方澤翰、陳建錦（2017），『基於搜尋日誌及排序學習之新式台灣景氣狀態監測系統』，*中華民國資訊管理學報*，第二十四卷，第四期，頁435-454。

基於搜尋日誌及排序學習之新式台灣景氣狀態監測系統

方澤翰*
國立台灣大學資訊管理學系

陳建錦
國立台灣大學資訊管理學系

摘要

景氣狀態監測對於政府及企業是個重要的議題，多數研究使用經濟指標監測景氣狀態。由於經濟指標需由不同政府部門協力完成，往往需要經過漫長的處理時間，導致景氣狀態發佈的延遲，而發佈的延遲將會增加政府及企業決策的不確定性。為了克服這個問題，本研究基於搜尋日誌建構了新式的景氣狀態預測模型，此模型使用排序學習演算法，從搜尋引擎回傳的小量高排名文件中選取最能夠代表景氣狀態的詞彙，接著取得這些詞彙在網路上的搜尋日誌，結合先進的機器學習模型來預測景氣狀態的變化。由於搜尋引擎提供的文件及搜尋頻率具有高度的即時性與可用性，因此基於搜尋詞彙所制定的經濟指標可以有效降低景氣狀態發佈的延遲，進而降低決策上的不確定性。實驗結果顯示我們所提出的架構能夠準確的預測景氣狀態，且基於排序學習演算法所建構的模型其準確率也優於使用傳統特徵選取方法所建構的模型。

關鍵詞：景氣狀態監測、搜尋日誌、排序學習、特徵選取

* 本文通訊作者。電子郵件信箱：d03725003@ntu.edu.tw
2016/07/31 投稿；2017/01/22 修訂；2017/06/05 接受

Fang, Z.H. and Chen, C.C. (2017), 'A novel Taiwan prosperity surveillance system based on search log and learning to rank', *Journal of Information Management*, Vol. 24, No. 4, pp. 435-454.

A Novel Taiwan Prosperity Surveillance System Based on Search Log and Learning to Rank

Ze-Han Fang*
Department of Information Management, National Taiwan University

Chien-Chin Chen
Department of Information Management, National Taiwan University

Abstract

Purpose－Prosperity surveillance is an important issue for countries and organizations. Generally, the surveillance indicators are comprised of multiple economic variables which are compiled by different government departments. Compiling these variables involves a great deal of data processing, which delays the surveillance of prosperity.

Design/methodology/approach－In this paper, we propose a novel prosperity surveillance system that utilizes the search logs from search engine. The system employs learning to rank algorithm to identify discriminative terms that are representative of prosperity. Representative terms and their query frequencies are then applied to a state-of-the-art data mining model to enhance the effectiveness of prosperity surveillance.

Findings－The experimental results show that our prosperity surveillance system performs well and our feature selection method based on learning to rank outperforms other popular feature selection methods.

Research limitations/implications－This study focused only on using search log information, in our future work, we plan to investigate more information sources (e.g.,

* Corresponding author. Email: d03725003@ntu.edu.tw
2016/07/31 received; 2017/01/22 revised; 2017/06/05 accepted

news posting, internet forum) to enhance the proposed feature selection method.

Practical implications－In this paper, we have proposed an effective framework for predicting the status of prosperity in Taiwan, the proposed method can provide effective support for government officials and authorities in order to help them respond to fast-changing events and topics, and make appropriate decisions.

Originality/value－This study is, to the best of our knowledge, the first attempt to apply search log and learning to rank to predict the status of prosperity in Taiwan.

Keywords: prosperity surveillance, search log, learning to rank, feature selection

壹、緒論

景氣狀態長久以來都是各國政府及企業所關心的議題，透過監測景氣狀態，可以分析一個國家目前的經濟狀況及未來經濟發展的趨勢，進而制定適當的政策。例如：當景氣狀態步入衰退時，政府會採取寬鬆的財政政策及貨幣政策來刺激景氣復甦，以避免景氣持續低迷。Burns 與 Mitchell（1946）對景氣狀態下了如此定義：景氣狀態是國家總體經濟的波動，一個景氣循環是由許多經濟活動發生擴張，衰退、收縮、復甦，週而復始循環的現象。基於這個定義，過去多數研究（Jun & Joo 1993; Andersson et al. 2006; Vosen & Schmidt 2011）使用多項經濟指標預測景氣狀態；然而，許多經濟指標由政府所編制，編制這些經濟指標需由不同的政府部門協力完成，往往需要經過漫長的資料蒐集及處理時間，因此造成景氣狀態發佈的延遲。例如，Chen 與 Tsai（2012）指出在台灣由國家發展委員會所編制的景氣對策信號通常需要 1-2 個月的處理時間，意即當前的景氣狀態需要等到 1-2 個月後才會得知。由於景氣狀態可能在短時間內劇烈的變化（Jun & Joo 1993; Chauvet & Piger 2008; Kannan & Ramaraj 2010），發佈的延遲將會增加政府及企業決策上的不確定性。

隨著資訊科技快速的發展，搜尋引擎已成為群眾在網路上蒐集資訊的重要工具，使用者藉由搜尋引擎所獲得的資訊也影響使用者後續決策的過程。例如：Eysenbach（2002）觀察發現群眾會透過搜尋引擎查詢與傳染病相關的詞彙來了解傳染病散播的情形及獲得預防傳染病對策的資訊；Vosen 與 Schmidt（2011）觀察群眾會依賴搜尋引擎所獲得的資訊來決定是否購買某項商品，這些發現也支持了 Baeza-Yates 與 Tiberi（2011）的論點：如果群眾在搜尋引擎上的行為能夠反映使用者決策的過程，那麼存在於搜尋引擎中的搜尋日誌將會是研究社會活動的一項重要參考指標。Hamilton 與 Perez-Quiros（1996）指出人們日常中的經濟活動如消費商品、購買服務及儲蓄等都會受到景氣狀態的影響。例如：當景氣好時，人們的購買力將會提升，也更有意願購買銀行的理財服務；反之當景氣衰退時，人們會減少不必要的開銷，金融商品的買氣也會下降。基於以上討論內容，我們認為景氣狀態也會影響群眾在網路上的搜尋行為，因此可以利用群眾搜尋景氣狀態的詞彙，結合該詞彙在搜尋引擎中的搜尋日誌，來監測台灣景氣狀態的變化。

