驗證方式，本研究乃以南華大學中企管系、中文系、資管系、旅遊系與會計系等5所系所中各邀請5位學生（男女不限）作為試驗者。為能提供試驗者相似電子信件中收發信件之模擬環境，本研究乃以南華大學之校內電子郵件信箱為模擬環境（即模擬信箱），並以「李榮陸文本分析語料庫」中郵件文本（即樣本郵件）分別模擬使用者收發信件與郵件管理方式。當中模擬試驗分為兩大階段，分別為「階段一、電子郵件信箱使用訓練」、「階段二、系統推論類別分類模擬」，當中「階段一、電子郵件信箱使用訓練」主要協助試驗者於模擬試驗開始前電子信箱使用學習與訓練，待試驗者完成5次以上學習與訓練後則進行「階段二、系統推論類別分類模擬」之試驗，當中階段二乃以本系統所推論個人化郵件類別結果於模擬信箱中建立郵件類別，並請示驗者進行郵件整理與管理模擬，模擬方式如圖5.30所示。

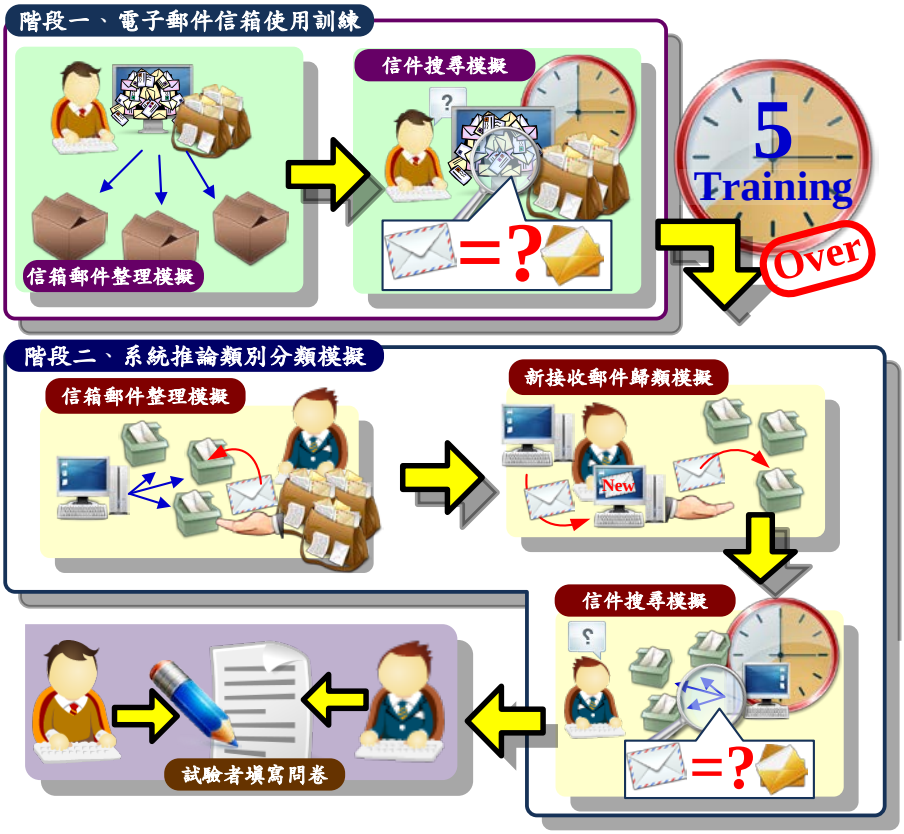

圖 5.30、郵件整理試驗模擬步驟示意圖

為讓試驗者比較郵件類別建立前後對郵件整理效益之差別，故本研究於將階段模擬中分為「信箱郵件整理模擬」、「新接收郵件歸類模擬」與「信件搜尋模擬」三階段模擬使用者於電子信箱中郵件管理與信件收發行為，以統計本研究方法論對使用者管理郵件上助益性。以下乃劃分為「階段一、電子郵件信箱使用訓練」、「階段二、系統推論類別分類模擬」兩部分詳細敘述「信箱郵件整理模擬」、「新接收郵件歸類模擬」與「信件搜尋模擬」此三階段模擬實驗說明：

階段一、電子郵件信箱使用訓練

於此階段乃主要模擬試驗者於尚未透過自動化方式進行人工管理之感受，並訓練試驗者學習電子信箱管理郵件方式與模擬軟體使用方式。故於此次模擬信箱皆不建立郵件類別，以供使用者以個人領域知識進行判定並歸類郵件。此外，為能促使試驗者完整學習電子郵件信箱使用與模擬軟體應用方式，本研究乃邀請試驗者反覆進行「階段一、電子郵件信箱使用訓練」達5次以上訓練，完成後才進入階段二之模擬。

- **信箱郵件整理模擬：**為能使試驗者感受人工分類與自動類別名稱建立差異，本研究乃先邀請試驗者進行郵件模擬分類，首先本研究乃於100份樣本郵件中隨機抽出20

份樣本郵件於模擬信箱中，並請試驗者根據自我解讀方式與個人知識將此20份郵件進行人工分類並將歸類郵件類別名稱命名。於試驗者分類結束後告知試驗者與系統分類差異處，供試驗者比較並評級系統推論郵件類別名稱與試驗者個人知識認知差異性。

- **信件搜尋模擬：**一般電子郵件使用者常於電子信箱中搜尋過往郵件，故本研究乃針對使用者搜索郵件進行模擬。本研究乃於模擬信箱中匯入20筆樣本郵件，並隨機挑選5筆郵件主旨邀請試驗者於模擬信箱中搜尋此5筆郵件且紀錄搜索時間。

階段二、系統推論類別分類模擬

此次模擬試驗乃根據系統推論郵件類別名稱與郵件群集建立模擬信箱中郵件分類資料夾，並邀請試驗者根據此管理方式與「階段一」進行比較，進而確認本研究於個人化類別名稱推論與使用者個人認知類別差異性。

- **信箱郵件整理模擬：**本研究乃根據系統分析樣本郵件所得類別名稱，於模擬信箱中建立系統推論郵件類別，接著，邀請試驗者將隨機抽出50份樣本郵件進行郵件分類。於分類結束後告知試驗者與系統分類差異處，接著試驗者根據個人定義郵件類別與系統推論類別名稱進行比較並回答相關問項之提問。
- **新接收郵件歸類模擬：**為模擬使用者接收新郵件時系統推論類別名稱適用性，本研究將模擬次數假設為5次，並每次以5封樣本郵件為數量透過不同電子信箱發送至試驗者模擬信箱中，試驗者於接收新郵件後可自行點擊閱讀郵件主旨與信件內容後，試驗者根據個人知識判定郵件所屬類別，並歸類於模擬信箱中郵件類別。於模擬結束後公佈試驗者與系統推論差別，以驗證系統推論類別名稱與使用者個人知識認知符合性。
- **信件搜尋模擬：**根據系統推論郵件群集與郵件類別名稱於模擬信箱中建立類別，並匯入各類別所含郵件共計100筆樣本郵件，且隨機挑選5筆郵件主旨邀請試驗者於模擬信箱中搜尋此5筆郵件且紀錄搜索時間。於模擬結束後試驗者比較搜尋時間與搜尋時自我感觀並回答問題，以驗證系統推論個人化郵件類別名稱及郵件分群方式對使用者之效益。

於上述模擬試驗結束後，試驗者檢閱模擬紀錄並根據模擬記錄判斷並回答問卷提問即可完成模擬試驗，階段二模擬試驗之對應提問項目如表5.4所示。此外，本研究透過試

驗中試驗者回答分數進行驗證指標計算，以統計本系統所推論個人化郵件類別名稱與使用者個人知識與個人認知符合性。

表 5.4、模擬試驗之對應提問項目彙整表

試驗類型	模擬試驗	對應問項
類別已建立信箱模擬	信箱郵件整理模擬	● 分類結構可靠度 ● 分群準確性 ● 過量郵件整理助益性 ● 減少電子郵件檢查頻率 ● 協助使用者判別郵件內容重要性 ● 減少使用者繁複信箱管理作業
	新接收郵件歸類模擬	● 分類結構變化性 ● 分類名稱靈活性 ● 過量郵件整理助益性 ● 減少使用者繁複信箱管理作業 ● 協助使用者回覆多筆郵件內容
	信件搜尋模擬	● 類別名稱解釋性 ● 郵件檢索方便性 ● 減少郵件錯誤通訊 ● 協助使用者快速瀏覽收件夾中所有信件 ● 協助使用者回覆多筆郵件內容

➢ **系統評估指標定義：**

針對郵件類別名稱推論結果合適性，本研究乃設計「郵件使用者滿意度指標」及「郵件類別名稱適應指標」作為驗證指標以檢視本系統郵件類別名稱推論效果。「郵件使用者滿意度指標」乃透過試驗中試驗者檢索系統推論郵件類別名稱與郵件群集，並勾選推論正確類別名稱與群集後，試驗者根據五項評級標準針對整體推論結進行評分，以驗證本研究對使用者實用價值，「郵件類別名稱適應指標」根據試驗者勾選推論正確數與系統推論數量進行計算，乃代表郵件類別名稱推論結果之準確性與系統將郵件歸類群集之適用性。以下分別針對「郵件使用者滿意度指標」以及「郵件類別名稱適應指標」進行說明。

「郵件使用者滿意度指標」乃以使用者觀感進行績效驗證，並依據本研究參考過去文獻所設定試驗提問項目包含「項目一、電子郵件組織結構」、「項目二、電子郵件檢索功能」、「項目三、郵件分群對郵件管理幫助」、「項目四、郵件類別名稱對郵件內容閱讀影響」與「項目五、整體推論結果對郵件應用影響」等提問項目，以作為郵件類別名稱推論結果評級之提問項目。待試驗者確定問項內容後再根據評級予以給分，並根據 Soucek 和 Moser（2010）定義與計算方式統計計算，再以數字方式呈現表示等級，如公式(5.1)所示：

$Sat_i(MC)$　第i位試驗者之郵件使用者滿意度

$Tes_i(Q_j)$　第i位試驗者對第j題問卷提問之評分，評分標準為非常認同7分、認同3分、不認同1分

$N(Q)$　提問問卷中結果評級之提問項目數量

$$Sat_i(MC) = \frac{\sum_{all\ j} Tes_i(Q_j)}{N(Q)} \tag{5.1}$$

「郵件類別名稱適應指標」乃以郵件分群之準確性進行驗證，針對郵件分群結果乃於試驗者問卷中統計試驗者所選正確數量，並與試驗者所填寫實際所需郵件類別數量進行分析。本研究乃以 Ugˇuz（2011）中適應函數（Fitness Function）為基礎發展「郵件類別名稱適應指標」。當中，乃以正確郵件群集數量與郵件群集總數計算進行適應函數計算，以作為本研究之「郵件類別名稱適應指標」。其各指標定義與計算方式如公式(5.2)至公式(5.4)所示：

$N(TMC_i)$　系統推論個人化郵件類別時，第i位試驗者認定推論正確類別數量

$N(RealMC_i)$　第i位試驗者認定測試樣本實際所需郵件類別數量

$R(Mail_i)$　系統推論個人化郵件類別時，第i位試驗者所得之類別推論召回率

$N(AllMC)$　系統推論個人化郵件類別之總數量

$P(Mail_i)$　系統推論個人化郵件類別時，第i位試驗者所得之類別推論正確率

$Fit(Mail_i)$　系統推論個人化郵件類別時，第i位試驗者所得之郵件類別名稱適應指標

$$R(Mail_i) = \frac{N(TMC_i)}{N(RealMC_i)} \tag{5.2}$$

$$P(Mail_i) = \frac{N(TMC_i)}{N(AllMC)} \tag{5.3}$$

$$Fit(Mail_i) = \frac{2 \times R(Mail_i) \times P(Mail_i)}{R(Mail_i) + P(Mail_i)} \tag{5.4}$$

- **個人化郵件類別推論系統驗證結果與分析**

本研究乃將個人化郵件類別推論系統驗證劃分為兩個階段共計七個週期，以驗證本研究方法論可行性。於第一階段中，針對個人化郵件類別推論驗證部份則以 800 份領域文件中隨機挑選 200 份作為第一階段驗證之訓練資料，並逐一匯入系統中以建立類別名稱詞彙庫。於第二階段本研究乃劃分為六個週期，每週期皆匯入 100 筆不重複之領域文件，共計 600 筆組合測試系統於不同數量資料下績效變化。以下即針對各階段說明系統驗證過程，並分析系統驗證之結果。

(I) 個人化郵件類別推論系統第一階段驗證結果分析

針對個人化郵件類別推論第一階段驗證部分，乃於 200 份領域文件作為訓練資料之基礎下，並將 100 筆郵件資料包含發信人、郵件標題、郵件時間與郵件內容等匯入系統，作為測試資料（如表 5.5 所示，以「談談老子」、「黨史研究中存在的問題」與「哲學的威力」…等等郵件為範例），於第一階段測試結果乃彙整於表 5.6 與表 5.7。以表 5.6 為例，於階層式郵件分群所得結果中，郵件「第一張人民幣」與郵件「黨史研究中存在的問題」皆屬相同群集，且推論群集名稱（即類別名稱）為「大陸」，於第二層級中歸類於「兩岸」類別，且屬於第一層級「中國」類別中次層級類別。此外，郵件使用者滿意度與郵件類別名稱適應指標之分佈與趨勢彙整如表 5.7、圖 5.31 與圖 5.32。

表 5.5、個人化郵件類別推論之測試郵件資料（其中 5 份）

發信人	郵件標題	郵件時間	郵件內容
蝸牛	談談老子	Sun May 30 21:48:40 1999	我不否認老子是位智者。但是這位智者的智慧的得來卻是以否定生命達至幻滅境界為代價的。人有各種欲求…
冷暖自知	哲學教材要重寫	Fri Sep 11 20:38:54 1998	「各種形式的信息技術正在改變哲學概念的涵義」，這是參加「第 20 屆世界哲學大會」的哲學教授…
賈磊磊	電影藝術理論研究的現狀	Sun May 30 21:48:40 1999	應當正視，中國電影長期以來主要承擔的是宣傳、教育的職能，自電影製片廠實行企業化管理以後，實際上等於宣告了電影…
GnG	什麼是流星和流星雨	Wed Oct 28 16:55:41 1998	在太陽系中除了九大行星和它們的衛星以外，還有彗星、小行星及一些更小的天體。小天體的體積雖小，但它們也和九大行星…
luxh	中國海軍的發展方向	Thu Aug 20 11:56:01 1998	三月份的英國「詹氏情報和評論」刊出一篇名為「中國海上戰略演變」 的報導，指出中國大陸人民解放軍…

表 5.6、個人化郵件類別推論第一階段實際結果呈現（部份資料）

階層一類別	階層二類別	階層三類別	郵件名稱
中國	兩岸	大陸	中國體育尚需擴大「塔基」
			黨史研究中存在的問題
			東北農大培養新一代北大荒人
			第一張人民幣
		經濟	AMD 推出橄欖球型計算機
			CORBA 結構
			Domino Web 服務器被查出存在安全漏洞
			美國計算機廠商寄希望於使用方便的互聯網設備
			英特爾大力投資互聯網設備公司
	政府	社會	《中國特色社會主義哲學觀》簡介
			正確評價清朝
			中國近代海軍的戰略佈局
			中國載人航天工程負責人談我國載人?
			在中國倡導親子遊戲的特殊意義
			在中國家庭內開展親子遊戲的策略
			談談老子
			騰飛,中國的航天事業 3:45r
		國家	台灣空軍遠距離空戰能力
			中國陸軍航空兵
			中國海軍的發展方向
			中國航天員登九重天指日可待
			評介《鄧小平辯證法思想研究》
			台灣學者贊大陸太空船 稱航天科技達新水準

