以模型融合配合社群網路資料進行流感趨勢預測

Building Fusion Model for Flu Trend Prediction by using Social Web Data

張昭憲
淡江大學資訊管理系
新北市淡水區英專路 151 號
jschang@mail.tku.edu.tw

周書任
淡江大學資訊管理系
新北市淡水區英專路 151 號
jason093351@hotmail.com.tw

摘要

根據世界衛生組織(WHO)統計，流感每年在全球約造成 300 萬個嚴重病例及 25 萬人死亡，對民生、經濟之影響有目共睹。為監測流感疫情，各國疾管局通常藉由臨床就診通報來彙整資料，但可能產生 1~2 週的延遲，顯然緩不濟急。考量網路社群已成為現代人生活一部分，若能從中蒐集資料並發展預測方法，應可更快了解流感現況，降低其負面影響。此外，流感流行變化快速、預測不易，但若能整合不同的預測方法，將可提升其準確性。有鑒於此，本研究將運用社群網路資料，以模型融合(Model Fusion)為基礎，建立有效的流感就診率預測方法。首先，我們由不同的網路資料來源蒐集資料，透過關鍵詞集統計建立資料集。接著，配合延遲概念，以線性迴歸建立多種不同特質的預測模型。最後，再透過模型融合整合各模型之預測結果，以提升總體準確性與穩定性。

為驗證提出方法之有效性，本研究蒐集英國地區約 82 週(2015/8–2017/3)超過 160 萬筆的 Twitter 發文，及同時期的 Google 關鍵字搜尋熱度資料，經處理後進行實驗。與六種單一預測模型相較，本研究提出之方法具有最高的預測關聯度，顯示方法之有效性。為了解各種模型之穩定性，再將流感資料依照流行程度分為「劇升降」區與「緩升降」區進行統計。結果顯示本方法在二個區域分別具有的最高與次高之關聯度，其他單一模型則呈現不一致之預測效果，驗證本方法確能產生較穩定之預測。綜合上述結果，我們相信透過本研究所提出之方法，能提供更有效之流感早期預警，建立更多元的防疫防線。

關鍵詞：流感監測(Flu Monitoring)、模型融合(Model Fusion)、線性迴歸(Linear Regression)、社群網路(Social Web)

Abstract

According to statistics of WHO, flu averagely causes 3 million serious cases of illness and 250 thousand deaths per year. Obviously, it is a constant and big threaten for global people. However, if we can discover the epidemic trend of flu in advance, it is still possible to reduce

the illness rate effectively. To monitor the epidemic situation of flu, the CDCs of countries usually collect weekly influenza-like illness (ILI) rate by gathering clinical reports from hospitals. However, it could cause about 1-2 weeks delay and therefore might miss the information about the peak period of flu epidemic. To remedy the above situation, it is necessary to develop new methods to discover the epidemic of flu in time. Because social webs have become part of our lives, it is a promise way to build prediction methods by mining the flu information from the webs. In addition, in view of the drastic change of flu epidemic trend, it is necessary to combine several prediction methods to provide a more accurate prediction. To this end, this paper tries to develop effective methods for flu trend prediction by model fusion and mining data from the social web. First, we collect the web data from different sources. Next, various prediction models are built by considering the delay of epidemic. Finally, those generated models are merged by model fusion to increase the accuracy and stability of prediction.

To demonstrate the effectiveness of the proposed method, we collected over 1.6 million posts from Twitter in England and the flu-related keywords search statistics from Google Trends for experiments. Compared with the six single prediction models, the proposed method has the highest predictive relevance that shows the effectiveness of the method. In order to understand the stability of various models, the data will be divided into "dramatic up-down" area and "slow up-down" area. The results show that the method has the highest correlation with the second highest in the two regions, and the other single models show inconsistent prediction effects, indicating that the method can produce more stable prediction results. Based on the above results, the proposed method of this study does contribute to the early warning of influenza surveillance and establish more anti-epidemic defense lines.

Keywords: Flu Monitoring、Model Fusion、Linear Regression、Social Web

一、緒論

流行性感冒(Influenza)是由流感病毒引起的急性感染，患者除身體不適外，更可能引起嚴重併發症，甚至導致死亡，對民眾健康帶來巨大威脅。由於流感傳染力強，透過頻繁國內、外大眾運輸，經常造成全球大規模流行。根據 WHO(World Health Organization)統計，依嚴重程度不同，流感在全球每年平均造成約 300~500 萬個嚴重病例，以及 25~60 萬人死亡[22][24]。若疫情控制不當，更可能爆發如 1918 年之全球性大流行。當時，H1N1 流感病毒在全球造成了五億個感染病例，估計有五千萬到一億名病患死亡，約為當時世界人口的 3%~5%[21]。由上可知，流感確實是全球性重要公共健康議題，各國需投注足夠的人力物力來防範。

流感可透過接種疫苗提高抵抗力，減少高風險族群感染的機會。若能及早發現流感散播趨勢，更可進一步降低感染人數。為監測流感疫情，各國疾管局通常藉由臨床就診

通報以了解疫情。但考量資料彙整與確認等流程，可能產生 1~2 週的延遲，使流感即時監控功能大打折扣。若因此造成大幅擴散，衍生的社會成本將十分龐大，因此需有更積極的早期預警作為。有鑒於網路社群(如 Twitter, Facebook 等)已成為現代人生活一部分，許多人將每日的經歷與感受發布於社群中，與親友或會員分享。因此，學者開始研究如何由網路社群獲取相關資訊並進行分析，了解事件的現況或民眾親身反應。例如，針對災害應變， Lampos 等人[17]蒐集群眾在網路上有關天氣之留言，預測氣象事件(大雪、暴雨等)的發展，以提供更即時、正確的應變訊息。Akay 等人[1]則分析醫藥專業社群的發言，即時了解民眾對特定藥品之用藥反應，做為開發或改良新藥的依據。Lampos 等人[18]又利用選民在 Twitter 上的留言，預測民眾對各主要政黨的投票傾向，其結果與實際投票統計頗為相符。針對食品工業，He 等人[12]則分析 Facebook、Twitter 上之顧客反應，以了解不同 Pizza 廠牌之顧客接受度與競爭關係。

根據上述文獻顯示，透過蒐集民眾在社群網路的活動，確有助於改善事件監測之即時性。因此，為提供更即時的流感監測，應可將其視為重要資訊來源。根據 Fox 的研究[10]，美國每年約有九千萬成年人在網路上搜尋與特定疾病相關醫學資訊，說明網路已成為現代人重要醫療資訊來源。配合此種改變，Eysenbach 以搜尋關鍵字廣告的點擊為根據，在加拿大進行流感監測[9]。Hulth 等人[14]則將網路的流感資訊限縮在醫學網站，透過解析網站內部所發生之查詢或頁面瀏覽，分析流感發生或散播狀況。為增加流感監測之準確性，學者開始也開始研究特定關鍵詞集與流感之間的關係。例如，Polgreen 等人[19]利用 Yahoo 搜尋關鍵字，驗證特定查詢字彙是否與病毒學、死亡率的統計資料具有關連性。Ginsberg 等人[7]則運用 Google 搜尋關鍵字比對 2003~2008 年美國類流感統計資料，並藉此建立每日更新之疫情趨勢資訊。

除網路關鍵字搜尋外，民眾在社群網路中的發文，也是流感監測另一重要資訊來源。Lampos 等人[16]分析英國 Twitter 社群發文，配合與流感相關關鍵詞集，以線性迴歸建立預測模型，其即時性明顯優於傳統通報體系。Bardak 和 Tan[2]則透過收集 Google Flu Trend 和 Wikipedia Logs 流感文章，整理後依關聯度挑選屬性集，並以線性公式產生預測模型。Prakash[20]則將流感病情分為健康期、隱藏期、發病期、康復期，藉由分析四種階段可能的發文情形，推斷每月發病之案例數。Byrd 等人的研究[3]則再次證實，對 Twitter 上的發文進行關鍵詞分析，確實可獲得精確的流感就診率預測結果，並提醒衛生當局能加以重視。

由上述前人研究可知，透過蒐集社群網路資訊，確有助於改善流感監測之即時性。然而，實際應用時，準確性則為另一重要課題。一般而言，為提升預測準確性，可挑選合適的學習方法，建立較準確的流感預測模型。但有鑑於各種預測模型有其特質，針對不同應之偵測效能亦有差異，為提升整體效能，便可採用模型融合(Model Fusion)概念加以整合。模型融合是一種由各種模型中擷取有利於預測效能的流程，並已被應用在各種不同領域中。例如， Huang 等人[13]以加權方式融合多種模型，並透過學習法調整權重，以產生更精確的商品推薦名單。針對電子商務詐騙偵測，Chen 等人[4]則利用 logistic regression 組合各種模型產生之可疑分數，以獲得較佳的偵測效能。Li 等人[6]則同時使

