垃圾郵件過濾技術之初步研究

邱炫盛
國立台灣師範大學資訊工程所
g93470240@csie.ntnu.edu.tw

陳柏琳
國立台灣師範大學資訊工程所
berlin@csie.ntnu.edu.tw

摘要

隨著網際網路的蓬勃發展，在過去幾年來電子郵件已成為社會大眾溝通的重要媒介之一。因此，許多廣告商利用網際網路這個方便且低成本的管道，大量地發送廣告信件，這些對一般民眾來說是沒有用的資料，我們稱之為垃圾郵件(Junk Mails)。關於垃圾郵件過濾的研究，在近幾年來逐漸受到重視，本論文提出以隱藏式馬可夫模型(Hidden Markov Model)作為垃圾郵件過濾模型，並且與近年來陸續被發展出的過濾模型，諸如貝氏分類器(Bayesian Classifier)、潛藏語意分析(Latent Semantic Analysis)等，做一系列的比較。實驗部份一共使用兩套語料，一套是目前國際上公開的英文郵件實驗語料，Ling-Spam，另一套是我們在台灣所收集的中文郵件實驗語料。我們將上述的垃圾郵件過濾模型分別實作在這兩套郵件實驗語料，實驗結果顯示我們所提出的隱藏式馬可夫模型可以在垃圾郵件過濾上可以得到相當不錯的效果。

關鍵詞：郵件過濾、自動分類、隱藏式馬可夫模型、貝氏分類器、潛藏語意分析

1. 前言

由於網際網路技術的持續發展，以及各國政府對於網路寬頻與資訊電子數位化的積極推動，網際網路早已經成為國民日常生活中不可或缺的一環，而電子郵件(Electronic Mails, E-Mails)也成為民眾普遍溝通的媒介之一。因此，許多圖利的不肖商人利用網路的方便性及低成本來發送廣告信件，所以民眾常常會收到大量的、非自願的電子郵件，我們稱之為垃圾郵件(Junk Mails)。垃圾郵件的問題日趨嚴重，不僅造成網路伺服器資源的浪費，也對民眾造成困擾。根據統計，國內網路服務提供者(ISP)處理垃圾電子郵件，其付出的費用包含頻寬、人力、軟硬體等等，每一封平均約花費0.02元，而一年下來需要多花費幾百萬元在處理垃圾電子郵件上面，這些額外的營運成本，最後仍是轉嫁在一般使用者身上。而當使用者收到郵件之後，同樣地也需要進行手動的過濾篩選動作，無形中卻也浪費了時間，並間接地降低了工作的效率與生活的品質。

近年來，相關的法令規範與垃圾電子郵件過濾技術研發等議題逐漸被重視與研究，例如行政院通訊傳播委員會籌備處和電信總局共同研擬「濫發商業電子郵件管理條例」，已經過行政院審查完畢，如果立法院三讀通過後，民眾將可對濫發垃圾郵件的業者提出賠償，法院亦可對匿名寄發者處以徒刑，這對抵制垃圾郵件不啻是為一股助力[1]。而除相關法令規範之外，自動化垃圾郵件過濾技術的發展，近幾年來，不論在國內或國外均有許多學者陸續在從事這方面的研究。

傳統的電子郵件過濾主要是以規則式(Rule-based)方法為基礎，可分成使用者端與伺服器端處理方式。使用者端利用收信軟體，如Outlook等的內建功能，使用者可以事先設定一些關鍵詞或者是欲排除信件的主旨與寄件者資訊，當如果收到一封新郵件符合部分的條件時，於是這封信可能就會被收信軟體認定為垃圾郵件而予以摒除。但是這種郵件過濾方法，卻已經不足以應付現在的垃圾郵件的多變性。譬如現在的廣告商已會巧妙地去迴避特定性的廣告詞語，而將廣告信件的主旨與內容以不相干或是用其他的用語代替，以避免被偵測出為垃圾郵件的可能性。還有一個會面臨到的問題就是，某些關鍵詞可能具有某種正當的語意，但是卻會受到一些自然語言處理(例如中文斷詞時產生的混淆問題)導致郵件無法傳達正確語意因而被過濾掉。然而對使用者來說，關鍵詞是否能定義明確又是另一個問題。另一方面，以伺服器端處理的郵件過濾主要是以標頭檔或是信件流量來作分析，伺服器會藉由同時間、同主機所大量湧進來的相同信件判斷是否為垃圾信，這種方式對於會員電子報的誤判率最高；或者是分析寄件者信箱是否有效，然後收集黑名單，以便日後拒絕掉由名單寄來的信。上述這兩種方式也沒有辦法有效地防制廣告信件，因為廣告商會透過多台的郵件主機定量定時的發送，以規避流量，同時也會捏造不同的寄件者等標頭以欺騙收信端主機。通常主機端與使用者端兩類型的規則導向過濾技術是可以並行採用的。