表 5.7、第一階段「使用者滿意度」與「類別名稱適應指標」指標績效彙整

第一階段試驗者問項回答分數與指標績效統計																									
試驗者編號	A01	A02	A03	A04	A05	B01	B02	B03	B04	B05	C01	C02	C03	C04	C05	D01	D02	D03	D04	D05	E01	E02	E03	E04	E05
試驗者系所	企管系					中文系					資管系					旅遊系					會計系				
項目一、電子郵件組織結構																									
分類結構可靠度	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	1	4	4	4	4	4	4	4	4	4
分類結構變化性	4	4	4	4	4	4	4	4	4	4	4	4	4	4	1	4	4	4	4	4	4	4	4	4	4
類別名稱解釋性	1	1	4	1	1	4	4	4	4	1	1	1	1	1	4	4	4	4	1	1	4	4	1	1	1
項目二、電子郵件檢索功能																									
分群準確性	4	1	4	4	4	1	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
分類名稱靈活性	1	4	4	1	4	4	4	4	4	4	1	1	4	1	4	1	4	4	1	4	1	4	4	4	1
郵件檢索方便性	4	4	1	4	4	1	4	4	4	1	4	4	4	4	4	4	4	4	4	4	4	4	4	4	1
項目三、郵件分群對郵件管理幫助																									
過量郵件整理助益性	4	4	4	4	4	1	4	7	4	4	4	4	4	1	4	4	4	4	1	4	4	4	4	4	4
減少電子郵件檢查頻率	1	1	4	1	4	1	4	4	1	1	4	1	4	4	1	4	4	1	1	4	4	1	4	4	1
項目四、郵件類別名稱對郵件內容閱讀影響																									
減少郵件錯誤通訊	1	4	1	1	1	1	4	1	1	1	4	1	1	4	4	1	4	1	4	1	4	1	1	1	1
協助使用者判別郵件內容重要性	4	4	1	4	4	1	4	4	1	4	1	1	4	1	1	4	4	1	4	4	4	1	1	4	1
項目五、整體推論結果對郵件應用影響																									
協助使用者快速瀏覽收件夾中所有信件	4	4	4	4	4	4	4	4	1	4	4	4	1	4	4	7	4	4	4	4	4	4	4	4	4
減少使用者繁複信箱管理作業	4	4	1	4	4	4	4	7	4	4	4	4	4	7	4	7	4	4	7	4	4	4	4	4	4
協助使用者回覆多筆郵件內容	1	4	1	1	4	1	1	1	4	1	1	4	1	1	1	4	1	1	4	1	4	1	1	4	1
項目六、正確類別名稱統計																									
正確郵件類別名稱與群集勾選	8	7	8	9	7	7	8	7	9	9	7	8	8	9	10	9	7	7	7	7	8	8	10	10	9
使用者期望郵件類別實際數量	11	8	10	9	13	12	11	9	9	8	11	11	13	9	11	13	12	9	8	12	12	11	11	11	9
第一階段推論類別實際數量	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
系統驗證指標																									
郵件使用者滿意度	2.85	3.31	2.85	2.85	3.54	2.38	3.77	4.00	3.08	2.85	3.08	2.85	3.08	3.08	3.08	3.77	3.77	3.08	3.31	3.31	3.77	3.08	3.08	3.54	2.38
類別名稱適應指標	42%	40%	43%	50%	35%	36%	42%	39%	50%	51%	37%	42%	40%	50%	53%	45%	36%	39%	40%	36%	41%	42%	53%	53%	50%
郵件使用者滿意度平均值	3.19																								
類別名稱適應指標平均值	43%																								

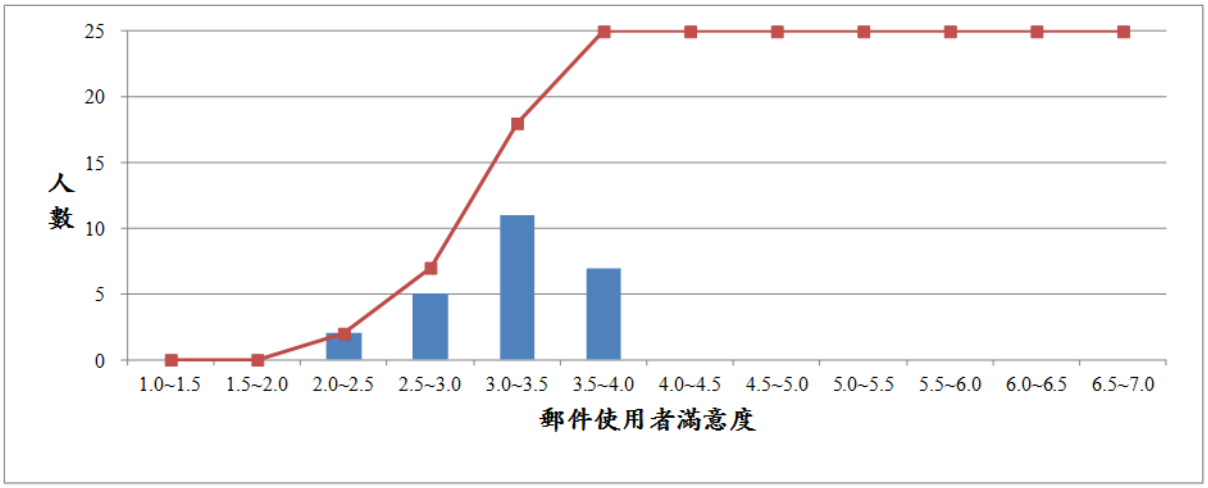

圖 5.31、第一階段郵件使用者滿意度之分佈趨勢

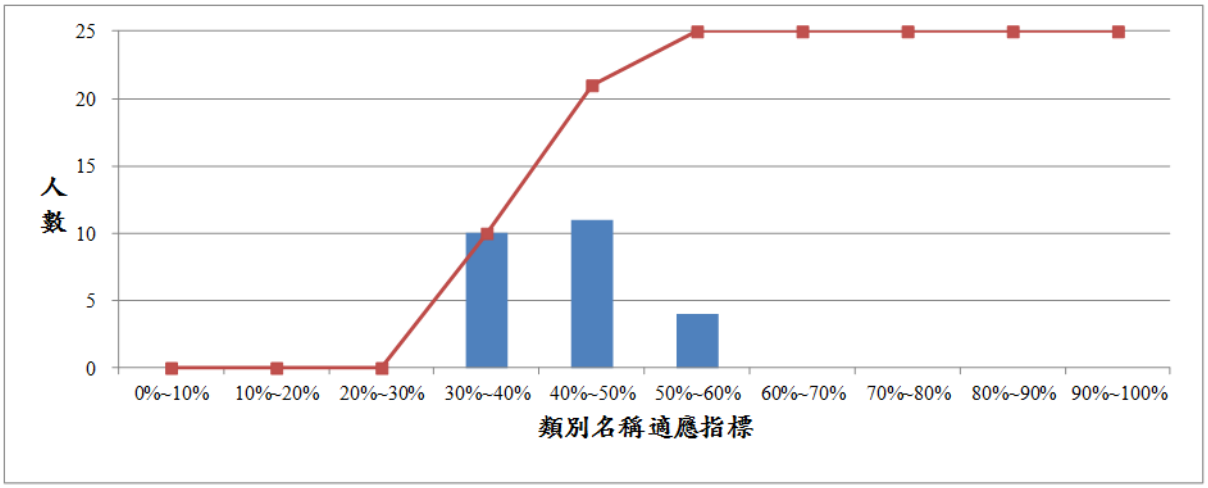

圖 5.32、第一階段郵件類別名稱適應指標之分佈趨勢

藉由表 5.7、圖 5.31、與圖 5.32 統計結果以得知，於第一階段之系統驗證結果中其郵件使用者滿意度約落於 2.38 與 3.77，其平均滿意度為 3.19。另外郵件類別名稱適應指標部分，約落於 30%至 50%間，其平均為 43%。整體而言，根據滿意度指標（滿意度 7、4、1）可得知郵件使用者滿意度皆未達良好滿意度，亦或是郵件類別名稱適應指標績效尚無法準確推論個人化郵件類別名稱。

(II) 個人化郵件類別推論系統第二階段驗證結果分析

第二階段系統驗證之作法乃以第一階段驗證中，針對個人化郵件類別推論部分將第二階段分為分 6 個週期（每週期皆匯入 100 筆不重複之領域文件，共計 600 筆組合），並以第一階段驗證時所選取之 100 筆郵件資料（如表 5.8 所示）重新進行系統績效測試，以瞭解系統於不同訓練資料數量基礎下進行個人化郵件類別名稱推論之績效變化趨

勢，進而分析本系統之學習成長能力。以下乃針對個人化郵件類別推論之各項指標說明系統第二階段各週期驗證過程，並分析系統第一階段與第二階段共七週期之驗證結果。

針對個人化郵件類別推論之第二階段驗證，其最後測試（第七週期）之結果如表 5.9、表 5.10 所示。以表 5.9 為例，於階層式郵件分群所得結果中，郵件「第一張人民幣」與郵件「黨史研究中存在的問題」皆屬相同群集，且推論群集名稱（即類別名稱）為「政府」，於第二層級中「政府」類別與「電腦」類別乃歸類於「生活」類別中，且屬於第一層級「兩岸關係」類別中次層級類別。而最後測試（第七週期）個人化郵件類別推論之「郵件使用者滿意度」與「郵件類別名稱適應指標」績效統計數據如表 5.10。此外本研究於各週期試驗者「郵件使用者滿意度」與「郵件類別名稱適應指標」之統計數據乃彙整如表 5.11、表 5.12-

根據表 5.10 中各試驗者之多數問項分數皆呈現 4 分（可以）以上滿意度，且於類別名稱適應指標中，針對各使用者之適應指標多數表現於 75%以上績效，因此於第二階段之驗證中可發現，本系統所推論之郵件類別名稱可有效針對不同使用者推論適當個人化郵件類別名稱，且於使用者滿意度調查「項目五、整體推論結果對郵件應用影響」中多數試驗者皆給予正向甚至 7 分（非常認同），故代表本研究所提之「個人化郵件類別推論模式」方法論於推論應用上準確性與實際使用之應用性皆具良好效果。

表 5.8、個人化郵件類別推論之測試郵件資料（其中 5 份）

發信人	郵件標題	郵件時間	郵件內容
蝸牛	談談老子	Sun May 30 21:48:40 1999	我不否認老子是位智者。但是這位智者的智慧的得來卻是以否定生命達至幻滅境界為代價的。人有各種欲求…
冷暖自知	哲學教材要重寫	Fri Sep 11 20:38:54 1998	「各種形式的信息技術正在改變哲學概念的涵義」，這是參加「第 20 屆世界哲學大會」的哲學教授…
賈磊磊	電影藝術理論研究的現狀	Sun May 30 21:48:40 1999	應當正視，中國電影長期以來主要承擔的是宣傳、教育的職能，自電影製片廠實行企業化管理以後，實際上等於宣告了電影…
GnG	什麼是流星和流星雨	Wed Oct 28 16:55:41 1998	在太陽系中除了九大行星和它們的衛星以外，還有彗星、小行星及一些更小的天體。小天體的體積雖小，但它們也和九大行星…
luxh	中國海軍的發展方向	Thu Aug 20 11:56:01 1998	三月份的英國「詹氏情報和評論」刊出一篇名為「中國海上戰略演變」 的報導，指出中國大陸人民解放軍…

表 5.9、個人化郵件類別推論第七週期實際結果呈現（部份資料）

階層一類別	階層二類別	階層三類別	郵件名稱
兩岸關係	生活	政府	中國體育尚需擴大「塔基」
			黨史研究中存在的問題
			東北農大培養新一代北大荒人
			第一張人民幣
		電腦	AMD 推出橄欖球型計算機
			CORBA 結構
			Domino Web 服務器被查出存在安全漏洞
			美國計算機廠商寄希望於使用方便的互聯網設備
			英特爾大力投資互聯網設備公司
	中國大陸	策略	《中國特色社會主義哲學觀》簡介
			正確評價清朝
			中國近代海軍的戰略佈局
			中國載人航天工程負責人談我國載人?
			在中國倡導親子遊戲的特殊意義
			在中國家庭內開展親子遊戲的策略
			談談老子
			騰飛,中國的航天事業 3:45r
		軍事	台灣空軍遠距離空戰能力
			中國陸軍航空兵
			中國海軍的發展方向
			中國航天員登九重天指日可待
			評介《鄧小平辯證法思想研究》
			台灣學者贊大陸太空船 稱航天科技達新水準
教育	學校	老師	開展「我要上學了」主題系列活動，激發幼兒強烈的
			美國父母怎樣看待「超常兒童」
			我國職業教育異軍突起
			發達國家職業教育的新趨勢
			關於「素質·能力·個性」教育模式的思考
			中國宋學與東方文明國際學術研討會
			全軍高級幹部學《鄧選》輪訓班開學
			創設良好的教育環境，培養幼兒人際交往能力
			利用動態觀察法，培養幼兒的觀察力
			哲學教材要重寫
			澳大利亞教師發展幼兒體力的做法
			應用開發人員結構演變
			容易忽視的小兒疾病

表 5.10、第七週期「使用者滿意度」與「類別名稱適應指標」指標績效彙整

第七週期試驗者問項回答分數與指標績效統計																									
試驗者編號	A01	A02	A03	A04	A05	B01	B02	B03	B04	B05	C01	C02	C03	C04	C05	D01	D02	D03	D04	D05	E01	E02	E03	E04	E05
試驗者系所	企管系					中文系					資管系					旅遊系					會計系				
項目一、電子郵件組織結構																									
分類結構可靠度	7	7	4	4	7	4	7	7	7	7	4	7	7	7	4	7	4	7	4	7	4	4	7	4	7
分類結構變化性	7	7	7	4	7	4	7	7	4	7	7	7	7	4	4	7	4	7	4	4	7	7	4	7	7
類別名稱解釋性	4	7	4	4	7	4	4	4	4	4	4	4	7	4	4	4	4	7	4	4	4	4	1	4	4
項目二、電子郵件檢索功能																									
分群準確性	7	4	7	4	7	4	7	4	4	7	4	7	7	4	4	4	4	7	4	7	4	7	4	7	4
分類名稱靈活性	4	4	7	4	7	7	4	4	4	4	7	4	7	4	7	7	4	4	4	4	1	7	4	7	7
郵件檢索方便性	7	4	7	7	7	4	7	4	4	7	7	4	4	7	7	4	7	4	7	4	4	4	4	4	1
項目三、郵件分群對郵件管理幫助																									
過量郵件整理助益性	7	4	7	7	7	7	4	7	4	7	4	7	7	4	7	4	7	4	4	4	4	7	4	7	7
減少電子郵件檢查頻率	4	4	4	4	4	4	4	4	4	4	4	4	4	4	7	4	4	4	4	4	4	4	4	4	4
項目四、郵件類別名稱對郵件內容閱讀影響																									
減少郵件錯誤通訊	4	4	1	4	4	4	4	4	4	1	4	4	4	4	4	4	4	4	4	4	4	1	1	1	1
協助使用者判別郵件內容重要性	4	4	4	4	4	4	4	4	7	4	4	4	4	4	4	4	4	4	4	4	4	1	1	4	1
項目五、整體推論結果對郵件應用影響																									
協助使用者快速瀏覽收件夾中所有信件	7	4	7	4	4	7	4	7	7	7	7	4	4	4	7	7	4	4	7	7	4	4	4	7	4
減少使用者繁複信箱管理作業	7	7	7	4	4	4	4	7	4	7	7	7	4	7	7	7	4	4	7	7	7	4	4	7	7
協助使用者回覆多筆郵件內容	1	4	1	4	4	4	4	1	4	1	1	4	1	1	1	4	4	4	4	1	4	1	1	4	1
項目六、正確類別名稱統計																									
正確郵件類別名稱與群集勾選	15	14	16	13	15	15	16	15	15	16	15	15	15	16	15	14	15	16	15	15	15	15	15	14	15
使用者期望郵件類別實際數量	11	9	9	9	11	9	12	9	9	9	11	9	11	9	10	11	11	9	11	11	9	11	9	9	9
第一階段推論類別實際數量	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
系統驗證指標																									
郵件使用者滿意度	5.38	4.92	5.15	4.46	5.62	4.69	4.92	4.92	4.69	5.15	4.92	5.15	5.15	4.46	5.15	5.15	4.46	4.92	4.69	4.69	4.23	4.23	3.31	5.15	4.23
類別名稱適應指標	79%	78%	89%	72%	79%	83%	82%	83%	83%	89%	79%	83%	79%	89%	81%	74%	79%	89%	79%	79%	83%	79%	83%	78%	83%
郵件使用者滿意度平均值	4.79																								
類別名稱適應指標平均值	81%																								

表 5.11、各週期郵件使用者滿意度指標彙整

週期	試驗者編號與使用者滿意度																									
	A01	A02	A03	A04	A05	B01	B02	B03	B04	B05	C01	C02	C03	C04	C05	D01	D02	D03	D04	D05	E01	E02	E03	E04	E05
	企管系					中文系					資管系					旅遊系					會計系				
第一週期	2.85	3.31	2.85	2.85	3.54	2.38	3.77	4.00	3.08	2.85	3.08	2.85	3.08	3.08	3.08	3.77	3.77	3.08	3.31	3.31	3.77	3.08	3.08	3.54	2.38
第二週期	2.85	3.54	3.54	3.08	3.77	2.62	3.77	4.23	3.31	3.08	3.31	3.08	3.31	3.31	3.08	4.00	3.77	3.08	3.31	4.00	4.00	3.08	3.08	4.00	2.62
第三週期	3.08	4.00	4.00	3.54	4.00	4.69	3.77	4.46	3.54	3.31	3.77	3.77	3.54	3.77	3.54	4.46	3.77	3.54	3.77	4.46	4.00	3.31	3.31	4.00	3.08
第四週期	4.69	4.46	5.15	3.77	4.23	4.00	4.69	4.46	3.77	4.23	4.23	4.69	3.54	4.00	4.23	4.46	4.00	4.00	4.69	4.69	4.23	3.31	3.31	4.23	3.77
第五週期	4.92	4.92	5.15	4.23	5.62	4.46	4.92	4.69	4.69	4.69	4.92	5.15	3.77	4.23	4.46	5.15	4.46	4.92	4.69	4.69	4.23	4.23	3.31	5.15	4.23
第六週期	5.38	4.92	5.15	4.46	5.62	4.46	4.92	4.69	4.69	5.15	4.92	5.15	4.00	4.23	5.15	5.15	4.46	4.92	4.69	4.69	4.23	4.23	3.31	5.15	4.23
第七週期	5.38	4.92	5.15	4.46	5.62	4.69	4.92	4.92	4.69	5.15	4.92	5.15	5.15	4.46	5.15	5.15	4.46	4.92	4.69	4.69	4.23	4.23	3.31	5.15	4.23
平均值	4.16	4.30	4.43	3.77	4.63	3.90	4.39	4.49	3.97	4.07	4.16	4.26	3.77	3.87	4.10	4.59	4.10	4.07	4.16	4.36	4.10	3.64	3.24	4.46	3.51

