以文件關係為基礎之文件分類方法

彭國彥
資訊管理學系
元智大學
s1006220@mail.yzu.edu.tw

陳煜仁
資訊管理學系
元智大學
s1006217@mail.yzu.edu.tw

林榆青
資訊管理學系
元智大學
s1006219@mail.yzu.edu.tw

吳娟
資訊管理學系
元智大學
s1019209@mail.yzu.edu.tw

邱南星
資訊管理學系
健行科技大學
nhchiu@uch.edu.tw

邱昭彰
資訊管理學系
元智大學
imchiu@saturn.yzu.edu.tw

摘要

隨著資訊科技的進步，電子資料的流通越來越快速且種類多樣化，為了讓使用者可快速且準確查詢到需要的文件，眾多學者運用機器學習 (Machine Learning) 的方式進行議題分類研究，取代傳統人工需要耗費大量時間成本的作業方式。本研究透過文字探勘 (Text Mining) 技術，結合社會網路分析 (Social Network Analysis) 中的 K-core、Degree 指標，提出一套自動化議題分類機制。研究以網路部落格「無名小站」中提及自殺相關字詞的文件為實驗對象，並且將其分為關注與不關注兩個類別，最後評估其分類模型的準確率與召回率。實驗結果顯示結合了社會網路分析，的確提升分類的準確性。

關鍵詞：文字探勘、社會網路分析、議題分類。

Abstract

The advancement in information technology, the electronic information flow be more quickly and diversify. In order to allow the user could inquiries to the need to file quickly and accurately, many scholars studied about machine learning topics classification method. Replace the traditional artificially operation mode need to spend a lot of time cost. This paper proposed a text mining techniques combined with social network analysis using the K-core Degree indicators, and raised the issue of an automated classification model. The data using the articles in weblog "Wretch" which include suicide-related words. The purpose is to classify all of the articles into suicidal or non-suicidal. Finally, the results proved the novel classification method can achieve better performance.

Keywords: Text Mining、Social Network Analysis、Topics classification

1. 緒論

資訊科技的演進快速，代表著資訊爆炸的時代來臨，造就不同形態網路資訊平台的林立，眾多網路使用者透過此平台進行知識分享與意見表達，針對各式各樣的疑難雜症提出解決方案，此外使用者亦藉此與人進行溝通互動，形成緊密的網路社群型態。而近年來網路平台上人們討論的內容，深入生活的各個面向，言論分析成為眾多領域研究者關注的對象。文件分類 (Text Classification) 的目的在於從大量文件中，透過機器學習的方式訓練分類模型，自動判別議題並將其分類。

在過去文字探勘 (Text Mining) 的研究流程中，從網路上擷取文件進行斷詞等前處理後，經常以 Term Frequency (TF)、Information Gain (IG) 等指標，計算字詞與議題間的關聯程度，再進行特徵選取以篩選出具代表性之特徵詞，並透過 Vector Space Model (VSM) 建立字詞對應文件的矩陣，建構文件的分類模型實驗，最後以評估分類的結果好壞。

但是在過去的分類模型建構流程中，尚未考慮到文件間的相關性以及其構成的社會網路結構，因此本研究的探討核心為社會網路分析 (Social Network Analysis) 中衡量節點間內聚緊密程度的 K-core，以及考量網路中各節點中心性的 Degree 指標，結合原有 TF-IDF 為基礎的矩陣，建構出新的分類模型。

在本研究的概念中，將每個文件視為網路中的單一節點 (Node)，在計算完字詞對應文件的 TF-IDF 矩陣後，使用餘弦相似度公式計算文件間的相似度，並設定其門檻值來定義文件間是否有連結，最後即可計算出文件的 K-core 與 Degree 指標。

為了評估本研究建構的分類模型，實驗以自殺相關字詞作為搜尋關鍵字，擷取網路部落格「無名小站」的文件，並透過人工標記將其區分為關注與不關注兩個類別，將文件經過分類模型進行分類處理，評估其實驗結果。