用貝式分類法與關聯規則探勘，歸納觀察詐騙帳戶之樣式，以利人類專家進一步判讀。由上可知，若能有效整合各種模型特點，對流感預測之準確性將有很大助益。

綜合上述討論，為提供有效、即時的流感就診率預測，協助相關單位及早因應，本研究將由不同的網路資料來源，建立多種不同特質之預測模型，並透過模型融合加以整合，以提升預測的準確性與穩定性。為此，我們蒐集 Twitter 社群發言與 Google 關鍵字搜尋熱度資料，配合關鍵字集整理後，產生二種不同資料集。之後，再根據官方公布之實際就診率，以線性迴歸分別建立流感就診率預測模型。此外，考量疫情因潛伏期可能造成的延遲，本研究亦將延遲因素納入，建立對應的延遲模型。為整合上述做法建立之各種模型，我們採用模型融合概念以線性迴歸再次加以組合，以產生更準確、穩定之預測結果。

為驗證提出方法之有效性，本研究蒐集英國地區約 82 週(2015/8–2017/3)超過 160 萬筆的 Twitter 發文，及同時期的 Google 關鍵字搜尋熱度資料，經處理後進行實驗。結果顯示，與 6 種單一預測模型相較，本研究提出之方法具有最高之預測關聯度。其次，為了解各種模型之穩定性，再將流感資料依程度分為「劇升降」區與「緩升降」區進行統計。結果顯示提出之方法在二個區域分別具有的最高與次高之關聯度，而其他單一模型則呈現不一致之預測效果，顯示本方法確能產生較穩定的結果。此外，當考慮延遲因素時，Google 資料集建立之模型效果明顯優於 Twitter 相關模型，說明不同的資料來源產生之模型具有不同的預測能力，若能善加利用，將可提升預測準確率。綜合上述結果，本研究之方法確有助於提供更即時、準確之預測結果，有效改善流感監測之預警效果。

本論文其餘章節之架構如下: 第二節為文獻探討與相關技術簡介；第三節介紹本研究提出之流感就診率預測方法；第四節為實驗結果，展示各種設定下，流感就診率之預測結果；第五節為結論與未來工作。

二、 文獻探討與相關技術

本節將介紹與本論文相關之文獻探討、背景知識及相關技術，以利後續章節之討論。

1. 流感監測系統

流感是由病毒引起的急性呼吸道感染，容易在人群間傳播，在發病前一天至症狀出現後三到七天皆具有感染能力。其主要症狀為發燒、頭痛、疲倦等，其中少數可能發生併發症，常見為肺炎等。截至目前，有效抑止流感的方法為施打疫苗，但流感病毒可能會發展出對疫苗的抗藥性[24]。

流感監測系統則提供早期示警功能，透過地方公共衛生部門、醫院、實驗室等合作，將病例資料進行整合，盼能及早掌握流感疫情變化，進而提出相對應行動，包括疫苗分配、高風險族群宣導、制定防疫政策等。以下為國內常見的流感監測方法[25]:

(1) 住院監測：將住院病患之病歷資料等進行統計，藉此分析每日住院趨勢，確定流感類型比例與了解流感併發症現況及嚴重程度。

(2) 病毒監測:將病歷檢體之檢驗結果及流感分型做為確診依據，並對流感重症病例、

流感群聚病例、流感輕症病例抽樣進行抗藥性檢測，確保疫苗不因病毒突變而降低效果。

(3) 門診監測：透過統計經過門診及急診之病例數，再計算出流感就診人次、年齡、流感型別比率等資料。

(4) 死亡監測：統計因流感併發症死亡之病例資料，計算流感死亡之比率。

傳統監測系統多用於統計當週統計病例資料與確認流感型別，少有對於下週流感趨勢進行預測。但若能發展早期預警監測系統，使政府及相關單位能及早擬定因應對策(疫苗接種、資源分配、政策宣導等)，藉此降低受感染人數，便可有效控制疫情。

2. 運用網路社群進行流感監測

雖有各種防範方式，但流感病毒株可迅速變種，讓人防不勝防。例如，2009 年 H1N1 新型流感疫情，即造成社會極大恐慌[22][24]。因此，我們需進一步發展全面性流感監控系統，方可有效降低流感造成的傷害。為此，學者開始研究各種更即時、準確的監測方法。有鑒於網路社群已經成為現代人生活一部分，許多人會將每日生活動態張貼於社群中，從中獲取流感資訊，便成為可行的解決之道。例如，Polgreen 等人[19]利用 2004 到 2008 年的 Yahoo 搜尋資料，透過分析類流感關鍵字佔所有搜尋的比率與 CDC 類流感陽性比率之相關度，建立了可用的預測模型，並以顯著性差異與決定係數檢定其實驗結果。其模型可預測一至三週後流感疫情是否升高，為流感監控提供一種有效即時的工具。Ginsberg 等人[7]則利用 Google 公司 2003 到 2007 年搜尋引擎的查詢資料，計算與 CDC 類流感監測關聯性最高 45 個關鍵字佔所有之搜尋比例。他們將這些比例加上權重後，建立了針對美國九個區域的監測模型。透過此方法可產生較 CDC 早一到兩周產出數據，有效提升流感監測的即時性。

除了搜尋引擎資料的分析外，如前所述，亦有學者利用社群網路來監測流感疫情。無論是透過 Twitter 取得流感關鍵字之發言，再進行預測[16]，或是以 Google Flu Trend 做為預測基礎[2]，其結果均進一步驗證運用網路社群於流感預測之可行性。這些預測雖然有效，但也經常產生意料之外的結果。例如，感染民眾因為試圖自行恢復，但一段時間後，若病情變得嚴重，才會求助於醫生，因而造成了預測延遲。當然，隨著民眾網路社群使用習慣的改變(例如改以貼圖顯示心情)，此類型作法也需隨之調整，才能提供即時正確的預測結果。

3. 皮爾森相關係數

為評估本研究所建立之預測模型效能，本研究使用皮爾森相關係數(Pearson's correlation coefficient)衡量預測數值與官方統計數值間的關聯度(Correlation)。皮爾森相關係數公式如下：

$$r_{ab} = \frac{\sum_{i=1}^{n}(a_i - u_a)(b_i - u_b)}{n\sigma_a\sigma_b} \qquad (1)$$

其中 n 為資料集涵蓋之週數(官方通常以週為單位公布流感就診率)，實際官方統計數值為 $\mathrm{A} = (a_1, a_2, \ldots, a_n)$，預測模型所產生數值為 $\mathrm{B} = (b_1, b_2, \ldots, b_n)$，

μ 為數列之平均值，σ 為標準差。由於相關係數僅為計算兩變量的線性相關程度，其結果 r 數值大小的意義根據所計算之對象而不同。例如，科學儀器校正的結果 $|r|$ 可能要超過 0.9 才算顯著相關，而行為科學上 $|r| > 0.5$ 即屬非常顯著。Cohen[5]對於皮爾森相關係數在社會科學上的意義做出了定義，當$|r| < 0.10$ 時可視為無相關性，$0.10 < |r| < 0.30$ 為弱相關，$0.30 < |r| < 0.50$ 為中度相關，$|r| > 0.50$ 為高度相關。

三、運用模型融合與網路社群資料建構流感就診率預測方法

本節將介紹本研究提出之流感就診率預測方法細節，包括如何由不同資料來源建構資料集，如何透過延遲概念建構多種預測模型，以及如何整合多個單一模型之預測值，以產生有效、穩定的預測結果。

為使後續的討論更清楚，在介紹提出方法之前，先說明所欲解決的問題:

問題陳述: 當疾病管制局尚未公布第 t 週流感就診率 $R(t)$之前，試圖由民眾在網路上之發文、搜尋等活動，建構一組預測函數$F(t) = \{f_1^t(X), f_2^t(X), \dots, f_n^t(X)\}$，每一函數$f_i^t \in F(t)$ 均以預測 $R(t)$為目標，其預測結果為$R_i'(t)$。之後，再以模型融合概念整合 $F(t)$中之函數，以$(R_1'(t), R_2'(t), \dots, R_n'(t))$為輸入，產生一綜合預測值$R_{Fusion}'(t)$，期能對實際的流感就診率 $R(t)$提供更佳的預測。

以下以問題陳述為基礎，說明本節之流程與架構:

(1) 在問題陳述中的 X 可由多個變數組成，也就是 $X=(x_1,x_2,\dots,x_m)$，x_i 則為與流感就診率預測相關之變數(因子)(產生過程將在 3.1 節中介紹)。

(2) 有關各預測函數 f，在本研究將採用線性迴歸來建立。也就是說，F 中之單一預測函數f_i^t可用如下公式來表示:

$$R_i'(t) = f_i^t \ \ (X) = w_{i,0}(t) + w_{i,1}(t)x_1 + w_{i,2}(t)x_2 + \cdots + w_{i,m}(t)x_m \qquad (2)$$

其中$w_{i,0}(t)$為公式之常數項，其餘$w_{i,j}(t)$為流感相關變數 x_i 在預測函數 f_i^t 中之係數(權重)。根據式(2)，我們將針對不同的資料來源，配合不同的設定，產生一組預測函數 $f_1, f_2, .., f_n$，並分別為其決定合適的係數(細節將在 3.2 節中介紹)。

(3) 據此，每週便可產生一組預測值$(R_1'(t), R_2'(t), \dots, R_n'(t))$。本研究將透過累積這些預測值向量，經由模型融合(Model Fusion)，產生更穩定、準確的流感就診率預測結果(細節將在 3.3 節介紹)。

