基於時間序列的網路輿情預測模型
An Online Public Opinion Forecast Model Based on Time Series

程輝 Hui Cheng, 劉雲 Yun Liu
北京交通大學通信與資訊系統北京市重點實驗室
07111013@bjtu.edu.cn, liuyun@bjtu.edu.cn

摘要

本文根據社會慣性理論和網路輿情在時間上的延續性，提出了一個從時間和數量的角度對網路輿情發展趨勢進行預測的模型。模型用網路上帖子數量隨時間變化的趨勢來表示網路輿情的發展趨勢，使用時間序列中常用的自回歸移動平均模型（Autoregressive Integrated Moving Average, ARIMA）對這個趨勢進行了預測。

關鍵字：網路輿情、時間序列、預測、自回歸移動平均模型。

Abstract

This paper presents a model to forecast the trend of online public opinion from the view of time and number, which bases on social inertia theory and time continuity of online public opinion. The model uses the trend of posts numbers with time in a forum to show the development trend of online public opinion, and forecasts this trend using ARIMA model which is a staple model in time series.

Keywords: Online Public Opinion, Time Series, Forecast, Autoregressive Integrated Moving Average Model (ARIMA Model).

1 緒論

網路發展到今天，網路的規模、使用者數量，特別是網路上的內容都在以爆炸性的速度增長。基於網路內容的網路輿情發展就呈現出了幾個特點：

(1)網路上的話題具有突發性，並可在短時間內產生大量的討論；
(2)網路上的話題傳播迅速，話題會通過不同的線民很快的傳播到各個網站；
(3)網路上的熱點話題會持續較長時間，大量線民會反覆的參與到討論中；
(4)網路討論影響的範圍不斷的擴大，並在一定程度上影響了人們的日常生活。

網路輿情的突發性和快速傳播的特性使其成為了社會輿論的一種快速的反應形式，網路輿情已經開始對現實社會產生一定的影響，因此提前預測網路輿情的發展趨勢，並對網路輿情的發展加以引導，對社會的穩定有著重要的意義。

2 相關研究

網路輿情預測屬於網路輿情分析方法的範疇，當前的網路輿情分析方法包括熱點話題、敏感話題識別、傾向性分析、主題跟蹤、自動摘要、突發事件分析、報警系統、統計報告等幾大類別，並且有了一定的研究成果，但是在網路輿情預測方面，尚未發現有研究提出成熟可靠的預測分析方法，因此儘快尋找一種成熟並且有一定可靠性的預測分析方法就顯得格外的重要。

人類社會普遍存在的一個規律是慣性規律，在任何事件的發生、發展的過程中，都存在一定的慣性，也可以稱為延續性，即大部分事件都不會憑空產生，同樣不會憑空消失，事件的發展都與事件已經存在的部分有某些聯繫，從另一個方面說，就是事件都是由歷史資料驅動著向前發展的，所以對歷史資料進行分析，研究其發展的規律，就能在一定程度上預測出事件的後續的發展趨勢。網路話題同樣具有這個特性，特別是網路上的熱點話題，熱點話題具有持續性，也就是上文所說的慣性，所以可以通過慣性原理進行網路熱點話題的預測。

經濟預測方面，經過多年的發展，各種相關理論都發展的比較成熟，其中時間序列是應用最廣泛的一種預測分析方法，時間序列是指同一種現象在不同時間的觀察值排列而成的一組數位序列。時間序列預測方法是指根據某個現象過去一段時間內的發展情況來預測其在未來的一定時間內的發展趨勢，即根據某個現象的歷史資料來揭示其隨時間的發展規律，並根據這個規律對該現象未來的發展做出預測。即時間序列的後續發展和前期資料有密切的聯繫，也就是前期資料決定了時間序列發展的大體方向，所以時間序列模型也是符合慣性原理的，所以可以使用事件序列的模型進行網路輿情發展趨勢的預測。由博克斯（Box）和詹金斯（Jenkins）於20世紀70年代提出的自回歸移動平均模型（Autoregressive Integrated Moving Average Model，簡稱ARIMA模型[1]）在眾多的基於時間序列預測方法中有著比較重要的地位，關於經濟量的預測方法大多使用ARIMA模型作為預測的基礎模型，然後進行預測。