在過去，提供搜尋引擎服務的公司並不會公開搜尋日誌紀錄，因此群眾無法取得搜尋日誌中紀錄的搜尋頻率。2006 年，Google 推出 Google Trend[1] 服務，此服務讓一般的使用者也可以取得搜尋日誌的紀錄。例如：圖 1 為台灣民眾在 2005 年到 2010 年搜尋基金這個詞彙的搜尋頻率，Chen 與 Tsai（2012）指出圖中的高

1 https://www.google.com.tw/trends/

峰表示在該期間使用者大量使用基金詞彙搜尋投資相關的資訊。Hameed、Kang 與 Viswanathan（2010）指出當景氣情況惡化時，人們傾向於不做過多投資的行為，該作者的論述也與圖 1 所顯示情況相符合，例如：由圖 1 可以觀察到，2007 年底時，基金詞彙搜尋頻率大幅下降，這是因為 2007 年 9 月美國爆發雷曼兄弟債務違約風暴，台灣景氣於此時受到了重創，因此這段期間人們的投資意願下降，連帶搜尋投資資訊的行為也跟著減少，故我們可以驗證景氣狀態確實會影響到群眾在網路上的搜尋行為。

圖 1：2005-2010 基金詞彙搜尋頻率

基於以上觀察，在本研究中我們提出一個新式的景氣狀態監測系統。在此系統中，我們首先使用搜尋引擎蒐集群眾所關注的景氣狀態資訊，而衡量群眾關注與否的準則乃是基於搜尋引擎回傳文件的排名所決定，接著我們使用排序學習演算法，挑選出使得文件排在高排名位置的詞彙，並下載所選詞彙的搜尋日誌，最後結合先進的機器學習模型來預測台灣各個時期景氣的狀態。

貳、文獻回顧

本章分為兩個主題進行文獻回顧。第一個主題探討使用搜尋日誌進行狀態監測的相關文獻。由於本研究架構的核心屬於一種特徵選取方法，第二個主題回顧

傳統被廣泛使用的特徵選取技術，並作為本研究後續實驗比較的基準。

一、使用搜尋日誌進行狀態監測之研究

搜尋日誌已被廣泛使用於許多領域當中，例如傳染病狀態監測領域以及經濟狀態監測領域。在傳染病狀態測領域中，Eysenbach 是第一位觀察到搜尋日誌可以用來監測傳染病狀態的學者（Eysenbach 2002）。他觀察到傳染病相關詞彙被搜尋的頻率與傳染病感染人數呈現高度正相關，因此推測傳染病詞彙搜尋日誌將會是監測傳染病重要且有效的指標。隨後他使用 Google 搜尋日誌及傳染病爆發人數驗證了搜尋日誌確實能夠準確預測傳染病在各時期的散播狀態。Ginsberg 等（2009）（檢驗了 Google 資料庫中 5 千萬個流感相關詞彙，他們從中篩選出 45 個詞彙建構出一個線性迴歸模型，並準確預測了美國疾病管制局發布的類流感人數。Fang 等（2010）透過領域專家制定多個與登革熱相關的詞彙，將預測登革熱傳染狀態視為一個分類問題，並檢驗多種生成式及判別式機器學習模型的成效，最後驗證生成式機器學習模型（e.g.,簡單貝式分類器）能夠準確預測登革熱傳染狀態的趨勢。

在經濟狀態監測領域中，也有許多使用搜尋日誌進行研究的文獻。例如：Askitas 與 Zimmermann（2009）使用 Google 搜尋日誌預測德國失業率。他們首先透過領域專家選擇與失業率最相關的四個類別詞彙，接著將每一個詞彙的搜尋頻率視為一個時間序列，建構誤差修正模型。實驗結果證實他們提出的誤差修正模型能夠準確地預測德國的失業率。Vosen 與 Schmidt（2011）觀察消費者購買商品之前習慣上網搜尋商品的資訊，因此使用搜尋日誌預測美國消費者信心指數。實驗結果顯示，使用搜尋日誌的預測模型顯著優於使用傳統經濟指標的模型（e.g.,密西根大學消費者信心指數）。Choi 與 Varian（2012）使用搜尋日誌預測零售銷售量、汽車銷售量、房地產銷售量以及旅遊套票銷售量。他們的實驗證實了使用搜尋日誌的成效優於沒有使用搜尋日誌的成效。Chen 與 Tsai（2012）將台灣景氣預測視為一個分類的問題，並使用爬文技術大量的擷取了網路上的 11228 篇的文件，萃取了 5 萬 8 千個詞彙。接著，他們擷取此 5 萬 8 千個詞彙的搜尋日誌，並結合簡單貝式演算法進行景氣預測模型的訓練以及預測，他們的實驗證實了使用搜尋日誌能夠有效的預測台灣的景氣預測狀態。Li 等（2014）發展基於本體論的架構預測美國失業率。作者首先諮詢勞工經濟領域專家建立勞工經濟知識本體，接著透過過濾式特徵選取方法從勞工知識本體中篩選最能代表失業率的詞彙，最後使用這些詞彙的搜尋日誌結合支持向量迴歸模型（Support Vector Regression; SVR）預測失業率，實驗結果顯示他們提出的架構能夠得到更加準確的預測美國失業率。