1. 利用網路社群歸納流感預測之相關因子

在疾管局公布就診率之前，身體不適的民眾可能在網路上留下不同的數位足跡(digital footprint)。例如，在網路社群中與朋友分享其個人或家人之病症、就診等細節，又如，透過入口網站搜尋與流感相關訊息，以了解或減緩身體不適。因此，這些訊息若經過整理，便可能從中找出與流感預測相關之因子(變數)。據此，本節將說明如何歸納

民眾在網路社群(如 twitter)發文以及 google 關鍵字搜尋等活動，以歸納與流感就診率預測相關之因子。

(一) 以網路社群發文歸納相關因子

為從網路社群之發文歸納預測相關因子，本研究以 Twitter 為對象，蒐集民眾在特定地區之發文，經過以下整理後，產生分析塑模所需之資料集。

(1) 定期蒐集發文: 利用 Twitter 提供之 API 定期由特定地區下載民眾的發文，儲存後以備後續分析。分析前需設定一關鍵詞集 $K=\{k_1, k_2,\ldots, k_m\}$，其中 k_i 為與流感相關之用詞(如 fever, headache, lung 等)。對此，本研究採用 Lampos 等人[16]所整理流感關鍵字集，內共含 74 個關鍵詞。

(2) 以關鍵詞正規化詞頻，建立資料集: 令 $t_j(k_i)$為第 j 週所有發文中包含關鍵詞 k_i 之文章數，以所有的關鍵詞為基礎進行統計，便可產生一向量$<t_j(k_1), t_j(k_2),\ldots, t_j(k_m)>$。直覺上，似可用向量來做為判別當週之流感熱度。但由於每週社群發文總數不同，分析時亦可能使用不同數量之關鍵詞，因此將利用以下公式再加以正規化:

$$s_j(k_i) = \frac{t_j(k_i)}{m \times n} \qquad (3)$$

其中$t_j(k_i)$為第 j 週所有發文中包含關鍵詞 k_i 之文章數，m為關鍵詞之總數(=|K|)，n為當周所有發文之總數(=|P_j|)。例如，當週發文總數共有 2000 則(n=2000)，關鍵字‘fever’在第 j 週共出在 894 篇文章中，則 t_j(‘fever’)=894。又假設關鍵詞總數 m=74，則’fever’在本週之正規化頻率為 s(‘fever’) = 894/(74*2000)。根據上述說明，便可以關鍵詞集 K 正規化詞頻為基礎，產生 *Twitter* 在第 j 週發文之流感預測資料向量D_j^T(如(4)所示)，其中$Rate_j$為官方所提供之當週流感就診率。

$$D_j^T(K) =< s_j(k_1), s_j(k_2) \ldots, s_j(k_m), Rate_j > \qquad (4)$$

假設資料蒐集期間共有 q 週，根據式(4)對所有蒐集資料進行分週統計，產生如表 1 之資料集$D^{Twitter} = \{D_1^T, D_2^T, \ldots, D_q^T\}$。此整理後的資料集將做為後續塑模時之訓練、測試之用。

表 1: 透過關鍵詞正規化詞頻產生 1~q 週之資料集($D^{Twitter}$)

#week\變數值	x_1	x_2	…	x_m	R
1*	$s_1(k_1)$	$s_1(k_2)$	…	$s_1(k_m)$	R_1**
2	$s_2(k_1)$	$s_2(k_2)$	…	$s_2(k_m)$	R_2
…	…	…	…	…	…
q	$s_q(k_1)$	$s_q(k_2)$	…	$s_q(k_m)$	R_q

*: 該列所對應之資料由$D_1^T(K)$表示，其餘各列依此類推

**: R_j 表示由官方公布之第 j 週實際流感就診率

(二) 以 Google 關鍵字搜尋歸納流感預測相關因子

預測時，若僅由單一資料來源(Twitter)建立模型，可能因某時期資料代表性不足，無法準確預測。由於使用者亦會透過其他管道(如 Google 關鍵字搜尋)，了解流感相關治療、預防等事宜，因此可將其做為預測之資料來源。除用以產生預測模型外，之後亦可與其他資料建立之模型結合，產生更佳之預測結果。圖 1 所示便為 Google Trend 對於 fever(發燒)、flu(流感)等字詞所提供之搜尋熱度統計(地區:英國，時間: 2012/5~2017/5)，可看出各種字詞在不同時間具有不同搜尋熱度。搜尋熱度之範圍介與 0~100 分，計算方式如下:將特定地區在特定時間內最熱門之搜尋字詞設為 100 分，50 分為 100 分熱門程度的一半，而 0 分則表示熱門程度不到 100 分字詞的 1%。

為歸納民眾有關流感的搜尋，本研究採用世衛組織公布之流感八個主要症狀做為搜尋關鍵詞集，K_g={發燒, 咳嗽, 頭痛, 疲倦, 喉嚨痛, 畏寒, 肌肉痠痛, 打噴嚏}。之後，透過 Google 蒐集每個關鍵詞(g_i)在指定時間內之搜尋熱度。因此，第 j 週蒐集的關鍵字搜尋熱度便可以式(5)來表示，其中$Rate_j$為官方所提供之當週流感就診率，$h_j(g_i)$為關鍵詞 g_i 在第 j 週的搜尋熱度。假設資料蒐集區間共有 q 週，便可產生一資料集$D^{Google} = \{D_1^G, D_2^G, \dots, D_q^G\}$，做為後續塑模預測之用。

$$D_j^G(K_g) =< h_j(g_1), h_j(g_2) \dots, h_j(g_8), Rate_j > \qquad (5)$$

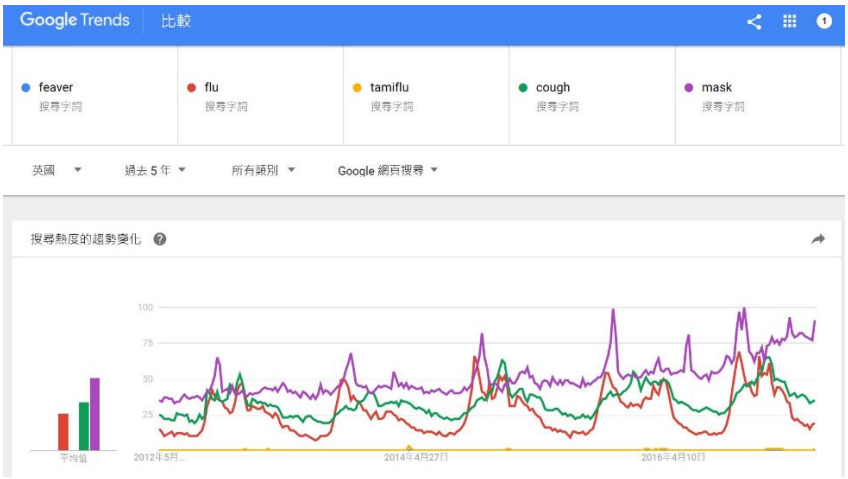


圖 1: 關鍵字 flu 在英國(2012/5~2017/5)之 Google 搜尋熱度變化
(資料來源: https://trends.google.com)

2. 建立流感就診率之預測模型

以前一節產生之資料集為基礎，本節將介紹如何建立各種流感預測模型。內容涵蓋如何以滑動視窗(sliding window)建立單一模型、如何考量延遲效應建立延遲模型，以及以模型融合概念組合多種不同預測模型，以產生更穩定的預測結果。

(一) 以線性迴歸建立單一預測模型

假設以週為時間單位，欲預測第 t 週之流感就診率($R(t)$)，訓練集資料大小為 b 週，則將使用第 t-1,t-2, …,t-b 週之資料來塑模。以 Twitter 資料集為例，將使用

$\{d_{t-1}^T, d_{t-2}^T, \ldots, d_{t-b}^T\}$做為訓練集，以產生預測公式。據此，此過程等同於求解以下之聯立方程式中之係數，其中 $w_i(t)$為第 t 週預測公式中變數 x_i 之係數，$w_0(t)$為預測公式中的常數項，$x_i(t\text{-}u)$為第 i 個自變數在第 $t\text{-}u$ 週的值，$R(t\text{-}u)$則為第 $t\text{-}u$ 週的實際就診率:

$$\begin{aligned} R(t-1) &= w_0(t) + w_1(t)x_1(t-1) + w_2(t)x_2(t-1) + \cdots + w_n(t)x_n(t-1) \\ R(t-2) &= w_0(t) + w_1(t)x_1(t-2) + w_2(t)x_2(t-2) + \cdots + w_n(t)x_n(t-2) \\ &\cdots\cdots \\ R(t-b) &= w_0(t) + w_1(t)x_1(t-b) + w_2(t)x_2(t-b) + \cdots + w_n(t)x_n(t-b) \end{aligned} \quad (6)$$

根據求得之 w_i 值，則本週之預測就診率 $R'(t)$便可以如下公式來求得:

$$R'(\mathrm{t}) = w_0(t) + w_1(t)x_1(t) + w_2(t)x_2(t) + \cdots + w_n(t)x_n(t) \quad (7)$$

