基於協同架構之混合式Splog偵測系統

簡校呈[1]　　蘇建郡[2]
[1]南台科技大學資訊管理系 m9990205@stut.edu.tw
[2]南台科技大學資訊管理系 ccsu@mail.stut.edu.tw

摘要

隨著科技的進步與網路的普及，人們開始使用網路建立屬於自己的網站，因而帶起了部落格(Blog)的興起，卻也連帶產生所謂的「Splog」的出現，利用Blog簡易申請、快速發佈的特性，大量的複製其他Blog的文章，誤導使用者點擊，希望透過這些「Splog」藉以賺取廣告費用。Splog不只影響使用者在搜尋時的不便，甚至導致了一些資安的問題發生。目前在國外已經出現許多偵測Splog的方法，但是由於語言特性的關係，這些偵測方法並無法有效的防止Splog的增加與擴散，因此本研究提出一種新的方法與架構，透過協同合作的方式，結合使用者與網路的力量，進而減少電腦誤判的機率，並有效的提升偵測的準確度與偵測速度，希望藉由本研究能夠減少使用者點擊Splog的機率，因而減少Splog的產生。

關鍵字：部落格、垃圾部落格、垃圾部落格偵測

1.前言

1.1 研究背景

在網際網路的蓬勃發展之下，網際網路提供了快速獲取資訊的便利，這也使得使用者越來越傾向使用搜尋引擎進行搜尋、處理資訊、進行社交活動。根據圖1資料顯示，台灣地區的上網人口呈現每年逐漸增加的趨勢，在2011年三月份，統計上網的人數已達到約2555萬戶，網路普及率高達了8成，這也代表網際網路在一般民眾的生活中越來越普遍，對於企業而言，更是不容忽略、日益重要的市場，如此可觀的數據，代表著民眾的日常生活習慣，終究脫離不了網際網路的世界(資策會FIND)。

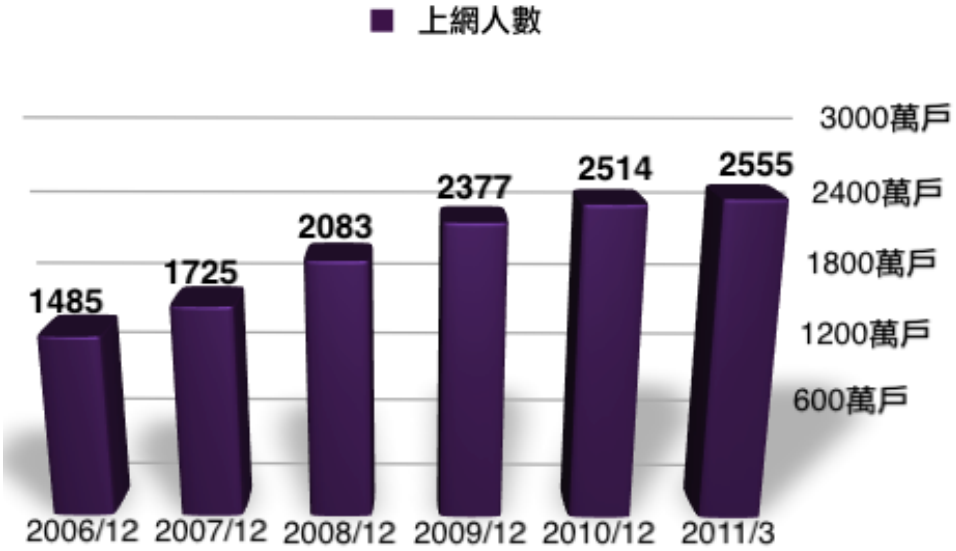

圖 1.1 上網人口統計

1.2 研究動機

因為Web2.0網站的興起，人們可以在網路上互相溝通討論，像是Blog、Wiki、討論區，但是這也讓廣告廠商發現了新的商業模式：網路廣告以及搜尋引擎最佳化(Search Engine Optimization,SEO)，廣告廠商為了使產品能夠在如此龐大的網路資料量下提高產品能見度，因而延伸出了黑帽(Black Hat)SEO，利用作弊的方式來快速提高網頁的能見度，例如：大量垃圾連結、隱藏網頁、關鍵字等等。根據研究(Jansen, B. J., Spink, A., & Saracevic, T, 2000)指出，當使用者進行網路搜尋時，有58%的使用者只瀏覽第一個搜尋頁面，而會進而瀏覽第三個搜尋頁面的使用者只剩下9%，也就是說，大多數的使用者通常在第一個搜尋頁面就獲取到所想要的資訊，或者變更關鍵字進行下一次的搜尋。所以一個網站的成功與否通常取決於網站在搜尋引擎中的排名，也因為如此，部落格(Blog)簡易申請、快速發佈、易用性等的特性，造成了這些利用不肖手段的業者作弊的溫床。Splog不同於一般的E-mail Spam，目的不一定是為了讓使用者點擊網站內某個超連結，而是為了提高Splog在搜尋引擎中的排名，用來吸引更多不知情的訪客進入，趁機賺取廣告費用，甚至在網頁當中植入木馬或偽裝成釣魚網站，騙取使用者帳號密碼。根據圖二的資料顯示，大多數民眾在網路上的行為主要是以搜尋資訊與瀏覽網頁為主，搜尋資訊上，在民國100年1月份高達了49.18%，而大量的Splog不僅造成了使用者在進行搜尋行為時的不便，降低搜尋時的效率，也很容易成為資安上的漏洞，因此要如何保護使用者，避免不知情的使用者進入Splog，即是本研究所想達成的目標。