本文與 Chen 與 Tsai（2012）皆屬景氣預測之研究，並且將景氣預測視為一個分類的問題。然而，Chen 與 Tsai（2012）在實驗設計上偏向於窮舉法，並未針對使用者的使用行為作進一步的觀察;本文則受到 Dou 等（2010）的啟發，考量使用者使用搜尋引擎時的心理以及行為，僅透過選擇搜尋引擎上小量高排序的 3 篇文件，萃取 565 個詞彙並結合排序學習演算法，即可達到不亞於 Chen 與 Tsai（2012）大量文件集訓練及測試的預測準確度。這顯示了小量高排名文件集即可代表群眾所關注的景氣議題。此外，由於爬文以及處理 3 篇小量高排序文件集的複雜度遠低於處理 11228 篇大量文件集的成本，故本研究提出的方法能夠有效降低所需的人力及運算成本。

二、特徵選取

特徵選取（Feature Selection）在文件探勘中扮演了重要的角色，能夠透過篩選出重要詞彙建構模型達到良好的監測成效。本研究使用排序學習演算法由搜尋引擎回傳的小量高排名文件中挑選出具代表性的詞彙建構景氣狀態預測模型，同屬於特徵選取範疇，因此在本節中回顧 Term Frequency- Inverse Document Frequency（TF-IDF）、Jaccard Index 及 Pointwise Mutual Information and Information Retrieval（PMI-IR）等傳統被廣泛使用的特徵選取方法，並作為本研究實驗比較的基準。

（一）TF-IDF

TF 假設若某一詞彙出現頻率很高，則表示此詞彙對於該文件集具有相當程度的重要性，因此可選擇高頻詞彙來建構模型。但是部分高頻詞彙，例如停止詞（stop words），其詞頻代表性不足，故發展出 Inverse Document Frequency（IDF）方法協同濾除如停止詞等高頻詞彙。TF-IDF 可表示為 $w_{t,d} = tf_{t,d} \times \log(N / df_t)$，其中，$w_{t,d}$ 表示詞彙 t 文件 d 中 TF-IDF 的權重，$tf_{t,d}$ 表示詞彙 t 出現在文件 d 中的次數，$\log(N / df_t)$ 表示詞彙 t 的 IDF 的權重，N 表示文件集中的文件總數，df_t 表示詞彙 t 在文件集中出現的文件數量。本研究計算文件集中每個候選詞彙的 TF-IDF，並挑選數值前 K 高的詞彙建構模型。

（二）Jaccard Index

Jaccard Index（Jaccard 1901）主要用以衡量兩個詞彙之間的相似度。給定詞彙 t_i 及 t_j，此兩個詞彙的 Jaccard Index 可表示成 $J(t_i,t_j)=|S_i \cap S_j|/|S_i \cup S_j|$，其中，$S_i$ 及 S_j 分別代表詞彙 t_i 及 t_i 在文件集中出現的文件數量，Jaccard Index 數值越高表示詞彙間的相似程度越高。本研究計算主題詞彙與候選詞彙之間的 Jaccard Index 並排序，並選出相似度前 K 高的候選詞彙建構模型。

（三）PMI-IR

PMI-IR（Turney 2001）是用來衡量詞彙之間關聯程度的指標。給定兩個詞彙，PMI-IR 透過搜尋引擎回傳此兩個詞彙共同出現的文件數量，再除以這兩個詞彙分別出現的文件數量得出 PMI-IR 值。PMI-IR 數值越高表示兩個詞彙關聯性越高。本研究計算主題詞彙與候選詞彙的 PMI-IR 值並排序，並選出關聯性前 K 高的候選詞彙建構模型。

參、研究方法

一、問題定義

在本研究中，我們選取與景氣狀態相關的詞彙，並取得這些詞彙的搜尋日誌建構景氣狀態預測模型。我們將預測景氣狀態視為一個分類問題，以下式表示：

$$\hat{\alpha}_l = \Gamma(\underline{Q}_l)$$

其中，$\hat{a}_l$ 表示在時間時模型預測的景氣狀態，$\underline{Q}_l$ 表示在時間 l 時由搜尋日誌中取得的景氣詞彙搜尋頻率向量，Γ 表示景氣狀態預測模型。