電子郵件過濾亦可以機械學習式(Machine-learning-based)方法為基礎，針對垃圾(廣告)電子郵件與一般正常郵件建立模型，從郵件內容(Content)來判斷郵件的類別屬性，機械學習式郵件過濾亦可以施行於使用者端、伺服器端或者兩者並行 [16]。本論文所將探討的垃圾郵件過濾技術即屬於此類，我們提出以隱藏式馬可夫模型(Hidden Markov Model, HMM)作為垃圾郵件過濾模型，並且與近年來陸續被發展出的過濾模型，諸如貝氏分類器 (Bayesian Classifier)、潛藏語意分析(Latent Semantic Analysis, LSA)等，做一系列的比較。實驗

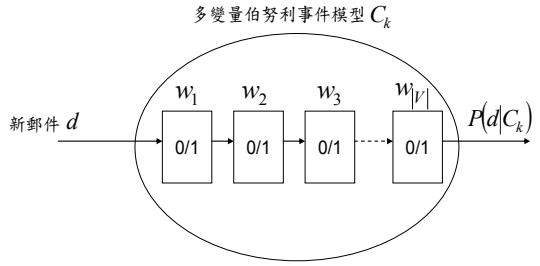

圖 1 多變量伯努利事件模型

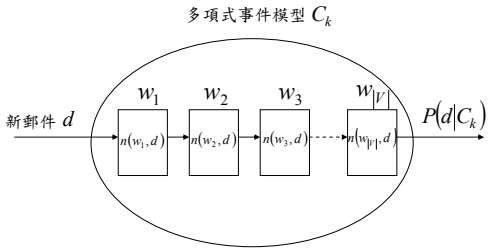

圖 2 多項式事件模型

部份一共使用兩套語料，一套是目前國際上公開的英文郵件實驗語料，Ling-Spam，另一套是我們在台灣所收集的中文郵件實驗語料。我們將上述的垃圾郵件過濾模型分別實作在這兩套郵件實驗語料，實驗結果顯示我們所提出的隱藏式馬可夫模型可以在垃圾郵件過濾上得到相當不錯的效果。

論文接下來的安排如下：第二章將介紹貝氏分類器及其變形；第三章介紹潛藏語意分析的過濾應用；第四章說明我們所提出的隱藏馬可夫模型及其變形；第五章為實驗與討論；第六章是結論與未來展望。

2. 貝氏分類器

貝氏分類器從貝氏網路延伸而來[13]，其假設某一個類別擁有多種特徵(或索引，像是詞、字等)，所以屬於該類的樣本(Samples)或是郵件可由不同權重的特徵組合而成，以本論文探討的主題來說，屬於合法的電子郵件是由合法郵件類別的特徵組成，垃圾郵件是由垃圾郵件類別的特徵所組成。在貝氏分類器中我們是以某些詞的出現與否或次數作為特徵，因此合法與垃圾郵件的特徵有可能會重疊。我們使用固定的共用詞典 $V=\{w_1,w_2,\cdots,w_{|V|}\}$ 當作兩類特徵的來源，如果該類沒有出現過某個詞，其權重為零，仍符合原來的定義。

現在假設要對一封新進的電子郵件 d 做分類。我們會分別計算給定郵件 d 情況下，屬於某類的機率值 $P(C_k|d)$，其中 $k=l$ 表示合法郵件，$k=j$ 表示垃圾郵件；而因為我們沒辦法記錄所有特徵可能組合的郵件 d 產生 C_k 的機率，所以無法直接求得 $P(C_k|d)$，因此透過貝氏定理做轉換，可以進一步表示成：

$$P(C_k|d)=\frac{P(d|C_k)P(C_k)}{P(d)} \tag{1}$$

對於合法與垃圾郵件兩類而言，因 $P(d)$ 並不會影響最後郵件類別判斷結果，可以省略不計。最後分類器的輸出可表示成：

$$k^*=\arg\max_{k\in\{l,j\}} P(d|C_k)P(C_k) \tag{2}$$

我們需要估算 $P(C_k)$ 與 $P(d|C_k)$ 的機率分佈，其中 $P(C_k)$ 可以使用訓練郵件語料以最大相似度估算(Maximum Likelihood Estimation, MLE)[11]求得：

$$P(C_k)=\frac{N(C_k)}{\sum_{k'\in\{l,j\}}N(C_{k'})} \tag{3}$$

其中 $N(C_k)$ 代表某一類電子郵件在訓練語料中的數量，即郵件的封數。接下來要估算 $P(d|C_k)$ 的機率，由於是由許多特徵(詞)所組成，要直接估算 $P(d|C_k)$ (也就是 d 的所有特徵同時聯合出現在 C_k 機率)並不容易，因此通常會作簡單貝氏假設(Naïve Bayesian Assumption)[11]，也就是假設特徵之間是相互獨立的。在這個基本假設下，又有不同觀點及假設，其中機率估測方式也不同，可以分成多變量伯努利事件模型(Multivariate Bernoulli Event Model)與多項式事件模型(Multinomial Event Model)[12]。