2.文獻探討

2.1 何謂 Splog

Splog全名為Spam Blog，根據維基百科的定義，Splog其主要的目的在於創造流量借此賺取廣告金，這些Splog文章內容大多是竊取其他網站中的文章，且文章內容中的超連結並非是有關於超連結文字(Anchor texts)中所描述的內容，最主要目的是在增加其網頁的PageRank。因為大多數的搜尋引擎在決定搜尋結果排名時多以超連結參照的方式來給予積分，被連結到越多次的網頁則在搜尋結果排名也會也會有較佳的能見度，也因為如此，Splogger通常以Link Farm的方式，利用超連結將所有的Splog互相串連起來，借此提昇所有Splog的能見度，進而提昇廣告收入。在現今，越來越多的廣告廠商是根據廣告的點擊數來支付廣告金，因此Splogger透過各種方法在內容頁中欺騙或引誘使用者點擊廣告，甚至植入木馬或釣魚網站，這種行為不只降低使用者的搜尋效率以及網路資源的浪費，更造成一些資安上的問題。

目前已有集體式的Splog業者透過出售Link Farm來賺取金錢，利用自動化的方式將所有頁面串連在一起，一旦不知情的使用者進入某個Splog，Splog很容易透過超連結文字引誘使用者連結至另外的Splog，一再而再的引誘使用者點擊，因而提高能見度以及賺取廣告金。

2.2 常見的 Blog Spam 類型

根據(Chien-Yu Kuo, Hsien-Tsung Chang , 2010)(Saeed Abu-Nimeh, Thomas M.Chen, 2010)研究中，一般常見的Blog spam可分為以下幾種類型：

(1)、Comment spam：一般常見於可以隨意自由發言的網頁，如：Blog、Wikis或留言板，目的是為了讓其他使用者感到不適，貶低網站的聲譽，進而減少網站的流量。

(2)、Term spam：在網頁中大量插入與本文無關的詞語，使網頁在搜尋引擎的關鍵字中擁有搜尋結果。

(3)、Link spam：文章內容中包含著URL超連結至Spammer的網頁，進而增加網頁的能見度。

(4)、Spam page：假造的網頁，主要是用來誤導搜尋引擎，每一個假造的網頁都會有一個PageRank值，當這些假造的網頁建立超連結至目標網頁時，將會大幅提昇目標網頁的PageRank值，提昇在搜尋引擎中的排名。

(5)、Spam Blogs：假造的Blog，目的在於賺取廣告金與提昇搜尋引擎排名，吸引使用者進入網站。

(6)、Trackback Spam：Trackback為一種Blog機制，用來追蹤Blog文章，可以讓文章作者知道有哪些人引用了你的文章，Spammer利用這種機制來提高網頁的PageRank或Backlink至Splog。

2.3 Support Vector Machine

Support Vector Machines (SVM) (Boser B.E.,Guyon,I., & Vapnik, V. M., 1992)為一種建立於統計計算理論上的機器學習方法，被廣泛的運用在資料的分類(Classification)問題中，其主要的概念是在一個多維度的空間中找出一個超平面(Hyper-Plane)，希望透過此Hyper-Plane能夠將資料準確的切割成兩個不同的集合，以圖2.1二維資料為例：