圖 2 為本論文之研究架構及流程，針對流程中各模組初步解釋如下：

流程 1. 定義一個與景氣狀態相關的主題詞彙

流程 2. 將主題詞彙鍵入搜尋引擎

流程 3. 取得搜尋引擎將回傳的前 N 篇景氣相關文檔

流程 4. 使用排序學習演算法計算文檔中各詞彙的權重

流程 5. 選擇權重最高的前 K 個詞彙集合當作景氣代表詞彙

流程 6. 取得景氣代表詞彙的搜尋趨勢

流程 7. 結合歷史景氣狀態建構景氣狀態預測模型

流程 8. 透過景氣狀態預測模型預測景氣狀態並驗證成效

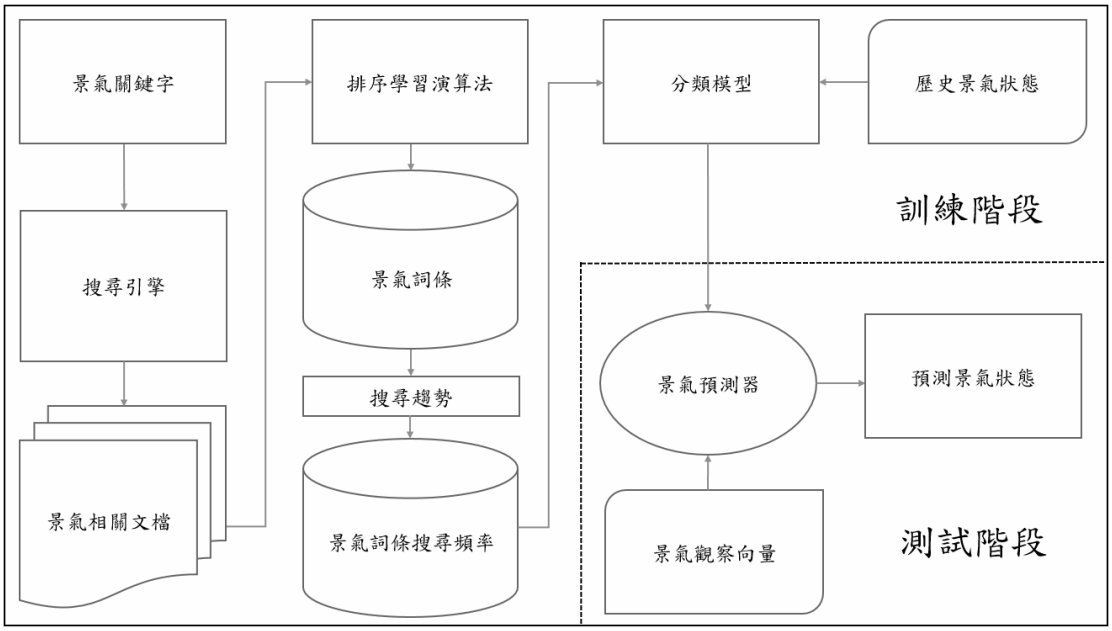

圖 2：研究流程及架構

我們將上述景氣狀態預測流程架構歸納為由兩階段所組成，分別是詞彙選取階段以及模型建構階段。在詞彙選取階段，我們使用排序學習演算法從搜尋引擎回傳的小量高排名文件中選取最能代表景氣狀態的詞彙，並下載所選詞彙搜尋日誌取得搜尋頻率。在預測模型建構階段，我們使用支持向量機（Support Vector Machine; SVM）訓練模型，並基於此模型監測景氣狀態。我們在接下來的兩節詳細說明此兩階段執行的流程。

二、詞彙選取階段

如前所述，詞彙的搜尋頻率能夠反映一個事件的狀態。在過去，提供搜尋引擎服務的公司並不會公開搜尋日誌紀錄，因此群眾無法取得搜尋日誌中紀錄的搜尋頻率。2006 年，Google 推出 Google Trend[2] 服務，此服務將某一時期內特定搜尋詞彙被搜尋的頻率劃分為 101 個等級（0-100），使得使用者能夠取得並了解在此一時期中，此搜尋詞彙在各時間點中被查詢的頻繁程度，故在此階段我們要挑選出能夠代表景氣狀態的詞彙。

隨著網路資訊科技的發展，群眾經常透過搜尋引擎提供的資訊了解一個事件的發展及狀態，而搜尋引擎提供的文件排名也隱含了重要的資訊。例如：Dou 等

2 https://www.google.com.tw/trends/

（2010）指出了搜尋引擎文件排名的重要性。作者基於心理學的促發理論（priming theory），驗證商品在搜尋引擎的排名會影響使用者對於商品的品牌形象，描述商品的文件排名越高，使用者對於商品的印象也會越好，進一步影響使用者的購買決策。由於搜尋引擎文件的排名是基於使用者點擊文件的回饋調整而得（Doan et al. 2011），因此文件的排名能夠反映出群眾對於一個事件主要關注的內容。又文件是由多個詞彙所組成，故我們認為可由群眾關注度較高的景氣相關文件中，取得代表景氣狀態的詞彙。

文獻指出當使用者在使用搜尋引擎時，僅會關注高排名的文件（Joachims 2002; Joachims et al. 2007），因此本研究使用小量高排名文件作為選取代表性詞彙的主要來源文件集。決定詞彙的主要來源文件集後，由於文件內容由多個詞彙所組成，因此我們認為存在具有代表性的詞彙使得文件能夠獲得較高的排名。先前文獻（Dou et al. 2010）雖驗證文件排名的重要性，但是卻沒有指出哪些詞彙能夠代表群眾關注的內容，使得文件在搜尋引擎中獲得高的排名位置，這當中存在了一個研究缺口。由於排序學習演算法能夠發掘獲得高排名的特徵（Liu 2009），故我們使用排序學習演算法選取使得文件獲得高排名的詞彙克服這個問題。在本研究中，我們使用知名的成對式（pairwise）排序學習演算法 RankSVM（Joachims 2002）選取具有代表性的詞彙，且據我們所知，本研究是第一個使用排序學習演算作為特徵選取詞彙的研究。

詞彙選取詳細步驟如下：

步驟 1：為由搜尋引擎取得高排名文件，首先需定義一個與景氣狀態相關的主題詞彙，記為 q_{topic}。由於主題詞彙的選擇將會影響取得文件的品質，我們參考 Fan 等（2010）的建議，使用主題式查詢詞彙原則（topic-based query term principle）定義主題詞彙。

步驟 2：將主題詞彙鍵入搜尋引擎，搜尋引擎將回傳一個經過排序的文件序列，選取排序在此文件序列的前 N 篇文件作為詞彙來源文件集，記為 D，$D = \{d_1, d_2, ..., d_N\}$。

步驟 3：使用詞性標註系統將文件轉換為詞彙向量。由於文獻指出使用者傾向使用名詞當作搜尋的詞彙（Barr et al. 2008），因此我們選取文件集 D 中的名詞作為候選詞彙，記為 V，$V = \{v_1, v_2, ..., v_M\}$。