2.1 多變量伯努利事件模型

多變量伯努利事件模型假設一封電子郵件 d 是由連續 $|V|$ 次的伯努利試驗所產生，V 是詞典。每一次試驗 t 的結果只有兩種，代表詞 w_t 的出現有無。所以一封電子郵件可以想成是 $|V|$ 維的向量，每一維度 t 對應到詞典的某一個詞 w_t，其值 $I_{d,t}$ 是二元的，$I_{d,t}=1$ 表示電子郵件 d 有出現詞 w_t，反之，$I_{d,t}=0$ 表示沒有出現。根據簡單貝氏假設與使用多變量伯努利模型，我們可以用下面的式子來表示當給定某類別 C_k 時郵件 d 發生的機率：

$$P(d|C_k)=\prod_{t=1}^{|V|}\left[\left(I_{d,t}\cdot P(w_t|C_k)\right)+\left(1-I_{d,t}\right)\cdot\left(1-P(w_t|C_k)\right)\right] \tag{4}$$

由式(4)可看出，在多變量伯努利模型的假設下，即使是沒有出現在郵件中的詞也代表某種資訊而有所貢獻。其中 $P(w_t|C_k)$ 代表屬於 C_k 的電子郵件訓練語料中有出現詞 w_j 的郵件篇數的比例：

$$P(w_t|C_k)=\frac{\sum_{d_i\in C_k}I(w_t,d_i)}{N(C_k)} \tag{5}$$

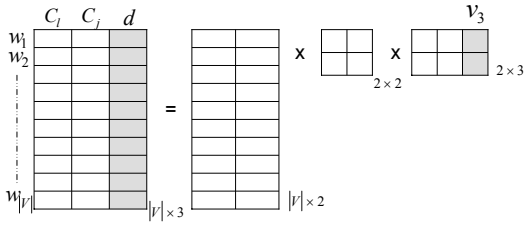

圖 3 潛藏語意分析示意圖

其中 $I(w_t, d_i)$ 表示某一個詞 w_t 是否出現在郵件 d_i，值是 0 或 1；$N(C_k)$ 代表某一類電子郵件在訓練語料中的數量。圖 1 是多變量伯努利事件模型的示意圖。

2.2 多項式事件模型

多項式事件模型假設一封電子郵件是由一連串的詞所組成，且這些詞來自詞典 V。所以與多變量伯努利模型相同，每一封郵件 d 也可表示成長度為 $|V|$ 的向量；但不同的是向量中每一維度 t 的值不再是 0 與 1，而是對應的詞 w_t 出現在郵件 d 的次數。故在給定某一個類別 C_k 的情況下，產生郵件 d 的機率可以用多項式分佈表示：

$$P(d|C_k) = P(|d|)\frac{\left(\sum_{t=1}^{|V|} n(w_t,d)\right)!}{\prod_{t=1}^{|V|} n(w_t,d)!}\prod_{t=1}^{|V|} P(w_t|C_k)^{n(w_t,d)} \quad (6)$$

其中 $P(|d|)$ 是郵件 d 長度的事前機率，$p(w_t|C_k)$ 可以用屬於 C_k 的電子郵件訓練語料中有出現詞 w_t 的頻率來評估：

$$P(w_t|C_k) = \frac{\sum_{d_i\in C_k} n(w_t,d_i)}{\sum_{k'\in\{l,j\}}\sum_{d_{i'}\in C_{k'}} n(w_t,d_{i'})} \quad (7)$$

圖 2 是多項式事件模型的示意圖。

3. 潛藏語意分析

另外一個方法是潛藏語意分析(LSA)，這個方法原先使用在資訊檢索上，現在推廣應用在垃圾郵件過濾上[3,4]。除了不同於貝氏分類器以機率生成模型來執行郵件分類，潛藏語意分析強調的是語意上的分類，根據隱藏在文字內的語意來區分郵件是屬於垃圾郵件或是正常郵件。潛藏語意分析應用於分類中可分成三個步驟：建立特徵矩陣、奇異值分解運算(SVD)與產生語意指標(Semantic Anchor)及文件分類。