2. 文獻探討

2.1 議題分類之研究

近年來網路科技的快速發展，使得各式各樣的資訊平台如微型網誌、部落格、論壇等紛紛建立，人們透過這些平台彼此交換意見或知識分享，構成一個社群網路。但這些資料的數量龐大且主題繁多，使用者不易從中蒐集相關的討論議題。因此學者開始針對這些資料來源，進行議題分類之研究。

而議題分類的方法有很多種，如支援向量機 Support Vector Machine (SVM) (Zheng, 2009), Naive Bayes (NB) (Qu et al., 2006；Phuc et al., 2007；Phan et al., 2008；Vidhya & Aghila, 2010)。

2.2 社會網路分析

社會網路通常是指獨立個體間所形成的某種鏈結關係，而獨立個體可以是人、群體或組織，組成網路最基本的元素就是節點與連線（Node and Link），節點就代表著獨立個體，而連線則用來表示個體與個體或個體與群體之間的某種關係或是聯繫。衡量社會網絡的分析指標大致可分為三種：社會網絡規模(size)、社會網路密度(density)、網路中心性(centrality) [2] [9]。社會網絡規模是指特定社會網路中所有節點的數量，社會網路的規模是影響節點間關係的重要因素，因為節點彼此間關係的建立會受到資源多寡的限制，若是節點越多表示可用的資源可能也越多。社會網路密度是用來表示節點與節點間是否緊密或表示社會網路中節點之間的連結程度，通常密度越高代表節點與節點之間的關係越緊密。網路中心性則可用來衡量節點在網路中影響力的大小，而一般而言分析網路中心性指標可分為三種：程度中心性 (degree centrality)、親近中心性(closeness centrality)、中介中心性 (between centrality) [3]。以下針對這三種測量網路影響力的計算做相關的說明：

(1)程度中心性 (degree centrality)

一個網路的中心節點為網路中與其他節點之互動最為頻繁之節點，其中互動的頻繁程度定義為該節點的值，當節點的值愈高，則表示該節點愈可能是網路的中心。其公式如下：

$$C_D(n_i) = \frac{d(n_i)}{g-1} \qquad \text{公式 1}$$

其中，n_i 為欲計算之節點相鄰的節點數，$d(n_i)$ 為計算與其他節點連帶關係，g 為整個網路的節點數量。

(2)親近中心性 (closeness centrality)

主要是測量節點與其他節點的接近緊密程度，接近中心性值越高表示該節點影響其他節點的速度快也很強烈。其公式如下：

$$C_c(n_i) = \frac{(g-1)}{\sum_{j=1}^{g} d(n_i, n_j)} \qquad \text{公式 2}$$

其中，n_i 為欲計算之節點，$d(n_i, n_j)$ 為兩的節點間距離，g 為整個網路的節點數量。

(3)中介中心性 (between centrality)

衡量某一節點存在於其他任兩點路徑上的重要程度。當仲介中心性值越高時，表示該節點是位於溝通與橋梁的重要地位。其公式如下：

$$C_B(n_i) = \frac{\sum_{j<k} g_{jk}(n_i)/g_{jk}}{[(g-1)(g-2)]} \qquad \text{公式 3}$$

其中，$g_{jk}(n_i)/g_{jk}$ 為通過節點 i 連接節點 j 及 節點 k 最短路徑數除以連接節點 j 及節點 k 的最短路徑數，g 為整個網路的節點數量。

3. 研究方法

研究方法主要分為三大部分：第一部分為實驗資料的採集與前處理，包含資料文件之網頁原始碼從部落格擷取後，如何解析成所需的資訊並儲存於資料庫中。第二部分為第一階段文字前處理完後，進行特徵詞選取，選擇出具

代表性字詞，再利用向量空間模型把文件透過字詞維度轉化成向量空間，最後計算文件與文件間的相似程度，並利用社會網路分析的概念建立關連性指標。第三部分為評估加入了關連性指標之準確率。本研究提出一研究流程，如圖 1 所示：

