建構一個具有自動分類能力之電子郵件管理系統

游坤明[1]　黃智群[1]　吳穎朋[2]
中華大學資訊管理系[1]、中華大學資訊工程系[2]
{yu;wise;ypwu}@pdlab.csie.chu.edu.tw

摘要

電子郵件對於現今的企業而言,提供了非常方便的溝通工具，而為了幫助企業解決電子郵件的管理問題，本論文提出並且實作了一個以 SMTP 為基礎的電子郵件管理系統，其功能包括流量控管、郵件備份、內容過濾、郵件分類、自動回覆和統計分析等。在我們的架構下，無論企業使用何種電子郵件伺服器，此郵件管理系統都能夠正常運作，不需依賴特定的郵件伺服器。而針對郵件分類的部份，我們以 N-gram 表示法為基礎，設計了一個中英文郵件皆可適用的分類演算法，不但能用在電子郵件分類上，亦可用於一般文件分類，經過測試後，證明此分類演算法準確率可達 80%以上，並且具有學習機制，來修正和調整其分類結果。

關鍵詞：電子郵件管理、內容過去、郵件分類、自動回饋

Abstract

The phenomenal growth of Internet accesses has made e-mail an essential communication tool of an enterprise. In this paper, we propose and implement an e-mail management system, which is based on SMTP protocol. The system is designed to resolve the content filtering, mail classification, automatic reply, analysis, auto feedback and statistics. Every company can utilize the proposed system without changing their mail server infrastructure. Moreover, we devise a N-gram based algorithm, which can be applied in both Chinese and English mails. This algorithm not only can be used in classifying e-mails, but also can be used in classifying other documents. The result of our simulation shows the accuracy of this algorithm can reach up to 80 percent and above. Also, the classification results can be revised or adjusted by the learning mechanism.

Keywords: Mail management、Content filtering、Mail classification、Auto feedback

1. 前言

隨著網際網路的快速發展,電子郵件已經成為企業日常運作不可或缺的一部分，其在企業中所扮演的角色，除了是內部的溝通橋樑之外，同時也是企業與外部進行各種業務往來的重要管道。在許多企業之中，電子郵件已漸漸具備正式公文的性質，使其從單純的通訊工具進而轉變成企業的一項重要資產。在這樣的演變下，企業之檔案資料的管理重點，已不僅限於原來的紙張文件，更包括所有進出企業的電子郵件。然而電子郵件的應用越廣泛，其潛在的問題越多，舉例來說：

- 電子郵件的流量過大，影響企業的網路效能。
- 企業的機密文件容易洩露。
- 電腦病毒的危害。
- 電子郵件伺服器易成安全的漏洞。
- 垃圾郵件的各種問題。
- 管理電子郵件的耗時費力。

以上都是當前企業管理者必須面對的重要課題，其中以如何有效管理企業的電子郵件資產最為困難，類似的問題不只存在於企業的電子郵件管理，一般的文件管理也面臨同樣的困境。

另外，文件分類是根據文件的內容或是主題給定類別的一種工作，其目的在於對文件進行分門別類的加值處理，使得文件易於管理和利用。文件分類亦可將非結構化的資料轉換成結構化的資訊，可進行進一步的分析。傳統的分類工作需要用到大量的人力，不僅耗去了許多時間，不同的人也會做出不同的結果。在大部分電子郵件管理的相關研究[2,6,7,9]中，其郵件分類的功能通常針對垃圾郵件、管理個人郵件或是自動回覆等特殊目的，不易協助企業對往來之電子郵件做更詳細的分類。且並未考慮中文郵件的分類問題，故本論文彙整電子郵件管理系統與中文文件自動分類方法[8,10,11]，使企業處理電子郵件更加便利。

在中文分類問題中，由於中文書寫單位是「字」，故中文文章只有「字」界線，而沒有「詞」的界線。中文斷詞方法通常有詞庫(Lexicon)比對、語料庫(Corpus)分析和 N-gram 表示法等。