首先，假設以詞典 V 中的詞當作兩個郵件類別的共同特徵，我們需要先利用訓練語料產生一個 $|V|\times 2$ 的稀疏矩陣 $\boldsymbol{M}$，$\boldsymbol{M}$ 的第一行向量表示合法郵件訓練群，第二行向量表示垃圾郵件訓練群，其中第 t 列第 k 行的值 $g_{t,k}$ 可表示成：

$$g_{t,k} = (1-\varepsilon_t)\cdot P(w_t|C_k) \quad (8)$$

其中 $P(w_t|C_k)$ 如式(7)所示為詞 w_t 在類別 C_k 出現的機率；ε_t 是詞 w_t 的正規化熵值 (Normalized Entropy)，可表示成：

$$\varepsilon_t = -\frac{1}{\log 2}\sum_{k\in\{l,j\}}\frac{n(w_t,C_k)}{\sum_{k'\in\{l,j\}} n(w_t,C_{k'})}\log\frac{n(w_t,C_k)}{\sum_{k'\in\{l,j\}} n(w_t,C_{k'})} \quad (9)$$

其中 $n(w_t,C_k)$ 是 w_t 出現在所有類別為 C_k 之郵件的次數，可以表示成：

$$n(w,C_k) = \sum_{d_i\in C_k} n(w_t,d_i) \quad (10)$$

當詞 w_t 出現在不同類別之郵件次數越多，ε_t 將會越大。在式(9)中乘上 $(1-\varepsilon_t)$ 是為了區分擁有相同次數 $g_{t,k}$ 的詞 w_t，使其有不同的資訊，給予不同的權重，也就是出現在不同類別之郵件次數越高的詞其重要性將越低。

建立矩陣 $\boldsymbol{M}$ 之後，我們需要對 $\boldsymbol{M}$ 做奇異值分解(Singular Value Decomposition, SVD):

$$\boldsymbol{M} \approx \boldsymbol{USV}^T \quad (11)$$

其中 $\boldsymbol{U}$ 是 $|V|\times 2$ 的矩陣，由維度為 2 的列向量 $\vec{\boldsymbol{u}}_t$ $(1\le t\le|V|)$ 組成；$\boldsymbol{S}$ 是 2×2 的對角矩陣，其對角線上的值 $\sigma_1\ge\sigma_2$；$\boldsymbol{V}$ 是 2×2 的矩陣，包含了 2 個列向量 $\vec{\boldsymbol{v}}_l$ 與 $\vec{\boldsymbol{v}}_j$；T 是矩陣轉置符號。經過奇異值分解後產生了兩個語意上的映射：合法郵件映射與垃圾郵件映射，$\bar{\vec{\boldsymbol{v}}}_l = \vec{\boldsymbol{v}}_l\boldsymbol{S}$ 表示合法郵件，$\bar{\vec{\boldsymbol{v}}}_j = \vec{\boldsymbol{v}}_j\boldsymbol{S}$ 表示垃圾郵件。然後使用這兩個向量來當作類別的語意指標。經過 SVD 與映射之後所產生的特徵向量與貝氏分類器的特徵向量不同，特徵不再是特定的詞，而是考慮整體的語意。

再者，我們想要對新郵件 d 做分類的動作，會先把 d 轉換成向量的表示方式，然後判斷 d 比較靠近哪一個語意指標，以達到分類的效果。把 d 轉換成一個 $|V|$ 維的向量 $\vec{\boldsymbol{d}}$ 的作法如式(8)，與建立矩陣 $\boldsymbol{M}$ 時的方式相同。所以現在我們要對原來的 $|V|\times 2$ 的矩陣 $\boldsymbol{M}$ 做延伸，變成一個 $|V|\times 3$ 的矩陣 $\hat{\boldsymbol{M}} = [\boldsymbol{M}\quad \vec{\boldsymbol{d}}]$，而將 $\vec{\boldsymbol{d}}$ 用式(11)奇異值分解表示可得到 $\vec{\boldsymbol{d}} = \boldsymbol{US}\vec{\boldsymbol{v}}^T$，所以：

$$\bar{\vec{\boldsymbol{v}}} = \vec{\boldsymbol{v}}\boldsymbol{S} = \vec{\boldsymbol{d}}^T\boldsymbol{U} \quad (12)$$

$\bar{\vec{\boldsymbol{v}}}$ 是語意空間 $\boldsymbol{S}$ 延伸出來的新郵件 d 的向量，不過式(12)的計算只是一個近似結果，因為 $\bar{\vec{\boldsymbol{v}}}$ 並非真正經過 SVD 運算產生。如果新郵件 d 與訓練語料的詞兩者代表的意思不同， SVD 延伸就會有問題。如果新郵件中的詞代表的意思維持相同，可以想像成郵件 d 是訓練語料的一部份，所以這樣做就是一個合理的方式。

最後我們採用餘弦測量來計算兩個向量之間

的夾角，夾角越小表示兩者相似度越高，反之夾角越大則相似度越低：

$$R(\vec{v},\vec{v}_k)=cos(\vec{v}S,\vec{v}_kS)=\frac{\vec{v}S^2\vec{v}_k^T}{\|\vec{v}S\|\|\vec{v}_kS\|} \quad (13)$$