同樣地，當欲預測第 t+1 週之流感就診率時，便可依照前述步驟求得另一組 w_i(t+1)值，並以如下公式進行預測:

$$R'(\mathrm{t}+1) = w_0(t+1) + w_1(t+1)x_1(t+1) + w_2(t+1)x_2(t+1) + \cdots + w_n(t+1)x_n(t+1) \quad (8)$$

假設預測週數由第 t 週至 t+r 週，則可產生一預測值序列($R'(t), R'(t+1),\ldots,R'(t+r)$)，再與實際就診率(($R(t), R(t+1),\ldots,R(t+r)$)進行關聯度分析(如使用皮爾森相關係數)，即可了解其關聯度。以圖 2 為例，圖中顯示以 Twitter 資料集建立預測公式之預測結果(訓練集資料範圍為 4 週)。其中之虛線為預測結果，實線為官方公布之實際就診率，二者之關聯度可達 0.866，顯示由 Twitter 做為資料來源具有可行性。

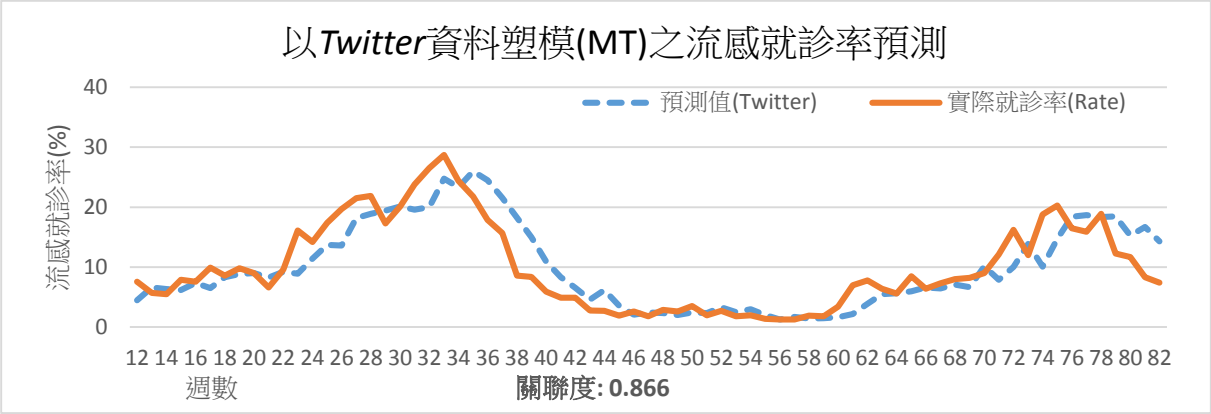


圖 2: 以 Twitter 資料集建立迴歸預測公式之預測結果

(二) 建立具有延遲特性之預測模型

由於流感具有潛伏期，且病人發病後可能延遲就醫，導致當週的 Twitter 發文內容以及 Google 關鍵字搜尋熱度，未必能反應至當週之就診率。因此，除上一節提出的預測方法，本研究將進一步考量延遲因素，以如下方式建立延遲預測模型：

(1) 考量延遲因素產生資料集: 假設當週之 Twitter 發文與搜尋熱度可能反應在 d 週後之就診率，則需將資料集 $D^{Twitter}$ 與 D^{google} 中之每一週 j 之資料向量$D_j^T(K)$與$D_j^G(K_g)$分別修改如下(參考公式(9),(10))。其中，$s(k_i)$及 $h(g_i)$分別代表第 j 週關鍵詞 k_i 出現的

文章數與關鍵詞 g_i 的搜尋熱度。值得注意的是，公式中各週資料向量 D_j 最後一元素並非當週實際流感就診率(R_j)，而是 d 週後(因延遲)的 R_{j+d}。

$$D_j^T(K) =< s_j(k_1), s_j(k_2) \dots s_j\left(k_{|K|}\right), R_{j+d} > \qquad (9)$$

$$D_j^G(K_g) =< h_j(g_1), h_j(g_2) \dots h_j\left(g_{|G|}\right), R_{j+d} > \qquad (10)$$

(2) 依照延遲週數與回溯週數建立預測模型：

若預測時間點為第 t 週之流感就診率，延遲週數為 d，訓練資料集大小為 b 週，由資料集(此處以 Twitter 資料集為例)中取出D_{t-1-d}^T, D_{t-2-d}^T,…, D_{t-b-d}^T，再透過線性迴歸計算預測公式之係數。

以下舉例說明延遲模型之資料集建立方式: 參考表 2，若本週為第 6 週(t=6)，訓練資料集大小為 4 週，延遲週數 d=1，則塑模時將採用D_{6-1-1}^T, D_{6-2-1}^T,…, D_{6-4-1}^T，也就是取出第 1 週至第 4 週的資料進行訓練。此外，由於延遲因素，第 j 週之資料向量會採用第 $j+1$ 週之實際就診率做為其最後一個元素。

表 2: 在延遲因子 d=1、訓練集大小為 4 週之狀況下，預測第 6 週流感就診率

#week	資料向量名稱	x_1	x_2	…	x_m	R
1	D_1^{T*}	$s_1(k_1)$	$s_1(k_2)$	…	$s_1(k_m)$	$\boldsymbol{R_2}$
2	D_2^T	$s_2(k_1)$	$s_2(k_2)$	…	$s_2(k_m)$	$\boldsymbol{R_3}$
3	D_3^T	$s_3(k_1)$	$s_3(k_2)$	…	$s_3(k_m)$	$\boldsymbol{R_4}$
4	D_4^T	$s_4(k_1)$	$s_4(k_2)$	…	$s_4(k_m)$	$\boldsymbol{R_5}$
5	D_5^T	$s_5(k_1)$	$s_5(k_2)$	…	$s_5(k_m)$	-
6	D_6^T	$s_6(k_1)$	$s_6(k_2)$	…	$s_6(k_m)$	**待測**
…		…	…	…	…	**…**

*: 此處假設使用 Twitter 資料集

(三) 模型融合(Model Fusion)

使用單一模型進行預測時，因資料集或學習演算法之特性，其效能可能受到限制。因此，本研究將使用模型融合方式，整合多種預測模型，將其預測結果以線性方式加以組合，並透過迴歸分析，找出其最佳的組合係數，相關做法詳述如下。

(1) 將多種模型的預測結果做為資料集：參考表 3，未進行模型融合前，本研究產生六種不同的預測模型。其中，前三個模型($M_{Twitter\_d0}$, $M_{Twitter\_d1}$, $M_{Twitter\_d2}$)為使用 Twitter 發文資料所建立，並配合不同的延遲週數(d=0, d=1, d=2)；後三個(M_{Google_d0}, M_{Google_d1}, M_{Google_d2})則採用 Google 關鍵字搜尋熱度，使用不同之延遲週數所建立。綜合以上六種模型之預測結果，配合官方提供之實際就診率，產生每週之資料向量如下：

$$D_j^{Fusion}(MF) =< mt_{d0}, mt_{d1}, mt_{d2}, mg_{d0}, mg_{d1}, mg_{d2}, R_j > \qquad (11)$$

其中 MF={$M_{Twitter\_d0}$, $M_{Twitter\_d1}$, $M_{Twitter\_d2}$, M_{Google_d0}, M_{Google_d1}, M_{Google_d2}}, 而$(mt_{d0}, mt_{d1}, mt_{d2}, mg_{d0}, mg_{d1}, mg_{d2})$分別對應前述六種模型在第 j 週之預測結果，R_j

則是當週實際的流感就診率。假設資料蒐集區間為 1~q 週，便可建立融合資料集 $D^{Fusion} = \{D_1^{Fusion}, D_2^{Fusion}, \dots, D_q^{Fusion}\}$。

表 3: 模型融合所使用之六種不同流感預測模型

模型編號	模型名稱	延遲週數(d)	說明
1	$M_{Twitter\_d0}$	0	以 $D^{Twitter}$ 資料集，配合迴歸所建立之預測模型, 無延遲週數。
2	$M_{Twitter\_d1}$	1	同上，但延遲週數為 1
3	$M_{Twitter\_d2}$	2	同上，但延遲週數為 2
4	M_{Google_d0}	0	以 D^{Google} 資料集，配合迴歸所建立之預測模型，無延遲週數。
5	M_{Google_d1}	1	同上，但延遲週數為 1
6	M_{Google_d2}	2	同上，但延遲週數為 2

(2) 建立融合模型(M_{Fusion})：

根據步驟(1)所建立資料集，若預測時間點為第 t 週之流感就診率，資料蒐集區間為第 s 週至第 t-1 週(s≤t-1)，則可由資料集 D^{fusion} 中取出$D_s^{Fusion}, D_{s+1}^{Fusion}, \dots, D_{t-1}^{Fusion}$，產生如下之融合預測模型：

$$R'(t) = w_{t_{d0}}(t) * mt_{d0}(t) + w_{t_{d1}}(t) * mt_{d1}(t) + \cdots + w_{g_{d2}}(t) * mg_{d2}(t) + w_0(t) \quad (12)$$