詞庫比對和語料庫分析都需要大型的詞庫和複雜的斷詞系統，需耗費大量的時間做斷詞的工作，且有可能因詞庫的不完全而造成其無法被擷取出來的情形。因此我們選擇了 N-gram 表示法來擷取郵件中的中文詞彙，但 N-gram 擷取方式會有無意義之詞彙數量過多的缺點，我們的演算法也針對此問題做了調整和改善，並配合類似 tf-idf 的權重計算方式找出文章中的關鍵詞。

2. 中（英）文郵件分類演算法

我們參考了各種中文和英文的文件分類方法，加以改進與合併後，提出了一個不論中文或英文文件皆可適用的分類演算法，而本系統之郵件分類功能亦可由此演算法加以實現。這個演算法共可分成四個主要的部分－（一）從郵件的主旨、內文和所有附件檔之檔名中擷取其關鍵詞。（二）根據已分類之訓練資料找出各類別的關鍵詞。（三）比對郵件和類別的關鍵詞來替郵件分類。（四）新增訓練資料時類別的重新訓練。

2.1 郵件關鍵詞

我們使用 N-gram 表示法來擷取郵件的中文詞彙，並配合類似 tf-idf 的權重計算方式來找出其中的關鍵詞。而使用 N-gram 表示法有以下幾個優點：

- 不需使用詞庫比對或是複雜的斷詞系統來做中文斷詞，因此減少了許多在中文斷詞上所耗費的時間。
- 可以節省建立、擴充和維護詞庫或語料庫時所需的人力及時間。
- 避免有重要詞彙因為未被詞庫網羅而造成其無法被擷取出來的情形發生。

但 N-gram 的擷取方式最大的缺點便在於無意義之詞彙的數量過多，所以我們的演算法也針對這方面的問題做了一些調整與改善。

中文詞彙的擷取

假設現在有一個如圖 2.1（a）所示的中文語句要處理，首先我們會根據一些常見且較不可能為關鍵詞之代名詞、介係詞或連接詞（例如：“我們”、“以及”、“對於”、“他”、“的”、“之”…）等將其分割，如圖 2.1（b）所示，以期能減少一些無用的詞彙。接著我們便設定一個最大N值（Max N），然後從N=2開始，依序取其分割後各段之 2-gram、3-gram…，直到N=Max N為止，而所有這些拆解出之詞彙的聯集，便是我們初步的中文詞彙集合，如圖 2.1（c）至圖 2.1（e）所示。也就是說，如果以TS_i表示Max N=i的詞彙集合，而S_j表示其j-gram的詞彙集合，則：

$$TS_i = \{S_j \mid 2 \le j \le i\}$$

一封郵件的中文部分在經過上述之初步拆解後，會存在著許多無意義的詞彙，所以我們另外訂定了一個最小詞頻的限制（Min F），然後從$TS_{MAX\ N}$之內把在全部郵件中出現之總次數（包含在目前

郵件中出現之次數）小於 Min F 的所有項目刪除，這樣做的目的便是希望能盡量過濾掉那些無意義的詞彙，而其剩餘的項目也才是最後的中文詞彙集合。

圖 2.1 中文詞彙的初步拆解範例

詞彙在郵件中的權重

若我們以 f 代表某個詞彙在一封郵件中出現的次數，以 d 代表有出現過這個詞彙的郵件數，而 D 為全部郵件的總數，w 為此詞彙在這封郵件中的權重，則 w 的 tf-idf 計算法可用下列式子來表示：

$$tf = f$$
$$idf = \log\frac{D}{d}$$
$$w = tf \cdot idf = f \cdot \log\frac{D}{d}$$