其中當 $k=l$ 表示合法郵件，$k=j$ 表示垃圾郵件。與貝氏分類器相同的作法，觀察新郵件向量與兩類郵件語意指標的相似度，如果 $\vec{v}$ 與 $\vec{v}_l$ 較相似，會把新郵件標記成合法郵件。反之，如果 $\vec{v}$ 與 $\vec{v}_j$ 較相似，則會標記成垃圾郵件。圖 3 是潛藏語意分析示的示意圖。

4. 隱藏式馬可夫模型

隱藏馬可夫模型[6]假設合法與垃圾郵件分別為一個生成模型，其中包含了其類別特徵的機率分佈，在此方法中，我們同樣以詞為特徵，所以可將內部的生成模型機率分佈視為屬於該類郵件的語言模型分佈。假設新郵件 d 是一個長度為 m 的詞序列 $d=w_1w_2..w_i..w_m$，而我們可以利用合法與垃圾郵件語料及整個語料分別訓練出三個語言模型 C_l、C_j、G，再分別計算詞序列 W_1^m 在這三個語言模型產生的機率 $P(w_1^m|C_k)$。根據語言模型的不同，可分為三種型式，如式(14)至(16)表示：

型 I：單連郵件模型與單連通用模型

$$P(d|C_k)=\prod_{i=1}^{m}[\lambda_1\cdot P(w_i|C_k)+\lambda_2\cdot P(w_i|G)] \quad (14)$$

型 II:二連郵件模型與單連通用模型

$$P(d|C_k)=[\lambda_1\cdot P(w_1|C_k)+\lambda_2\cdot P(w_1|G)]\cdot \prod_{i=2}^{m}[\lambda_1\cdot P(w_i|C_k)+\lambda_2\cdot P(w_i|G)+\lambda_3\cdot P(w_i|w_{i-1},C_k)] \quad (15)$$

型 III：二連郵件模型與二連通用模型

$$P(d|C_k)=[\lambda_1\cdot P(w_1|C_k)+\lambda_2\cdot P(w_1|G)]\cdot \prod_{i=2}^{n}[\lambda_1\cdot P(w_i|C_k)+\lambda_2\cdot P(w_i|G)+\lambda_3\cdot P(w_i|w_{i-1},C_k)+\lambda_4\cdot P(w_i|w_{i-1},G)] \quad (16)$$

其中 $\lambda_z, z=1,2,3,4$ 是比重參數；$k=l$ 為合法郵件，$k=j$ 為垃圾郵件；G 表示通用模型，即將法與垃圾郵件語料合起來($\{C_l,C_j\}$)所訓練成的語言模型。與其他方法類似，計算出 $P(d|C_l)$ 與 $P(d|C_j)$ 後，比較其機率值，如果是 $P(d|C_l)$ 大，則認定新郵件 d 是合法的郵件，反之 $P(d|C_j)$ 較大則判斷為垃圾郵件。例如在型 I，我們可假設 $\sum_{z=1}^{2}\lambda_z=1$，$\lambda_z$ 可以透過 EM(Expectation-Maximization) 演算法 [7] 求得，例如型 I 的 λ_1 可透過式(17)更新：

$$\hat{\lambda}_1=\frac{\sum_{d\in C_k}\sum_{w\in d}n(w_t,d)\left(\frac{\lambda_1\cdot P(w_t|C_k)}{\lambda_1\cdot P(w_t|C_k)+\lambda_2\cdot P(w_t|G)}\right)}{\sum_{d'\in C_k}\sum_{w\in d'}n(w,d')} \quad (17)$$

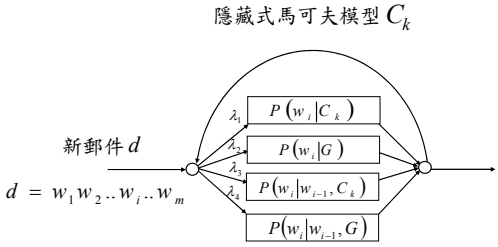

圖 4 隱藏式馬可夫模型

上述所使用的語言模型，是採用 SRI Language Modeling Toolkit(SRILM) [15]訓練，SRILM 是一套方便且容易使用的語言模型訓練工具圖 4 是隱藏馬可夫模型的示意圖。

5. 實驗與討論

5.1 文字語料

我們的實驗主要做在兩個語料上，一個是公開可獲得的電子郵件語料：Ling-Spam[2]，屬於英文的語料；另外一套是我們在台灣從新聞伺服器上收集到的垃圾郵件及幾位使用者的電子郵件當作合法郵件，屬於中文的語料。