表 5.12、各週期郵件類別名稱適應指標彙整

週期	試驗者編號與類別名稱適應指標																									
	A01	A02	A03	A04	A05	B01	B02	B03	B04	B05	C01	C02	C03	C04	C05	D01	D02	D03	D04	D05	E01	E02	E03	E04	E05
	企管系					中文系					資管系					旅遊系					會計系				
第一週期	42%	40%	43%	50%	35%	36%	42%	39%	50%	51%	37%	42%	40%	50%	53%	45%	36%	39%	40%	36%	41%	42%	53%	53%	50%
第二週期	51%	56%	41%	56%	45%	51%	42%	56%	47%	51%	53%	42%	61%	50%	53%	45%	56%	50%	51%	56%	42%	42%	57%	56%	50%
第三週期	61%	68%	68%	51%	63%	51%	63%	42%	56%	62%	47%	56%	47%	68%	51%	60%	47%	61%	63%	56%	53%	72%	53%	50%	50%
第四週期	61%	74%	78%	68%	63%	67%	63%	50%	79%	63%	79%	70%	58%	72%	67%	55%	62%	72%	63%	67%	72%	68%	83%	78%	72%
第五週期	68%	78%	89%	68%	78%	83%	82%	83%	79%	82%	79%	83%	74%	89%	67%	74%	77%	89%	79%	77%	83%	74%	83%	74%	83%
第六週期	79%	78%	89%	68%	78%	83%	82%	83%	79%	82%	79%	83%	79%	89%	77%	74%	77%	89%	79%	77%	83%	79%	83%	74%	83%
第七週期	79%	78%	89%	72%	79%	83%	82%	83%	83%	89%	79%	83%	79%	89%	81%	74%	79%	89%	79%	79%	83%	79%	83%	78%	83%
平均值	63%	67%	71%	62%	63%	65%	65%	62%	68%	69%	65%	66%	63%	72%	64%	61%	62%	70%	65%	64%	65%	65%	71%	66%	67%

此外，本研究除了針對系統推論結果進行正確性績效之驗證外，於第二階段乃觀察系統之學習趨勢與不同數量領域文件對系統績效影響。以下乃針對「郵件使用者滿意度」與「郵件類別名稱適應指標」兩指標之成長趨勢進行說明。

- **郵件使用者滿意度績效分佈趨勢**

郵件使用者滿意度乃代表郵件使用者對推論結果進行實用性評估所得分數。本研究乃將各週期之郵件使用者滿意度彙整如表 5.13 及圖 5.33。由表 5.12 中可發現於第一週期使用者滿意度平均為 3.19，而多數使用者滿意度不佳且系統所匯入領域文件為 200 份文件，亦代表系統於此階段中因匯入訓練資料（即領域文件）上未足夠導致系統推論結果缺乏準確性，故多數使用者認為推論結果不符使用者期望；於第四週期過後多數使用者之滿意度平均提升至 4 分以上，則代表系統匯入 500 份以上領域文件後，其系統推論類別名稱結果符合多數使用者需求。此外，由圖 5.33 中得知郵件使用者滿意度之趨勢分佈於第五週期逐漸平緩呈現收斂狀態，故代表系統匯入 600 筆領域文件後（即第五週期）推論結果具一定準確性，且多數使用者認同系統推論結果能幫助使用者歸類電子郵件。

表 5.13、郵件使用者滿意度績效彙整

個人化郵件類別推論		各週期個人化郵件類別推論驗證－領域文件匯入數量							平均
		第一階段	第二階段						
		第一週期	第二週期	第三週期	第四週期	第五週期	第六週期	第七週期	
		200 筆	300 筆	400 筆	500 筆	600 筆	700 筆	800 筆	
郵件使用者滿意度	平均值	3.19	3.39	3.78	4.19	4.64	4.72	4.79	4.10
	標準差	0.42	0.45	0.43	0.46	0.50	0.51	0.48	0.46
	成長率	--	6.37%	11.43%	10.99%	10.56%	1.79%	1.56%	7.12%

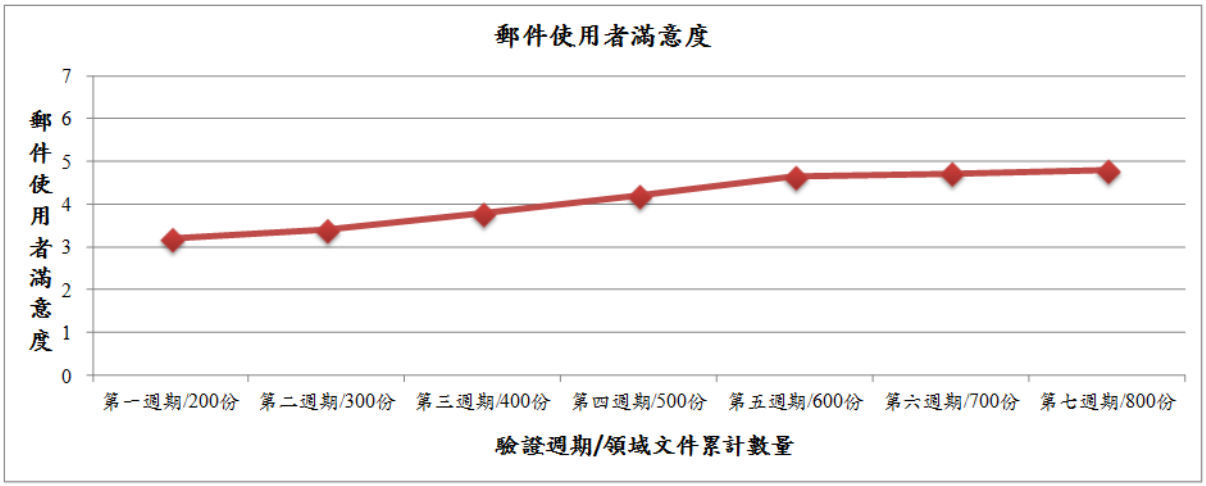

圖 5.33、各驗證週期之郵件使用者滿意度指標分佈趨勢

此外，本研究觀察各系所試驗者滿意度指標並彙整如表 5.14，當中，企管系試驗者平均滿意度於第七週期達 5 分以上水準，而會計系試驗者平均滿意度於第七週期呈現 4.23 分之水準，且於七週期總平均值中發現，各系所接坐落於 4 分上下，且無明顯差距；此外根據各週期平均成長率各系所於每週期平均成長幅度坐落於 4%~8%區間，各系所間成長率無較大落差間隔，意表示系統推論個人化郵件類別於訓練資料數量增加下，其系統推論結果趨於準確，且意符合知識背景不同使用者所期望郵件類別。

表 5.14、各系所郵件使用者滿意度指標各週期平均值彙整表

週期	各週期系所試驗者滿意度指標平均值				
	企管系	中文系	資管系	旅遊系	會計系
第一週期	3.08	3.22	3.03	3.45	3.17
第二週期	3.36	3.40	3.22	3.63	3.36
第三週期	3.72	3.95	3.68	4.00	3.54
第四週期	4.46	4.23	4.14	4.37	3.77
第五週期	4.97	4.69	4.51	4.78	4.23
第六週期	5.11	4.78	4.69	4.78	4.23
第七週期	5.11	4.87	4.97	4.78	4.23
總平均值	4.26	4.16	4.03	4.26	3.79
平均成長率	8%	6%	7%	5%	4%

- **郵件類別名稱適應指標績效分佈趨勢**

郵件類別名稱適應指標乃代表系統推論郵件類別名稱適用性與正確性。本研究乃將各週期之郵件使用者滿意度彙整如表 5.15 及圖 5.34。由表 5.15 中可發現於第一週期適應指標平均為 43.37%，系統所匯入領域文件為 200 份文件時，則表示第一週期時推論類別名稱多數適用性不佳，亦無法表達郵件群集內所含括郵件類型；於第四週期過後適應指標平均雖提升至 68.17%，但績效表現普通亦表示多數郵件類別名稱無法適用，於第五週期系統匯入 600 份文件後，適應指標平均即達到 79%以上準確性，且於圖 5.34 中得知郵件類別名稱適應指標之趨勢分佈於第五週期逐漸平緩呈現收斂狀態，亦代表本系統於匯入 600 筆訓練文件後則可有效針對不同郵件使用者推論個人化郵件類別名稱。

表 5.15、郵件類別名稱適應指標績效彙整

個人化郵件類別推論		各週期個人化郵件類別推論驗證－領域文件匯入數量							平均
		第一階段	第二階段						
		第一週期	第二週期	第三週期	第四週期	第五週期	第六週期	第七週期	
		200 筆	300 筆	400 筆	500 筆	600 筆	700 筆	800 筆	
郵件類別名稱適應指標	平均值	43.37%	50.44%	56.93%	68.17%	79.01%	80.26%	81.40%	65.66%
	標準差	6.01%	5.71%	8.00%	8.05%	6.21%	4.87%	4.42%	6.18%
	成長率	--	16.29%	12.87%	19.74%	15.90%	1.59%	1.42%	11.30%

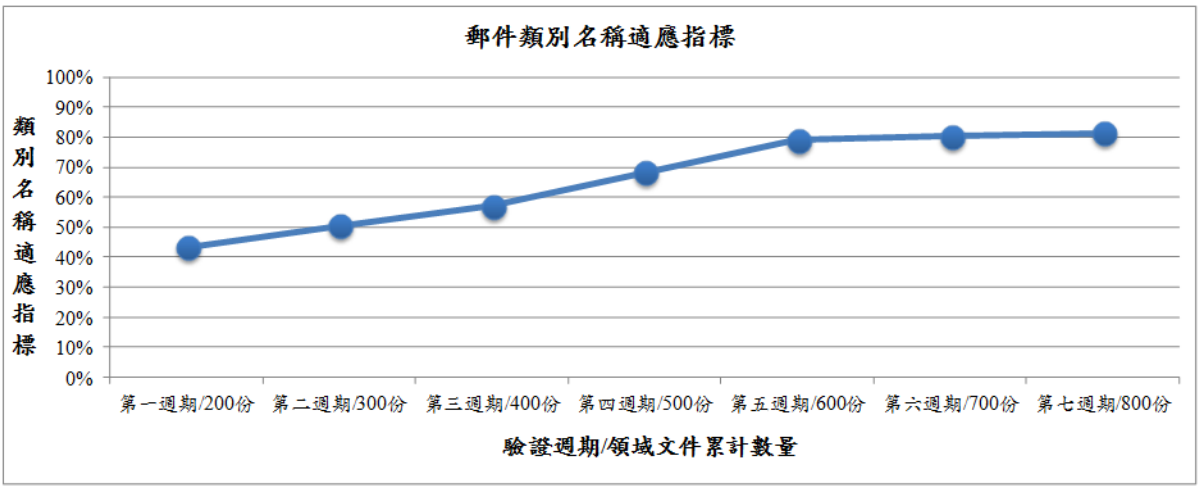

圖 5.34、各驗證週期之郵件類別名稱適應指標分佈趨勢

5.3個人化郵件類別推論整體驗證結果分析

綜合兩階段之驗證成效，各項驗證指標之相關結果整理如表5.16所示。由表5.16整理結果可得知，各項驗證指標之「收斂前每週期平均成長率」及「整體每週期平均成長率」皆為正數成長，且各項驗證指標皆於第五週期內呈現收斂狀態，於表5.16中可發現兩者指標平均值皆屬績效分數中位數，但以績效分數最大值比較下，郵件類別名稱適應指標表現較郵件使用者滿意度為佳，此種結果乃因郵件使用者滿意度以使用者主觀之意見進行評估，故每位使用者給分方式皆具差異，而郵件類別名稱適應指標則單獨針對推論類別正確數量進行計算，因此，若單以郵件類別名稱適應指標檢視系統績效則可有效且正確推論郵件類別名稱。而郵件使用者滿意度之平均值（4.10分）雖未到達7分（非常認同）之分數，但分數座落於4分（可以）以上，故代表多數使用者認同系統推論郵件類別名稱能幫助使用者管理電子郵件信箱。

表 5.16、個人化郵件類別推論綜合兩階段之驗證績效彙整

驗證指標	整體平均值	績效分數最大值	收斂週期	收斂週期領域文件累計數量	整體每週期平均成長率
郵件使用者滿意度	4.10	4.79	第五週期	600筆	7.12%
郵件類別名稱適應指標	65.66%	81.40%	第五週期	600筆	11.30%

整體而言，由第一階段與第二階段之驗證分析中，可得知當使用者進行「個人化郵件類別推論」時，訓練資料（即領域文件）與週期乃影響系統績效之主要原因，系統之績效將隨著不同週期及訓練量之增加，呈現持續成長之狀況，並於最終達到穩定且良好之績效水平。針對兩驗證指標之分析結果彙整如下：

- 針對郵件使用者滿意度部分，由驗證結果以得知於領域文件筆數為200筆時，其滿意度平均值為3.19未達4分（可以）之分數，亦無法取得多數試驗者之認同，待匯入約600筆領域文件時達到收斂，並於第七週期時郵件使用者滿意度平均值由3.19成長至4.79。有鑑於上述之數據，即表示多數試驗者認同系統推論之結果能有效協助使用者管理電子郵件信箱。
- 針對郵件類別名稱適應指標部分，由驗證結果以得知於領域文件筆數為200筆時，其郵件類別名稱適應指標平均為43%。待匯入約600筆領域文件時即可達到收斂，平均每週期郵件類別名稱適應指標由43%成長至79%，且於第七週期乃成長至81%。綜合上述數據，即表示本系統之個人化郵件類別推論具可行性。

第六章、結論與未來發展

現行電子信箱為協助使用者歸類過量郵件，因此提供使用者以人工方式建立郵件類別（亦即郵件資料夾），以協助使用者透過自訂郵件類別檢索郵件。然而使用者所接收郵件數量過於龐大，導致使用者需耗費大量時間詳細瀏覽郵件內容並建立類別；然而，過去研究所發展郵件分類技術皆以特定領域或公共郵件為分析對象，因此所區分類別較為單純且一致性。但一般使用者所使用之私人郵件大多隱含著使用者專業領域知識或個人生活經歷，故郵件使用領域較為複雜，導致現有郵件分類技術容易產生失效問題，且容易隨著時間、身分等使用情境改變而不同，而需針對不同使用者產生適合使用者個人之郵件類別。

為解決上述問題，本研究乃針對電子郵件內容先行解析郵件使用領域特徵，並取得電子郵件之代表性特徵詞彙，進而發展一套「個人化郵件類別推論模式」，本模式將解析電子郵件之郵件主題、文章內容以及相關性語彙，並以此為基歸納使用者特徵資訊，進而與使用者寄發郵件之特徵字詞進行語意關聯，以推論使用者接收郵件類型。另外，本研究亦建立個人件類別推論法則，以分群法則為基礎並透過郵件內容特徵區分郵件類型並賦予郵件類型名稱，以作為個人化郵件類別，協助使用者解讀該類別所歸納之郵件類型，並協助電子郵件使用者辨識郵件類型且提高郵件整理效率。以下將於第 6.1 小節總結本研究所完成之工作與任務，並於第 6.2 小節歸納本研究未來發展之議題與方向。

6.1 論文總結

依據第 1.2 小節之「研究步驟」所述，本論文可將完成工作分為三項任務，以下乃分別說明各項任務之成果。

1. 發展「個人化郵件類別推論模式」方法論

本研究乃針對一般使用者私人郵件進行解析，並發展一套「個人化郵件類別推論模式」方法論。當中方法論可包含「郵件語意關鍵字擷取模組」與「個人化郵件類別推論模組」。透過「郵件語意關鍵字擷取模組」以判斷並擷取郵件中具語言含意之詞彙。再由「個人化郵件類別推論模組」透過分析郵件中語意詞彙擷取郵件中代表性特徵詞彙，並藉由特徵詞彙進行郵件階層劃分群，接著，為協助使用者解讀群集內所含郵件，本研究乃將群集中所含郵件特徵詞彙與本研究所制定之類別名稱詞彙庫解析詞彙關聯，以取

得使用者之個人化類別名稱。

2. 建置「個人化郵件類別推論系統」

本研究乃以「個人化郵件類別推論模式」方法論為依據，運用 JSP(Java Server Pages) 語法進行開發，並利用 SQL Server 2008 來存取資料庫，以建置一套「個人化郵件類別推論系統」，並使系統之使用者能進行上傳與查詢電子郵件、分析郵件中代表性特徵、推論使用者個人之郵件類別名稱、查詢郵件類別名稱推論結果以及設定系統中相關預設值等維護作業。

3. 方法論與系統績效驗證

為驗證本研究所提出方法論與系統之正確性與有效性，本研究乃以「李榮陸文本分析語料庫」內蒐集郵件資料為例，並以行政院發行之「台灣光華雜誌」新聞內容作為領域文件（即訓練文件），並針對「個人化郵件類別推論」議題進行驗證，並以「郵件使用者滿意度」與「郵件類別名稱適應指標」兩指標驗證與評估本研究之個人化郵件類別名稱之實用性。