其中，$R'(t)$為模型產生之流感就診率預測值，w_i為模型 i 預測結果之權重。

3. 以模型融合為基礎之流感就診率預測方法運作流程

由上述各節可知，本研究提出方法之特點為: 結合不同資料來源，並產生多種不同的單一模型，最後透過模型融合來加以整合，期能產生更穩定、準確的預測結果。圖 3 總結了本研究提出之流感就診率預測方法之運作流程:

步驟 1: 首先，透過不同資料來源(Twitter 發文, Google 熱度搜尋)蒐集與流感相關資訊。

步驟 2: 配合不同關鍵詞集整理擷取的資料，以產生訓練、測試所需之資料集($D^{Twitter}$, D^{Google})。

步驟 3: 經過迴歸分析，以不同之延遲設定產生 6 種不同偵測模型($M_{Twitter\_d0}$, $M_{Twitter\_d1}$, $M_{Twitter\_d2}$,M_{Google_d0}, MG_{Google_d1}, MG_{Google_d2})。

步驟 4: 最後，透過模型融合概念，將各種模型之預測結果建立之資料集，以線性迴歸進行融合，產生最後的預測結果(R'_{Fusion})。

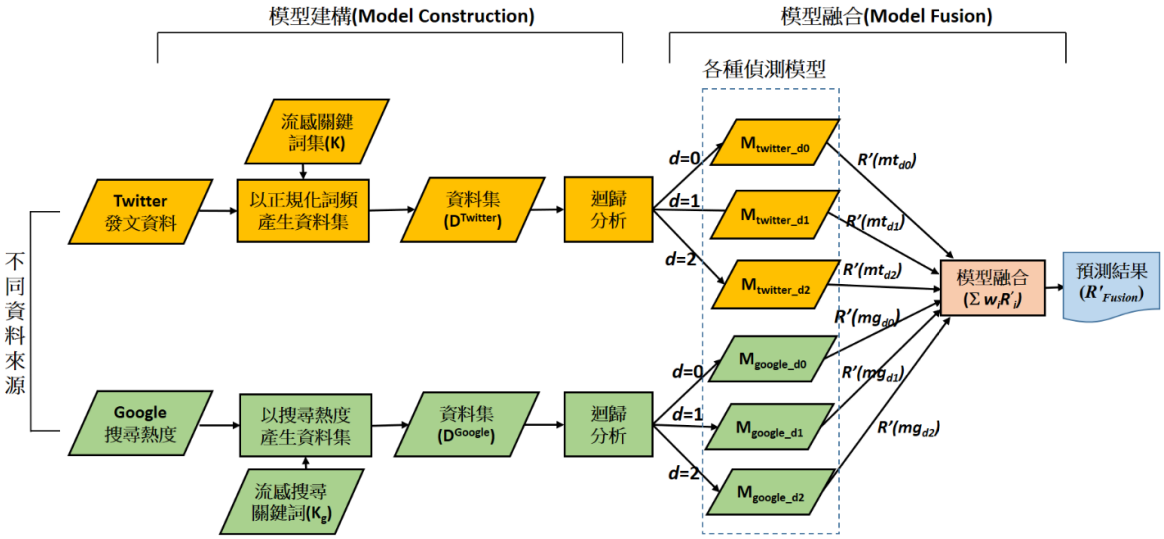


圖 3: 本研究提出之流感就診率預測方法流程圖

理論上，在應用上述方法時，應以最後的預測結果(R'_{Fusion})作為主要參考依據。但考量網路社群的多變與流感流行的不確定性，在實際應用時，人類專家可依照各種單一模型之預測特性，配合融合結果一起判讀未來可能的流行趨勢。

四、實驗結果

為驗證提出方法之有效性，本節使用實際下載資料進行實驗，分別針對 Twitter 流感預測模型、Google 搜尋熱度模型、延遲模型與模型融合等做法，進行效能評估與結果討論。

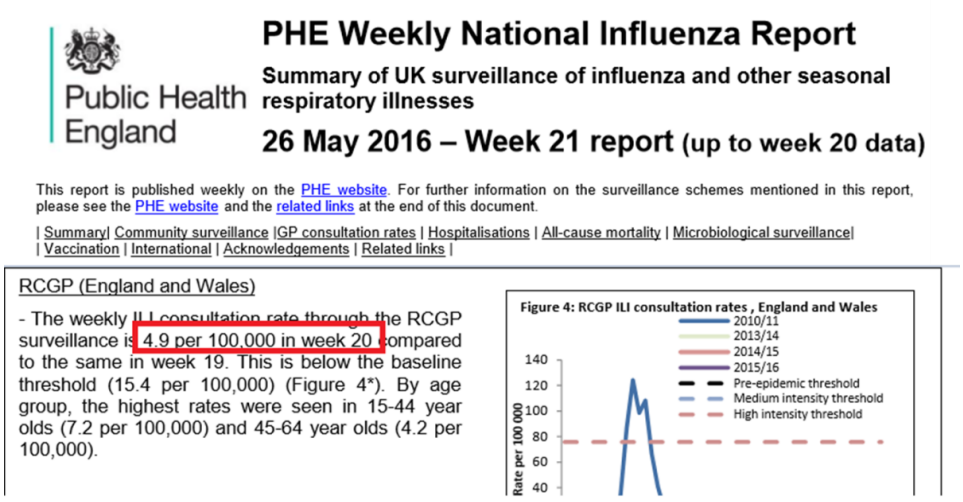
1. 實驗設定

為進行實驗，本研究擷取 Twitter 發文進行分析，並以英國地區為對象。Twitter 為英國當地主流網路社群之一，其中的發文應可適當反映民眾的心情，做為流感預測依據。為建立資料集，本研究由以下三種不同來源蒐集資料，期間為 2015/8 – 2017/3 (共 82 週):

(1) 蒐集 Twitter 發文: 透過呼叫 Twitter Streaming API，取得每日英國地區之發文(tweets)。由於官方公布之流感就診率以週為單位，因此本研究將 Twitter 上每日發文以週為單位累計(每日約 3 千餘筆，每周可蒐集約 2 萬餘筆)。上述資料經過一年多的蒐集(2015/8/13 – 2017/3/15)，共蒐集超過 160 萬筆的 tweets 發文。

(2) 蒐集 Google 關鍵字搜尋熱度: 透過 Google 網站取得與(1)相同時間與地區之關鍵字搜尋熱度。

(3) 蒐集流感就診率: 由英國官方(Public Health England)網站下載每週實際之流感就診率(如圖 4 所示)，做為訓練與測試時之實際值。


Public Health England

PHE Weekly National Influenza Report

Summary of UK surveillance of influenza and other seasonal respiratory illnesses

26 May 2016 – Week 21 report (up to week 20 data)

This report is published weekly on the PHE website. For further information on the surveillance schemes mentioned in this report, please see the PHE website and the related links at the end of this document.

| Summary| Community surveillance |GP consultation rates | Hospitalisations | All-cause mortality | Microbiological surveillance| | Vaccination | International | Acknowledgements | Related links |

RCGP (England and Wales)

- The weekly ILI consultation rate through the RCGP surveillance is 4.9 per 100,000 in week 20 compared to the same in week 19. This is below the baseline threshold (15.4 per 100,000) (Figure 4*). By age group, the highest rates were seen in 15-44 year olds (7.2 per 100,000) and 45-64 year olds (4.2 per 100,000).

圖 4: 英國官方公布之流感就診率統計數據

(資料來源: https://www.gov.uk/government/statistics/weekly-national-flu-reports)

2. 單一模型與融合模型之流感就診率預測結果

本節將介紹以 Twitter 發文、Google 搜尋熱度為基礎之各種單一預測模型之效能，其中包含無延遲 (d=0)、延遲一週(d=1)、延遲 2 週(d=2)等三種模型。

(一) 以 Twitter 發文資料建立流感就診率預測模型

本小節以 Twitter 發文資料做為資料來源，配合不同的延遲週數，建立流感就診率預測模型($M_{Twitter\_d0}$, $MT_{Twitter\_d1}$, $MT_{Twitter\_d2}$)。實驗時，採用 Lampos 等人[16]所提出的關鍵詞集(共含 74 個關鍵詞)進行統計，塑模訓練集資料大小均設為 4 週。此外，為與後續模型融合之實驗結果時間區間一致，將針對 12-82 週的結果進行比較(與實際值)。為使結果比較更為清楚，我們將全部資料(12-82 週)再分為二個時期，分別是區間 1(12-47 週)，與區間 2(48-82 週)，其中分別包含流感就診率大幅的上升、下降時期，唯區間 1 較為劇烈(劇升降)，區間 2 較為緩和(緩升降)。

(a) $M_{Twitter\_d0}$ 模型: 圖 5 所示為無延遲模型之偵測結果，其與實際就診率之總體關聯度為 0.866(參考圖下方數據)，屬高度相關。顯示以 Twitter 發文資料進行流感預測，具有實用價值。此模型在「區間 1」與「區間 2」之關聯度則有差異，分別為 0.868 與 0.828，顯示此模型在區間 2(緩升降)的預測結果不如區間 1(劇升降)。上述結果顯示 Twitter 發文與流感之相關性，在劇烈流行期較一般流行期更具有參考性。