Ling-Spam 的語料包括了 2,412 封合法的郵件及 481 封垃圾郵件，語料包括了兩種處理方法，詞項化 (Lemmatization) 與去除虛字 (Stopword Removal)，我們參考 Schneider[14]的設定，也採用僅做詞項化處理的語料，並以出現過的詞當作詞典，含有 59,829 詞，詳細如表 1 所示。中文語料部份，我們於郵件收集完畢後將具有相同標題的信僅保留一份，即合法郵件僅保留原始信件，排除回信；一模一樣的垃圾郵件只取一個樣本，再將檔案小於 50 位元組的信件去除。然後我們移除非中文字元，再根據含有 66,290 詞的詞典做斷詞，並排除外詞彙(Out-of-Vocabulary, OOV)詞。我們總共收集了 1,359 封合法郵件與 481 封垃圾郵件。

由於 Ling-Spam 英文語料已經平均分成十群，每一群約 241 封合法郵件與 48 封垃圾郵件，所以我們可以使用十次交互驗證(10-fold Cross Validation)，九群當作訓練語料，剩下一群當作測試資料，執行十次後取其平均值當作最後的結果；而第二個中文語料，我們根據收集時期將語料分成訓練語料與測試語料，如表 2、表 3 所示。

表 1 Ling-Spam 郵件語料資料表

類別	合法郵件	垃圾郵件
數量	2412	481
總詞數	1,531,209	439,040
平均長度	634.83	912.77
收錄時期	~July 17, 2000	

表 2 中文訓練郵件語料資料表

類別	合法郵件	垃圾郵件
數量	1188	430
總詞數	157,861	74,197
平均長度	132.88	172.55
收錄時期	~March 2005	~June 2005

表 3 中文測試郵件語料資料表

類別	合法郵件	垃圾郵件
數量	171	83
總詞數	11,844	5,174
平均長度	69.26	62.34
收錄時期	April 2005	July 2005

5.2 特徵選擇

我們可以透過交互資訊(Mutual Information)選出部份詞當新的特徵[18]；使用類別中含有較多資訊的部份特徵，可以減少過濾時的計算量，交互資訊計算方式如式(17)：

$$MI(C_k;W_t)=\sum_{C_k\in\{C_l,C_j\}}\sum_{f_t\in\{0,1\}}P(C_k,f_t)\log\frac{P(C_k,f_t)}{P(C_k)P(f_t)} \quad (17)$$

對於不同的貝氏機率模型，其交互資訊的計算方式也有所不同，以多變量伯努利事件模型來說，$P(C_k,f_t)$ 是訓練語料中某一類別 C_k 中包含詞 w_t 的郵件數除以所有類別郵件的總數；$P(C_k)$ 是某一類別 C_k 的郵件數除以所有類別的郵件總數；$P(f_t)$ 是所有類別中包含詞 w_t 的郵件數除以所有類別的郵件總數。

另一方面，對於多項式事件模型來說，$P(C_k,f_t)$ 是訓練語料中某一類別 C_k 中詞 w_t 的詞頻數數除以兩類別 C_l 、C_j 的總詞數；$P(c)$ 是某一類別 c 的總詞數除以所有類別的總詞數；$P(f_t)$ 是有類別中詞 w_t 的詞頻數除以有類別的總詞數[12]。

5.3 實驗結果與討論

我們首先以貝氏分類器為模型，實驗語料採用 Ling-Spam，對特徵數量做一個初步實驗，觀察特徵數量與正確率的關係。如圖 5 所示，對於多項式模型來說，當特徵量越多時，雖然非嚴格遞增，但是都有正確率提昇(即正相關)的趨勢，且在使用最多的特徵時有最高的正確率，所以接下來的實驗設定，初步我們都會採用全部的特徵去評估。

接著是潛藏語意分析模型部份，語料同樣是 Ling-Spam。Bellegarda[4]提出的作法是將訓練語料中所有合法郵件合併成一個大郵件，垃圾郵件也如此處理，所以做完 SVD 運算後，語意空間 S 只有兩維，即語意指標是二維向量。我們嘗試不將郵件合併，建立"詞-郵件矩陣"，在對此矩陣執行 SVD 分解，分別取 25、100、250 等維度，再分別針對

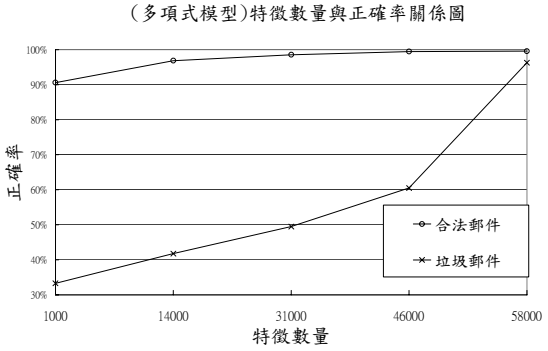

圖 5 特徵數量與正確率關係圖(使用 Ling-Spam 語料)