但在我們的演算法中，中文詞彙是以 N-gram 的拆解法取得，所以可能會有許多無用卻常常出現的詞彙存在。為了解決這樣的問題，我們把 idf 取 log 的部分去除，希望藉此增加 idf 的強度，使得出現機率較小的詞彙能有更高的權重，其修改後的權重計算方式如下：

$$w = f \cdot \frac{D}{d}$$

擷取郵件關鍵詞的流程

郵件關鍵詞的擷取可分成以下幾個步驟：

i 將郵件的主旨、內文和各附件檔之檔名分成中文和英文兩個部分，同時去除其中的標點符號和空白，且順便利用中英文的不同和這些特殊符號把郵件內容截斷，若有單一中文字或是單一英文字母的片段，便將其一並去除。這個步驟完成後，中文部分會是一個個的語句，而英文部分則已是個別的詞彙。

ii 用中文詞彙的擷取的初步拆解方式處理所有的中文語句，然後統計其初步拆解結果之中文詞彙以及前一步驟分出之英文詞彙在目前郵件出現的次數，並隨之更新郵件資訊庫中各詞彙在所有郵件出現的總次數和總郵件數，更新完這些資訊後，如果是中文詞彙則還必須做如中文詞彙的擷取所述之最小詞頻檢查。

iii 以詞彙在的郵件中的權重計算方式，計算郵件

之全部中英文詞彙的權重，然後依據我們事先定好的門檻值（Threshold），將小於門檻值的詞彙去除，而剩下的就是此封郵件的關鍵詞集合。

2.2 類別關鍵詞

在擷取類別關鍵詞之前，必須要有已分類好的訓練資料，以供各類別收集相關的詞彙資訊，然後才能從這些類別的詞彙資訊中找出其關鍵詞。我們用前一節所描述之方法，找出每一筆訓練資料的所有關鍵詞以及其在此訓練資料中出現的次數，接著依圖 2.2 的方式歸納整理出各類別的詞彙集合。在圖 2.2 中，關鍵詞下方的數字為其在所屬訓練資料中出現的次數，而一個類別的詞彙集合，便是由屬於此類別之所有訓練資料的關鍵詞所構成，對每個類別來說，我們必須統計其各個詞彙在此類別中出現的次數和郵件（訓練資料）數，以供接下來的類別關鍵詞擷取之用。

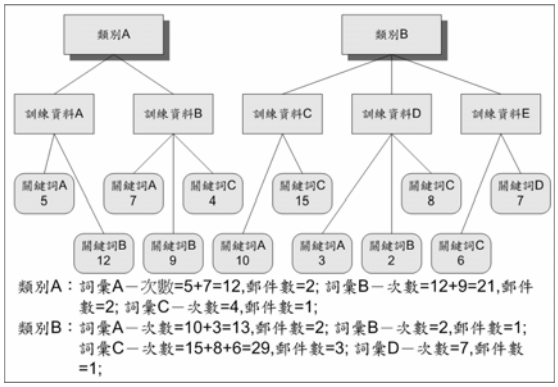

圖 2.2 類別詞彙的收集範例

假設T_i代表一個詞彙，C_j代表一個類別，且共有n個類別。我們用f_{ij}表示T_i在C_j出現的次數，d_{ij}表示T_i在C_j出現的郵件數，D_j表示C_j的總郵件數，而w_{ij}則為T_i在C_j的權重。在On-line模式的權重計算上，我們考慮詞彙出現的集中度，若T_i集中出現於某幾個類別，而不是平均出現於所有類別，則T_i在其較常出現的類別便會有較高的權重，其算法如下：

$$w_{ij} = \frac{f_{ij}}{\sum_{j=1}^{n} f_{ij}}$$

（On-line 模式）

而對於Off-line模式，我們除了考慮以上的集中度之外，還參考詞彙在類別中出現的平均度，若T_i平均出現於C_j的所有郵件，並非只集中出現於C_j的某幾封郵件，則T_i在C_j便會有較高的權重，

如以下式子所示：

$$w_{ij} = \frac{f_{ij}}{\sum_{j=1}^{n} f_{ij}} \cdot \frac{d_{ij}}{D_j}$$

（Off-line 模式）

最後，我們為類別關鍵詞也訂定一個門檻值，希望能夠藉此留下較有分類價值的詞彙，避免分類結果受到不重要之詞彙的干擾。在各類別中，其權重符合門檻值要求之詞彙的集合，便是此類別的關鍵詞集合。