(b) $M_{Twitter\_d1}$ 模型:當將塑模延遲週數設定為 1 之實驗結果如圖 6 所示，區間 1、區間 2、全區間與實際值之關聯度分別 0.847, 0.729, 0.822。由數據可看出，與無延遲模型相較，其總體關聯度(0.822)明顯下降。此外，區間 1 的關聯度明顯優於區間 2，此點則與無延遲模型一致。

(c) $M_{Twitter\_d2}$ 模型:當延遲週數增加為 2 時，其預測效果則更進一步下降(參考圖 7)，總體關聯度僅為 0.804，區間關聯度則均低於 0.8。此結果顯示，Twitter 發文資料與流感就診率，並無明顯的延遲現象，也就是當週的發文最能反應當週的就診率。

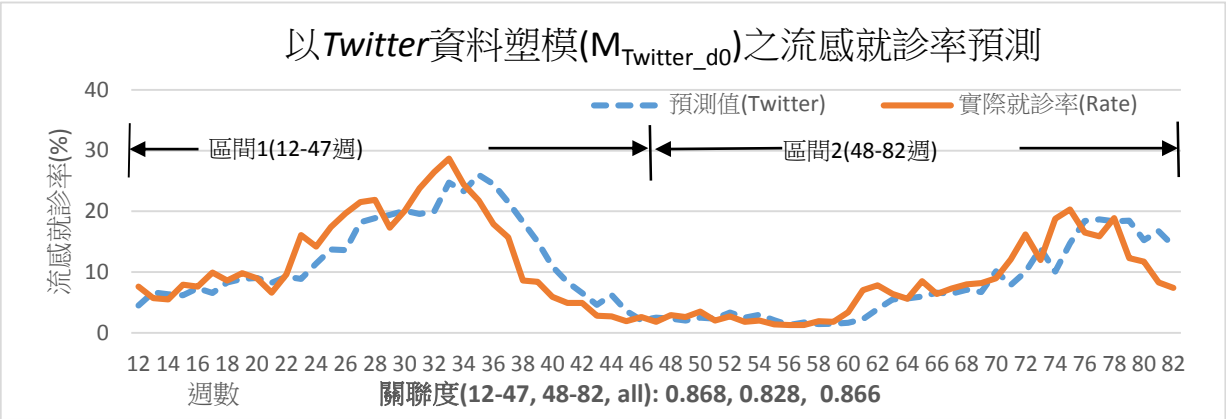


圖 5: 以 Twitter 發文資料建立預測模型(無延遲週數, d=0)之流感就診率偵測結果

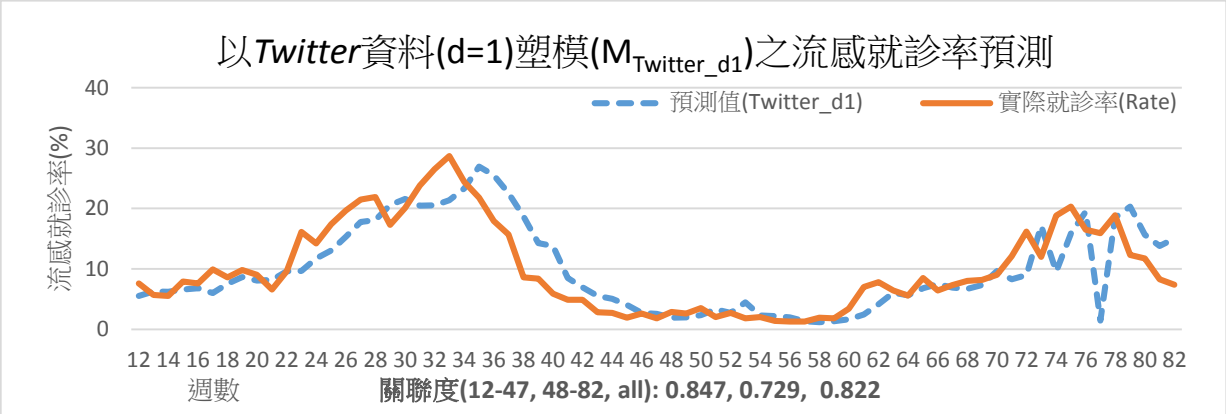


圖 6: 以 Twitter 發文資料建立預測模型(延遲週數 d=1)之流感就診率偵測結果

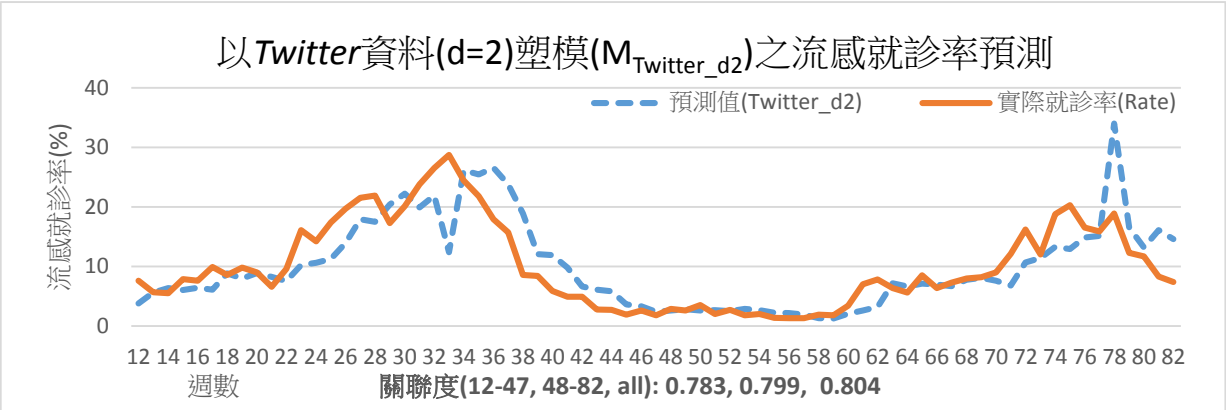


圖 7: 以 Twitter 發文資料建立預測模型(延遲週數 d=2)之流感就診率偵測結果

(二) 以 Google 搜尋熱度建立流感就診率預測模型

本小節則以世衛組織公布之 8 個流感關鍵詞(參考 3-2 節)在 Google 之搜尋熱度為資料來源，配合不同的延遲週數，建立流感就診率預測模型(MG, MG_{d1}, MG_{d2})，其餘實驗設定則與上一節之 Twitter 模型相同。

(1) MG 模型: 圖 8 所示為無延遲模型(MG)之偵測結果，與實際就診率之總體關聯度僅為 0.772，雖仍屬高度相關，效果明顯不如 Twitter 模型。但值得注意的是，此模型在

「區間 1」與「區間 2」之關聯度分別為 0.714 與 0.823，此結果與 Twitter 模型相反，顯示 Google 模型在區間 2(緩升降)的預測結果明顯優於區間 1(劇升降)。換言之，在預測效果上，Twitter 模型與 Google 模型可產生互補效果。

(2) MG_{d1} 模型: 當將塑模延遲週數設定為 1(MG_{d1})之實驗結果如圖 9 所示，區間 1、區間 2、全區間與實際值之關聯度分別 0.823, 0.734 與 0.801。由數據可看出，區間 1 的關聯度明顯優於區間 2，此結果與 Twitter 模型類似。

(3) MG_{d2} 模型: 當延遲週數增加為 2 時(MG_{d2})，其預測效果則明顯提升(參考圖 10)，總體關聯度高達 0.859，明顯優於 MG 與 MG_{d1}。值得注意的是，其在區間 2 的關聯度高達 0.950，為前述所有模型之最優。此結果顯示，Google 流感關鍵字搜尋熱度對於緩升降區間具有優良預測效果，且此結果在配合延遲塑模時才能展現，顯示延遲塑模的必要性與效果。