屬於合法與垃圾郵件的郵件向量取平均值，成為新的語意指標，但效果不如預期的好，因為每一封郵件長度雖長，但隱藏的語意不夠明確，如果合併在一起，語意反而變得較集中且清楚，結果如表 4 所示。整體正確率的計算方式是根據訓練語料中合法郵件與垃圾郵件比例乘上其正確率加總得到。

接著，我們比較本論文提出的馬可夫模型三種型式的正確率，可以發現型 III 的效果最好，因為比起型 I 與型 II，使用了更多資訊，如二連郵件模型與二連通用模型；而根據設定的參數，類別的權重比較通用模型高，表示類別模型比較能表示該類的資訊，且二連模型權重比單連模型高，表示二連模型給予的資訊更多，實驗結果如表 5 所示。

表 4 LSA 維度與正確率實驗結果(使用 Ling-Spam 語料)

	合法郵件正確率(%)	垃圾郵件正確率(%)	整體正確率(%)
d=2604, *r*=25	85.49	66.34	82.31
d=2604, *r*=100	85.66	68.42	82.79
d=2604, *r*=250	85.74	68.42	82.86
d=2604, *r*=500	85.78	68.42	82.89
d=2, *r*=2	94.36	91.07	93.81

d 表示訓練郵件總數，*r* 表示 SVD 的維度

表 5 隱藏式馬可夫模型各類型比較(使用 Ling-Spam 語料)

	合法郵件正確率(%)	垃圾郵件正確率(%)	整體正確率(%)
型 I	99.42	96.68	98.96
型 II	99.96	97.30	99.52
型 III	99.96	98.34	99.69

最後以每種模型最好的結果做比較，發現本論文所提出的隱藏式馬可夫模型(HMM)的正確率比貝式模型，也就是多變量伯努利事件模型(MVB)與多項式事件模型(MN)，或潛藏語意分析(LSA)效果要好，如表 6 所示。多變量伯努利事件模型雖然有考慮到沒有出現的詞的影響，但是光看出現與否是不夠的，因為如先前所提到的，合法郵件與垃圾郵

件可能擁有部分相同的特徵，且垃圾郵件訓練資料較少，影響較大。而根據式子(7)，多項式模型假設 $P(|d|)$ 與郵件類別無關，且進行過濾時是使用同一封郵件 d 對兩類計算機率，所以只與郵件 d 有關的階層都可忽略不計，所以式子會簡化成郵件 d 中有出現的詞的機率連乘，可以看成是隱藏式馬可夫模型型 I 中 $\lambda_1 = 1$、$\lambda_2 = 0$ 的一種，能夠調整的參數有限，所以效果稍差。對於潛藏語意分析來說，可能語料隱含的語意不夠明確，所以採用語意而非特徵詞的結果稍差了點。

表 6 各種模型正確率(使用 Ling-Spam 語料)

	合法郵件正確率(%)	垃圾郵件正確率(%)	整體正確率(%)
MVB	99.50	64.67	93.71
MN	99.42	97.73	99.14
LSA	94.36	91.07	93.81
HMM	99.96	98.34	99.69

我們在中文語料上的實驗也有相似的結果，如表 7 所示，多變量伯努利事件模型的效果最差，但是潛藏語意分析的效果提升，比多項式事件模型好，而隱藏式馬可夫模型仍然比其他的模型好一點。

表 7 各種模型正確率(使用中文語料)

	合法郵件正確率(%)	垃圾郵件正確率(%)	整體正確率(%)
MVB	99.42	49.40	86.12
MN	98.25	60.02	88.15
LSA	99.42	67.47	90.93
HMM	98.25	74.70	91.99

6. 結論與未來展望

現在的垃圾郵件變化演進相當地迅速，傳統的規則式過濾法已經不敷使用，透過數學模型的方式自動學習與過濾已成為必然的趨勢，本論文提出隱藏式馬可夫模型的方式且與已經實際被應用的方法，如貝氏分類器的多變量伯努利事件模型、多項式事件模型與潛藏語意分析等做一個比較，發現在英文或中文語料上的實驗，隱藏式馬可夫模型都能夠更有效地過濾郵件。

統計式的模型仍然有可以改進的地方，例如有些不肖的廣告商(spammer)會使用屬於被分類器認定為合法的特徵組成垃圾郵件，或是加入雜訊破壞應有的統計值，造成過濾器的誤判[17]，所以未來希望能夠針對特徵的選取，歸納出更多強健性的特徵，例如網址、圖片、甚至語音、影像等特性；或是針對分類過濾的方法作研究，例如隱藏式馬可夫模型的參數訓練從原來的最大相似度估算(Maximum Likelihood Estimation, MLE)方式改用最小化分類錯誤(Minimum Classification Error, MCE)方式訓練或是使用機率式潛藏語意分析(Probabilistic Latent Semantic Analysis, PLSA)模型、機率主題混合模型(Topical Mixture Model, TMM)[5]或最大熵值(Maximum Entropy, ME)模型[19]等。