綜合上述內容，本研究所提出之「個人化郵件類別推論模式」方法論與「個人化郵件類別推論」系統可有效將電子郵件解析並取得代表性特徵詞彙，以進行郵件分群，並能將電子郵件群集推論適合類別名稱，以協助使用者檢索郵件。以下即以「理論方法」、「技術開發」與「實務應用」等三種層面說明本研究之具體貢獻與成效。

理論方法層面

本研究乃以使用者之私人郵件為基礎，發展一套「個人化郵件類別推論」模式，並根據使用者個人所接收郵件類型自動劃分郵件類別，並針對郵件敘述內容自動推論郵件類別名稱，郵件使用者即可透過郵件類別稱檢索所需郵件，以達到自動化與個人化郵件類別建立。此方法論之相關重點成效乃歸納如下：

- 本研究乃解析私人郵件內容之詞彙，並針對詞彙之語言含意進行分析，以取得郵件中具代表性語意詞彙並建立郵件中代表性特徵。
- 本研究提出之個人化郵件類別即可針對不同使用者以分群方式歸類郵件，以取得適合使用者之郵件整理模式，此外為方便使用者檢索群集內郵件內容，亦推論郵件群

集分析類別名稱，以取得針對使用者製訂之個人化郵件類別與郵件分類管理方式。

技術開發層面

本研究乃以「個人化郵件類別推論」方法論為依據，並以JSP（Java Server Pages）以及SQL Server 2008等系統開發工具建置「個人化郵件類別推論」系統，其系統具體成效乃歸納如下：

- 本系統所開發之「個人化郵件類別推論」乃建置於網際網路中，並可執行「郵件語意關鍵字擷取」與「個人化郵件類別推論」以取得郵件中代表性詞彙、郵件群集，以及針對使用者推論之個人化郵件類別名稱。
- 由驗證結果得知，本系統之各項驗證指標（即「郵件使用者滿意度」與「郵件類別名稱適應指標」）皆具學習能力；當使用至一定訓練資料數量系統即可有效推論使用者之個人化郵件類名稱。

實務應用層面

本研究可有效針對私人郵件進行類別劃分，並根據使用者個人需求自動建立專屬個人郵件類別，以及協助使用者管理郵件，故本系統可有效應用於一般使用者之私人郵件管理問題等任務。其具體貢獻與成效乃歸納如下：

- 本系統可應用於網頁電子郵件信箱網站，以有效減少使用者人工分類時間，並協助使用者檢索與郵件篩檢等郵件管理任務。
- 本系統可應用於電子郵件管理軟體，以自動化方式建立郵件類別，並協助使用者檢索與郵件篩檢等郵件管理任務。

6.2 未來展望

依據第 6.1 小節所述，本研究乃完成研究步驟所規劃之各項任務，並提出「個人化郵件類別推論」方法論與「個人化郵件類別推論」系統理論層面、技術層面以及實務應用層面之成效與貢獻。而後續發展方面，綜合本論文之研究成果與既有文獻再結合未來資訊發展之拓展，發現本論文尚有若干研究主題具有深入研究之價值，歸納未來發展方向如下：

1. 發展郵件回覆內容自動生成與推薦模式

本研究主要發展個人化與自動化之使用者個人郵件類別推論，故主要乃著重於協助

使用者管理並區分電子郵件，進而方便使用者檢索所需郵件。但本研究尚未針對使用者之信件回覆等相關郵件管理動作進行深入探討，故期望未來能透過使用者電子信箱中郵件內容進行解析，並根據使用者欲回覆郵件自動生成回覆內容，進而協助使用者快速回覆電子郵件訊息，以增加訊息傳遞便利性與速度。

2. 發展多語言郵件內容解析模式

本研究主要解析郵件中內容敘述，並分析敘述內容代表特徵與類型，進一步推論使用者個人之郵件類別名稱。現階段本研究主要針對繁體中文郵件之文本內容進行解析，然而中文文字之詞彙用法與各國語言不同，因此導致本研究無法分析非中文語系之電子郵件，故期望藉由多語言郵件解析技術發展出一套多語言郵件內容解析系統，以協助系統進行多國語言郵件之資料解析。

3. 發展非純文字格式郵件解析系統

由於本研究方法論以分析郵件文字內容為主，然而多數一般使用者大多以網頁式電子信箱為主要使用信箱，而網頁式電子信箱大多主要以 HTML 格式編寫郵件，但本研究乃以文字內容為解析對象，故無法解析以 HTML 網頁語言所編寫郵件，故本研究期望未來可發展支援多種格式郵件解析技術，並融入本研究方法論中，促使本研究可解析格式更加多元。

參考文獻

1. Alberts, I. and Forest, D., 2012, “Email pragmatics and automatic classification: A study in the organizational context,” *Journal of the American Society for Information Science and Technology*, Vol. 63, No. 5, pp. 904-922.
2. Andreas, N. and Marcin, D., 2006, “Externally growing self-organizing maps and its application to e-mail database visualization and exploration,” *Applied Soft Computing*, Vol. 6, No. 4, pp. 357-371.
3. Appavu, S., Rajaram, R., Muthupandian, M., Athiappan, G. and Kashmeera, K. S., 2009, “Data mining based intelligent analysis of threatening e-mail,” *Knowledge-Based Systems*, Vol. 22, No. 5, pp. 392-393.
4. Asuncion, U. A., Smyth, P. and Welling, M., 2011, “Asynchronous distributed estimation of topic models for document analysis,” *Statistical Methodology*, Vol. 8, No. 1, pp. 3-17.
5. Atkinson, J., Ferreira, A. and Aravena, E., 2009,“Discovering implicit intention-level knowledge from natural-language texts,” *Knowledge-Based Systems*, Vol. 22, No. 7, pp. 502-508.
6. Banday, M. T., Mir, F. A., Qadri, J. A. and Shah N. A., 2011, “Analyzing Internet e-mail date-spoofing,” *Digital Investigation*, Vol. 7, No. 3-4, pp. 145-153.
7. Baruch, Y., 2005, “Bullying on the net: adverse behavior on e-mail and its impact,” *Information & Management*, Vol.42, pp. 361-371.
8. Bhor, M. and Mason, H. L., 2006, “Development and validation of a scale to assess attitudes of health care administrators toward the use of e-mail communication between patients and physicians,” *Research in Social and Administrative Pharmacy*, Vol. 2, No. 4, pp. 512-532.
9. Bouguila, N. and Amayri O., 2009, “A discrete mixture-based kernel for SVMs: Application to spam and image categorization,” *Information Processing & Management*, Vol. 45, No. 6, pp. 631-642.
10. Buffett, S. and Geng, L., 2010, “Using classification methods to label tasks in process mining,” *Journal of Software Maintenance & Evolution: Research & Practice*, Vol. 22, No. 6/7, pp. 497-517.

11. Burgess, A., Jackson T. and Edwards J., 2004, “Email training significantly reduces email defects,” *International Journal of Information Management*, Vol. 25, pp. 71-83.
12. Cai, D. M., Gokhale, M. and Theiler J., 2007, “Comparison of feature selection and classification algorithms in identifying malicious executables,” *Computational Statistics & Data Analysis*, Vol. 51, No. 6, pp. 3156-3172.
13. Chang, M. and Poon, C. K., 2009, “Using phrases as features in email classification,” *Journal of Systems and Software*, Vol. 82, pp. 1036-1045.
14. Chundi, P., Subramaniam, M. and Vasireddy, D. K., 2009, “An approach for temporal analysis of email data based on segmentation,” *Data & Knowledge Engineering*, Vol. 68, No. 11, pp. 1253-1270.
15. Çıltık, A. and Güngör, T., 2008, “Time-efficient spam e-mail filtering using n-gram models,” *Pattern Recognition Letters*, Vol. 29, No. 1, pp. 19-33.
16. Cornwall, A., Moore, S. and Plant, H., 2008, “Embracing technology: Patients’, family members’ and nurse specialists’ experience of communicating using e-mail,” *European Journal of Oncology Nursing*, Vol. 12, pp. 198-208.
17. Coussement, K. and Poel, D. V., 2008, “Improving customer complaint management by automatic email classification using linguistic style features as predictors,” *Decision Support Systems*, Vol. 44, No. 4, pp. 870-882.
18. Couzenza, T., 2009, “Using electronic mail to motivate students,” *Teaching and Learning in Nursing*, Vol. 4, pp. 76-78.
19. Crawford, E., Koprinska, I. and Webster, P. J., 2004, “Phrases and feature selection in e-mail classification,” *Proceedings of the 9th Australasian Document Computing Symposium, Melbourne, Australia.*
20. Duan, D., Guo, S. Z., Li, Z. B. and Liu, S., 2007, “Community mining based on email classification,” *Journal of Computer Applications*, Vol. 27, No. 12, pp. 3039-3041.
21. Duan, Z., Dong, Y. and Gopalan, K., 2007, “DMTP: Controlling spam through message delivery differentiation,” *Computer Networks*, Vol. 51, No. 10, pp. 2616-2630.
22. Gains, J., 1999, “Electronic Mail—A New Style of Communication or Just a New Medium?: An Investigation into the Text Features of E-mail,” *English for Specific Purposes*, Vol. 18, No. 1, pp. 81-101.

23. Gomez, J. C. and Moens, M. F., 2012, “PCA document reconstruction for email classification,” *Computational Statistics & Data Analysis*, Vol. 56, No. 3, pp. 741-751.
24. Guzella, T. S., Mota-Santos, T. A., Uchôa, J. Q. and Caminhas, W. M., 2008, “Identification of spam messages using an approach inspired on the immune system,” *Biosystems*, Vol. 92, No. 3, pp. 215-225.
25. Hadjidj, R., Debbabi, M., Lounis, H., Iqbal, F., Szporer, A. and Benredjem, D., 2009, “Towards an integrated e-mail forensic analysis framework,” *Digital Investigation*, Vol. 5, No. 3-4, pp. 124-137.
26. Hassini, E., 2006, “Student–instructor communication: The role of email,” *Computers & Education*, Vol. 47, pp. 29-40.
27. Herzberg, A., 2009, “DNS-based email sender authentication mechanisms: A critical review,” *Computers & Security*, Vol. 28, No. 8, pp. 731-742.
28. Hobbs, J., Wald, J., Jagannath, Y. S., Kittler, A. Pizziferri, L. Volk, L. A., Middleton, B. and Bates, D. W., 2003, “Opportunities to enhance patient and physician e-mail contact,” *International Journal of Medical Informatics*, Vol. 70, pp. 1-9.
29. Hu, C., Wong, A. F. L., Cheah, H. M. and Wong, P., 2009, “Patterns of email use by teachers and implications: A Singapore experience,” *Computers & Education*, Vol. 53. pp. 623-631.
30. Huang, E. Y. , Lin, Sh. W. and Lin, S. C., 2011, “A quasi-experiment approach to study the effect of e-mail management training Original Research Article,” *Computers in Human Behavior*, Vol. 27, No. 1, pp. 522-531.
31. Hung, S. Y., Huang, A. H., Yen D. C. and Chang C. M., 2007, “Comparing the task effectiveness of instant messaging and electronic mail for geographically dispersed teams in Taiwan,” *Computer Standards & Interfaces*, Vol. 29, No. 6, pp. 626-634.
32. Iqbal, F., Binsalleeh, H., Fung, B. C.M. and Debbabi, M., 2010, “Mining writeprints from anonymous e-mails for forensic investigation,” *Digital Investigation*, Vol. 7, No. 1-2, pp. 56-64.
33. Iqbal, F., Hadjidj, R. B., Fung, C. M. and Debbabi M., 2008, “A novel approach of mining write-prints for authorship attribution in e-mail forensics,” *Digital Investigation*, Vol. 5, pp. S42-S51.

34. Irena, K., Josiah, P., James, C. and Jason, C., 2007, “Learning to classify e-mail,” *Information Sciences*, Vol. 177, pp. 2167-2187.
35. Islam, M. R., Zhou, W., Guo, M. and Xiang, Y., 2009, “An innovative analyser for multi-classifier e-mail classification based on grey list analysis,” *Journal of Network and Computer Applications*, Vol. 32, No.2, pp. 357-366.
36. Itakura, H., 2004, “Changing cultural stereotypes through e-mail assisted foreign language learning,” *System*, Vol. 32, pp. 37-51.
37. Jerejian, A. C. M., Reid, C. and Rees, C. S., 2013, “The contribution of email volume, email management strategies and propensity to worry in predicting email stress among academics,” *Computers in Human Behavior*, Vol. 29, No. 3, pp. 991-996.
38. Joung, Y. J. and Yang, C. J, 2009, “Email licensing,” *Journal of Network and Computer Applications*, Vol. 32, No. 3, pp. 538-549.
39. Kadoya, Y., Fuketa, M., Atlam, E., Morita, K., Kashiji, S. and Aoe, J. I., 2004, “An efficient e-mail filtering using time priority measurement,” *Information Sciences*, Vol. 166, No. 1-4, pp. 213-229.
40. Laorden, C., Santos, I., Sanz, B., Alvarez, G. and Bringas P. G., 2011, “Word sense disambiguation for spam filtering,” *Electronic Commerce Research and Applications, In Press, Corrected Proof, Available online.*
41. Laura, B. and Maria, C. C., 2001, “*Models of mail server workloads*,” *Performance Evaluation*, Vol. 46, No. 2-3, 2001, pp. 65-76.
42. Lewis, C. E., Thompson, L. F., Wuensch, K. L., Grossnickle, W. F. and Cope, J. G., 2004, “The impact of recipient list size and priority signs on electronic helping behavior,” *Computers in Human Behavior*, Vol. 20, No. 5, pp. 633-644.
43. Li, C. H. and Huang J. X., 2012, “Spam filtering using semantic similarity approach and adaptive BPNN,” *Neurocomputing*, Vol. 92, pp. 88-97.
44. Li, Y. and Shawe-Taylor, J., 2007, "Advanced learning algorithms for cross-language patent retrieval and classification," *Information Processing & Management*, Vol. 43, No. 5, pp. 1183-1199.
45. Mano, R. S. and Mesch, G. S., 2010, “E-mail characteristics, work performance and distress,” *Computers in Human Behavior*, Vol. 26, No. 1, pp. 61-69.

46. Mao, C. H., Lee, H. M. and Yeh, C. F., 2011, “Adaptive e-mails intention finding system based on words social networks,” *Journal of Network and Computer Applications*, Vol. 34, No. 5, pp. 1615-1622.
47. Marsono, M. N., El-Kharashi, M. W. and Gebali, F., 2009, ”Targeting spam control on middleboxes: Spam detection based on layer-3 e-mail content classification,” *Computer Networks*, Vol. 53, No. 6, pp. 835-848.
48. Meng, J., Lin, H. and Yu, Y., 2011, “A two-stage feature selection method for text categorization,” *Computers & Mathematics with Applications*, Vol. 62, No. 7, pp. 2793-2800.
49. Merisavo, M. and Raulas, M., 2004, “The impact of e-mail marketing on brand loyalty,” *Emerald Group Publishing Limited*, Vol. 13, pp. 498-505.
50. Mohammad, A. H. and Zitar, R. A., 2011, “Application of genetic optimized artificial immune system and neural networks in spam detection,” *Applied Soft Computing*, Vol. 11, No. 4, pp. 3827-3845.
51. Nagabhushan, P., Angadi, S. A. and Anami, B. S., 2009, “A soft computing model for mapping incomplete/approximate postal addresses to mail delivery points,” *Applied Soft Computing*, Vol. 9, No. 2, pp. 806-816.
52. Neese, W. T., Ferrell, L. and Ferrell, O. C., 2005, “An analysis of federal mail and wire fraud cases related to marketing,” *Journal of Business Research*, Vol. 58, No. 7, pp. 910-918.
53. O’Kane, P. and Hargie, O., 2007, “Intentional and unintentional consequences of substituting face-to-face interaction with e-mail: An employee-based perspective,” *Interacting with Computers*, Vol. 19, pp. 20-31.
54. Okolica, J. S., Peterson, G. L. and Mills, R. F., 2007, “Using author topic to detect insider threats from email traffic,” *Digital Investigation*, Vol. 4, No. 3-4, pp. 158-164.
55. Phan, R. C.-W., 2008, “Cryptanalysis of e-mail protocols providing perfect forward secrecy,” *Computer Standards & Interfaces*, Vol. 30, No.3, pp. 101-105.
56. Poon, C. K. and Chang, M., 2003., “An email classifier based on resemblance,” *Proceedings of the 14th International Symposium on Methodologies for Intelligent Systems*, Vol. 4, No. 1, pp. 334-338.