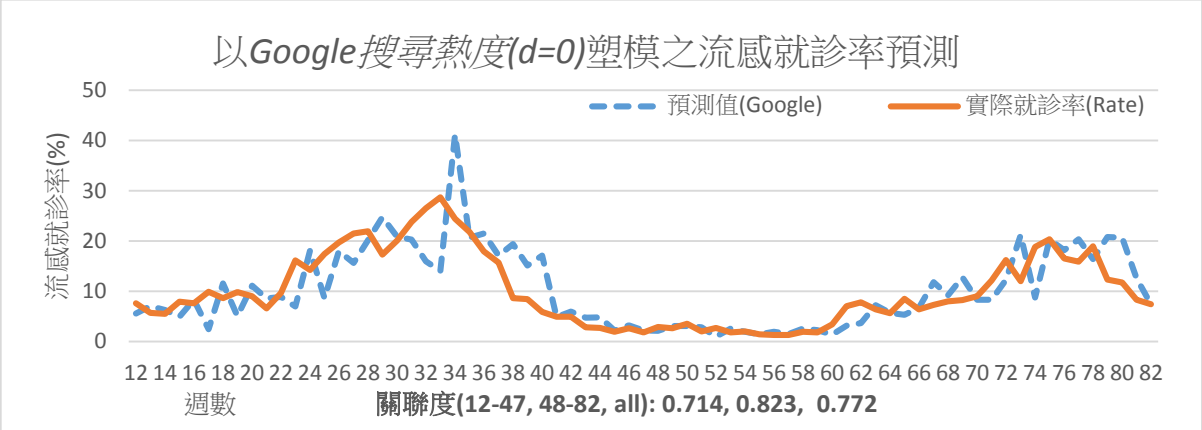


圖 8: 以 Twitter 發文資料建立預測模型之流感就診率偵測結果

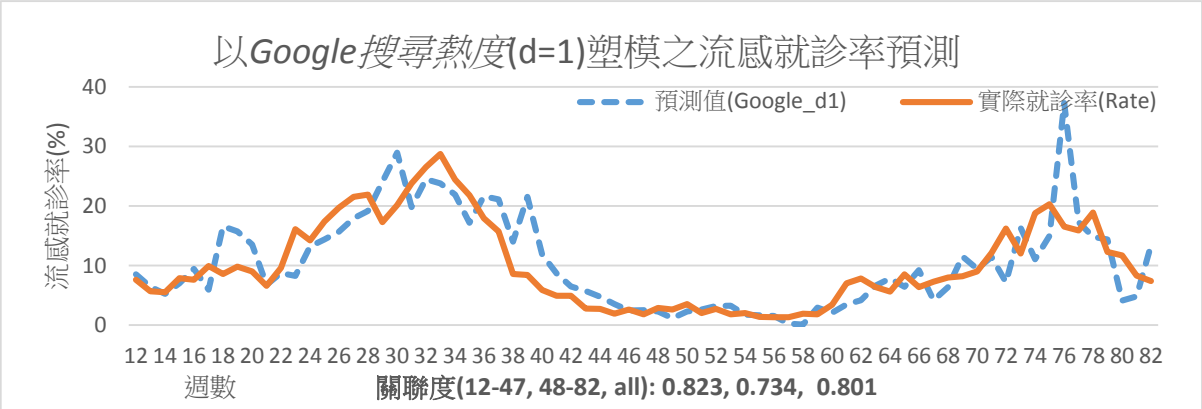


圖 9: 以 Google 搜尋熱度建立預測模型(延遲週數 d=2)之流感就診率偵測結果

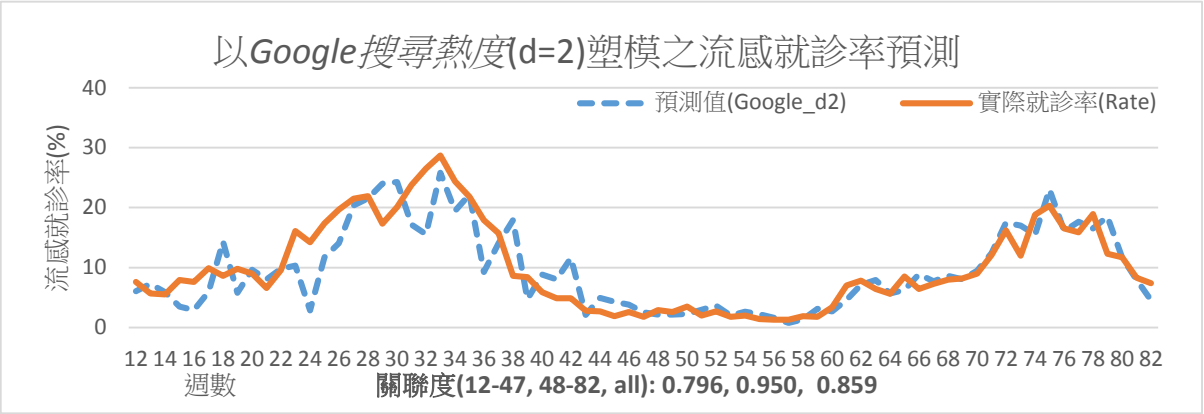


圖 10: 以 Google 搜尋熱度建立預測模型(延遲週數 d=2)之流感就診率偵測結果

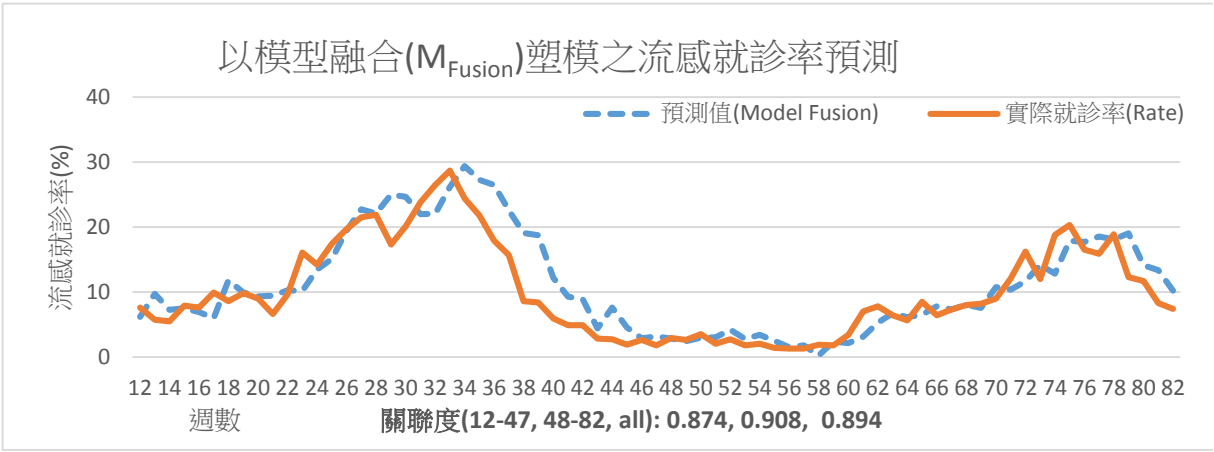


圖 11: 以模型融合塑模(參考 3-2 節)之流感就診率偵測結果

3. 模型融合(Model Fusion)之實驗結果

本節將驗證 3.2 節所介紹之模型融合之預測效果，並與單一預測模型進行比較。由於本研究以線性迴歸方式組合各種模型，若迴歸結果之預測公式為常數，則顯示當週以模型融合預測並無意義，因此將其結果予以排除。據此，本實驗只針對 12 週以後的結果進行統計。圖 11 所示為運用模型融合方法之流感就診率預測效果，其總關聯度達 0.894，優於之前所介紹之六種單一模型。此外，區間 1 與區間 2 之關聯度亦達 0.874 與 0.908，表現均衡，並無如前述 Twitter 或 Google 預測模型在此二區間產生的明顯差異。上述結果顯示，本研究所運用之模型融合方法，確能整合各種模型的特質，提供更精準、穩定的預測結果。

4. 結果討論

總結上述實驗結果，表 4(與圖 12)所示為各種模型在不同時間區間內之預測關聯度比較，以下將對此進行討論:

(1) 由表中可知，與實際流感就診率相較，融合模型(M_{Fusion})具有最高的總關聯度(0.894)，驗證了本研究提出方法之有效性。次高的則為 Twitter 無延遲模型($M_{Twitter\_d0}$)之 0.867，顯示英國地區當週的 Twitter 發文與流感就診率確有明顯關係。

(2) 當加上延遲因子後，Twitter 系列模型之預測關聯度反而逐次下降(0.867→ 0.822→ 0.804)，顯示運用 Twitter 發文資料塑模時，應無需特意加上延遲因子。上述現象在以 Google 關鍵字搜尋熱度建立之模型中則相反，其關聯度依照延遲週數遞增(參考表 4 第 2 列後 3 個欄位)，依序為 0.772→0.801→0.859。此結果顯示不同資料來源之特質不同，需以不同方式來運用，不宜一視同仁。

(3) 在具有劇烈升降特質之區間 1(12-47 週)，融合模型 M_{Fusion} 仍具有最高關聯度 0.874，次佳的為 $M_{Twitter\_d0}$。在緩升降的區間 2(48-82 週)中，延遲週數 2 週的 Google 搜尋熱度模型 M_{Google_d2} 具有最高的關聯度(0.950)，融合模型(M_{Fusion})的 0.908 為次高，二者均超過 0.9，之前表現較優良之 $M_{Twitter\_d0}$ 則僅有 0.828。由上可知，與其他模型相較，融合模型具有相對的穩定性，可應用於不同的流感流行期間。

(4) 比較 Twitter 發文資料與 Google 關鍵字熱度二種資料來源，大體而言 $M_{Twitter}$ 系列模型在劇烈升降的區間 1 有較佳表現，M_{Google} 系列的三種模型則較有利於緩升降區(區間 2)之預測。由上結果可知，使用多種不同資料來源塑模確有其必要性，若在配合模型融合，將可產生更準確的預測結果。

表 4: 融合模型(M_{Fusion})與其他 6 種單一模型之流感就診率預測關聯度比較

關聯度/預測模型	M_{Fusion}	$M_{Twitter\_d0}$	$M_{Twitter\_d1}$	$M_{Twitter\_d2}$	M_{Google_d0}	M_{Google_d1}	M_{Google_d2}
全部週數(12-82 週)	**0.894**	0.867	0.822	0.804	0.772	0.801	0.859
區間 1(12-47 週)	**0.874**	0.868	0.847	0.783	0.714	0.8237	0.796
區間 2(48-82 週)	0.908	0.828	0.729	0.799	0.823	0.734	**0.950**