2.3 郵件與類別之比對

我們的分類演算法使用線性分析的方式來比對郵件與各類別間的相似度，以找出與要分類的郵件最相近之類別。首先，假設有m個詞彙，其分別為$T_1, T_2, \ldots, T_m$，以T_i表示之，還有n個類別，其分別為$C_1, C_2, \ldots, C_n$，以C_j表示之，並以D_j代表C_j的總郵件數，以E代表要分類的郵件。接下來，如果T_i屬於C_j的關鍵詞集合，則c_{ij}等於T_i在C_j的權重，否則c_{ij}便為 0，同樣地，若是T_i屬於E的關鍵詞集合，則e_i等於T_i在E的權重，否則e_i便為 0。此時，E和C_j可分別用向量（Vector）$e = (e_1, e_2, \ldots, e_m)$以及向量$c_j = (c_{1j}, c_{2j}, \ldots, c_{mj})$來表示，而E與$C_j$間的相似度Sim$(E, C_j)$之計算方式如下：

$$Sim(E, C_j) = e \cdot c_j \cdot \frac{\sqrt{D_j}}{\|c_j\|} = \left(\sum_{i=1}^{m} e_i \cdot c_{ij}\right) \cdot \frac{\sqrt{D_j}}{\sqrt{\sum_{i=1}^{m} {c_{ij}}^2}}$$

以上之相似度的計算方式主要是以郵件向量（e）和類別向量（c_j）之內積（Dot Product）為基礎，但由於各類別向量的長度差異可能會影響到比較時的公平性，所以我們決定將其做正規化（Normalization）之處理，使得所有類別向量的長度皆相同。可是另一方面，每個類別的郵件數通常都不同，也就是說，郵件出現在各類別的機率本就不一樣，因此若是每個類別向量的長度皆相同，則亦有不公之處，於是我們再另外以類別的郵件數之函數（$\sqrt{D_j}$）為其加權，讓郵件數較多的類別能有較高的優先權。在這樣的計算方式下，其得到的數值愈大，則表示相似度愈高，所以對郵件E而言，只要求得Sim(E, C_1)至Sim(E, C_n)之間全部的值，然後取其最大者即為E所屬之類別。

2.4 類別的重新訓練

當類別被刪除或是其訓練資料有增加時，我們就必須更新類別的詞彙資料並執行類別的重新訓練。在本系統之中，類別的重新訓練可分為兩種－（一）刪除所有的類別關鍵詞，然後重新計算各類別所有詞彙之權重。此種方式主要用於刪除類別或是加入多筆訓練資料（包含手動加入整理好的訓練

資料以及 Off-line 模式的學習機制）之時。（二）僅刪除可能變動的類別關鍵詞，接著針對這些可能變動的詞彙重新計算其權重。其主要應用在只新增一筆訓練資料（On-line 模式的學習機制）之時。

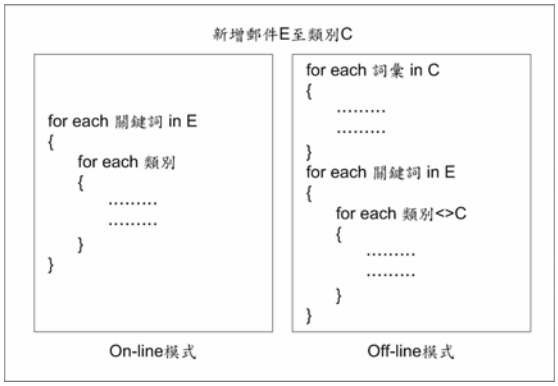

圖 2.3 新增單筆訓練資料之類別重新訓練

圖 2.3 是在只新增一筆訓練資料（E）至某一類別（C）的情況下，On-line 模式和 Off-line 模式兩種不同的權重計算方式之類別重新訓練的比較。On-line 模式的計算方式只需要重算新增的訓練資料之所有關鍵詞在各類別的權重，而 Off-line 模式則除了要重算跟 On-line 模式相同的資料外，還要重算此訓練資料所屬類別之全部詞彙的權重，兩者相較之下，On-line 模式要更新的資料比 Off-line 模式要少得多。