57. Rao, H., Cheng, Y. H., Chang, K. H. and Lin, Y. B., 2003, “iMail: A WAP mail retrieving system,” *Information Sciences*, Vol. 151, pp. 71-91.
58. Sakurai, S. and Suyama, A., 2005, “An e-mail analysis method based on text mining techniques,” *Applied Soft Computing*, Vol. 6, No. 1, pp. 62-71.
59. Salcedo-Campos, F., Díaz-Verdejo, J. and García-Teodoro, P., 2012, “Segmental parameterisation and statistical modelling of e-mail headers for spam detection,” *Information Sciences*, Vol. 195, pp. 45-61.
60. Scheffer, T., 2004, “Email answering assistance by semi-supervised text classification,” *Intelligent Data Analysis*, Vol. 8, No. 5, pp, 481-493.
61. Schuff, D., Turetken, O. and D'Arcy, J., 2006, “A multi-attribute, multi-weight clustering approach to managing ‘e-mail overload’,” *Decision Support Systems*, Vol. 42, No. 3, pp. 1350-1365.
62. Shih, D. H., Chiang, H. S. and Yen, C. H., 2005, “Classification methods in the detection of new malicious emails,” *Information Sciences*, Vol. 172, No. 6, pp. 241-261.
63. Šmídl, V. and Quinn A., 2007, “On Bayesian principal component analysis,” *Computational Statistics & Data Analysis*, Vol. 51, No. 9, pp. 4101-4123.
64. Soucek, R. and Moser, K., 2010, “Coping with information overload in email communication: Evaluation of a training intervention,” *Computers in Human Behavior*, Vol. 26, No. 6, pp. 1458-1466.
65. Stuit, M. and Wortmann, H., 2012, “Discovery and analysis of e-mail-driven business processes,” *Information Systems*, Vol. 37, No. 2, pp. 142-168.
66. Sumecki, D., Chipulu, M. and Ojiako U., 2011, “Email overload: Exploring the moderating role of the perception of email as a ‘business critical’ tool,” *International Journal of Information Management*, Vol. 31, No. 5, pp. 407-414.
67. Sun, P. and Dong, U. A., 2010, “Automatic E-mail Classification Using Dynamic Category Hierarchy and Semantic Features,” *IETE Technical Review*, Vol. 27, No. 6, pp. 478-492.
68. Sung, S. W. and Chih, K. L., 2004, “Using text classification and multiple concepts to answer e-mails,” *Expert Systems with Applications*, Vol. 26, No. 4, pp. 529-543.
69. Szóstek, A. M., 2011, “Dealing with my emails: Latent user needs in email

management," *Computers in Human Behavior*, Vol. 27, No. 2, pp. 723-729.

70. Ug˘uz H., 2011, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowledge-Based Systems*, Vol. 24, No. 7, pp. 1024-1032.
71. Wang, J., Chen, R., Herath, T. and Rao, H. R., " Visual e-mail authentication and identification services: An investigation of the effects on e-mail use Original Research Article," *Decision Support Systems*, Vol. 48, No.1, 2009, pp. 92-102.
72. Wang, M. F., Tsai, M. F., Jheng, S. L. and Tang, C. H., 2012, "Social feature-based enterprise email classification without examining email contents," *Journal of Network and Computer Applications*, Vol. 35, No. 2, pp. 770-777.
73. Wei, C. P., Chen, H. C. and Cheng, T. H., 2008, "Effective spam filtering: A single-class learning and ensemble approach," *Decision Support Systems*, Vol. 45, No. 3, pp. 491-503.
74. White, C. B., Moyer C. A., Stern D. T. and Katz, S. J., 2004, "A content analysis of e-mail communication between patients and their providers: patients get the message," *Journal of the American Medical Informatics Association*, Vol. 11, pp. 260-267.
75. Yang, J., Liu, Y., Liu, Z., Zhu, X. and Zhang, X., 2011, "A new feature selection algorithm based on binomial hypothesis testing for spam filtering," *Knowledge-Based Systems*, Vol. 24, No. 6, pp. 904-914.
76. Yang, X., 2009, "Managing email overload with an automatic nonparametric clustering system," *Journal of Supercomputing*, Vol. 48, No. 3, pp. 227-242.
77. Ying, K. C., Lin, S. W., Lee, Z. J. and Lin, Y. T., 2010, "An ensemble approach applied to classify spam e-mails," *Expert Systems with Applications*, Vol. 37, No. 3, pp. 2197-2201.
78. Yu, B. and Xu, Z. b., 2008, "A comparative study for content-based dynamic spam classification using four machine learning algorithms," *Knowledge-Based Systems*, Vol. 21, No. 4, pp. 355-362.
79. Yu, B. and Zhu, D. H., 2009, "Combining neural networks and semantic feature space for email classification," *Knowledge-Based Systems*, Vol. 22, No. 5, pp. 376-381.
80. Zajic, D. M., Dorr, B. J. and Lin, J., 2008, "Single-document and multi-document

summarization techniques for email threads using sentence compression," *Information Processing & Management*, Vol. 44, No. 4, pp. 1600-1610.

81. Zhou, G.. W., Cheng, J. and Ping, X. J., 2007, "Intelligent Email Classification System based on PIM and Keywords," *Jisuanji Gongcheng / Computer Engineering*, Vol. 33, No. 15, pp. 199-201.
82. Zorkadis, V., Karras, D. A. and Panayotou, M., 2005, "Efficient information theoretic strategies for classifier combination, feature extraction and performance evaluation in improving false positives and false negatives for spam e-mail filtering," *Neural Networks*, Vol. 18, No. 5-6, pp. 799-807.
83. 張云濤與龔玲，2007，「資料探勘原理與技術」，台北市：五南圖書出版股份有限公司。
84. 法務部全國法規資料庫，http://law.moj.gov.tw/。

附錄、系統功能操作說明

本研究所發展之個人化郵件類別推論系統乃以本文中第 4.4.1 節所提出系統功能流程為依據，並開發系統之「郵件資料維護模組」、「郵件語意關鍵字擷取模組」、「個人化郵件類別推論模組」、「類別名稱詞彙庫維護模組」及「系統參數設定模組」等五大模組，以下分別介紹此五大模組中各功能詳細說明。

A. 郵件資料維護模組

為使權限內使用者方便維護各郵件相關資料，本研究乃開發「郵件資料維護模組」。本系統乃提供使用者於線上新增、查詢、刪除及修改各郵件資料。郵件資料維護模組包含「郵件上傳」、「郵件查詢」、「郵件修改」與「郵件刪除」等四大功能；其中，「郵件上傳」功能乃提供權限內使用者將未分類郵件匯入系統資料庫內，同時系統亦擷取郵件內文及寄件者信箱等個人化郵件類別推論所需之分析資料，藉以擷取郵件特徵詞彙並推論郵件所屬類別。「郵件資料查詢」功能乃提供權限內使用者查詢所有郵件資料之內容，以方便使用者瞭解系統內各項郵件資料之維護結果。此外，本系統亦提供「郵件資料修改」與「郵件資料刪除」功能乃提供權限內使用者進行修改與維護錯誤或缺乏時效性之郵件資料，進而保持郵件資料之正確性與即時性。

A.1 郵件上傳功能使用說明

權限內使用者可透過「郵件上傳」功能將郵件資料匯入並維護於系統資料庫中。當權限內使用者執行郵件上傳功能時，系統乃提供郵件資料上傳之系統介面予權限內使用者（如圖 A.1 所示），而權限內使用者可於該介面輸入欲上傳之郵件名稱及郵件語言等郵件資料，並上傳郵件。其中，系統時間乃以下拉式選單方式提供權限內使用者直接點選合適之時間等項目。待權限內使用者完成郵件資料後，待權限內使用者點選介面之「上傳」按鈕，系統亦擷取郵件中收件者、寄件者、郵件主旨及郵件內容等郵件類別判定之分析資料。即完成郵件資料之新增作業，系統並將郵件資料與郵件分析資料以表格訊息提供予使用者（如圖 A.2 及圖 A.3 所示）。此外，待權限內使用者所上傳郵件數量過多，可於點選系統介面之「增加檔案」按鈕，則系統亦增加郵件上傳之輸入區塊，以供權限內使用者輸入並增加上傳郵件（如圖 A.4 所示）。

舉例而言，當權限內使用者執行「郵件上傳」功能時，系統乃提供郵件資料新增介

面予使用者（如圖 A.1 所示），此時權限內使用者依序輸入此郵件名稱為「亞洲電視劇搜查線 89 期」、郵件上傳時間為「2012 年 11 月 5 日」、郵件語言為「中文」，並點擊瀏覽選擇上傳郵件之所在位置，如「C:\Users\Leo\Documents\亞洲電視劇搜查線 89 期 .eml 」；最後，使用者按下「上傳」鍵後，並擷取郵件中收件者為「ren239106@pchome.com.tw」、郵件主旨為「亞洲電視劇搜查線 89 期」及寄件者為「亞洲電視劇搜查線」等郵件資料，即完成郵件資料之上傳作業（如圖 A.2 及圖 A.3 所示），如欲上傳多封郵件則點擊「增加檔案」按鈕，並如上述步驟輸入郵件資料並選擇欲上傳郵件所在位置，即可進行多封郵件上傳之功能（如圖 A.4 所示）。

圖 A.1、郵件上傳(1)

圖 A.2、郵件上傳(2)

圖 A.3、郵件上傳(3)

圖 A.4、郵件上傳(4)

A.2 郵件查詢功能使用說明

為方便權限內使用者查詢所需之郵件資料，本研究乃開發「郵件資料查詢」功能，以提供權限內使用者查詢已上傳之郵件資料。當權限內使用者選擇郵件資料查詢功能時，於下拉式選單選取查詢類型輸入郵件搜尋與選擇上傳時間範圍搜尋，並運用邏輯運算子（AND）與（OR）組合各種可能之搜尋條件。待權限內使用者輸入郵件搜尋、選取查詢條件、選擇上傳時間範圍及點選邏輯運算子，並按下「查詢」鍵後（如圖 A.5 所示），系統將符合該查詢條件之郵件資料顯示於系統頁面下方（如圖 A.6 所示）。此外，權限內使用者亦可於此介面中點選「更多資訊」之連結功能，系統以彈跳視窗方式顯示郵件之所有詳細資料（如圖 A.7 所示）。

當權限內使用者執行「郵件資料查詢」功能時，系統乃提供條件查詢欄位供使用者

輸入（如圖 A.5 所示）。當使用者以「郵件名稱」為指定查詢方式，輸入其郵件搜尋條件為「電視」，並選擇上傳時間範圍為「2006 年 1 月 1 日」至「2012 年 12 月 31 日」執行查詢，且指定運用邏輯運算子為「AND」之搜尋條件組合後（如圖 A.5 所示），即可取得郵件名稱名為「亞洲電視劇搜查線 89 期」、「八大電視娛樂百分百」等符合查詢條件之郵件資料（如圖 A.6 所示）；此外，若權限內使用者點選「更多資訊」連結功能，如點選「亞洲電視劇搜查線 89 期」之「更多資訊」連結，系統以彈跳視窗方式顯示郵件之名稱為「亞洲電視劇搜查線 89 期」、郵件主旨為「亞洲電視劇搜查線 89 期」等各項詳細資料（如圖 A.7 所示），完成查詢郵件資料之步驟。

圖 A.5、郵件查詢(1)

圖 A.6、郵件查詢(2)

圖 A.7、郵件查詢(3)

A.3 郵件修改功能使用說明

郵件資料修改功能乃提供權限內使用者修改缺乏時效性或錯誤之郵件資料，進而保持郵件資料之正確性與即時性。當權限內使用者選擇郵件資料修改功能時，於下拉式選單選取查詢條件輸入查詢字串與選擇上傳時間範圍搜尋，待權限內使用者完成輸入查詢資料後（如圖 A.8 所示），並按下「查詢」鍵後，系統將符合該查詢條件之郵件資料顯示於系統頁面下方（如圖 A.9 所示）。此外，權限內使用者亦可於此介面中點選「資料修改」之連結功能，系統以彈跳方式顯示修改郵件基本資料視窗（如圖 A.10 所示），使用者於郵件修改視窗更改缺乏時效性或錯誤資料，並於詳細資料視窗下方按下「確定修改」鍵後，系統自動將該封郵件於系統資料庫中修改，同時執行訊息「郵件資料修改成功」顯示於系統介面上（如圖 A.11 所示），即完成修改郵件資料之步驟。

舉例而言，當權限內使用者執行「郵件資料修改」功能時，系統乃提供條件查詢欄位供使用者輸入。如圖 A.8 所示，當使用者以「主旨」為指定查詢方式，輸入其查詢字串為「電視劇」，選擇上傳時間範圍為「2006 年 1 月 1 日」至「2012 年 12 月 31 日」執行查詢，即可取得郵件主旨為「亞洲電視劇搜查線 89 期」、「電影電視劇最新娛樂報」等郵件資料（如圖 A.9 所示）；之後，權限內使用者針對郵件「電影電視劇最新娛樂報」按下「資料修改」之連結功能，系統以彈跳視窗方式顯示修改郵件基本資料介面，如圖 A.10 所示，使用者將郵件名稱「電影電視劇最新娛樂報」修改為「亞洲華人最新娛樂報」並按下「確定修改」鍵後，系統自動將該封郵件於系統資料庫中修改，同時執行訊息「郵件資料修改成功」顯示於系統介面上，如圖 A.11 所示，郵件名稱即修改為「亞洲華人最新娛樂報」，接著按下「關閉視窗」鍵即完成郵件資料修改。

圖 A.8、郵件修改(1)

圖 A.9、郵件修改(2)

圖 A.10、郵件修改(3)

圖 A.11、郵件修改(4)

A.4 郵件刪除功能使用說明

郵件資料刪除功能乃提供權限內之使用者刪除缺乏時效性或錯誤之郵件資料，進而保持郵件資料之正確性與即時性。當權限內使用者選擇郵件資料刪除功能時，於下拉式選單選取查詢條件輸入查詢字串與選擇上傳時間範圍搜尋並選擇運算式。待權限內使用者輸入查詢字串、選取查詢條件、選擇上傳時間範圍及點選邏輯運算子（如圖 A.12 所示），並按下「查詢」鍵後，系統將符合該查詢條件之郵件資料顯示於系統頁面下方（如圖 A.13 所示）。接著，權限內使用者可於此介面中核選方塊勾選欲刪除郵件詳細資料（如圖 A.13 所示），且於詳細資料視窗下方按下「確定刪除」鍵後，系統自動將被勾選之郵件於系統資料庫中刪除，同時顯示「資料已成功刪除」訊息與此次勾選刪除郵件之資訊於系統介面上（如圖 A.14 所示），即完成刪除郵件資料之步驟。

當權限內使用者執行「郵件資料刪除」功能時，系統乃提供條件查詢欄位供使用者輸入。如圖 A.12 所示，當使用者以「郵件名稱」為指定查詢方式，輸入其查詢字串為「電視」，選擇上傳時間範圍為「2006 年 1 月 1 日」至「2012 年 10 月 31 日」執行查詢，並指定運用邏輯運算子為「AND」之搜尋條件組合後，即可取得郵件聯絡人為「娛樂週報」、「美劇小週報」等郵件資料（如圖 A.13 所示）；之後，權限內使用者勾選「是否刪除」之核選方塊，如圖 A.13 所示，使用者勾選郵件主旨為「電影電視劇最新娛樂報」、「Jolin 最新專輯 MV 大首播」等郵件，並於視窗下方按下「確定」鍵，系統自動將權限內使用者所勾選郵件於系統資料庫中刪除，同時執行「資料已成功刪除」與勾選刪除郵件之資訊於系統介面上（如圖 A.14 所示）。

圖 A.12、郵件刪除(1)

圖 A.13、郵件刪除(2)

圖 A.14、郵件刪除(3)

B. 郵件語意關鍵字擷取模組

本系統開發「郵件語意關鍵字擷取模組」乃以使用者所上傳之私人郵件擷取關鍵字，並進行關鍵字語意主題分析以確實擷取具有語言含意性之詞彙。「郵件主題詞彙集合建立」功能乃考量郵件內容敘述主題多樣性，且各主題皆有相關領域關鍵字與不同語言含意，因此，透過郵件中所包含各式談論主題中分析特定詞彙，並以詞彙於郵件中出現機率分辨詞彙語意性程度。此功能主要以郵件中詞彙出現機率與詞彙主題關聯進行擷取與分析。詞彙出現機率即代表該詞彙於此封郵件敘述內容之重要程度，為考量各詞彙於各主題中代表語言含意程度皆不同，故透過該筆詞彙出現機率與該主題亦同時於郵件中發生機率計算詞彙主題關聯，再透過詞彙主題關聯判別該詞彙於主題中具有語言含意程度，並作為該封郵件類別推論所需詞彙。「詞彙後驗機率近似值計算」功能乃分析郵件中內容主題發生機率，並透過此主題比例強化詞彙主題關聯性。首先，系統進行郵件主題抽樣模擬，當中系統管理者可透過「模擬抽樣參數設定」功能設定模擬抽樣次數，並於模擬抽樣中取得郵件中內容主題發生機率，並將各組模擬抽樣後數據取平均值後與「郵件主題詞彙集合建立」功能所得詞彙主題關聯性進行計算，即可強化詞彙主題關聯性，並根據強化後關聯係數篩選詞彙。本系統乃提供權限內使用者進行郵件主題詞彙集合建立、詞彙後驗機率近似值計算等功能。