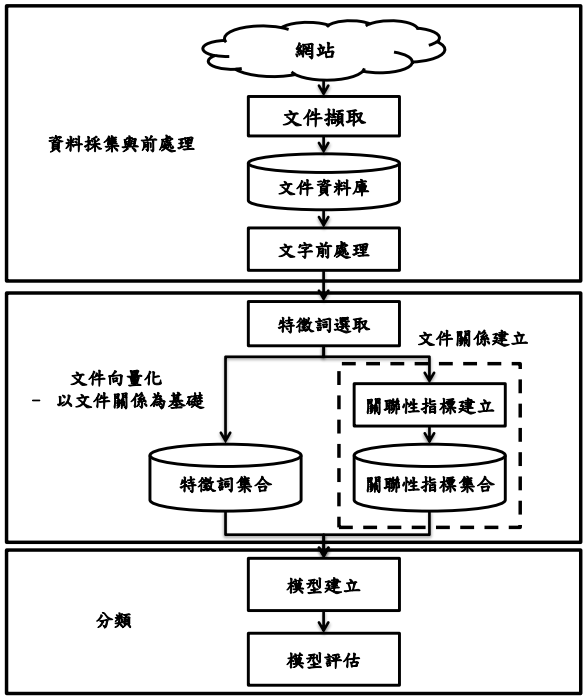

圖 1 研究流程圖

3.1 資料採集與前處理

3.1.1 網頁擷取

本研究以自殺為研究議題，利用數個與自殺相關的字詞，如：「想死」、「想自殺」、「自殺方式」等作為關鍵字，從網路部落格「無名小站」搜尋含有關鍵字的公開發表文件，並藉由網路爬蟲程式 (Web Crawler) 從中擷取有用的資訊，如：文件日期、文件標題、文件內容等資料，存入資料庫，並給予每筆資料一個索引值，利用此方法以達到快速搜尋原文以及後續研究方便性等目的。

3.1.2 人工標記

為探討自殺議題的關注與不關注，在擷取資料後，請領域專家的人進行人工標記的方式，標記文件內容是否有自殺意念。本研究透過專業人員間交叉標記驗證的方式來提升標記品質。

3.1.3 文件前置處理 (Data Preprocessing)

文件在經由人工標記後進行斷詞處理，利用中央研究院所開發的中文斷詞系統 (Chinese Knowledge and Information Processing, CKIP)，將所擷取的文件標題及內文進行斷詞與詞性標註 (POS Tagging)。由於在中文的文件中，名詞和動詞較能表達出文件的主旨，故本研究在針對文件作斷詞處理後，經由詞性過濾只保留標註為名詞和動詞的字詞。

3.2 文件向量化 – 以文件關係為基礎

3.2.1 特徵詞選取 (Feature Selection)

為降低與內容無關的特徵詞，研究中藉由不同的方法選擇出代表性的片語或詞彙成為特徵詞集合在不影響辨識力的前提下，刪除多餘的特徵維度，挑選出最佳的部分特徵，如此不僅能有效降低往後資料辨識所需花費的運算量及時間，亦可提升文件的辨識度 [1]。本研究利用 Information Gain (IG) 作為選詞標準。IG 主要的意義是用來定義為測試前的資訊，減去測試後的資訊，其概念就是計算有特徵詞跟沒有特徵詞的熵 (Entropy) 值差異，熵可視為資訊亂度，就是代表該資訊的不確定性的指標，亦代表該資訊的辨別度上不夠明確，因此當熵的數值越大，則代表資訊亂度越高，辨別度上越不明確。其數學公式 [4] 如下：

$$IG(X)=\frac{a}{N}\times\log\frac{a\times N}{(a+c)\times(a+b)}+\frac{b}{N}\times\log\frac{b\times N}{(b+d)\times(a+b)}+\frac{c}{N}\times\log\frac{c\times N}{(a+c)\times(c+b)}+\frac{d}{N}\times\log\frac{d\times N}{(b+d)\times(c+b)} \quad \text{公式 4}$$

其中，N 表所有的文件總數，a 表有分到正確類別內的文件數且有包含 X 特徵詞，b 表有分到正確類別內的文件數但沒有包含 X 特徵詞，c 表沒有分到正確類別內的文件數但有包含 X 特徵詞，d 表沒有分到正確類別內的文件數且沒有包含 X 特徵詞。