3 郵件分類與訓練學習機制

3.1 郵件分類與自動回覆之運作流程

郵件的分類與回覆可分成 8 個步驟，如圖 3.1 所示：

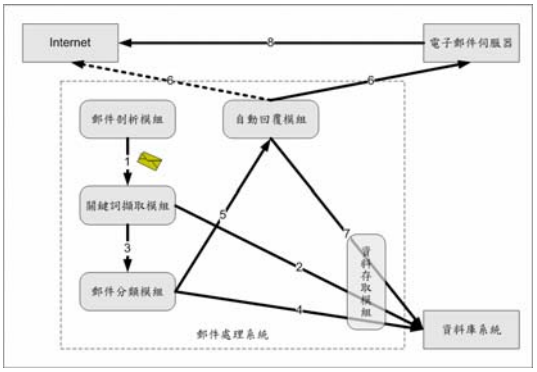

圖 3.1 郵件的分類與回覆之運作流程

郵件處理系統之自動回覆模組可以選擇其回覆信件的方式，而回覆方式共有兩種，一是透過電子郵件伺服器，一是自行回覆。若是透過電子郵件伺服器，自動回覆模組就可以省下查詢 DNS 和失敗重試等工作的時間。如果是自行回覆，則可確實知道信件的回覆狀態，讓管理者能得到更多的資訊。

3.2 郵件分類的訓練與學習之機制

圖 3.2 可分成兩個部分來看，下半部為用整理好之訓練資料做類別的訓練，上半部為管理者回饋的學習機制。這兩種方法都可以協助我們調整類別的特徵，讓分類結果能夠更加符合管理者的需求。

圖 3.2 的下半部是屬於郵件處理系統的工作範圍，共有 5 個步驟：

- 步驟 1－由郵件剖析模組讀取所有的訓練資料檔案並解析其內容。
- 步驟 2－郵件剖析模組將解析出的全部訓練資料皆轉交給關鍵詞擷取模組處理，而這些解析出的內容不會被儲存到資料庫。
- 步驟 3－關鍵詞擷取模組把每一筆訓練資料的所有關鍵詞都依其類別儲存至資料庫之中。
- 步驟 4－關鍵詞擷取模組通知類別訓練模組進行重新訓練的工作。
- 步驟 5－類別訓練模組從資料庫中取得重新訓練全部類別所需的資訊，訓練完後並將結果存回資料庫。

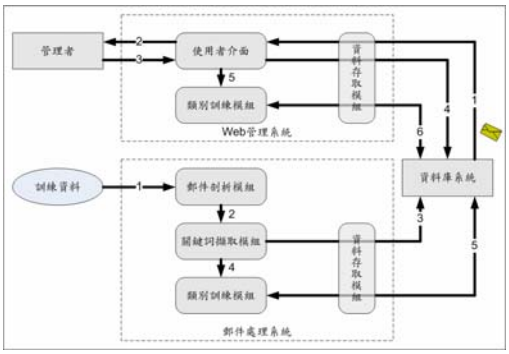

圖 3.2 類別的訓練與學習之運作流程

圖 3.2 的上半部是屬於 Web 管理系統的工作範圍，共有 6 個步驟：

- 步驟 1 至步驟 2－使用者介面自資料庫中讀取一封已分類但未確認類別之郵件，然後將此封郵件交由管理者進行類別確認的工作。
- 步驟 3 至步驟 4－管理者確認其類別之後，使用者介面便將結果存到資料庫，並將此封郵件的關鍵詞全都依其所屬類別歸類。
- 步驟 5－使用者介面通知類別訓練模組進行加入單筆訓練資料的類別訓練工作。
- 步驟 6－類別訓練模組從資料庫中取得加入此封郵件之類別訓練工作會用到的相關資訊，訓練完後並將結果存回資料庫。