ARIMA方法又可以表示為*ARIMA*(*p, d, q*)過程，它可以分為兩部分：*AR*(*p*)過程和*MA*(*q*)過程[2][3]。

MA(*q*)過程稱為移動平均過程，如公式(1)：

$$y_t=\mu+\varepsilon_t+\theta_1\varepsilon_{t-1}+...+\theta_q\varepsilon_{t-q} \quad (1)$$

其中 ε_t 為隨機干擾項，μ 為 y_t 的均值，$\theta_1,\theta_2,\cdots,\theta_q$ 為移動平均參數，y_t 為一個平穩的隨機序列。

$AR(p)$過程又稱為自回歸過程，如公式(2)：

$$y_t=\varphi_1 y_{t-1}+\ldots+\varphi_p y_{t-p}+\delta+\varepsilon_t \qquad (2)$$

其中 δ 為常數項（和 y_t 的均值有關），$\Phi_1,\Phi_2,\ldots,\Phi_p$ 為自回歸參數。

然而許多隨機過程不能純粹的用 $AR(p)$ 或者 $MA(q)$ 來表示，並且隨機序列也不一定是平穩的，因此提出了 $ARIMA(p, d, q)$ 過程，它包含了 $AR(p)$ 和 $MA(q)$ 兩個過程，並且在計算之前做了 d 次差分，使序列變為一個平穩的隨機序列，如公式(3)：

$$\varphi(B)w_t=\delta+\theta(B)\varepsilon_t \qquad (3)$$

其中 $\Phi(B)=1-\Phi_1B-\Phi_2B^2-\ldots-\Phi_pB^p$ 為自回歸運算元，$\theta(B)=1+\theta_1B+\theta_2B^2+\ldots+\theta_qB^q$ 為移動平均運算元，$w_t=\Delta d y_t$ 為 y_t 的 d 階差分序列。

3 預測模型詳細設計

模型是在我們的網路輿情分析系統中實現的，網路輿情分析系統分為幾個部分，首先是網路爬行器，也就是通常提到的網路爬蟲，網路爬蟲將抓取的互聯網網頁交給網路內容分析器，分析得到後續處理所需要的內容，比如作者、時間、標題、內容、閱讀人數、回覆人數等等，並將所有的資料存入資料庫；接下來是對所有內容的聚類、熱點話題發現模組，這也是預測模組的兩個重要的預處理過程，此外網路輿情分析系統還有敏感話題發現、資訊搜索、資訊查詢、話題追蹤等一系列的功能，網路輿情預測是分析系統的最後一個模組，是負責預測網路上的熱點話題在接下來一段時間內發展的趨勢。