步驟 4：使用 RankSVM 來選取代表景氣狀態的詞彙。文件集 D 中的每一篇文件 d_n 可以表示為一個 M 維度的向量 $\underline{X}_n$，$\underline{X}_n = <x_{n1}, x_{n2}, ..., x_{nM}>$，其中，$x_{nm}$ 表示詞彙 v_m 在文件 d_n 中出現的次數，接著我們使用 y_{ij} 表示文件 d_i 與 d_j 之間排名的關係，$y_{ij} = 1$ 表示文件 d_i 排名在文件 d_j 之前，反之 $y_{ij} = -1$。經由下式計算出每一個詞彙在文件集 D 中的權重：

$$\min_{W} \frac{1}{2} \underline{W}^T \underline{W} + C \sum_{p=1}^{P} \xi_p$$

上式條件限制為

$$\underline{W}^T (\Phi(\underline{X}_i) - \Phi(\underline{X}_j)) \geq 1 - \xi_p, \xi_p \geq 0, p = 1, ..., P, \forall y_{ij} = 1$$

其中，$\underline{W}$ 表示詞彙的權重向量，Φ 為核函數，C 為懲罰參數，ξ_p 為遲滯參數。計算出 $\underline{W}$ 後，我們選擇權重最高的前 K 個詞彙集合當作代表性詞彙，記為 T，$T=\{t_1, t_2, ..., t_K\}$。

三、預測模型建構階段

步驟 1：下載詞彙集合 T 在時間 L 的搜尋日誌並取得搜尋頻率。在時間點 l，詞彙集合 T 的搜尋頻率可以一個高維度向量表示，記為 $\underline{Q}_l$，$\underline{Q}_l=<q_{1l}, q_{2l}, ..., q_{Kl}>$，$q_{kl}$ 表示詞彙 t_k 時間點 l 的搜尋頻率，每一個搜尋頻率 $\underline{Q}_l$ 對應到一個當期的景氣狀態 a_l，故可以得到一個景氣狀態訓練集，記為 R，$R = \{(\underline{Q}_1, a_1), (\underline{Q}_2, a_2), \ldots (\underline{Q}_L, a_L)\}$。

步驟 2：使用 SVM 訓練模型，經由求解下式得到模型參數：

$$\min_{W'} \frac{1}{2} \underline{W'}^T \underline{W'} + C' \sum_{l=1}^{L} \xi_l$$

上式條件限制為

$$a_l (\underline{W'}^T \Phi'(\underline{Q}_l) + b) \geq 1 - \xi_l, \xi_l \geq 0, l = 1, ..., L$$

其中，$\underline{W}'$ 表示詞彙的權重向量，Φ' 為核函數，C' 為懲罰參數，ξ_l 為遲滯參數。

肆、實驗設計與結果

一、實驗資料說明及實驗評估準則

本研究旨在建立台灣景氣狀態監測系統，因此我們參考 Chen 與 Tsai（2012）的研究，使用景氣對策信號代表台灣各期景氣狀態。我們按照國家發展委員會景氣指標查詢系統[3]之定義，將景氣狀態分為低迷（藍燈）、轉向（黃藍

3 http://index.ndc.gov.tw/n/zh_tw

燈）、穩定（綠燈）、轉向（黃紅燈）、熱絡（紅燈）5 種景氣狀態，並下載取得台灣 2009 年到 2013 共 5 年 60 期之景氣對策信號資料進行實驗。為取得詞彙來源文件集，首先需要定義主題詞彙。在本研究中我們將主題詞彙定義為*景氣*，接著將主題詞彙鍵入 Google 搜尋引擎，Google 搜尋引擎會回傳與*景氣*相關且經排序的文件序列。由於文獻指出使用者通常僅關注高排名文件，因此我們從中選取排名前 10 名的文件當作本研究的實驗文件集，選擇排名前 10 名文件的原因是搜尋引擎（e.g., Google）在一個頁面中僅回傳 10 個文件。取得文件後需要對文件進行詞性標註，本研究使用中研院開發之 CKIP 中文斷詞系統[4]對文件進行詞性標註，由於文獻指出使用者習慣以名詞作為搜尋詞彙，因此僅保留名詞詞彙當作候選詞彙。接著使用成對式排序學習演算法 RankSVM，從中挑選出具足夠代表性的景氣狀態名詞詞彙，最後由 Google Trend 下載這些名詞詞彙的搜尋日誌結合支持向量機訓練模型。

在本研究中，為了驗證使用小量高排名文件選取詞彙建立預測模型的實驗成效，我們分別檢驗了 3 篇、5 篇、10 篇文件當做詞彙來源文件集，意即 N=3、5、10；而詞彙 K 值則分別設定為 5、10 及 15 個詞彙以進行不同特徵選取演算法之比較。表 1 列出本研究所蒐集之 3 篇、5 篇、10 篇文件集的基本統計資訊。

表 1：景氣狀態文件集統計資訊

研究主題	景氣狀態
主題詞彙	景氣
時間	2009－2013
# 景氣狀態	60
# 文件數 /# 候選詞彙	3 / 565 5 / 780 10 / 1053

在成效計算部分，本研究使用準確率（Accuracy）評估模型的成效，準確率之定義即為成功預測之次數 / 總預測次數（Fang et al. 2010）。為了更準確的計算成效，本研究採取了 Leave-One-Out 方法（Cawley & Talbot 2004）對實驗結果進行交叉驗證，也就是說，每 1 次的實驗使用 1 期的景氣狀態作為測試集，剩餘的景氣狀態作為訓練集，並預測 1 期的景氣狀態。由於實驗蒐集了 5 年共 60 期景氣狀態，實驗將重複執行 60 次蒐集 60 期的預測狀態，最後計算準確率驗證模型的成效。Leave-One-Out 方法已被證明能夠避免模型產生過度擬合（overfitting）

4 ttp://ckipsvr.iis.sinica.edu.tw/

的現象。

二、實驗結果與討論

表 2 列出本研究之實驗結果，Random 為隨機猜測之期望機率，在以不同的文件集 *N* 值的設定為分組，以及不同的詞彙 *K* 值設定之下，準確率分別可以達到 0.68（*N*=3，*K*=15）、0.53（*N*=5，*K*=10）及 0.55（*N*=10，*K*=15）。結果顯示，基於 RankSVM 所選取的詞彙其所建構模型的預測準確率皆優於使用 TD-IDF、Jaccard Index 及 PMI-IR 等傳統特徵選取技術所建構模型的預測準確率。由於本研究所做達到之最佳準確率為 0.68（*N*=3，*K*=15）。