3.2.2 向量空間模型 (Vector Space Model)

在經過特徵詞選取後，利用向量空間模型把文件透過字詞維度轉化成向量空間。向量空間模型可以將文件集以文字為基礎的非結構化格式轉換為以數值表示的結構化格式，以二維矩陣的格式呈現。將每篇文件視為一個向量，並根據其在文件中的重要程度賦予一權重

值。在向量空間模式中，為刪除出現次數過多或過少的詞彙，一般在計算字彙的權重上，常以 Term Frequency & Inverse Document Frequency (TFIDF) 的方式來做計算，其公式如下：

$$tfidf_{i,j} = \frac{n_{i,j}}{\sum_i \sum_j n_{i,j}} \cdot \log \frac{|D|}{|\{d : d \ni t_j\}|} \quad \text{公式 5}$$

其中 d 為文件數，總數為 i，n 為詞的數量，總數為 j 個。$tfidf_{i,j}$ 指在文件 d_i 中 $n_{i,j}$ 的詞頻，$n_{i,j}$ 為該詞在文件 d_i 中的出現次數，$|D|$ 為語料庫中的文件總數，$\{d : d \ni t_i\}$ 指包含詞語 tj 的文件數目，即 $n_{i,j} \neq 0$ 的文件數目。

TF-IDF 主要用於過濾常用字詞，保留重要的詞語。若該字詞在一文件中出現的頻率高，則其值會增加，但隨著它在語料庫中出現頻率增加，則其值會減少。

3.2.3 文件關係與關連性指標建立

在自殺的議題分類中，文件間的用詞頻率相近，表示彼此間的關係密切，進而找出需要被關注的文件，並且加入了社會網路分析的概念，根據文件間的用詞特性，計算其相似度。計算公式如下：

$$\cos(D_i, D_j) = \frac{\sum_{p=1}^{t} tfidf_{pi} \cdot tfidf_{pj}}{\sqrt{\left(\sum_{p=1}^{t} tfidf_{pi}^{2}\right)\left(\sum_{p=1}^{t} tfidf_{pj}^{2}\right)}} \quad \text{公式 6}$$

其中 D_i 為第 i 篇文件，D_j 為第 j 篇文件。$tfidf_{p,i}$ 指第 p 個字詞在第 i 篇文件中的權重，$tfidf_{p,j}$ 指第 p 個字詞在第 j 篇文件中的權重。

將計算完相似度的文件，進行關連性指標值之計算，包含 K-core、Degree 等兩個指標，其中，加入了相似度門檻值，定義文件間的連結，接著進行關連性指標計算，建立關連性指標。故本研究以社會網路分析概念為基礎，考量文件間的用詞特性，利用餘弦相似度 (cosine similarity) 計算其相似度，並利用社會網路指標建構關連性指標後，經由 Weka 資料探勘工具中三種不同的演算方法進行自殺關注與不關注的預測模型之建立。

3.3 分類

本研究利用 C4.5、貝式定理 (Naïve Bayes, NB) 以及支援向量機 (Support Vector Machine, SVM) 進行分類。

(1) C4.5

C4.5 是由 Ross Quinlan 在 1993 年所發展的一種分類決策樹，以樹狀資料結構其為基礎做分析，產生易被解讀與運用的決策法則。且利用監督式的學習法，從訓練範例集合中建構模型。

(2) Naïve Bayes

以貝氏定理 (Bayesian Theorem) 為基礎計算出該類別資料與各類別之間的機率，以機率方式來推薦分類結果。而貝氏分類器是採用監督式學習方式，透過訓練樣本的訓練學習，以處理未來欲分類的資料。

(3) 支援向量機 (Support Vector Machine, SVM)

支援向量機為一種以統計理論為基礎所發展出來的方法，乃根據統計學中結構風險最小化 [10]，其目的在於避免過度訓練資料而導致研究結果正確率下降。主要方法是將分布在特定空間中的資料，找出一個超平面 (Hyper plane)，使兩個不同的集合分開，達到資料分類的目的。超平面意義即為高維度中的平面，為一個多項式或是三角函數所組成。