上述之管理者回饋的學習機制我們稱之為 On-line 模式，除此之外，本系統還另外提供了一種 Off-line 模式的回饋機制，兩者間的差異有二：

I. 類別訓練的時間點和工作模組不同－Online

模式的回饋機制如圖 3.9 的上半部所示，它在步驟 5 至步驟 6 中利用 Web 管理系統的類別訓練模組，針對管理者已確認完的郵件做即時的處理，也就是說這封郵件的特徵會立刻反映在下一次的分類行為上，但這樣的運作模式也相對增加了郵件確認所需的時間和本系統的負荷；而所謂 Off-line 模式的回饋機制，其郵件類別的確認方式就如 Online 模式之步驟 1 至步驟 4 一樣，只是它省略了步驟 5 至步驟 6 的過程，並將類別的重新訓練工作改由郵件處理系統中的類別訓練模組定時地去整批執行，管理者可將執行的時間點設定在電子郵件進出的離峰時段，以避免影響本系統的效率，但這種運作模式就無法在下次分類立即反映出類別已確認之郵件的特徵。

II. 類別關鍵詞的權重計算方式不同－在 On-line 的學習模式中，由於必須立即做類別特徵的更新，因此我們用了跟 Off-line 模式不同的權重計算方式，使得在每加入一封郵件時，需要重新計算的資料量較 Off-line 模式少，但分類的準確率也較差。而 Off-line 的學習模式所用的權重計算方式則相反，分類的準確率較高，但每加入一封郵件時所需重新計算的資料量也較多。

4 實驗結果

為了進行郵件分類的實驗，我們從 Yahoo 奇摩的新聞中收集了 2449 篇平均長度為 550 個中文字(即 1100 個字元)的報導，作為分類實驗中的郵件樣本。

針對此樣本進行測試實驗，每筆測試資料之分類結果的前三名都會被記錄下來，並且測量其實驗的統計數據。在我們的分類實驗中有以下幾種參數：訓練資料數=2449-測試資料數=1224；Max N(N-gram 表示法之最大 N 值)=2；Min F(郵件的中文詞彙之最小詞頻)=5；郵件關鍵詞門檻值=100；類別關鍵詞門檻值=0.7(On-line 模式)或 0.02(Off-line 模式)；學習機制=No；類別郵件數加權=Yes。以上各種參數之等號後面的值，是它們各自的預設值。我們分別測試各種情況下 On-line 模式和 Off-line 模式的分類結果。

由表 4.1 和表 4.2 的實驗數據我們可以知道，較大的 Max N 對分類的準確率並沒有太大的幫助，且會使得資料量大幅增加(原始詞彙代表中文部分尚未機過 Min F 篩選的詞彙集合)，郵件分類所需的時間也會增加，故以 Max N=2 來擷取郵件的中文詞彙較為恰當。

表 4.1 Max N 對郵件分類之影響-1

Max N	郵件詞彙總數	郵件平均原始詞彙數	郵件平均詞彙數	郵件平均關鍵詞數
4	1216779	821.9	261.6	1121.8
3	698607	578.5	245.4	111.5
2	217812	297.8	190.5	78.4

表 4.2 Max N 對郵件分類之影響-2

Max N	On-line 模式第一名命中率 (%)	Off-line 模式第一名命中率 (%)	On-line 模式前三名命中率 (%)	Off-line 模式前三名命中率 (%)
4	84.08	82.04	94.78	95.35
3	83.67	83.02	94.69	95.18
2	82.37	82.61	94.53	94.94

表 4.3 列出了類別郵件數加權對分類結果的影響，就準確率方面來說，此方法的結果跟取較大的 Max N 類似，只使準確率略為提升，但因為類別郵件數加權並不會增加分類時的負擔，因此我於郵件分類演算法中仍然繼續採用此種加權方式。

表 4.3 類別郵件數加權對郵件分類之影響

類別郵件數加權	On-line 模式第一名命中率 (%)	On-line 模式前三名命中率 (%)	Off-line 模式第一名命中率 (%)	Off-line 模式前三名命中率 (%)
Yes	82.37	94.53	82.61	94.94
No	82.29	94.37	81.96	95.1

表 4.4 為起始訓練資料=612 時，學習機制使用與否的命中率比較。在使用學習機制的情況下，確實能夠提高分類的準確性，雖然提高的程度有限，但最重要的是這種學習的方式可隨時根據管理者之行為模式調整其類別的特徵，使分類結果更具可信度。