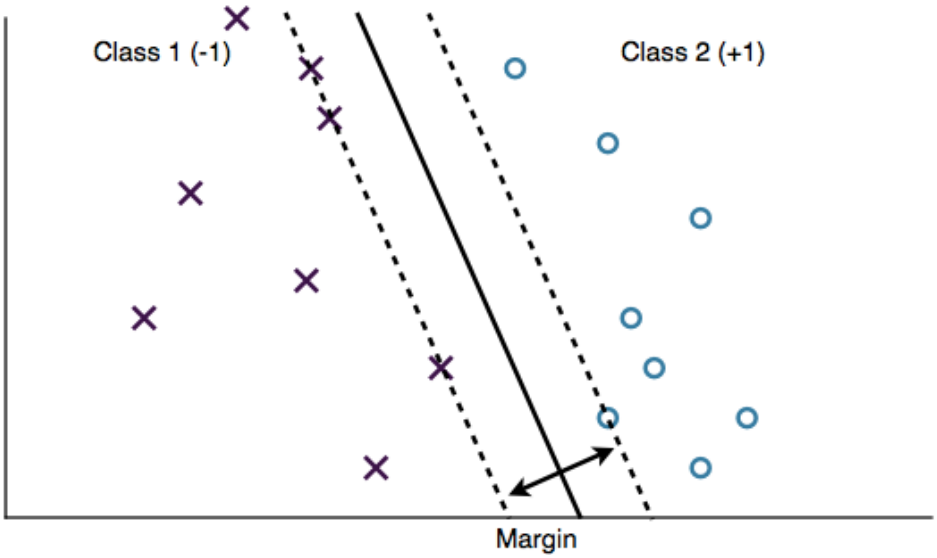

圖 2.1 SVM 分類範例

在二維的範例中，SVM 希望找尋一條直線能夠有效的將 Class 1 與 Class 2 分開，並且希望兩集合的邊界(Margin)越大越好，如此才能夠明確的分辨資料是屬於哪個集合，否則容易在計算中出現誤差。其中能夠將邊界最大化的超平面稱之為 Optimal separating hyper-plane(OSH)，而為了找出此超平面必須先找出 Support Hyper-plane，Support hyper-plane 為一與 OSH 平行的超平面，如上圖虛線部分。根據前述，我們希望找出 Optimal separating hyper-plane 即是找出相距最遠的 Support hyper-plane，使其有最大的邊界。換句話說，OSH 即是在兩個 Support hyper-plane 中間的平行線。

圖 2.2 不好的分界例子

然而大多數實際問題都是非線性可分割的狀況，為此，Boser 等學者提出了 Non-linear classifier，透過核心函數Φ，將原始資料從低維度空間轉換成高維度空間，讓 SVM 能夠順利的使用線性分割來進行分類(Boser B.E.,Guyon,I., & Vapnik, V. M., 1992)。

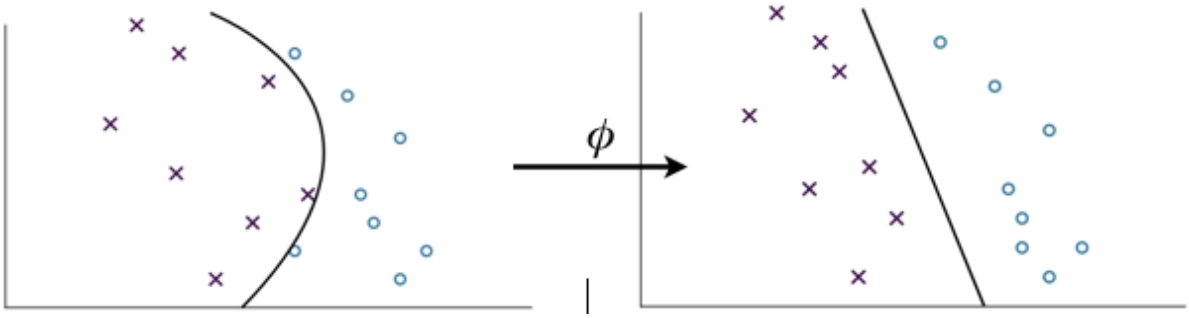

圖 2.3 非線性分割範例

目前常見的核心函數類型如下：

一、Linear： $K(x_i,x_j)=x_i^T x_j$ (1)

二、Polynomial： $K(x_i,x_j)=(\gamma x_i^T x_j+r)^d, \gamma>0$ (2)

三、Radius basis function： $K(x_i,x_j)=exp\left[-\frac{|x_i-x_j|^2}{\sigma^2}\right]$ (3)

四、Sigmoid： $K(x_i,x_j)=tanh(\gamma(x_i \times x_j)+c)$ (4)

SVM藉由不同的核心函數來映射到特徵空間中，不同的核心函數將會有不同的分類效果，所以選擇核心函數是SVM重要的一個環節，透過這些核心函數，造就了SVM的應用領域的廣泛，例如：資料探勘、影像識別(Image Recognition)、文件分類(Text Categorization)等，皆能獲得相當好的成效。