除了網路上的垃圾郵件外，近年來犯罪集團透過電話或語音留言來騙取民眾財產的事件越來越多，這些語音對民眾來說不只是一種垃圾訊息，更是直接侵害到民眾的生活，未來也希望透過結合語音辨識與郵件過濾的技術，研究電話語音過濾的相關應用，進而解決這些問題。

參考文獻

[1] 交通部電信總局－濫發商業電子郵件管制條例草案相關議題：http://www.dgt.gov.tw/chinese/ncc/mail-requlation/ncc-SPAM-Q&A.shtml

[2] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, Georgios Sakkis, C. D. Spyropoulos, and P. Stamatopoulos, “Learning to filter spam e-mail: A comparison of a Naive Bayesian and a memory-based approach”, in H. Zaragoza, P. Gallinari, and M. Rajman, editors, Proc. Workshop on Machine Learning and Textual Information Access, 4th European Conference on Principles and Practice of Knowledge Discovery in Databases, pages 1–13, Lyon, France.

[3] J. R. Bellegarda, “Latent Semantic Mapping: Dimensionality Reduction via Globally Optimal Continuous Parameter Modeling”, in Natural Language and Speech Processing Colloquium, 2005.

[4] J. R. Bellegarda, D. Naik, and K.E.A. Silverman, “Automatic Junk E–Mail Filtering Based on Latent Content,” in Proc. 2003 IEEE Aut. Speech Recog. Understanding Workshop, St. Thomas, U.S. Virgin Islands,pp. 465–470, December 2003.

[5] B. Chen, “Exploring the Use of Latent Topical Information for Statistical Chinese Spoken Document Retrieval,” to appear in Pattern Recognition Letters, 2005.

[6] B. Chen, H.M. Wang, L.S. Lee, “A Discriminative HMM/N-Gram-Based Retrieval Approach for Mandarin Spoken Documents”, in ACM Transactions on Asian Language Information Processing, Vol. 3, No. 2, June 2004, pp. 128-145.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm”, Journal of the Royal Statistical Society, Series B, Volume 39, no. 1, pages 1-38, 1977.

[8] J. Goodman, D. Heckerman and R. Rounthwaite, ”Stopping SPAM”, in Scientific American April, 2005

[9] A. Gray, M. Haahr, “Personalised, Collaborative Spam Filtering”, in First Conference on Email and Anti-Spam, Mountain View, CA, USA, 2004

[10] D. D. Lewis, “Naive (Bayes) at forty: The independence assumption in information retrieval”, in Proc. 10th European Conference on Machine Learning, volume 1398 of Lecture Notes in Computer Science, pages 4–15, Heidelberg. Springer.

[11] C. D. Manning, H. Schutze, Foundations of Statistical Natural Language Processing, 1999, pp.197.

[12] A. McCallum and K. Nigam, “A comparison of event models for Naive Bayes text classification”, in Proc. AAAI-98 Workshop on Learning for Text Categorization, 1998, pages 41–48. AAAI Press.

[13] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, “A Bayesian Approach to Filtering Junk E-Mail”, in Learning for Text Categorization: Papers from the AAAI Workshop, 1998. pages 55–62, Madison Wisconsin. AAAI Press. Technical Report WS-98-05.

[14] K.-M. Schneider, "A Comparison of Event Models for Naive Bayes Anti-Spam E-Mail Filtering," in Pmc. 11th Conf.Euro. Chop. ACL, Budapest, Hungary, 2003.

[15] A. Stolcke, “SRILM -- An Extensible Language Modeling Toolkit” in Proc. Intl. Conf. on Spoken Language Processing, vol. 2, pp. 901-904, Denver, 2002.

[16] J.H. Wang, L.F. Chien. “Toward Automated E-mail Filtering – An Investigation of Commercial and Academic Approaches”, in TANET 2003.

[17] G. L. Wittel, S. F. Wu, “On Attacking Statistical Spam Filters”, in First Conference on Email and Anti-Spam, Mountain View, CA, USA, 2004.

[18] Y. Yang and J. O. Pedersen, “A Comparative Study on Feature Selection in Text Categorization,” in Proc. ICML 1997, pp. 412-420, Jul. 1997.

[19] L. Zhang and T.S Yao, “Filtering Junk Mail with A Maximum Entropy Model”, in n proceeding of 20th International Conference on Computer Processing of Oriental Languages, ShenYang, P.R.China.