B.1 郵件主題詞彙集合建立功能使用說明

「郵件主題詞彙集合建立功能」之主要功能乃擷取郵件中語意詞彙。首先使用者以搜尋字串尋得系統中欲分析之已上傳郵件，並取得郵件內容之「郵件主旨」、「郵件內文」及「附加檔名」等文字資料，並將此些資料進行二至六字詞斷詞取得該郵件詞彙，接著計算該詞彙之詞彙頻率與詞彙出現機率，取得詞彙之詞彙出現機率後再與郵件中各主題之主題發生機率計算兩者同時發生機率，即可取得郵件中各筆詞彙及各詞彙與主題關聯係數，透過詞彙與主題關聯係數即可篩選出該郵件中所含語意詞彙。

舉例而言，當使用者進入「郵件主題詞彙集合建立功能」畫面後，使用者先行輸入「電視劇」作為搜尋字串，並選擇以「主旨」作為查詢條件進行郵件資料搜索（如圖 B.1 所示），再以郵件主旨「亞洲電視劇搜查線 89 期」之電子郵件進行解析（如圖 B.2 所示），系統即可取得該郵件之郵件主旨、附件檔名、郵件內容等資料，如圖 B.3 所示之郵件內容為「不讓貝克漢專美於前，韓國人氣足球明星安貞桓也到日本拍廣告，...」，接著系統於背後則進行非關鍵字去除並進行二至六字詞斷詞，待郵件內容經系統完成斷

詞後，系統將所得詞彙即分析詞彙頻率並換算為詞彙於郵件中出現機率，並與該筆郵件所含主題計算兩者同時發生機率作為詞彙與主題關聯係數，待系統計算完成後系統並將計算結果以主題進行區分彙整所得詞彙，並於系統畫面中呈現郵件所含主題如「運動」、「演員」等（如圖 B.3 所示），使用者將滑鼠指標指向主題即可展開與該主題較具語意關聯之詞彙，如圖 B.4 所示，當使用者指向「演員」主題時，則系統展開「足球」、「明星」等詞彙，但單筆主題所關聯詞彙眾多，因此系統以分頁方式顯示該筆主題關聯之詞彙，當使用者點擊畫面中左右向箭頭時，系統根據點擊箭頭方向切換前後 8 筆詞彙（如圖 B.5 所示），並將滑鼠指標指向詞彙「足球明星」時系統畫面則呈現詞彙「足球明星」與各主題之關聯係數（如圖 B.6 所示），如詞彙「足球明星」與「演員」主題之關聯係數為「0.076」，並相較於其他主題如「足球」之關聯係數還高，則代表詞彙「足球明星」於該封郵件中所代表語言含意與「演員」主題較具關聯。

圖 B.1、郵件主題詞彙集合建立功能(1)

圖 B.2、郵件主題詞彙集合建立功能(2)

圖 B.3、郵件主題詞彙集合建立功能(3)

圖 B.4、郵件主題詞彙集合建立功能(4)

圖 B.5、郵件主題詞彙集合建立功能(5)

圖 B.6、郵件主題詞彙集合建立功能(6)

B.2 詞彙後驗機率近似值計算功能使用說明

「詞彙後驗機率近似值計算」功能之主要功能乃強化「郵件主題詞彙集合建立功能」所得詞彙之詞彙主題關聯係數，以篩選該郵件中最具代表性詞彙。權限內使用者選擇欲分析郵件後（如圖 B.7、圖 B.8 所示），本系統乃根據該郵件之敘述主題進行模擬抽樣；由於詞彙主題關聯係數主要以兩者於郵件中同時發生機率與主題後驗機率計算而得，但該郵件內容敘述主題之後驗機率皆為未知，因此本系統透過模擬抽樣方式計算主題後驗機率近似值，以強化詞彙與主題間關聯差異；首先以隨機方式給定各主題一筆整數且小於該郵件主題數量，以作為第一次抽樣之主題發生次數，並以該次數計算發生機率，再依此發生機率再次進行抽樣，並依系統參數設定中所設抽樣次數反覆進行數次抽樣，系統於完成若干次抽樣後，根據系統參數設定中所抽樣擷取參數，擷取若干組抽樣，由於第一組抽樣初始值為自行設定因此不進行後驗機率計算，待抽樣擷取完成後進行各主題發生機率之平均計算，即可取得主題後驗機率近似值作為該主題最終發生機率（如圖 B.9、圖 B.10 所示），並與各詞彙重新計算詞彙主題關聯係數，以獲得強化後詞彙主題關聯係數（如圖 B.11 所示）。

舉例而言，權限內使用者進入「詞彙後驗機率近似值計算」功能後以「電視劇」為查詢關鍵字並以「主旨」查詢類型進行查詢，權限內使用者勾選「亞洲電視劇搜查線 89 期」之郵件並點擊解析按鈕（如圖 B.7、圖 B.8 所示）取得分析郵件後，本功能乃以蒐集該郵件之主題進行模擬抽樣，於模擬抽樣前需假設該郵件中各主題分別被敘述次數，以分析該主題於郵件中發生機率，故本系統先以假設方式給定各主題一筆隨機整數做為

第一次模擬抽樣，並根據假設所給定之整數計算該主題於此次抽樣中所發生機率，即完成一次模擬抽樣，並反覆進行多次抽樣取得多組不同機率組合，當中，此些組合則別代表各次抽樣時主題發生機率，本功能將根據「模擬抽樣參數設定」功能所設定預設值「15」擷取適用模擬抽樣組合進行平均值計算，即可取得後驗機率近似值，如「演員」主題之後驗機率近似值為「0.125」（如圖 B.9 所示）則代表該郵件所敘述內容涉及「演員」主題之機率約為「0.125」，此外，權限內使用者亦可透過系統介面中針對此次分析修改「模擬抽樣參數」中所設定抽樣組數進行重新分析，如圖 B.10 所示，權限內使用者將「模擬抽樣參數」修改為「18」後點擊重新分析按鈕後，由於修改後抽樣組數增加亦影響各主題之機率，如「演員」主題經重新計算後後驗機率近似值提升至「0.384」，接著點擊詞彙主題關聯修正系統即以此後驗機率近似值重新計算詞彙與主題間關聯（如圖 B.11 所示），權限內使用者亦可透過點擊「原始係數」連結，本功能即以新視窗方式呈現該筆詞彙重新計算前之「詞彙主題間關聯係數」供使用者參考，如圖 B.11、圖 B.12 所示，由於詞彙主題關聯係數透過後驗機率近似值修正計算後提升該詞彙與各主題間關聯之差異，如詞彙「足球明星」與「廣告」主題之關聯由「0.317」提升至「0.634」，並與其他主題相較下關聯性較高，則代表該詞彙於該郵件中之語言含意與「廣告」主題相關。

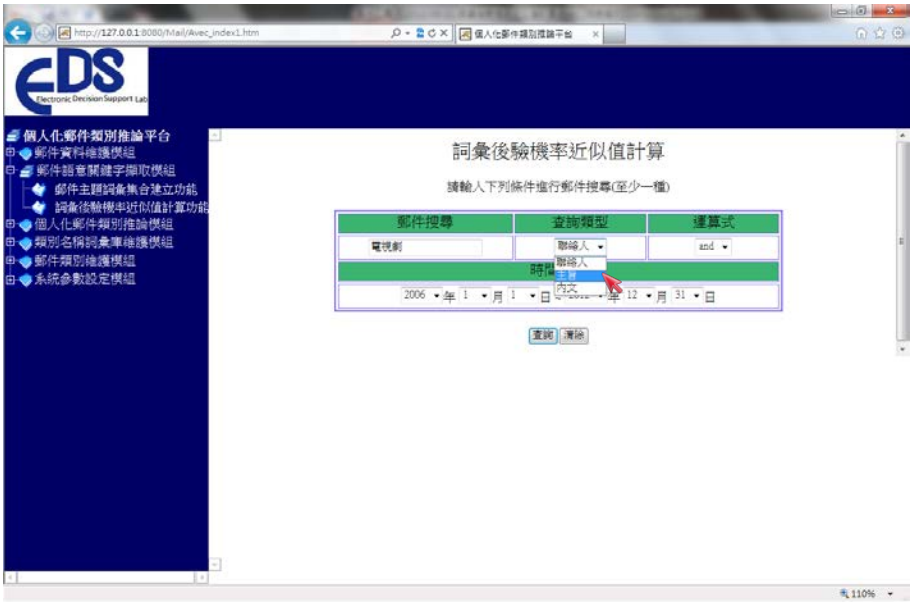

圖 B.7、詞彙後驗機率近似值計算(1)

圖 B.8、詞彙後驗機率近似值計算(2)

圖 B.9、詞彙後驗機率近似值計算(3)

圖 B.10、詞彙後驗機率近似值計算(4)

圖 B.11、詞彙後驗機率近似值計算(5)

圖 B.12、詞彙後驗機率近似值計算(6)

C. 個人化郵件類別推論模組

本研究所開發「個人化郵件類別推論模組」乃根據「郵件語意關鍵字擷取模組」之語意詞彙分析郵件代表性特徵，並透過分析郵件特徵方式將郵件以階層化方式分群並推論各群集之名稱，以作為使用者個人化郵件類別。

「郵件主成份計算」功能乃分析「郵件語意關鍵字擷取模組」所得之語意詞彙，以取得該郵件之郵件特徵。其中，郵件之郵件特徵乃以主成份分析方式為基礎進行分析，當中，本功能以詞彙與主題關聯進行計算詞彙共變異矩陣，再透過共變異矩陣計算詞彙與該郵件關聯程度並從中篩選，以取得最具代表性語意詞彙作為郵件特徵。詞彙共變異矩陣即將各主題所含詞彙之詞彙主題關聯計算共變異數，並彙整為共變異數矩陣而得，以分析出該主題中主要影響郵件之詞彙並計算詞彙與郵件關聯，並根據詞彙與郵件關聯篩選出該郵件最具代表性詞彙作為郵件分群所需之郵件特徵。

「郵件關聯程度與類別推論」功能透過郵件特徵詞彙計算郵件相似度，並根據相似度進行郵件階層式分群，再將郵件群集進行分析，並透過類別名稱詞彙庫取得符合群集內容之名稱作為類別名稱。於郵件階層式分群中，本系統乃根據各封郵件中郵件特徵詞彙計算相似度，再根據計算所得結果合併相似度較高郵件為同一群集，並反覆進行使所有郵件形成階層式樹狀群集，即可取得郵件群集。但取得郵件群集後，群集並非含有代表性名稱幫助使用者解讀群集所包含郵件，因此，本系統以群集內郵件詞彙與「類別名稱詞彙庫維護模組」中之類別名稱詞彙庫分析語意關聯，並根據語意關聯尋得所有詞彙中最具代表性與解釋性詞彙作為群集名稱，即完成個人化郵件類別推論。

本系統乃提供權限內使用者進行「郵件主成份計算」功能、「郵件關聯程度與類別推論」功能等功能，且為能供使用者查詢系統所推論名稱，本模組乃建立「郵件類別查詢」功能供使用者查詢系統推論之個人化郵件類別。

C.1 郵件主成份計算功能使用說明

本研究所開發之「郵件主成份計算」功能乃取得郵件中代表性郵件特徵詞彙，以作為郵件分群之分群依據；而郵件特徵詞彙乃根據郵件中所有詞彙主題關聯性差異進行彙整並篩選而得，因此，本功能乃以詞彙主題關聯係數計算詞彙共變異矩陣，並比較各詞彙之詞彙主題關聯係數差異，再根據詞彙共變異矩陣中各詞彙之差異大小將詞彙彙整為特徵向量，而特徵向量則將各筆關聯差異較小之詞彙與詞彙主題關聯繫進行線性組合而得，故特徵向量即代表一組詞彙含意相近之特徵詞彙組合，接著再透過特徵解釋比例門檻值篩選特徵向量，即可獲得該郵件之郵件特徵詞彙。故權限內使用者尋得欲分析之郵件後（如圖 C.1、圖 C.2 所示），系統則根據該郵件之詞彙主題關聯係數計算詞彙共變異數矩陣，接著詞彙再根據詞彙共變異數矩陣中詞彙間關聯差異大小組合成特徵向量，於取得郵件中各組特徵向量後，根據系統所制定之特徵解釋比例門檻值篩選出郵件中最具代表性特徵向量組合並彙整，進而取得郵件特徵詞彙（如圖 C.3 所示）。

舉例而言，權限內使用者進入「郵件主成份計算」功能後以「電視劇」為查詢關鍵字並以「主旨」查詢類型進行查詢，權限內使用者勾選「亞洲電視劇搜查線 89 期」之郵件並點擊解析按鈕（如圖 C.1、圖 C.2 所示）取得分析郵件後，本功能乃蒐集該郵件之語意詞彙與詞彙主題關聯係數進行分析，首先本功能於系統背後先行將詞彙主題關聯係數計算詞彙共變異數矩陣，並根據詞彙共變異數矩陣中詞彙關聯差異將詞彙組合成各組特徵向量組合，即可獲得郵件特徵詞彙，如圖 C.3 所示，目標郵件「亞洲電視劇搜查線 89 期」之郵件特徵詞彙包含「日本」、「足球」、「貝克漢」...等詞彙，由於郵件特徵詞彙需透過使用者設定特徵解釋比例門檻值進行篩選，因此，權限內使用者亦可點擊特徵解釋比例門檻值之選項，如「平均值」、「中位數」與「四分位數」等門檻值，並點擊「重新計算」即可重新篩選特徵詞彙（如圖 C.4、圖 C.5 所示），以取得更精確之郵件特徵詞彙。

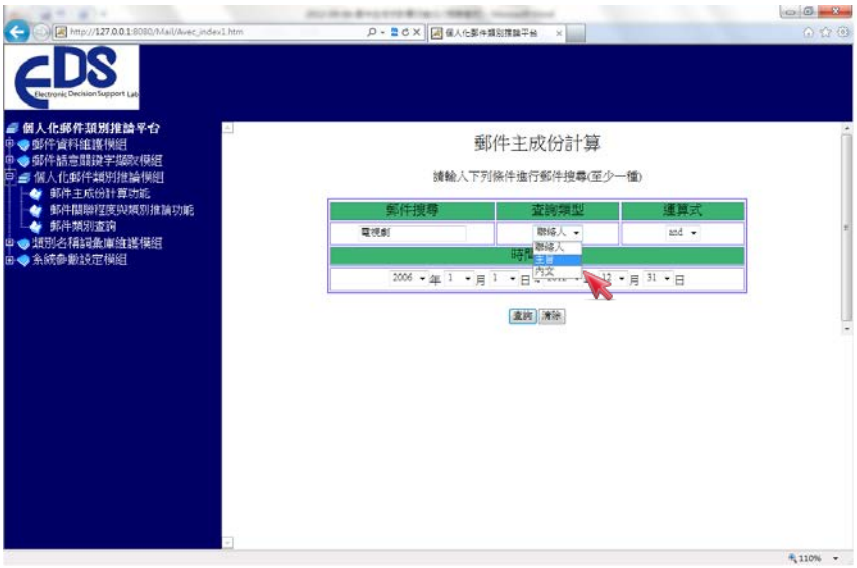

圖 C.1、郵件主成份計算(1)

圖 C.2、郵件主成份計算(2)

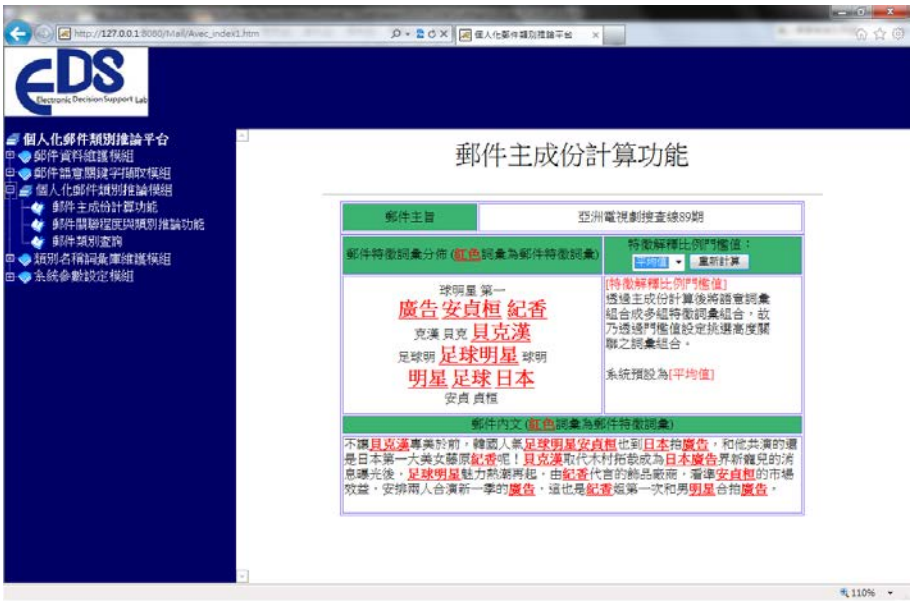

圖 C.3、郵件主成份計算(3)