2.4 TF-IDF

TF-IDF(Term frequency-inverse document frequency)為一種基於統計方法用來衡量字詞重要程度的加權技術，常見於資訊擷取與文字探勘中，其中最常被應用於搜尋引擎上，其主要概念可以分為兩個部分，TF(Term frequency)與IDF(inverse document frequency)，TF為衡量一個詞彙在某特定文件中所占的重要性，詞彙的重要性隨著出現程度成正比，計算方式如(5)所示，其中$n_{i,i}$為某詞彙出現的次數，而分母$\Sigma k^n{}_{k,j}$則是代表單一文件中所有詞彙出現次數之總和。而IDF(inverse document Frequency)則是用來判斷詞彙在所有文件中所占的重要性，如(6)，|D|為文件的總數，分母部分則是代表包含詞彙t_i出現的文件總數，也就是說當某一詞彙在其他文件中出現次數越高，則重要性越低。從(5)與(6)進行乘法運算，可以得出TF-IDF權重(7)。

$$tf_{i,j}=\frac{n_{i,j}}{\Sigma_k n_{k,j}} \quad (5)$$

$$idf_i=log\frac{|D|}{|\{j:t_i\in d_j\}|} \quad (6)$$

$$tf\text{-}idf_{i,j}=tf_{i,j}\times idf_i \quad (7)$$

透過TF-IDF我們可以得出一個詞彙的重要程度，若是單一個字彙在單一文章中出現次數越高且在其他文章中出現次數越低則代表此字彙的重要性越高，反之，若一字彙在單一文章出現越高但在其他文章中出現次數也高則代表此字彙的重要性越低。舉例來說，在英文裡某些詞彙出現的次數高且幾乎每篇文章裡都會出現，如：”a” 、“the” 、”it” ，那麼雖然TF計算中得出相當高的值，也因為幾乎每篇文章中都會出

現，則IDF將會得出越小的值，經由TF-IDF計算之後，重要性將會降低，也就是說越普遍出現的常用詞彙，其重要性越低。

2.5 中文斷詞

詞為一能表達意義的最小單位，因為中文的語言特性，字可獨立或合併表意，只透過標點符號區隔語句，在中文句子中往往只有字的界線，詞與詞之間並無明顯分界，要了解句子之中的意義必須依靠前後文以及經驗來判斷(陳治宸，民96)。因此任何中文自然語言處理，如：文件檢索、機器翻譯等，都必須將句子中的詞進行斷詞的動作，才能進行下一步的動作。目前中文的斷詞方法普遍以統計式斷詞與法則式斷詞以及混合式斷詞為主(陳稼興、謝佳倫、許方誠，2000)。統計式斷詞法主要是藉由語料庫(corpus)的訓練來判別，透過機率統計來決定斷詞的位置。其優點為執行效率高，且不必先定義詞彙，然而在處理較長的詞彙或語料庫的訓練不足時，斷詞效率與正確性將會降低，導致表達句子意義可能會有偏差。而法則式斷詞法主要是搭配詞庫或字典，比對輸入的文件，根據一些規則逐步排除不可能的詞語組合，其主要缺點為斷詞的效果受詞庫的品質影響很大，對於一些專有名詞以及新生詞彙並無法正確的判別，使得整體斷詞正確性下降。至於混合式斷詞法則是上述兩種斷詞法的混合，先利用規則式斷詞法搜尋所有可能的斷詞組合，再透過統計式斷詞法依照機率排列所有可能的斷詞組合，再逐一過濾確認此斷詞組合是否合於文法。

目前正體中文斷詞系統大多使用詞庫斷詞法來進行斷詞，例如：中研院資科所的CKIP中文斷詞系統以及蔡志浩先生所開發的MMSEG，且正確率都高達95%以上(CKIP,MMSEG)。在本文中我們將使用MMSEG來做為本系統中的斷詞系統，其以最大匹配法(Maximum Matching Algorithm)為核心演算法，且支援兩種斷詞模式，分別為Simple Maximum Matching 與 Complex Maximum Matching，其中以Complex Maximum Matching的斷詞正確率最為顯著，正確率高達98%(MMESG)，斷詞正確率相當理想。

3. 系統設計

根據過去所提出的Splog相關檢測方法，幾乎沒有辦法透過單一的檢測方式來正確分辨所有的Splog。因此我們提出了一協同混合式的架構來檢測Splog，期望透過此混合式架構能夠優於單一檢測法，取得更好的檢測效果。

在本系統中我們透過Client Plug-in來與Server協同偵測Splog，希望透過Client Plug-in保護使用者進入Splog，藉此防止Splog的擴張，以及根據使用者的回報資訊來了解目前Splog的趨勢與提昇偵測準確度，具體而言，Client Side Plug-in的目標如下：