在系統進行測試的初期，我們主要是針對互聯網上的論壇進行分析，下面就以互聯網上論壇為例，介紹預測模型。預測過程分為兩個階段，流程如圖一所示：

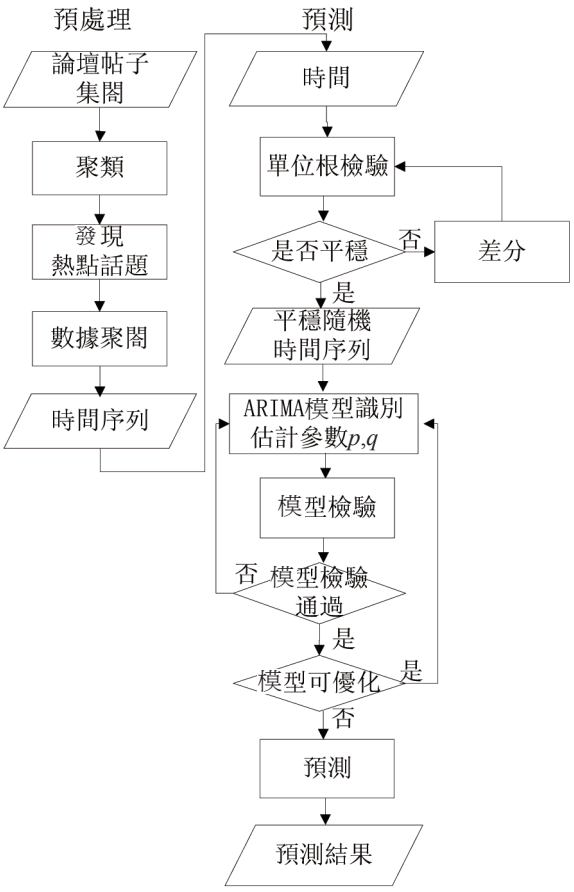
對於論壇，我們預測的物件為未來一段時間內論壇中的發帖數量，包括帖子的回覆數量。由於論壇上的帖子數量眾多，且沒有規律，因此我們首先要對所有的帖子進行預處理，將其轉換為預測所需的時間序列的格式。預處理過程分為以下三個步驟：

(1)對論壇上的帖子進行聚類。單個帖子時間上具有突發性，不具有普遍的規律，所以我們對論壇上所有討論同一個話題的帖子總量進行預測。聚類過程就是將所有描述同一個話題的帖子聚合到同一個類別中[4]。

(2)發現熱點話題。論壇上的帖子數量多，並且所對應的話題也不盡相同，大部分話題涉及的範圍很小，持續時間很短，對於這種話題沒有預測的必要，因此需要找到論壇中的熱點話題進行預測[5]。

(3)對論壇上的帖子進行資料聚合。使用時間序列模型進行預測，輸入為時間序列，因此需要將論壇上的帖子進行資料聚合，得到一個時間序列，每個時刻的值為到當前時刻為止，論壇上所有關於某個話題的帖子及其回覆的總量。

預處理過程得到的結果就是使用ARIMA模型預測所需要的時間序列。



圖一　應用時間序列進行網路輿情預測的流程

在預測的過程中使用的是標準的ARIMA模型，使用ARIMA模型進行預測分為以下四步驟，過程如下：

(1)對時間序列進行單位根檢驗。主要目的是判斷時間序列的平穩性，同時可以判斷時間序列的差分階數 d 和週期。週期的判斷和預處理過程的資料聚合過程有密切的關係，如果資料聚合的時間間隔小於一天，那麼週期很可能為一天；如果資料聚合的時間間隔大於一天，那麼週期為進行單位根檢驗得到的週期。

(2)通過差分和週期差分得到一個平穩的隨機序列。對得到的平穩隨機序列進行模型參數識別，得到 p 和 q。p 是自回歸運算元，指的是時間序列受到歷史資料影響的個數，q 是移動平均運算元，指的是時間序列受到隨機干擾項的干擾程度，對於單獨的 $AR(p)$ 和 $MA(q)$ 過程，p 和 q 可以通過時間序列的自相關和偏自相關函數來檢驗得到，當偏自相關函

數在p步後截尾，而自相關函數拖尾，此時時間序列為$AR(p)$過程，p就是$AR(p)$過程中的自回歸運算元，而當自相關函數在q步後截尾，而偏自相關函數拖尾，此時時間序列為$MA(q)$過程，q就是$MA(q)$過程中的移動平均運算元，若自相關和偏自相關函數都是拖尾的，那麼此過程為$ARIMA(p, d, q)$過程，此時中的p和q只能使用測試的辦法進行檢驗，選定(p, q)的值，進行參數估計，並且進行模型檢驗，檢驗合格的(p, q)即為模型的參數值。一般經過差分的隨機序列的自回歸和移動平均參數的階數都小於3，即我們取$max(p, q)<=3$的(p, q)值。