表 2：基於不同特徵選取方法之景氣狀態預測準確率

# 文件集	特徵選取	準確率（*K*=5）	準確率（*K*=10）	準確率（*K*=15）
3	RankSVM	0.55	0.62	0.68
	TFIDF	0.32	0.37	0.38
	PMI-IR	0.50	0.57	0.57
	Jaccard	0.48	0.53	0.60
	Random	0.20	0.20	0.20
5	RankSVM	0.48	0.53	0.51
	TFIDF	0.28	0.32	0.35
	PMI-IR	0.38	0.45	0.43
	Jaccard	0.40	0.45	0.45
	Random	0.20	0.20	0.20
10	RankSVM	0.53	0.60	0.62
	TFIDF	0.33	0.42	0.41
	PMI-IR	0.53	0.56	0.56
	Jaccard	0.45	0.48	0.50
	Random	0.20	0.20	0.20

為了進一步驗證此實驗結果於統計基礎上確實勝過基於傳統特徵選取演算法所建構之模型，表 3 列出成對 t 檢定結果，結果顯示在 95%信心水準下，RankSVM 確實顯著優於傳統之特徵選取技術。

表 3：基於最佳預測實驗集之成對 t 檢定結果

特徵選取方法	p-value
TF-IDF	0.0162**
Jaccard	0.0195**
PMI-IR	0.0145**
Random	0.0074***

*, **, *** 分別表示 $\alpha = 0.1, 0.05, 0.01$

本研究與 Chen 與 Tsai（2012）同屬景氣狀態預測之研究，並且同將景氣狀態預測視為一個分類的問題。然而，Chen 與 Tsai（2012）在實驗設計上偏向於窮舉法，他們使用爬網之技術大量的擷取了網路上的 11228 篇的文件集，萃取了 5 萬 8 千個詞彙。接著，他們擷取此 5 萬 8 千個詞彙的搜尋日誌，並結合簡單貝式演算法進行景氣預測模型的訓練以及預測，證實了使用搜尋日誌能夠有效的預測台灣的景氣預測狀態。在文件集合的選擇上，該文並未針對使用者的網路使用行為作進一步的觀察，僅是將網路上大量的文件蒐集下來並進行實驗；本文則受到 Dou 等（2010）的啟發，考量使用者使用搜尋引擎時的心理以及行為，僅透過選擇搜尋引擎上小量高排序的 3 篇文件，從 565 個候選詞彙中再透過排序學習演算法萃取使用者真正感到興趣的詞彙來進行景氣狀態預測研究。

Chen 與 Tsai（2012）使用 11,228 篇文件建構景氣狀態預測模型，其預測準確率為 0.67；本研究基於 3 篇文件集所建構模型的預測準確率為 0.68。我們的實驗結果驗證了基於小量高排名文件集所建構的模型其預測準確率不亞於使用大量文件集建構的模型，並且顯示小量高排名文件集即可代表群眾所關注的景氣議題，不需要將所有網路上的文件都蒐集下來進行處理。也由於處理小量高排名文件集的運算成本低於處理大量文件集的成本，故本研究提出的方法能夠降低運算所需的資源。

我們於圖 3 列出本研究最佳實驗結果各期之實際景氣狀態以及預測景氣狀態，圖 4 為本研究實際狀態與預測狀態之混淆矩陣。由圖 3、圖 4 可以觀察到，我們所提出的方法能夠準確的預測景氣狀態變化的趨勢，即使無法準確的預測，其預測結果也不會大幅的偏離真實的景氣狀態。