一、透過黑名單機制防止不知情的使用者進入Splog。

二、透過回報機制提供資訊至Server端進行判斷。

其Client Side Plug-in偵測流程如圖3.1表示：

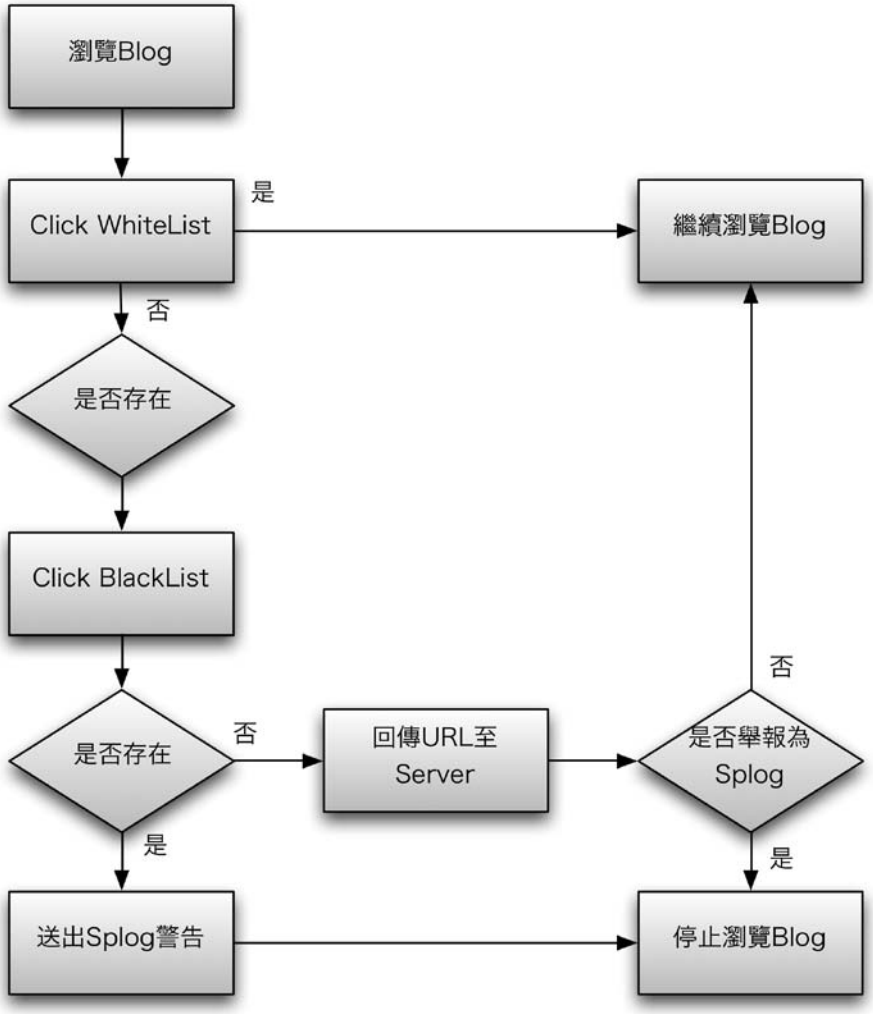

圖 3.1 Client Side Plug-in 偵測流程圖

在Client端，使用者可以透過Client Plug-in與Server系統協同合作偵測，一旦使用者進入Blog，程式將讀入白名單與黑名單判別此Blog的網域是否存在於名單中，若符合黑名單則送出Splog警告並停止瀏覽。若不存在於黑、白名單之中的Blog，Client Plug-in將回傳URL資訊交由Server端判斷，使用者也可以透過回報機制舉報Splog，幫助提昇Server端偵測的準確率。透過黑、白名單機制我們可以很快速的過濾一部分的Blog，並不需要再將資訊回傳給Server判斷，如此一來將可以降低Server端的負擔，也可以其他不知情的使用者再次進入。

至於Server端，將根據Client所回傳的資訊進行Splog偵測，本研究中，我們主要透過三種偵測方式在加以進行加權計算，若得出之權重值高於設定門檻則將其判斷為Splog並加入至黑名單之中。其架構如圖3.2所示：