表 4.4 學習機制對郵件分類之影響

On-line 模式			Off-line 模式		
學習機制	第一名平均命中率 (%)	前三名平均命中率	學習機制	第一名平均命中率(%)	前三名平均命中率 (%)
Yes	80.89	92.92	Yes	82.85	93.36
No	77.68	91.07	No	80.29	92.92

5 結論

有鑒於電子郵件對企業而言的重要性日益增加，且其每日進出的數量又太過龐大而使得管理不易，所以本研究提出了一套完整的管理架構及方法，主要目的便在於協助企業將郵件管理的工作自動化，以節省在時間和人力上可能耗費的成本。在經過系統實作與實驗模擬的過程之後，我們亦證明了本研究所提之架構和方法在實際應用上確實可行。

我們的電子郵件管理系統具備了跨平台的能力，不會受限於用戶端作業系統及電子郵件伺服器的種類，這樣的特性可彌補許多相關研究在這方面的不足之處。除此之外，本系統還結合了諸多與郵件管理有關的功能，例如流量控管、郵件備份、內

容過濾、和統計分析等，讓企業管理者可以更直接地監控各種往來的電子郵件。

另一方面，為了更有效地幫助企業管理電子郵件，我們設計了一個擁有學習能力的郵件(文件)分類演算法，已藉此達到郵件自動分類和回覆的目的。透過這個演算法，本系統可將接收到的郵件一管理者所訂定的類別執行自動分類的工作，然後再根據各個類別之行為定義，決定是否要自動回覆以及回覆的內容。當分類結果不正確時，本系統亦可隨時針對各類別的特徵進行校調，使未來的分類結果能夠更加符合企業的要求。

本研究往後還有以下幾個可發展的方向：

- 整合病毒掃描的功能，以便清除電子郵件中帶有病毒的附件檔以及具有高度危險性的Script。
- 建立分散式的資料庫架構，即使是一個擁有多個子公司的企業，我們仍然可以將其所有的電子郵件加以集中管理。
- 利用資料探勘（Data Mining）的技術分析已分類的電子郵件，從中發掘出有價值的知識以供企業參考。

參考文獻

[1] N. Borenstein, N. Freed, “MIME (Multipurpose Internet Mail Extensions): Mechanisms for Specifying and Describing the Format of Internet Message Bodies ”, June 1992.

[2]William W. Cohen, “Learning Rules that Classify E-Mail”, Proc. AAAI-1996 Spring Symposium on Machine Learning in Information Access, pp. 124-143, March 1996.

[3] David H. Crocker, “Standard for the Format of ARPA Internet Text Messages ”, RFC 822, August 1982.

[4] Li Cheng, Wang Weinong, “Internet Mail Transfer and Check System Based on Intelligence Mobile Agents”, Proceedings 2002 Symposium on Applications and the Internet (SAINT ‘02), pp. 2-3, 2002.

[5] Jonathan B. Postel, “Simple Mail Transfer Protocol”, RFC 821, August 1982.

[6] Jason D. M. Rennie, “ifile: An Application of Machine Learning to E-Mail Filtering”, Proceedings of the KDD-2000 Workshop on Text Mining, August 2000.

[7] Chang-Jiun Tsai, Shian-Shyong Tseng, Her-Tsaan Cheng, “Intelligent E- mail Management System”, IEEE International Conference on System, Man, and Cybernetics (SMC 1999), pp. 824-829, 1999.

[8] Jyh-Jong Tsay, Jing-Doo Wang, “Design and Evaluation of Approaches for Automatic Chinese Text Categorization”, International Journal of Computational Linguistics and Chinese Language Processing (CLCLP), Vol. 5, No. 2, pp. 43-58, August 2000.

[9] 陳瑞順、胡駿彥，“以代理人為基礎之客服電子郵件自動回覆系統”，國立交通大學資訊管理研究所，碩士論文，民國 91 年。

[10] 楊允言、謝清俊、陳淑美、陳克健，”中文文件自動分類之探討”，大漢學報，第 13 期，頁 241-256，民國 88 年 10 月。

[11] 蔡志忠、邱聖斌，”中文文件表示法在文件分類中之比較”，國立中正大學資訊工程研究所，碩士論文，民國 90 年 7 月。