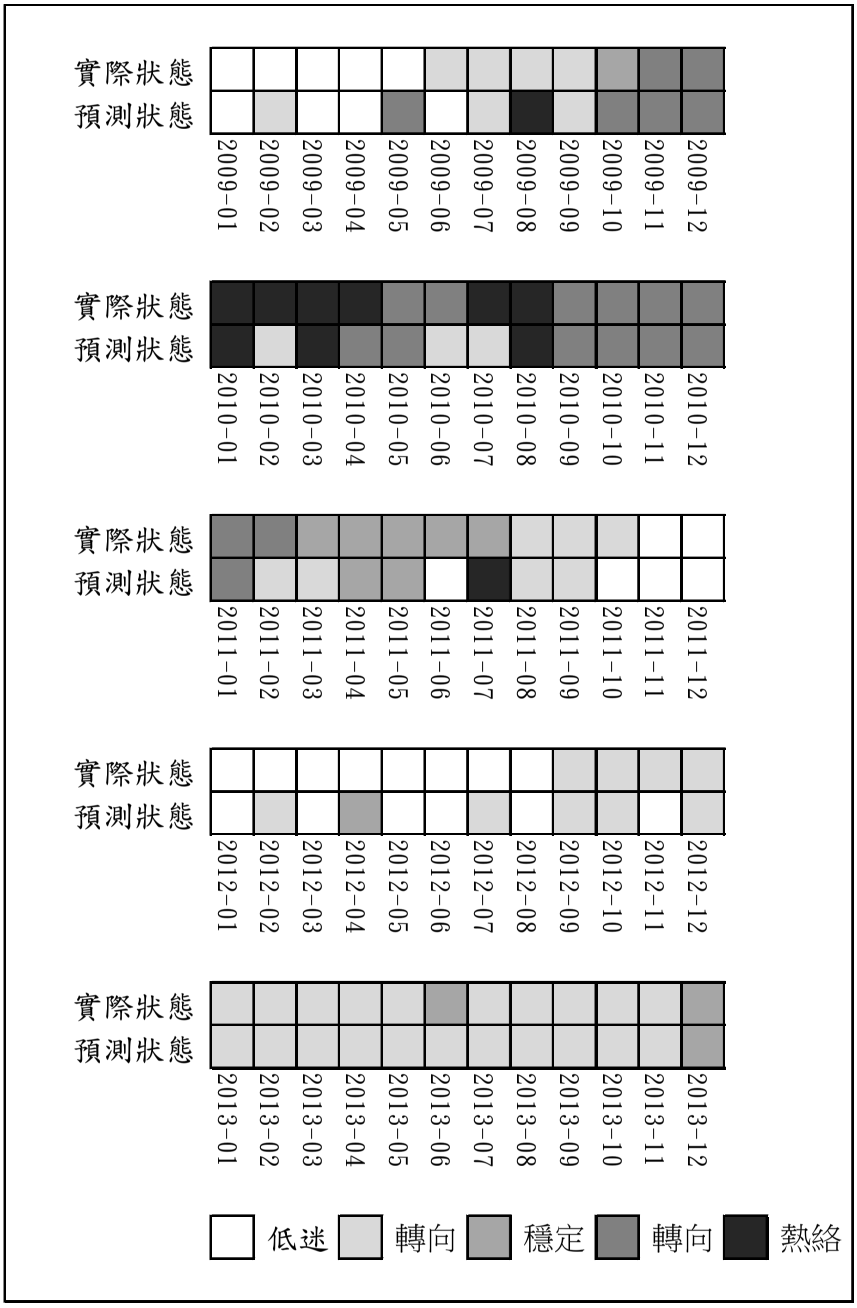

圖 3：模型預測狀態及真實狀態

	藍燈	黃藍燈	綠燈	黃紅燈	紅燈	總和
藍燈	10	3	1	1	0	15
黃藍燈	3	17	0	0	1	21
綠燈	1	2	3	1	1	8
黃紅燈	0	2	0	8	0	10
紅燈	0	2	0	1	3	6
總和	14	26	4	11	5	60

圖 4：實際狀態與預測狀態之混淆矩陣

表 4 列出本研究基於各特徵選取技術選出之詞彙。由表 4 可以發現傳統特徵選取技術所選出的詞彙多為一般經濟詞彙，例如：股市、基金；而成對式排序學習演算法 RankSVM 所選取的詞彙除了一般經濟詞彙外，尚能挑選出如政府、政策等詞彙，顯示群眾在關注景氣狀態時除了關注股市、基金表現外，同時也關注政府機關所提出的經濟政策，也因為這些詞彙的出現，使得文件能夠在搜尋引擎中回傳的序列中取得較高的排名。

表 4：基於不同特徵選取方法選出之代表性詞彙

TF-IDF	Jaccard	PMI-IR	RankSVM
權證	股市	4 月	股價
指數	股數	交易	因素
平均數	脫節	低廉	股市
基期	股息	國際	經濟
承銷價	股利	分析	力量
瓊斯	股份	上市	政府
標的	聰明	年度	本益比
樣本	群眾	1 月	趨勢
資訊	缺失	10 年	影響
代表性	突發性	基金	價格
交易所	衡量	分享	政策
市價	能力	市場性	利率
階層	缺點	報告	物價
香港	經濟	低價	市場
需求	策略	問題	貨幣

此外，由於本研究使用小量高排序文件集進行實驗，所以一些重要的景氣詞彙會經常出現於高排序文件集中，在這樣的條件設定之下，TF-IDF 演算法非常容易將這些重要的景氣詞彙當作停止詞濾除，這也是在本研究中，TF-IDF 演算法其實驗成效不好的重要原因。

伍、討論

一、意涵

景氣狀態監測對於政府及企業非常的重要，政府及企業的決策者透過取得的景氣資訊得以在快速變化的經濟形勢中制定出適當的政策以因應環境的變化。本研究提出一個新式的景氣狀態監測系統，我們將景氣狀態監測視為一個分類的問題，使用排序學習演算法從小量高排名文件集中取得使用者關注的景氣詞彙，接著下載所選詞彙的搜尋日誌，從中取得詞彙的搜尋頻率，結合支持向量機訓練景氣預測模型。

實驗結果顯示我們提出的方法能夠準確的預測景氣狀態，且基於排序學習演算法所選詞彙建構的模型其準確率也優於使用傳統特徵選取方法所選詞彙建構的模型。由於搜尋日誌具有高度的即時性及可用性，因此基於搜尋日誌的監測指標相較於傳統經濟指標能夠更快速的被制定，故可以有效克服景氣狀態發佈延遲的議題。此外，使用小量高排名文件集作為詞彙來源文件集也較先前文獻使用大量文件集更能夠節省運算的資源以及成本。

根據以上說明，本研究所提出的景氣狀態監測系統可以有效提升政府及企業的決策能力，幫助政府及企業在快速變化的經濟形勢中制定即時且正確的政策，降低資訊延遲對政府及企業所造成的損失。

二、研究限制與未來展望

本文主要之假設為基於搜尋詞彙所制定的新式經濟指標可以有效降低景氣狀態發佈的延遲。故在本研究中我們提出結合排序學習之新式特徵選取演算法，並於本文中與傳統的特徵選取演算法進行比較，實驗的結果證明了我們所提出方法的有效性。

然而，本研究具有下列限制：(1)本實驗之設計乃採用單一時間點建構之關鍵字集進行實驗，且搜尋引擎的搜尋結果也會隨著時間而有所變動，此變動將導致不同的時間點所建構出之關鍵字集有所不同。當我們擴增的實驗的區間，我們所選擇的詞彙其代表性有可能隨著時間的增長而逐漸的下降，這同時也意味在不同的時間點之下，使用者在搜引擎上所關注的景氣詞彙也會有所不同，因此如何選