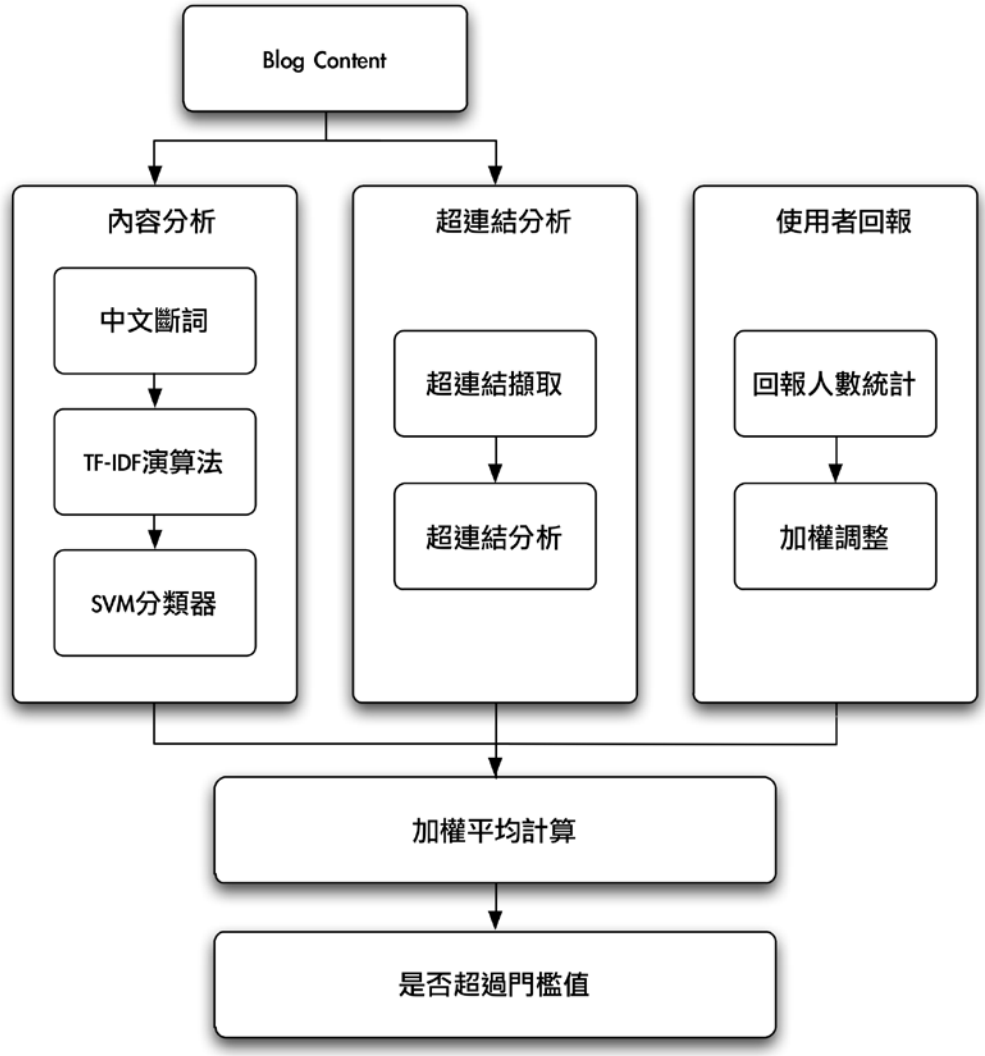

圖 3.2 Server Side 偵測架構圖

主要可以分為內容分析、超連結分析以及使用者回報三個部分。在內容分析中，因為中文語系的特性，我們必須先使用中文斷詞程式來將文章內容進行前置處理，在這裡我們使用蔡志浩先生所開發的MMSEG來做為本研究的中文斷詞程式，並將其斷詞模式設定為 Complex Maximum Matching，以求最佳的斷詞正確率，接著為了提高偵測的準確度，根據過去的相關研究(Akshay Java, Tim Finin, Tim Otates, Anupam Joshi, Pranam Kolari , 2006)(Pranam Kolari*,Tim Finin and Anupam Joshi , 2006)中，認為Splog的文章都是由系統產生，所以Splog的文章都有其相似的特徵，因此我們結合TF-IDF演算法篩選其Splog特徵並透過SVM進行分類，判別文章內容是否屬於Splog。

在超連接分析上，我們將擷取並分析其內文中的超連結以及Blog的URL，如果URL包含商業產品、性等相關詞，或者將其關鍵字重複排列，我們認為它有可能是一個Splog，以及透過Server分析資料庫中其他Blog中是否有相同連結出現或是已在黑名單中的連結，經由啟發式規則(Heuristic Rules)將特徵給予不同的分數，再依據這些規則給予所指定的分數進行加總，在本研究中我們將對文章所有超連結中符合規則的多寡給予總分，判別Blog中擁有的Splog特徵值的高低。

至於使用者回報，我們將統計使用者回報的次數並且針對回報的人數總數給予不同的權重值，例如：當回報次數為0時，其使用者回報機制權重值將調整為5%，而回

報人數在1~99之間時，將其權重值調整為10%。並且為了避免使用者惡意回報，我們將限制使用者回報的次數，透過Client Side Plug-in限制使用者短時間內大量回報，以及限制同一網站只能回報一次，以確保回報的品質。

接著我們將SVM分類結果、超連結分析結果以及使用者回報結果分別給予一個權重值，為了找出最佳的權重，我們進行多次的實驗，希望能夠最大限度的減少誤報率與提昇準確性。最後透過加權平均計算出最後的總分數，如果總分高於設定的門檻值時，我們將認為此Blog為Splog，並將其加入至黑名單。