4. 實驗結果

4.1 研究資料來源

本研究選定「無名小站」作為此資料的蒐集來源，其使用之關鍵字乃藉由網路上搜尋引擎中較熱門、且與自殺相關的「關鍵字」，如表 1 所示。

表 1 網路搜尋關鍵字

No.	關鍵字	No.	關鍵字
1	想死	13	自殺的方法
2	我想死	14	如何燒炭自殺
3	好想死	15	自殺遊戲
4	完全自殺手冊	16	燒炭自殺方法

5	自殺手冊	17	如何自殺
6	自殺俱樂部	18	最不痛苦自殺方式
7	燒炭自殺	19	青少年自殺
8	安眠藥自殺	20	循環自殺
9	自殺防治	21	上吊自殺
10	自殺防治中心	22	自殺方式
11	自殺方法	23	想自殺
12	不痛苦自殺的方法有幾種		

經由表 1 關鍵字，逐步搜尋並擷取「無名小站」部落格上的文件，日期範圍為西元 2007 年 1 月 1 日至 2009 年 12 月 31 日，所蒐集之文件篇數共有 1322 筆。透過人工方式先以有無自殺意念為依據，將文件分成兩類，如表 2 所示。

表 2 自殺類別

類別	筆數	總筆數
自殺關注	275	1322
自殺不關注	1047	

利用 CKIP 斷詞的範例如表 3，分別顯示在斷詞前的原始文件及斷詞後具詞性標籤之範例說明。斷詞後會做詞性過濾，將詞性為動詞和名詞以外的詞剔除，最後剩下 17911 個字詞。

表 3 CKIP 斷詞範例

CKIP 中文斷詞前
我再呆在這裡我一定會發瘋
CKIP 中文斷詞後
我 (Nh) 再 (D) 呆 (VH) 在 (P) 這裡 (NCD) 我 (Nh) 一定 (D) 會 (D) 發瘋 (VH)

將過濾完的字詞利用 IG，找出詞與類別間的特徵，將類別分為「自殺關注」與「自殺不關注」兩類，每個字詞在兩類中各有所屬的 IG 值，若該字詞在某一類的 IG 值越高，表其屬於該類別，如：該字詞在有自殺關注類別中值高，表其屬於需要自殺關注之類別。本研究透過選擇 IG 值為前 10% 高的字詞，剩 111 個字詞。

4.2 評估結果

將計算完的餘弦相似度，利用相似度門檻值建立文件間的連結，其相似度門檻值設定為 0.5，其利用社會網路指標建立的關聯性指標，與自殺關注與自殺不關注兩類中 IG 值較高的前 111 筆字詞，建構成向量空間模型，並利用 Weka 所提供的三種分類方法以自殺關不關注兩類做正確率的評估，實驗中以單純只有字詞做分類與加入了關連性指標做比較，並以 10-ford 測試交叉驗證分類結果，由表 4 可看出加入了關連性指標，三種分類結果均達水準以上，表示實驗所用之計算文件間相似度，建構關連性指標，在辨識自殺文件是否需要被關注是有明確效果的。

表 4 10-ford 測試交叉驗證分類結果之準確率

	字詞	字詞+關連性指標
SVM	82.55%	89.03%
NB	66.07%	88.50%
C4.5	80.82%	93.57%

為了評估加入了關連性指標確實在分類能力上有明確的效果，因此，使用 F-Measure 來評估分類能力上的成效 (Classification effectiveness)，結果於表 6、表 7 所示。本研究採用 F-Measure 的評估方式來針對文件在議題分類之後的成效做評估。其中，F-Measure 包含了召回率 (Recall) 及精確度 (Precision)，表 5 為準確率與召回率之事件對照表。

表 5 準確率與召回率之事件對照表

		系統預測文件	
		議題 C_i	議題 $\overline{C_i}$
實際文件	議題 C_i	True Positive (TP)	False Negative (FN)
	議題 $\overline{C_i}$	False Positive (FP)	True Negative (TN)