(3)檢驗模型的有效性，包括模型的顯著性檢驗和參數的顯著性檢驗。模型的顯著性檢驗通過殘差序列的LB統計量來確定，檢驗擬合殘差項中是否還蘊含相關資訊，如果不再蘊含任何相關資訊，即殘差序列為白色雜訊序列，此時的模型有效。參數的顯著性檢驗就是要檢驗每一個未知參數是否顯著非0。如果某個不顯著，即表示該參數所對應的那個自由變數對因變數的影響不明顯，該引數就可以從擬合模型中刪除。最終模型將由一系列參數顯著非0的引數表示。

(4)利用擬合模型，預測序列的未來走勢。

4 系統實現及實驗結果

預測模組是使用JAVA進行實現的，是一個基於WEB的系統，採用第3部分中所描述的過程進行預測，預測的結果通過圖片的形式展示給使用者，並且會留有數值序列的備份。

本次實驗主要針對人民網強國論壇和北京交通大學特思論壇上的資料進行了分析，從發現的熱點話題中選取了5個具有代表性的話題，這5個話題就是實際生活中，在對應的時間段，社會上或者在校園內討論比較多，比較熱門的話題，其預測結果如表1所示：

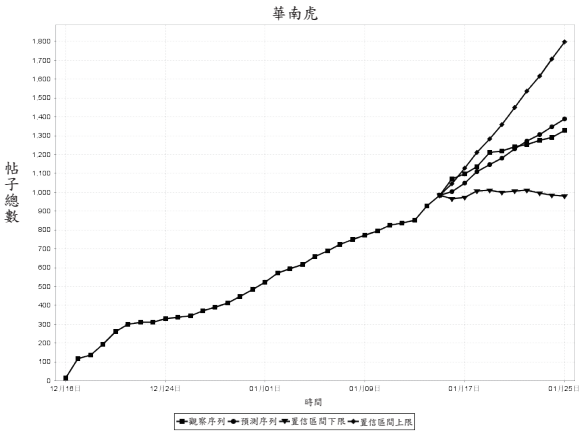
表 1　部分熱點話題預測結果

事件	帖子增量	預測時間	準確率
華南虎	258	2008年1月18日－2008年1月27日	89.86%
家樂福	48	2008年4月24日－2008年5月3日	90.67%
藏獨	372	2008年5月5日－2008年5月14日	79.05%
食堂	109	2008年5月23日－2008年6月1日	92.71%
奧運	98	2008年5月26日－2008年6月4日	87.93%

其中帖子增量是指在預測的時間段內帖子實際的增加量，預測時間分為兩個部分，分別是預測的開始和結束時間，準確率是自己定義了一個計算的方法，如公式(4)所示。

$$準確率=1-\frac{E(預測值-實際值)}{帖子增量} \tag{4}$$

從表1中可以看出，預測的準確率大都在87%-93%之間，只有少數的話題偏離了這個範圍，說明預測模型在整體趨勢的預測上有不錯的效果。下面就以具體熱點話題為例，說明預測的效果。



圖二　強國論壇關於華南虎的帖子數量的預測

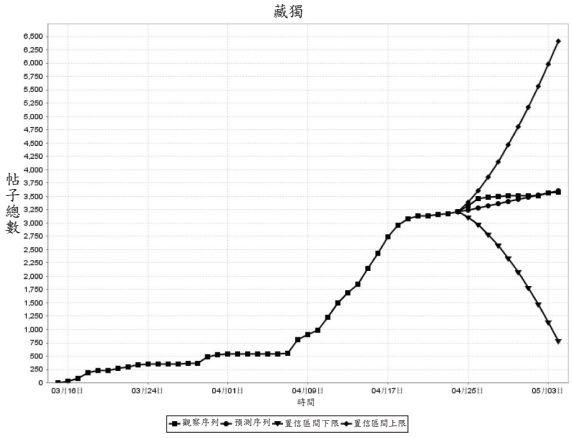
圖二所示為人民網強國論壇關於「華南虎」話題的帖子的預測結果。輸入為從2007年12月18日至2008年1月17日，時間間隔1天的觀察序列，輸出為此後10天的預測結果序列，