4.實驗

我們使用了UMBC ebiquity所提供的Splog Blog dataset(Pranam Kolari, Akshay Java, Anupam Joshi, 2006)，其資料集中共收集了3000篇Blog文章，其中有700個Splog以及700個正常Blog，但是此資料集中收錄的資料為英文語系Blog，為了測試本偵測系統對於中文語系的Blog偵測效果，我們另外從各大Blog服務提供商例如：Yahoo blog、google Blogger、無名小站等，收集共632篇Blog文章，包含了110個Splog與130個正常Blog來進行實驗。

表4.1：資料集內容

	Spam Blog	正常Blog
英文語系Blog	700	700
中文語系Blog	110	130
總計	810	830

在內容分析中，我們藉由SVM與TF-IDF的結合來檢測其文章內容是否為Splog，首先透過TF-IDF演算法篩選出關鍵字並將其轉換為SVM特徵向量，接著為了訓練SVM分類器，我們採用linear kernel作為我們的SVM核心函數並從資料集中隨機挑選了820個Blog來進行訓練(Training)，其中60%為正常Blog而40%為Splog。在參數設定中，由於本研究主要以LIBSVM作為分類工具，我們可以透過其內建的grid.py程式自動化的反覆測試來找出最佳的參數-c與-g值，因而產生最佳的模型(Model)以達到最低的誤判率(false-positive rate)。

藉由grid.py程式我們找出最佳參數-c=28 -g=0.002869，接著為了驗證Model的準確性，我們剃除訓練用的資料，針對另外的820筆資料進行預測(predict)，其結果如下表所示：

表4.2：SVM分類器偵測結果

項目	數值
Total number of Blog	820
Number of hits	749
Number of misses	71
Accuracy	0.913
False positives	42
False negatives	29

在超連結分析中，我們結合了啟發式規則來偵測Splog，首先我們比對超連結中是否有存在於黑名單之中的網址，以及分析此超連結是否頻繁的出現在其他Blog之中，接著刪除超連結中www、com、info等無意義的字詞，我們將獲得在超連結中剩餘的關鍵字，透過分析其關鍵字是否符合在規則資料庫中出現的規則，並針對符合的規則給

予不同的分數，最後根據文章中所有的超連結所得的分數進行加總給予最後的總分，我們將根據分數的高低來判斷Blog是否為Splog，其偵測結果如下表：

表4.3：超連結分析偵測結果

項目	數值
Total number of Blog	1640
Number of hits	1426
Number of misses	214
Accuracy	0.869
False positives	71
False negatives	143

最後我們使用我們的系統架構來進行測試，假設所有的使用者回報行為都為有效且準確的情況下，進行有使用者回報情形以及無使用者回報情形的測試，其偵測結果如下表：

表4.4：無使用者回報之偵測結果

項目	數值
Total number of Blog	1640
Number of hits	1536
Number of misses	104
Accuracy	0.936
False positives	47
False negatives	57

表4.5：有使用者回報之偵測結果

項目	數值
Total number of Blog	1640
Number of hits	1544
Number of misses	96
Accuracy	0.941
False positives	44
False negatives	52

透過上述實驗結果顯示，我們僅透過SVM分類器可以達到約91%的偵測率，這代表了因為絕大多數的Splog文章都是透過自動產生而來，因而造成了其相似內容特徵。此外，超連結分析中我們觀察到，因為Splog文章中裡的超連結通常都連結至另外一個Splog或者是連結到某個商業網站，因此在本系統架構中我們可以利用中央處理的方式得知此超連結是否出現於多個網站之中，藉此分析其超連結結構並有效的降低其系統誤判率。在我們的偵測架構中，我們結合了內容分析、超連結分析以及使用者回報來提昇偵測準確率，在表四顯示了，透過我們提出的架構能夠將偵測準確率提昇至94%，並且根據使用者的回報降低其誤判數及漏報數，因而將準確率改善。

5.結論與未來工作

本研究提出一基於協同架構之混合式Splog偵測系統，利用Client端之瀏覽器Plug-in與其Server端協同合作偵測，並透過與其Server連線更新黑名單，以及提供資訊給Server進行偵測。具體而言，本研究達到目標如下：(1)當不知情的使用者進入Splog時，能夠即時提醒使用者，避免使用者進入危險的環境之中。(2)透過使用者提供資訊，快速地建立起龐大的Splog資料庫，防止Splog的擴張。(3)能夠利用中央處理的架構優勢，分析Blog超連結是否大量出現於其他Blog中。(4)透過混合式偵測架構能夠有效的提昇偵測準確率。