擇實驗的長度是我們後續將要面臨及解決的挑戰。(2)本研究在比較各不同特徵選取演算法時，由於各比較對象皆使用相同演算法進行預測，故是否考慮景氣之間的時序性變化並不會影響比較結果。然而，景氣狀態之間時序性的考量對於整體系統設計乃是一項重要的考慮因素，故此，我們已規劃於下一階段考慮景氣之間之時序性變化並探討其對預測成效之影響，我們將建立新式的預測模型，並與傳統時間序列以及機器學習模型比較驗證成效。

除了以上所列之限制與後續改善之規畫外，再將來，我們將參考更多的資訊來源，例如：新聞文章，網路論壇，以及社群網站等對於景氣狀態的討論內容。接著，我們也將嘗試使用更多先進的機器學習模型來預測各期景氣狀態，期望能夠在此議題上持續達到更好的預測成效，幫助決策者做出更加即時且正確的判斷。

誌謝

本文接受行政院科技部專題研究計畫（MOST 105-2221-E-002-187）之補助研究經費，順利完成此篇著作之研究工作，謹此致謝。

參考文獻

Andersson, E., Bock, D. and Frisén, M. (2006), 'Some statistical aspects of methods for detection of turning points in business cycles', *Journal of Applied Statistics*, Vol. 33, No. 3, pp. 257-278.

Askitas, N. and Zimmermann, K.F. (2009), 'Google econometrics and unemployment forecasting', *Applied Economics Quarterly*, Vol. 55, No. 2, pp. 107-120.

Baeza-Yates, R. and Tiberi, A. (2011), 'Extracting semantic relations from query logs', *Google Patents*.

Barr, C., Jones, R. and Regelson, M. (2008), 'The linguistic structure of English web-search queries', *Proceedings of the conference on empirical methods in natural language processing*, Waikiki, Honolulu, Hawaii, October 25-27, pp. 1021-1030

Burns, A.F. and Mitchell, W.C. (1946), 'Measuring business cycles', *NBER Books*.

Cawley, G.C. and Talbot, N.L. (2004), 'Fast exact leave-one-out cross-validation of sparse least-squares support vector machines', *Neural networks*, Vol. 17, No. 10, pp. 1467-1475.

Chauvet, M. and Piger, J. (2008), 'A comparison of the real-time performance of business cycle dating methods', *Journal of Business & Economic Statistics*, Vol. 26, No. 1, pp. 42-49.

Chen, C.C. and Tsai, Y.T. (2012), 'A novel business cycle surveillance system using the query logs of search engines', *Knowledge-Based Systems*, Vol. 30, pp. 104-114.

Choi, H. and Varian, H. (2012), 'Predicting the present with google trends', *Economic Record*, Vol. 88, No. s1, pp. 2-9.

Doan, A., Ramakrishnan, R. and Halevy, A.Y. (2011), 'Crowdsourcing systems on the world-wide web', *Communications of the ACM*, Vol. 54, No. 4, pp. 86-96.

Dou, W., Lim, K. H., Su, C., Zhou, N. and Cui, N. (2010), 'Brand positioning strategy using search engine marketing', *Mis Quarterly*, Vol. 34, No. 2, pp. 261-279.

Eysenbach, G. (2002), 'Infodemiology: The epidemiology of (mis) information', *The American journal of medicine*, Vol. 113, No. 9, pp. 763-765.

Fan, J., Wu, H., Li, G. and Zhou, L. (2010), 'Suggesting topic-based query terms as you type', *Web Conference (APWEB), 2010 12th International Asia-Pacific*, April 06 - 08, pp. 61-67

Fang, Z.H., Tzeng, J.S., Chen, C.C. and Chou, T.C. (2010), 'A Study of Machine Learning Models in Epidemic Surveillance: Using the Query Logs of Search Engines', *PACIS*, pp. 137.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L. (2009), 'Detecting influenza epidemics using search engine query data', *Nature*, Vol. 457, No. 7232, pp. 1012-1014.

Hameed, A., Kang, W. and Viswanathan, S. (2010), 'Stock market declines and liquidity', *The Journal of Finance*, Vol. 65, No. 1, pp. 257-293.

Hamilton, J.D. and Perez-Quiros, G. (1996), 'What do the leading indicators lead?', *Journal of Business*, pp. 27-49.

Jaccard, P. (1901), Etude comparative de la distribution florale dans une portion des Alpes et du Jura, Impr. Corbaz.

Joachims, T. (2002), 'Optimizing search engines using clickthrough data', *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, Edmonton, AB, Canada, July 23-25, 2, pp. 133-142

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F. and Gay, G. (2007), 'Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search', *ACM Transactions on Information Systems (TOIS)*, Vol. 25, No. 2, pp. 7.

Jun, D.B. and Joo, Y.J. (1993), 'Predicting turning points in business cycles by detection of slope changes in the leading composite index', *Journal of Forecasting*, Vol. 12, No. 3-4, pp. 197-213.

Kannan, S.S. and Ramaraj, N. (2010), 'A novel hybrid feature selection via Symmetrical Uncertainty ranking based local memetic search algorithm', *Knowledge-Based Systems*, Vol. 23, No. 6, pp. 580-585.

Li, Z., Xu, W., Zhang, L. and Lau, R.Y. (2014), 'An ontology-based Web mining method for unemployment rate prediction', *Decision Support Systems*, Vol. 66, pp. 114-122.

Liu, T.Y. (2009), 'Learning to rank for information retrieval', *Foundations and Trends in Information Retrieval*, Vol. 3, No. 3, pp. 225-331.

Turney, P. (2001), 'Mining the web for synonyms: PMI-IR versus LSA on TOEFL', *Machine Learning*, pp. 491-502.

Vosen, S. and Schmidt, T. (2011), 'Forecasting private consumption: survey-based indicators vs. Google trends', *Journal of Forecasting*, Vol. 30, No. 6, pp. 565-578.