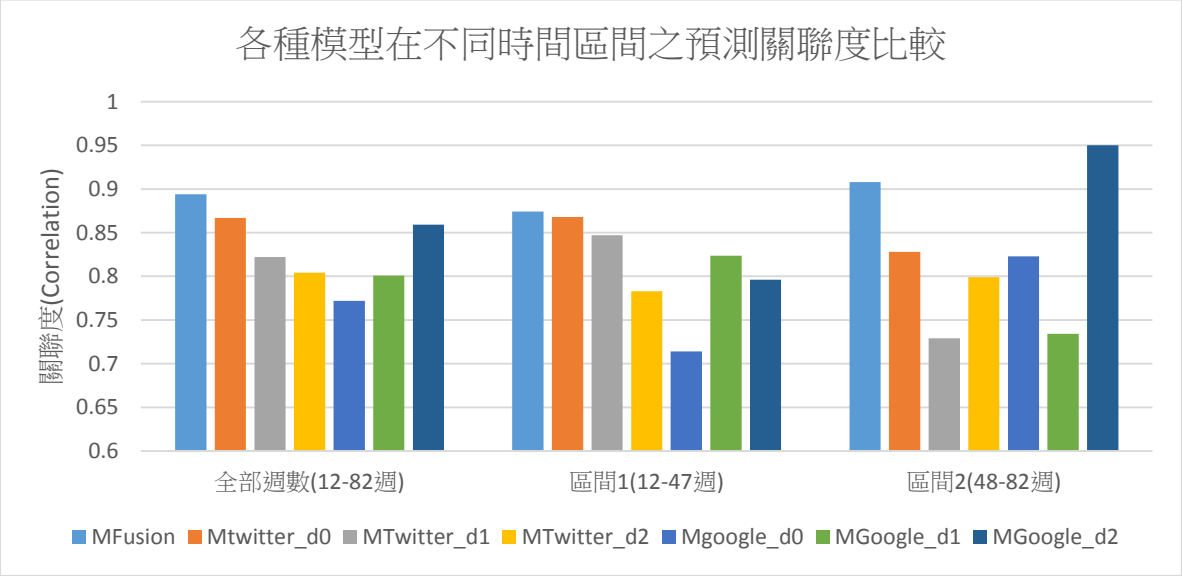


圖 12: 融合模型(M_{Fusion})與其他 6 種單一模型之流感就診率預測關聯度比較

五、結論與未來工作

由於交通便利，人口流動頻繁，導致流感每年均在各國造成流行，嚴重影響人們的生活。但若能及早示警，使政府及相關單位及早進行準備(防疫宣導、疫苗施打等預防措施)，便可有效減少大流行的發生。為監測流感疫情，各國疾管局通常藉由臨床就診通報來彙整資料，但可能產生一至二週的延遲，對於傳播快速的流感，顯然緩不濟急。有鑑於此，本研究運用 Twitter 發文與 Google 搜尋熱度，建立多種不同特質的預測模型，並透過模型融合加以整合，提升預測的準確性與穩定性。此外，我們也考量疫情因潛伏期將延遲因素納入，建立對應的延遲模型。

我們蒐集英國地區約 82 週(2015/8~2017/3)超過 160 萬筆的 Twitter 發文，及同時期的 Google 關鍵字搜尋熱度資料，經過處理後進行實驗。結果顯示，與各種單一預測模型相較，提出之方法具有最高的預測關聯度(0.894)，明白顯示其有效性。為了解各種預測模型之穩定性，再將資料依照流行程度分為「劇升降」區與「緩升降」區進行統計。結果顯示融合模型在二區域分別具有的最高(0.874)與次高(0.908)之關聯度，而其他單一模型則呈現不一致之預測效果，顯示模型融合能有效提升預測之穩定性。總結上述結果，本研究提出之方法確能補足傳統流感統計延遲之缺陷，提供即時、實用之預測結果。

本研究雖以英國地區為主要分析對象，在未來的發展中，若能克服相關技術困難(例如，網路社群業者對於外部搜尋之限制)，可將相同方法運用於其他地區之流感監測。其次，我們將嘗試多種可能提高準確度之方法，包括使用不同關鍵字集、使用不同學習方法來塑模、塑模時，採用不同區間資料等。此外，官方僅提供每週一次的數據，若能透過近似法提供每日提供就診率，將可大幅提高預測之即時性。最後，目前僅以 Twitter 及 Google 關鍵字搜尋熱度來進行分析，若能分析本地其他來源(如高風險族群感染人次)，應可進一步準確預測流感帶來的影響。

參考文獻

[1] Akay, A., et al., “A Novel Data-Mining Approach Leveraging Social Media to Monitor Consumer Opinion of Sitagliptin,” IEEE Journal of Biomedical and Health Informatics, Vol. 19, No. 1, Jan. 2015, pp. 389-396.

[2] Bardak, B., & Tan, M., ” Prediction of influenza outbreaks by integrating Wikipedia article access logs and Google flu trend data”. Bioinformatics and Bioengineering (BIBE), 2015 IEEE 15th International Conference.

[3] Byrd, K., Mansurov, A., Baysal. O., “Mining Twitter Data for Influenza Detection and Surveillance”. 2016 IEEE/ACM International Workshop, 2016, pp. 43-49.

[4] Chen, J., et al., "Big Data based fraud risk management at Alibaba," The Journal of Finance and Data Science 1 (2015) 1-10.

[5] Cohen, J., "Statistical Power Analysis for the Behavioral Sciences", Chapter 3, 1977, pp.77-81.

[6] Li, S.-H., et al., “Identifying the signs of fraudulent accounts using data mining techniques,” Computers in Human Behavior, Computers in Human Behavior, vol. 28, 2012, pp. 1002–1013.

[7] Ginsberg, J. et al, “Detecting influenza epidemics using search engine query data”, Nature, vol. 457, 2009, pp. 1012-1014.

[8] Grover, S., & Aujla, G. S., “Twitter data based prediction model for influenza epidemic”. 2015 the 2nd International Conference, 2015, pp. 879-879.

[9] Eysenbach, G., “Infodemiology: tracking flu-related searches on the web for syndromic surveillance”, AMIA: Annual Symposium Proceedings, 2006, pp. 244–248.

[10] Fox, S., “Online Health Search 2006” , Pew Internet & American Life Project, 2006.

[11] Google Trends, https://trends.google.com/trends/explore, last retrieved 2018/2/13.

[12] He, W., Zha S., and Li, L., “Social media competitive analysis and text mining: A case study in the pizza industry,” International Journal of Information Management, vol. 33, 2013, pp. 464-472.

[13] Huang, S., et al., "A Hybrid Multigroup Coclustering Recommendation Framework Based on Information Fusion," ACM Trans. on Intelligent Systems and Technology, Vol. 6, No. 2, Mar. 2015.

[14] Hulth, A., Rydevik, G. & Linde, A., “Web Queries as a Source for Syndromic Surveillance”, PLoS ONE 4(2): e4378. doi:10.1371/journal.pone.0004378, 2009.

[15] Johnson, H. A. et al., “Analysis of Web access logs for surveillance of influenza.”, MEDINFO, 2004, pp. 1202–1206.

[16] Lampos, V., & Cristianini, N., ” Tracking the flu pandemic by monitoring the Social Web”. The 2nd International Workshop on Cognitive Information Processing, 2010, pp. 411-416.

[17] Lampos, V., and Cristianini, N., “Nowcasting Event from the Social Web with Statistical Learning,” ACM Trans. On Intelligent Systems and Technology, Vol. 3, No. 4, Sep. 2012, pp. 72:1~72:22.

[18] Lampos, V., Preotiuc-Pietro, D., and Cohn, T., A user-centric model of voting intention from social media, The 51st annual meeting of the association for computational linguistics, 2013, pp. 993-1003.

[19] Polgreen, P. M., Chen, Y., Pennock, D. M. & Forrest, N. D., “Using internet searches for influenza surveillance”, Clinical Infectious Diseases, vol. 47, 2008, pp. 1443–1448.

[20] Prakash, B. A., ”Prediction Using Propagation: From Flu Trends to Cybersecurity”, IEEE Computer Society, 2016, pp. 84-88.

[21] Wikipedia, 1918 flu pandemic, https://en.wikipedia.org/wiki/1918_flu_pandemic, last retrieved 2018/2/13.

[22] Wikipedia, Influenza, https://en.wikipedia.org/wiki/Influenza, last retrieved 2018/2/13.

[23] Witten, I., Frank, E., and Hall, M., ” Data Mining: Practical Machine Learning Tools and Techniques (3rd Edition)”, 2011.

[24] World Health Organization (WHO), http://www.who.int/mediacentre/factsheets/fs211/en/ , last retrieved 2018/02/13.

[25] 賴淑寬等人, “我國 H1N1 新型流感四大監測體系簡介,” 疫情報導，第 26 卷，第 14 期，2010: P195 - 202。