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{公式 7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad \text{公式 8}$$

$$\text{F-Measure} = \frac{2RP}{R + P} \quad \text{公式 9}$$

TP 表實際文件隸屬議題 C_i 且預測文件為

議題 C_i 之筆數，FN 表實際文件隸屬議題 C_i 且預測文件為議題 $\overline{C_i}$ 之筆數，FP 表實際文件隸屬議題 $\overline{C_i}$ 且預測文件為議題 C_i 之筆數，TN 表實際文件隸屬議題 $\overline{C_i}$ 且預測文件為議題 $\overline{C_i}$ 之筆數。

表 6 原分類模型實驗結果

	字詞		
	Precision	Recall	F-Measure
SVM	74.64%	66.49%	70.33%
NB	64.30%	71.98%	67.92%
C4.5	70.31%	61.66%	65.70%

表 7 本研究分類模型實驗結果

	字詞+關連性指標		
	Precision	Recall	F-Measure
SVM	88.54%	76.44%	82.04%
NB	82.14%	85.74%	83.90%
C4.5	94.19%	86.02%	89.92%

5. 結論

本研究旨在提出一個議題分類模型的新方法，核心概念在於利用社會網絡分析建構關連性指標，並與自殺議題文件中的特徵詞結合做分類，先把特徵詞選取完的字詞轉化成向量空間，接著計算文件與文件間的相似程度，將文件視為整個網路中的一個節點，利用社會網絡分析的概念，分析出 K-core 與 Degree 兩個關連性指標，將此加入過去原有以詞頻為基礎的分類模型，因此新的分類模型考量到文件間的關聯性，多了兩個維度。研究結果顯示加入了關連性指標對於分類效果有明顯的提升，在未來在議題分類上，能夠準確的達到分類效果，如能建立自動分類系統，便能達到各種議題分類的目的。

參考文獻

[1] Changqiu, S., Xiaolong, W. and Jun, X., " Study on Feature Selection in Finance Text Categorization," ***Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, Art***, No. 5346030, PP. 5077-5082, 2009.

[2] Faust, K., "Centrality in affiliation networks," ***Social Networks***, Vol. 19, pp. 157-191, 1997.

[3] Freeman, L.C., "Centrality in Networks Conceptual Clarification," ***Social Networks***, Vol. 1, pp. 215-239, 1979.

[4] Lan, M., Tan, C. L., Su, J. and Lu, Y., "Supervised and Traditional Term Weighting Methods for Automatic Text Categorization," ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 31, Issue 4, pp. 721-735, 2009.

[5] Phan, X. H., Nguyen, L. M. and Horiguchi, S., "Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections," ***Proceeding of the 17th International Conference on World Wide Web 2008***, pp. 91-99, 2008.

[6] Phuc, D. and Phung, N.T.K., "Using Naive Bayes Model and Natural Language Processing for Classifying Messages on Online Forum," ***2007 IEEE International Conference on Research, Innovation and Vision for the Future, RIVF 2007*** , No. 4223081, pp. 247-252, 2007.

[7] Qu, H., La Pietra, A. and Poon, S., "Automated Blog Classification: Challenges and Pitfalls," ***AAAI Spring Symposium - Technical Report SS-06-03***, pp. 184-186, 2006.

[8] Quinlan, J.R., ***C4.5: Programs for machine learning, San Mateo***, CA, 1993.

[9] Scott, J., ***Social Network Analysis：A Handbook***, London, Sage, 2000.

[10] Vapnik, V.N., ***The Nature of Statistical Learning Theory***, Springer-Verlag New York, Inc., New York, 1995.

[11] Vidhya, K.A. and Aghila, G.,"Hybrid Text Mining Model for Document Classification," ***2010 The 2nd International Conference on Computer and Automation Engineering, ICCAE 2010 1, Art.*** No. 5451965, pp. 210-214, 2010.

[12] Zheng, W., "A SVM Text Classification Approach Based on Binary Tree," ***IFCSTA 2009 Proceedings - 2009 International Forum on Computer Science-Technology and Applications 3, Art.*** No. 5384927, pp. 455-458, 2009.