圖書與資訊學刊
1:1=68(Feb '09)12-29
ISSN 1023-2125

科技前瞻趨勢預測方法的成效驗證

Which Kinds of Trend Metrics Are More Effective for Emerging Trend Detection?

曾　元　顯
Yuen-Hsien Tseng
國立台灣師範大學資訊中心研究員
Research Fellow, Information Technology Center, National Taiwan Normal University
E-mail: samtseng@ntnu.edu.tw

洪　文　琪
Wen-Chi Hung
E-mail: wchung@mail.stpi.org.tw
財團法人國家實驗研究院科技政策研究與資訊中心副研究員
Associate Researcher, Science & Technology Policy Research and Information Center, National Applied Research Laboratories

李　宜　映
Yi-Yang Lee
E-mail: d25247@tier.org.tw
台灣經濟研究院生物科技產業研究中心副研究員
Associate Research Fellow, Biotechnology Industry Study Centre, Taiwan Institute of Economic Research

【摘要 Abstract】

科學計量學的科技前瞻方法常以各類參數觀察並預測趨勢，但是並未檢驗參數的有效性。本研究比較了數個趨勢觀察方法，利用資訊檢索評估相關排序的方法評估其排序效果，包括：趨勢呈現方式、趨勢公式以及時間區隔。以不同的領域、文件規模、以及主題來源進行成效比較。結果顯示時間序列線性迴歸斜率在各種情況下表現良好。本研究不僅提供科學計量學趨勢預測效果評估方法，對過去及未來的趨勢分析研究也提供了反思與洞察。

In scientometrics for trend analysis, parameter choices for observing trends are often made ad hoc in past studies. However, the effectiveness of these choices was hardly examined, quantitatively and comparatively.

This work provides clues to better interpret the results when a certain parameter choice is made. Specifically, by sorting research topics in descending order of interest predicted by a trend metric and then by evaluating this ordering based on information retrieval measures, we compare a number of trend metrics (percentage of increase vs. regression slope), trend formulations (simple trend vs. eigen-trend), and options (various year spans and durations for prediction) in different domains (safety agriculture and information retrieval) with different collection scales (72,500 papers and 853 papers) to know which one leads to better trend observation. Our results show that the slope of linear regression on the time series performs constantly better than the others. More interestingly, this metric is robust under different conditions and is hardly affected even when the collection is split into arbitrary periods. Implications of these results are discussed. Our work not only provides a method to evaluate trend prediction performance for scientometrics, but also offers insights and reflections for past and future trend observation studies.

關鍵詞 Keyword

集群　趨勢指標　特徵趨勢　線性迴歸　評估　資訊檢索

Clustering；Trend metrics；Eigen-trend；Linear regression；Evaluation；Information retrieval

壹、前言

科技前瞻（Technology foresight）或研發漸現趨勢（Emerging trends）的掌握，不但有助於預測未來技術發展，亦爲政府分配研發資源的重要參考資訊。因此追縱研發趨勢向爲科技政策決策者所矚目，特別是發掘正在發展中的熱門研究。這些研究被定義爲熱門主題（Hot topics）、上升趨勢（Upward trends），或是研發漸現趨勢（Emerging trends）。雖然這些術語之間有其意義上的些微差異（例如，可由其趨勢上升的急速程度，或主題出現的晚近程度來加以區分細別），在本研究中我們將其視爲同義詞。從其字面上的意義來看，這些術語都是隨著時間的演進而受到越來越多的矚目，因此依此原則來偵測熱門主題似乎是單純易成之事。然而，實際上熱門主題無法單純藉由論文發表的增加量來捕捉，因爲文章發表數量所能透露的訊息，並無法涵括熱門主題的真實意涵（如後續實證所示），還存在一些難以量化的因素，影響著一個研究主題是否真正能成爲熱門主題。

因此，領域的專家常被聘來徵詢研究趨勢的走向。藉由其專業知識，和累積的經驗，來協助制訂有發展潛力的研究主題。但是這個方式也有其缺點，亦即專家們無法有效推測其專精領域以外的研究主題。除此之外，當大量的研究主題需要被排定優先順序時，不同專家的不同觀點往往導致相左的結果。因此，一個能自動觀測趨勢的機制，特別是自動發掘漸現趨勢的方法，是值得發展、探究的研發課題。