察值序列作為參考。

從圖二中可以看出，自「華南虎」事件發生以來，人民網上關於華南虎的討論一直十分熱烈，並且可以預見，直到這個案件審判完畢為止，討論會持續不斷的進行下去。我們的預測結果和網路上討論數量的整體發展趨勢十分接近，但是由於網路輿情突發性等特徵，在某一個具體時刻的預測值有可能存在一定的偏差。

圖三所示為北京交通大學特思論壇關於「藏獨」話題帖子的預測結果，輸入為從2008年3月15日至2008年5月4日，時間間隔1天的觀察序列，輸出為此後10天的預測結果序列，並附上實際的觀察值序列作為參考。

從圖三可以看出，北京交通大學的同學們自西藏獨立分子發生暴動以來就開始積極關注，有了一定數量的討論，特別是在藏獨分子破壞奧運火炬傳遞事件發生之後，討論達到了一個高潮，在經歷一段高潮之後，由於藏獨分子活動的逐漸減少以及新的熱點話題討論的興起，關於藏獨分子的討論也漸漸減少，我們預測的結果也顯示出發展的大體方向。



圖三　北京交通大學特思論壇關於藏獨的帖子數量的預測

以上兩個例子分別代表了網路上熱點話題的兩個基本的類型，即持續不斷型和到達高峰停止型，分別代表了公眾長期關注話題和短期內突發話題，前者以「華南虎」話題為代表，引發話題的事件會持續很長時間，因此話題在長時間引起公眾的關注，後者以「藏獨」話題為代表，引發話題的事件為突發事件，持續事件短，因此話題的持續時間會比較短。

5 結論及展望

基於時間序列的網路輿情預測方法乃是依據人類社會普遍存在的慣性原理和網路輿情的持續性，根據時間序列中後續資料對歷史資料的依賴性來對其進行預測。從實驗的結果來看，使用標準的ARIMA模型能在一定程度上反映網路輿情發展的大體趨勢，但是針對具體話題在某一時間點上的突發性，預測模型並不能很好的進行預測，這是下一步工作要著重解決的問題。同時，根據第4部分最後提出的網路上熱點話題的兩種類型，可以嘗試將這兩種類型的特點參數加入到標準的ARIMA預測模型中，這樣就能得到專門針對互聯網論壇網路輿情預測分析的專門的預測模型，這也是接下來工作的一個重要內容。

而網路上其他可以用數量來表徵其發展趨勢的資訊類型，比如新聞及其評論、博客及其回復等，經過一定的調整也可以使用本文中提出的模型進行趨勢預測分析。

基金項目

教育部高等學校科技創新工程重大項目培育基金項目（707006），通信與資訊系統北京市重點實驗室資助項目，北京市教育委員會共建項目專項資助。

參考文獻

[1] George E., P. Box, Gwilym M., and Jenkins, "*Time Series Forecasting and Control [M]*," Prentice Hall, 1976.

[2] 馮文權，*經濟預測與決策技術[M]*，武漢大學出版社，2002。

[3] 中國人民銀行調查統計司，*時間序列X-12-ARIMA季節調整[M]*，原理與方法，中國金融出版社，2006。

[4] 劉遠超、王曉龍、徐志明等，*文檔聚類綜述[J]*，中文資訊學報，Vol. 20，No. 3，2005，pp.55-62。

[5] 周亞東、孫欽東、管曉宏等，*流量內容詞語相關度的網路熱點話題提取[J]*，西安交通大學學報，Vol. 41，No. 10，2007，pp.1142-1145、1150。

作者簡歷

程輝（Hui Cheng），中國北京交通大學電子資訊工程學院在讀博士生。研究方向為觀點動力學、複雜網路、自然語言處理。在北京交通大學獲得計算機科學與技術學士學位。

劉雲（Yun Liu），中國北京交通大學電子資訊工程學院教授。研究方向為觀點動力學、資訊／網路安全、計算機通信、智慧交通系統。在北京交通大學獲得通信與資訊系統博士學位。