經由實驗表明，本研究所提出之混合架構擁有比單一檢測方法更好的偵測結果，在往後的研究之中，為了能夠因應Spammer日新月異的手法，希望能夠找出更能區分出Splog的特徵，並加入至我們的偵測架構之中，並改善偵測效率與黑名單更新速度，期望獲得更好的偵測效果，進而提昇網路使用者的搜尋效率與資訊安全。

6.參考文獻

1. 林宗勳 , Support Vector Machines 簡介 , http://www.cmlab.csie.ntu.edu.tw/~cyy/learning/tutorials/SVM2.pdf
2. 陳治宸 , 民96 , 植基於個人郵件之雙層垃圾郵件過濾方法 , 國立台灣科技大學資訊工程系碩士論文
3. 陳稼興、謝佳倫、許方誠 , 2000 , 以遺傳演算法為基礎的中文斷詞研究 , 資訊管理研究 第二卷 第二期 pp27-44
4. 葉生正、蘇民揚、張僩鈞、陳相儒, 2007 , 結合SVM與Naïve Bayes 演算法防堵垃圾郵件的研究 , Journal of Informatics & Electronics vol.2 no.1 pp.1-7
5. 資策會FIND , 上網人口 http://www.find.org.tw/find/home.aspx?page=many&id=290
6. 廖婉淑 , 2009，利用整合式的機器學習方式提高垃圾網站偵測率 , 國立台灣科技大學資訊工程系碩士論文
7. Wikipedia , TF-IDF http://zh.wikipedia.org/wiki/TF-IDF
8. Akshay Java, Tim Finin, Tim Otates, Anupam Joshi, Pranam Kolari , 2006 ,"Detecting Spam Blogs: A Machine Learning Approach". Proceedings of the National Conference on Artificial Intelligence 2, pp. 1351-1356
9. B. E., Boser, I. M. Guyon, and V. N. Vapnik, 1992 ,“A training algorithm for optimal margin classifiers,” in Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, Pennsylvania, United States, pp. 144-152.
10. Chien-Yu Kuo, Hsien-Tsung Chang , 2010 , A Framework of Hybrid Methods for Splog Detection
11. CKIP , http://ckip.iis.sinica.edu.tw/CKIP/wordsegment.htm
12. Jansen, B. J., Spink, A., & Saracevic, T, 2000 , Real life, real users, and real needs: a study and analysis of user queries on the web. Information Processing & Management, 36(2), 207 – 227.
13. LIBSVM , http://www.csie.ntu.edu.tw/~cjlin/libsvm/
14. MMSEG:A Word identification system for mandarin chinese text based on two variants of the maximum matching algorithm , http://technology.chtsai.org/mmseg/
15. Pranam Kolari*,Tim Finin and Anupam Joshi , 2006 , "SVMs for the Blogosphere : Blog Identification and Splog Detection".AAAI Spring Symposium - Technical Report SS-06-03, pp.92-99.
16. Saeed Abu-Nimeh, Thomas M.Chen, 2010 , Proliferation and Detection of Blog Spam, IEEE Security & Privacy vol. 8 no.5 pp.42-47.

17. Pranam Kolari, Akshay Java, Anupam Joshi , 2006 ,Splog Blog Dataset , http://ebiquity.umbc.edu/resource/html/id/212/Splog-Blog-Dataset
18. Wei Liu, Songbo Tan, Hongbo Xu, Lihong Wang , 2008 , Splog filtering based on Writing Consistency , IEEE/WIC/ACM International Conference on Web Intelligence and Intellgent Agent Technology.
19. Wikipedia , Spam blog http://en.wikipedia.org/wiki/Splog

A Hybrid Splog Detection Method Based on the Collaboration Architecture

Hsiao-Cheng Chien[1] Chien-Chun Su[2]
[1]Department of Information Management, Southern Taiwan University
m9990205@stut.edu.tw
[2]Department of Information Management, Southern Taiwan University
ccsu@mail.stut.edu.tw

Abstract

With the technology advances and the popularity of Internet, people began to use the Internet to create their own Web site, which started a blog rise, but also produce the "Splog", use blog with feature of the easy to application and rapid release, a large number of copy other articles Blog, misleading users to click, hope that through these "Splog" to earn advertising fees. Splog not only affect the user in the search for the inconvenience, and even led to some of the problems of information security. Currently there have been many in foreign Splog detection methods, but because of the relationship between language features, and these detection methods can't effectively prevent the increase and diffusion Splog. In this paper proposes a new method and architecture, through the collaborative approach, combining the power of user and network, thereby reducing computer the probability of false positives, and effectively improve the detection accuracy and detection speed, this paper hoping that can reduce the chances of users clicking by the Splog, thus reducing the production of Splog .

Keywords: Blog、Splog、Splog Detection、Spam Blog