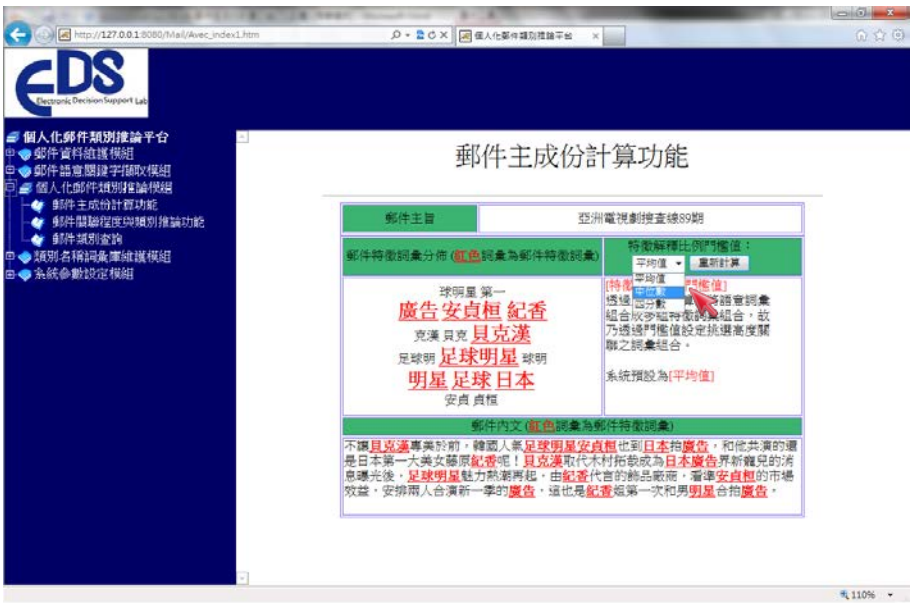

圖 C.4、郵件主成份計算(4)

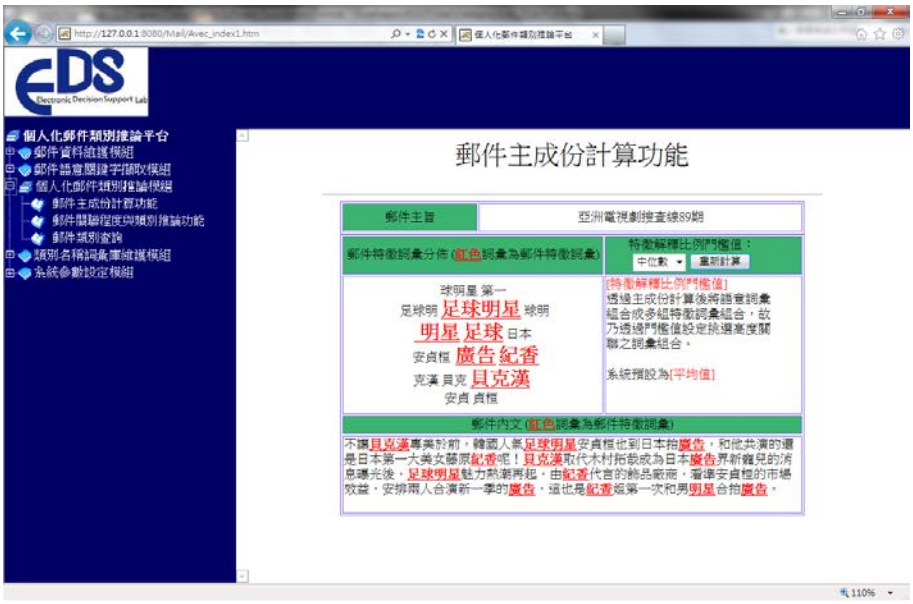

圖 C.5、郵件主成份計算(5)

C.2 郵件關聯程度與類別推論功能使用說明

「郵件關聯程度與類別推論」功能主要目的乃將權限內使用者所有上傳之郵件分析郵件群集與權限內使用者適用之個人化郵件類別。本系統以階層式分群方式並根據郵件間相似距離進行群集劃分，而相似距離乃透過郵件互相比對郵件特徵詞彙相似性進行計算，當郵件特徵詞彙相似性越小代表郵件相似距離越短，並將相似距離較短郵件區分為相同群集，並反覆進行至所有郵件具歸屬群集。由於所得郵件群集無實質名稱，導致使用者無法直覺化了解群集內郵件內容，亦無法成為郵件類別，因此，本系統於取得郵件群集後乃彙整郵件群集內所有郵件特徵詞彙，並與系統內類別名稱詞彙庫所含詞彙分析語意關聯與語意網路後，根據語意網路中尋找詞彙語意關聯之起點詞彙，由於語意網路中起點詞彙代表於單一領域所有詞彙中最具解釋性含意，因此，以該筆起點詞彙作為郵件群集之名稱，並以此名稱作為個人化郵件類別。以下乃分別說明本功能中個分析與使用方式。

Step(1) 彙整郵件特徵詞彙

於系統進行前，權限內使用者需先行完成「郵件語意關鍵字擷取模組」與「郵件主成份計算」功能後進入本功能，首先彙整使用者上傳之所有郵件與郵件中所含特徵詞彙，以做為後續分群進行依據。如圖 C.6 所示，系統先行彙整並分析系統內所有郵件之郵件特徵詞彙，當權限內使用者點擊郵件「亞洲電視劇搜查線 89 期」時，系統則顯示郵件特徵詞彙（如圖 C.7 所示），以供權限內使用者檢閱。

圖 C.6、郵件關聯程度與類別推論(1)

圖 C.7、郵件關聯程度與類別推論(2)

Step(2) 區分相似郵件

待系統完成「彙整郵件特徵詞彙」後，系統根據郵件中所含所有郵件特徵詞彙計算郵件相似性，並依據相似程度歸類郵件群集，如圖 C.8 所示，系統畫面中「相似郵件群A」所含相似郵件如「亞洲電視劇搜查線 89 期」、「華人娛樂新聞網」等郵件。完成相似郵件區分後，權限內使用者點擊「建立郵件群集樹」按鈕進入下一步驟。

圖 C.8、郵件關聯程度與類別推論(3)

Step(3) 建立郵件群集樹

進入此步驟後系統即以郵件特徵詞彙反覆計算郵件相似距離並分群，待系統完成郵件分群後，如圖 C.9 所示，系統畫面中「層級 2」之群集「相似郵件群 A」所含相似郵件如「亞洲電視劇搜查線 89 期」、「華人娛樂新聞網」等郵件。系統根據權限內使用者所設定之層級擷取參數擷取權限內使用者所需群集。

圖 C.9、郵件關聯程度與類別推論(4)

Step(4) 分析群集代表詞彙

完成郵件群集樹建立後，系統於此步驟乃分析群集內中各郵件之郵件特徵詞彙，並根據特徵詞彙之詞彙關聯取得詞彙等級較高詞彙，以作為群集代表詞彙。如圖 C.10 所示，「相似郵件群 A」中根據「亞洲電視劇搜查線 89 期」、「華人娛樂新聞網」等郵件

分析後，其郵件特徵詞彙關聯中「等級一」詞彙為「娛樂新聞」，使用者可點擊其他群集名稱，系統則展開該群集之郵件特徵詞彙關聯，如圖 C.11 所示。

圖 C.10、郵件關聯程度與類別推論(5)

圖 C.11、郵件關聯程度與類別推論(6)

Step(5) 推論郵件類別名稱

於最後，系統根據郵件群集中特徵詞彙關聯與郵件名稱詞庫進行關聯比對，以尋找郵件群集中代表性名稱，如圖 C.12 所示，「相似郵件群 A」之群集名稱經系統與名稱詞彙庫分析詞彙關聯後，該群集之名稱命名為「影視娛樂」，此外，使用者點擊其他群集系統則展現該群集之名稱（如圖 C.13 所示），於推論結束後，本系統供權限內使用者透過「郵件類別查詢」功能查詢系統推論結果。

圖 C.12、郵件關聯程度與類別推論(7)

圖 C.13、郵件關聯程度與類別推論(8)

C.3 郵件類別查詢功能使用說明

本研究之個人化郵件類別名稱乃經系統自動推論而得，因此為方便權限內使用者查詢系統內個人化郵件類別名稱，本研究乃開發「郵件類別查詢」功能，以提供權限內使用者查詢系統已推論之個人化郵件類別與類別所含郵件。當權限內使用者選擇「個人化郵件類別推論模組」並進入「郵件類別查詢」功能後，系統則展示經由本系統所推論之郵件類別層級（如圖 C.14 所示），由於系統所推論之個人化郵件類別乃以階層式樹狀展示，故使用者可點擊類別名稱展開次一層級之郵件類別（如圖 C.15 所示）。此外，當權限內使用者以滑鼠指標指向類別名稱時，系統則於系統畫面中顯示該類別所歸類之郵件，供使用者查看該類別所歸類郵件類型（如圖 C.15 所示）。

舉例說明，當權限內使用者執行「郵件類別查詢」功能時，本功能乃顯示系統所推

論之郵件類別層級，如圖 C.14 所示，系統則顯示「影視娛樂」、「旅遊」、「社會新聞」、「藝術文化」等郵件類別並以層級樹方式展示，如使用者點擊「影視娛樂」層級時則系統乃顯示「影視娛樂」層級中所包含次等層級「亞洲娛樂」、「電視劇」等次等層級（如圖 C.15 所示），且為方便使用者查看該類別所歸類郵件，因此，當使用者以滑鼠指標指向類別時顯示該類別所分類郵件，如圖 C.15 所示，當權限內使用者滑鼠指標指向「亞洲娛樂」類別時，系統則顯示「亞洲娛樂」類別所包含郵件為「電視劇搜查線 89 期」、「亞洲華人最新娛樂報」等郵件資訊。

圖 C.14、郵件類別查詢(1)

圖 C.15、郵件類別查詢(2)

D. 類別名稱詞彙庫維護模組

為能自動化進行郵件類別名稱推論，且推論名稱能具代表性含意與解釋性，故本研

究乃開發「類別名稱詞彙庫維護模組」，並提供權限內使用者進行領域文件上傳、名稱詞彙庫建立、名稱詞彙查詢與名稱詞彙刪除等功能。當中，「領域文件上傳」功能乃提供權限內使用者上傳領域文件，接著，權限內使用者再透過「名稱詞彙庫建立」功能解析領域文件中詞彙並進行潛在語意分析，以標記屬性與重要性彙整其詞彙關聯組合，此詞彙關聯組合乃表達各筆詞彙之上下層級關聯。於取得詞彙關聯組合後則整理成類別名稱詞彙庫，作為擷取類別名稱所需詞彙集。此外權限內使用者可透過「名稱詞彙查詢」、「名稱詞彙刪除」功能查詢或刪除系統內已彙整之名稱詞彙。

D.1 領域文件上傳功能使用說明

由於本研究所制定之名稱詞彙庫乃透過解析權限內使用者所上傳領域文件之文件詞彙，並彙整其詞彙關聯而得，因此本系統乃建立「領域文件上傳」功能作為權限內使用者上傳領域文件之管道。首先權限內使用者進入「領域文件上傳」功能畫面後（如圖 D.1 所示），首先權限內使用者輸入領域文件名稱、領域文件上傳日期、領域文件類別與領域文件內容等資料，並點擊瀏覽選擇欲上傳之領域文件；當中，上傳日期乃以下拉式選單供權限內使用選擇上傳時間；待權限內使用者完成輸入領域文件資料後，權限內使用者點擊上傳後系統則呈現領域文件之資料（如圖 D.2 所示），並完成領域文件上傳動作。

舉例說明，當權內使用者執行「領域文件上傳」功能時系統則顯示畫面，權內使用者欲上傳一份名稱為「尬出文化新生命──馮凱與《陣頭》」領域文件時，如圖 D.1 所示，於領域文件名稱輸入「尬出文化新生命──馮凱與《陣頭》」，且於領域文件類別輸入「大眾傳播」等相關資訊後，並點擊瀏覽選擇欲上傳檔案路徑「E:\Users\Leo\Downloads…」後，點擊「上傳」按鈕系統則顯示領域文件新增成功之系統畫面，如圖 D.2 所示，系統完成資料擷取並回饋系統後乃顯示領域文件之資料，如權限內使用者上傳名稱為「尬出文化新生命──馮凱與《陣頭》」領域文件後，系統則顯示領域文件名稱「尬出文化新生命──馮凱與《陣頭》」，並顯示領域文件之內容，如「今年春節期間，號稱「天龍八部」的 8 部國片…」等內容供權限內使用者確認。

圖 D.1、領域文件上傳(1)

圖 D.2、領域文件上傳(2)

D.2 名稱詞彙庫建立功能使用說明

「名稱詞彙庫建立」功能主要乃透過解析領域文件中詞彙並進一步彙整為名稱詞彙庫，以作為郵件類別名稱挑選依據。權限內使用者進入系統後，先行查詢欲解析之領域文件（如圖 D.3、圖 D.4 所示），本系統乃以二字六字詞方式拆解領域文件之文章內容，待系統完成詞彙擷取後，系統則以潛在語意分析方式分析詞彙機率，以分析出具有解釋含意之詞彙（如圖 D.5 所示），接著，系統為能判斷詞彙於該領域文件中代表性，故本研究乃以詞彙上下關聯進行分析該詞彙於此領域文件代表性，首先根據詞彙於文章中位置並與前一比詞彙建立串聯關係，接著，系統乃根據串聯後詞彙與詞彙機率計算此詞彙串聯於領域文件中同時出現機率，再根據詞彙串聯之機率分析各筆詞彙與不同詞彙串聯時之語意層級，並根據詞彙語意層級將重複詞彙進行整合型成詞彙層級樹，以作為兩者

之詞彙關聯建立名稱詞彙層級關聯（如圖 D.6 所示），取得領域文件之名稱詞彙層級關聯後，系統即與系統內既有領域文件之名稱詞彙層級關聯彙整為名稱詞彙庫，並回饋制系統中（如圖 D.7、圖 D.8 所示）。

舉例而言，權限內使用者進入「名稱詞彙庫建立」功能後，首先乃輸入領域文件搜尋關鍵字，如圖 D.3 所示，權限內使用者「陣頭」進行搜尋，則系統篩選出領域文件名稱為「尬出文化新生命──馮凱與《陣頭》」之領域文件（如圖 D.4 所示）。接著於系統背後乃根據該領域文件之文章進行斷詞並將斷詞後結果顯示於系統介面中，如圖 D.5 所示，當使用者點擊個段落時系統乃顯示個段落所擷取詞彙，如權限內使用者點擊「有人說，電影的情節張力與高票房...」段落，則系統於右側顯示「電影」、「導演」、「馮凱」、「電視劇」等該段落所擷取詞彙，接著，計算詞彙機率與詞彙關聯，再根據詞彙機率與詞彙關聯彙整為名稱詞彙集，如圖 D.6 所示，系統根據領域文件之「有人說，電影的情節張力與高票房...」等段落擷取「電影」、「導演」、「馮凱」等詞彙，且系統於畫面中乃根據詞彙之關聯等級進行樹狀排列，如「電影」詞彙標記為「LV1」則代表該詞彙為最高等級詞彙，亦為該領域文件中最具代表性與解釋性詞彙。最後於系統解析完成時，如圖 D.7 所示，本系統乃將「電影」、「導演」、「馮凱」等詞彙與系統內既有名稱詞彙庫分析詞彙與詞彙層級並整併相同詞彙，如圖 D.8 所示，使用者點擊名稱詞彙「影視娛樂」系統則顯示該筆詞彙來源之領域文件明「影像魔法師 ──王小棣」，即完成名稱詞彙庫建立功能。

圖 D.3、名稱詞彙庫建立功能(1)

圖 D.4、名稱詞彙庫建立功能(2)

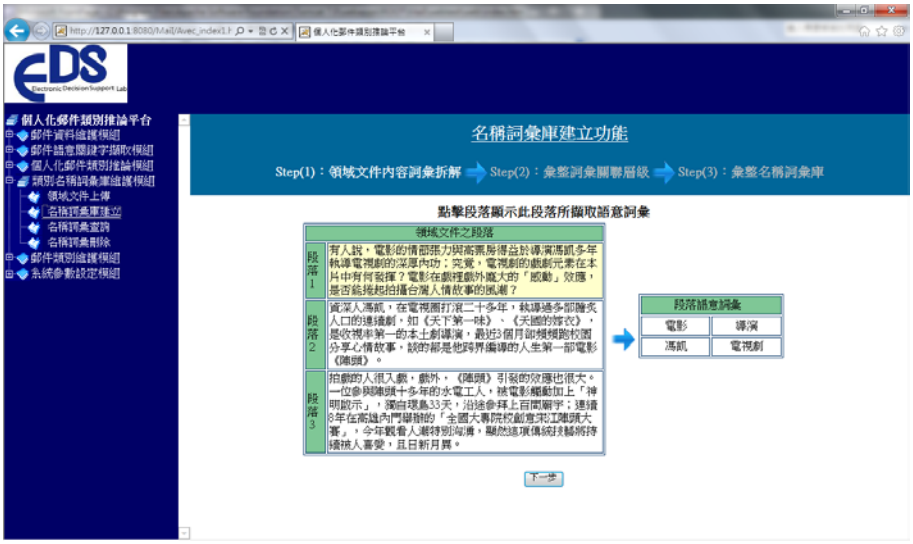

圖 D.5、名稱詞彙庫建立功能(3)

圖 D.6、名稱詞彙庫建立功能(4)

圖 D.7、名稱詞彙庫建立功能(5)

圖 D.8、名稱詞彙庫建立功能(6)

D.3 名稱詞彙查詢功能使用說明

為方便權限內使用者查詢系統之名稱詞彙與領域文件，本研究乃開發「名稱詞彙查詢」功能，以提供權限內使用者查詢已新增之名稱詞彙與領域文件。當權限內使用者選擇名稱詞彙查詢功能時，輸入查詢字串進行領域文件查詢（如圖 D.9 所示）。待權限內使用者輸入查詢字串、選取運算式與時間範圍等條件，並按下「查詢」鍵後，系統將符合該查詢條件之名稱詞彙及領域文件等資料顯示於系統頁面下方（如圖 D.10 所示）。

當權限內使用者執行「名稱詞彙查詢」功能時，系統乃提供查詢欄位於使用者輸入「陣頭」為查詢字串，同時限制文件時間範圍於「2006 年 1 月 1 日」至「2012 年 12 月 31 日」之間進行篩選（如圖 D.9 所示），即可取得名稱詞彙「電影」、「導演」、「馮凱」...等名稱詞彙與各名稱詞彙等級「LV1」、「LV2」、「LV3」，以及名稱詞彙所屬領域文件「尬

出文化新生命──馮凱與《陣頭》」與領域文件類別「大眾傳播」等資訊（如圖 D.10 所示）。

圖 D.9、名稱詞彙查詢功能(1)

圖 D.10、名稱詞彙查詢功能(2)

D.4 名稱詞彙刪除功能使用說明

名稱詞彙刪除功能乃提供權限內之使用者刪除不具代表性之名稱詞彙，進而保持資料之正確性與即時性。當權限內使用者選擇名稱詞彙刪除功能時，當權限內使用者選擇名稱詞彙查詢功能時，輸入查詢字串進行領域文件查詢。待權限內使用者輸入查詢字串、選取運算式與時間範圍等條件（如圖 D.11 所示），並按下「查詢」鍵後，系統將符合該查詢條件之名稱詞彙及領域文件等資料顯示於系統頁面下方（如圖 D.12 所示）。權限內使用者亦可於此介面中勾選欲刪除之名稱詞彙並點擊刪除（如圖 D.13 所示），系統

自動將該筆名稱詞彙於系統資料庫中刪除，同時執行訊息「詞彙刪除成功」與已刪除詞彙之資訊並顯示於系統介面上（如圖 D.14 所示），即完成名稱詞彙刪除之步驟。

當權限內使用者執行「名稱詞彙刪除」功能時，系統乃提供查詢欄位於使用者輸入「陣頭」為查詢字串進行篩選（如圖 D.11 所示），以取得名稱詞彙「電影」、「導演」、「馮凱」...等名稱詞彙（如圖 D.12 所示），如權限內使用者勾選名稱詞彙「馮凱」之「刪除詞彙」核選方塊（如圖 D.13 所示），並按下「刪除」按鈕，系統顯示「詞彙刪除成功」與名稱詞彙「馮凱」之資訊（如圖 D.14 所示），並完成名稱詞彙刪除。

圖 D.11、名稱詞彙刪除功能(1)

圖 D.12、名稱詞彙刪除功能(2)

圖 D.13、名稱詞彙刪除功能(3)

圖 D.14、名稱詞彙刪除功能(4)

E. 系統參數設定模組

為使權限內使用者方便維護各系統相關資料，本系統乃開發「系統參數設定模組」。本程式功能乃提供權限內使用者於線上修改系統中個模組參數資料。「系統參數設定模組」包含「郵件語意關鍵字擷取參數設定」與「個人化郵件類別推論參數設定」等二大功能；其中，「郵件語意關鍵字擷取參數設定」功能乃提供權限內使用者針對語意關鍵字之門檻值進行修改與維護，此外，權限內使用者亦可透過此功能設定「詞彙後驗機率近似值」之門檻值，進而保持門檻值之正確性。「個人化郵件類別推論參數設定」功能乃提供權限內使用者設定「郵件主成份計算」功能中郵件特徵詞彙篩選所需門檻值，權限內使用者並透過此功能重新設定「郵件關聯程度與類別推論」功能中郵件類別擷取參數，進而保持系統參數值與個人化郵件類別擷取之正確性。

E.1 郵件語意關鍵字擷取參數設定使用說明

「郵件語意關鍵字擷取參數設定」功能乃提供權限內使用者進行修改系統分析參數，以進而提升「郵件主題詞彙集合建立」與「詞彙後驗機率近似值計算」兩功能郵件語意詞彙分析準確性。當權限內使用者選擇「郵件語意關鍵字擷取參數設定」功能時，系統即呈現郵件主題詞彙集合建立功能之主題詞彙門檻值、詞彙後驗機率近似值計算功能之模擬抽樣參數等參數資訊，並提供各參數之設定說明以解釋參數值定義與用處（如圖 E.1 所示）。待權限內使用者點選設定連結後，系統即畫面下方顯示權限內使用者選擇修改設定參數之參數資料與修改欄位，當使用者更改數值完畢即可按下「設定」（如圖 E.2 所示），系統自動將系統參數數值於系統資料庫中修改，同時執行訊息「設定成功」顯示於系統介面（如圖 E.3 所示）。

權限內使用者進入「郵件語意關鍵字擷取參數設定」功能後，本系統則於系統畫面中顯示權限內使用者可修改參數之資訊，當中，本功能提供使用者「主題詞彙門檻值」與「模擬抽樣參數」等參數供使用者修改，如圖 E.1 所示，系統亦呈現此兩參數於系統中數值與定義說明，如「主題詞彙門檻值」於系統中設定為「40.0%」，且說明為「設定於郵件主題詞彙集合建立時，為能...」。當權限內使用者於「主題詞彙門檻值」點擊「設定」連結時，本系統於下方顯示使用者選擇修參數之設定範圍、預設值等資訊與修改欄位，如圖 E.2 所示，權限內使用者根據系統所提供資訊「可設定範圍為 1~100%之間」將「主題詞彙門檻值」之參數修改為「60.0%」並點擊設定後，本系統乃以紅色字體顯示「設定成功」訊息，並於訊息下方顯示修改後「主題詞彙門檻值」參數資料供權限內使用者確認。此外，權限內使用者亦透過本功能遵循上述說明修改「模擬抽樣參數」（如圖 E.3 所示）。

圖 E.1、郵件語意關鍵字擷取參數設定(1)

圖 E.2、郵件語意關鍵字擷取參數設定(2)

圖 E.3、郵件語意關鍵字擷取參數設定(3)

圖 E.4、郵件語意關鍵字擷取參數設定(4)

E.2 個人化郵件類別推論參數設定使用說明

「個人化郵件類別推論參數設定」功能乃提供權限內使用者進行修改系統分析參數，以進而提升「郵件主成份計算」與「郵件關聯程度與類別推論」功能兩者郵件特徵與個人化郵件等分析準確性。其中，「郵件主成份計算」乃透過設定「特徵解釋比例門檻」強化郵件特徵分析精確性，再透過「類別擷取層級設定」協助「郵件關聯程度與類別推論」功能推論使用者所需個人化郵件類別（如圖 E.5 所示）。以下乃分別敘述兩功能參數之設定方式。

圖 E.5、個人化郵件類別推論參數設定

➢ **特徵解釋比例門檻設定說明**

當權限內使用者選擇「個人化郵件類別推論參數設定」功能後，系統即呈現「特徵解釋比例門檻」、「類別擷取層級設定」等參數資訊，並提供各參數之設定說明以解釋參數值定義與用處。當權限內使用者點選「特徵解釋比例門檻」之設定連結後，如圖 E.6 所示，系統即畫面下方顯示權限內使用者選擇修改設定參數之參數資料與修改欄位，如圖 E.7 所示，權限內使用者根據系統所提供資訊「可設定範圍為 1~100%之間」將「特徵解釋比例門檻」修改為「60.0%」並點擊「設定」後，系統自動將系統參數數值於系統資料庫中修改，同時執行訊息「設定成功」顯示於系統介面（如圖 E.8 所示）。

圖 E.6、特徵解釋比例門檻設定(1)

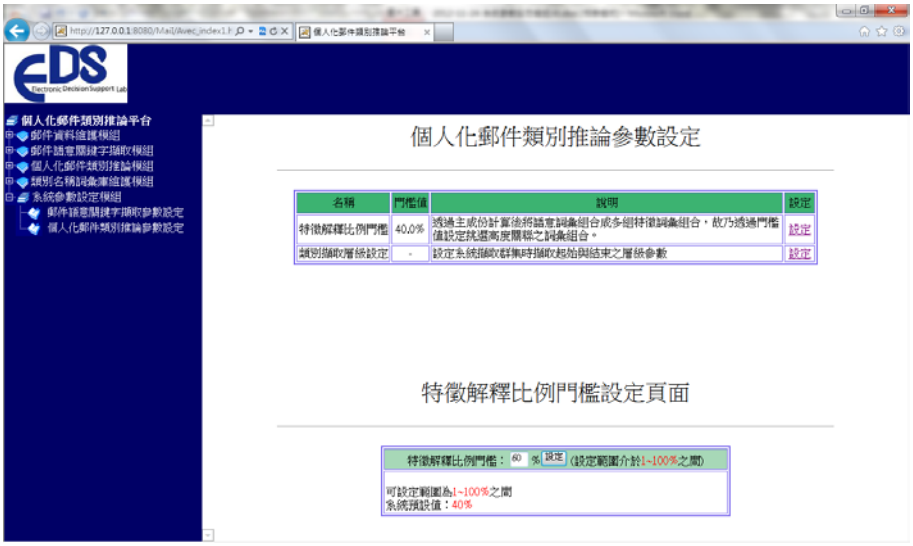

圖 E.7、特徵解釋比例門檻設定(2)

圖 E.8、特徵解釋比例門檻設定(3)

➢ **類別擷取層級設定說明**

當權限內使用者進入「個人化郵件類別推論參數設定」功能後，如圖 E.9 所示，系統即呈現「特徵解釋比例門檻」、「類別擷取層級設定」等參數資訊。當權限內使用者點選「類別擷取層級設定」之設定連結後，如圖 E.10 所示，本系統以彈跳新視窗顯示畫面，系統於視窗內提供權限內使用者「層級起始參數」、「層級結尾參數」等修改欄位與層級說明權限內使用者可透過點擊層級名稱方式顯示層級說明，如圖 E.11 所示，權限內使用者點擊系統畫面中「層級 1」時則系統於右方顯示「郵件分類規則詳細...」等說明訊。權限內使用者根據系統所提供資訊將「層級結尾參數」修改為「2.0」並點擊「設定」後，系統自動將系統參數數值於系統資料庫中修改，同時執行訊息「設定成功」與修改後參數資料顯示於系統介面（如圖 E.12 所示）。

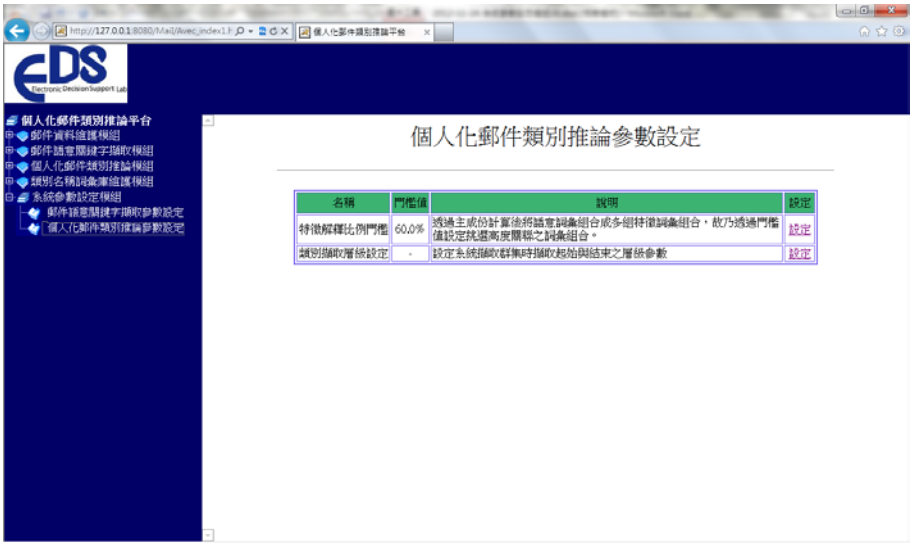

圖 E.9、類別擷取層級設定(1)

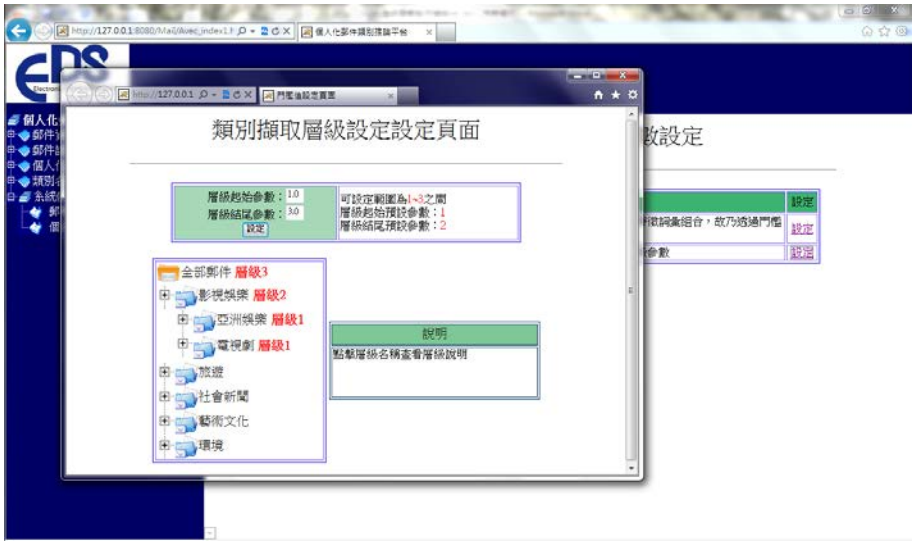

圖 E.10、類別擷取層級設定(2)

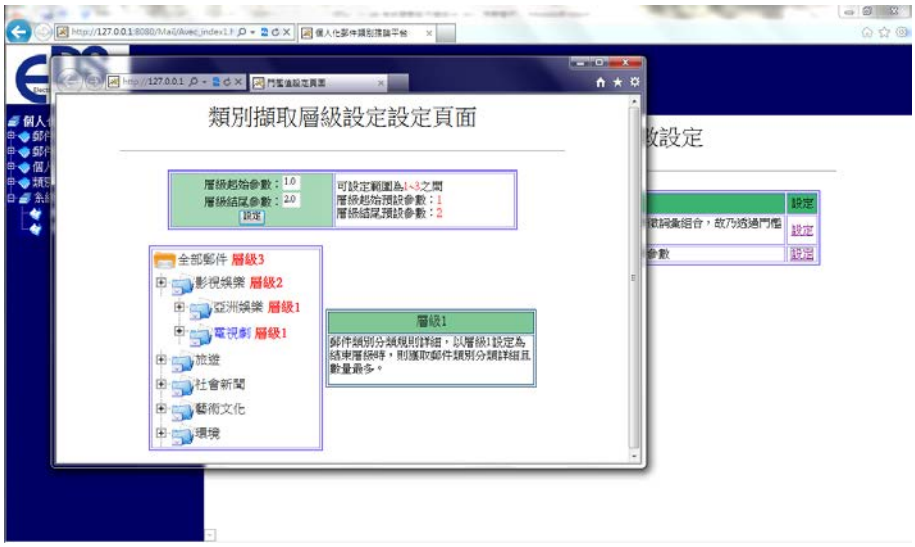

圖 E.11、類別擷取層級設定(3)

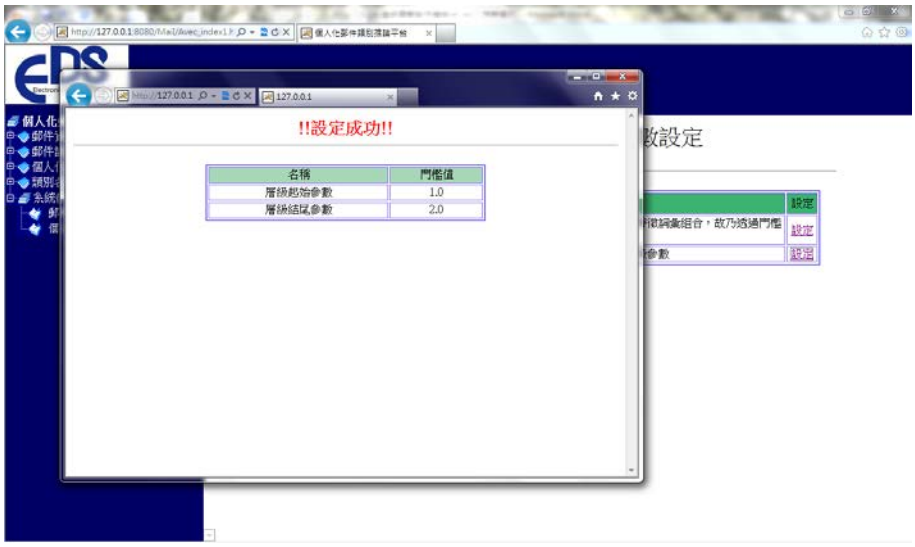

圖 E.12、類別擷取層級設定(4)