科學計量學以及其他領域的研究，已經發展了許多追蹤趨勢的方法。但大部分的方法缺乏專家的回饋加以佐證，在經過許多類似的研究後，Noyons 以及 van Raan 指出，找到合適的領域專家並不容易，因爲他們通常很忙，且對於科學計量學的方法並不熟悉。另一方面，政策制訂者往往對這類研究產生的大量資訊感到束手無策，他們在瞭解這些資訊並詮釋其結果上有所困難，最後往往歸因於他們對於該領域瞭解太少，以致於無法做出有用的解讀。(Noyons & van Raan, 1998a)除此之外，這類趨勢預測的有效性亦鮮少以量化的方式被討論。

在一項國家實驗研究院科技政策研究與資訊中心補助的研究計畫中，我們分析大量的農業學術論文，爲一群領域專家偵測漸現趨勢，並有機會獲得他們的回饋，確認了其中真正的漸現趨勢。爲了充分利用這份專家回饋的結果，本研究運用了資訊檢索品質衡量工具，以自動化的量化方式來評估與比較各種預測方法的有效性。

本文報告了此項研究與發現，其中我們檢驗並比較了常見於科學計量學及一般統計方法的預測指標。本文內容將重點式的回顧過去的相關研究，在研究方法的部分，將介紹科學計量學常用的趨勢指標計算方式，並介紹偵測漸現趨勢主題的方法，以及趨勢預測的成效評估方式；在測試資料的部分，將說明本研究採用的資料內容，以及專家對於主題趨勢的判讀和回饋方式；實驗結果的部份，將呈現各趨勢指標與選項的成效表現情形。最後將討論本研究結果的意涵與後續應用。

貳、相關研究

過去利用量化時間序列資料在趨勢觀測上的應用，已經發展了一系列的方法及多樣性的選擇。Noyon 等學者(Noyons & van Raan, 1998b; Noyons, Moed, & van Raan, 1999) 通常將待分析的文件資料集依據發表時間分爲兩個階段，以觀察第二階段相較於第一階段的成長(或減少)百分比，從而了解各主題的趨勢變化。

日本的科技政策研究所亦以類似的分析，來定義出急速發展的領域。所不同的是，成長率的計算是以每年爲基礎，並以所有觀察年度加以平準化。

陳超美以其發展的 Citespace 工具，利用互動

式的介面，允許使用者更動時間的區隔，將急速發展領域的趨勢加以圖像化。(Chen, 2006)

爲了系統性分析過去 25 年以來發表於 ACM SIGIR 會議的論文主題，Smeaton 等研究者(Smeaton, Keogh, Gurrin, McDonald, & Sodfing, 2003)將各主題趨勢，以主題文件篇數矩陣呈現之。其將文件群集化之後的主題置於矩陣中的列，每年的文件篇數置於行，以獲得各主題的出現時間點，以及其歷年來論文數量的資訊。利用這樣的二維視覺化呈現，他們嚐試預測未來會出現在 ACM SIGIR 會議的理想論文標題，該標題包含了所有他們認爲的熱門主題。

近來，Chi 等人(Chi, Tseng, & Tatemura, 2006)認爲按時期加總研究主題的出現篇數，作爲時間序列來進行趨勢預測，其有效性有待質疑。因爲他們認爲不一樣的來源（如不同的期刊或國家），在累計篇數時其貢獻的比重應該不同。爲了要處理貢獻度不一致的問題，他們提案採取奇異值分解法(Singular value decomposition)來進行更精緻的趨勢分析。

從上述文獻回顧可知，過去研究用了許多的方法來描述及分析趨勢，但卻缺乏方法間的互相參照及比較。實際上各研究的不同應用實例與情境，使得方法間的比較有其困難度。這些現象促使我們進行客觀、公平的趨勢預測方法的分析比較，以建議出各種情況下最佳的趨勢預測方式，希望能在科學計量學的趨勢分析課題上有所貢獻。

參、研究方法

一、趨勢的辨識

針對一文件集，如何找出其中所謂具有「漸現趨勢」的主題？本研究採取兩道程序，首先是將相關文件歸爲同一主題類別。本研究以多層次集群(Multi-stage clustering approach)的方式(Tseng, Lin, & Lin, 2007)自動辨識出知識階層架構(Knowledge structure)，亦即，詞彙或文件可以基於共用字(Co-word)或共引用(Co-citation)聚集成概念，這些概念可以更進一步群聚成主題，再進一步匯聚成領域。之後利用群集標題演算法(Tseng, Lin, Chen, & Lin, 2006)產生出各群集的描述詞，以協助人工後續的分析及詮釋。

其次，建立各階層(包含了概念、主題以及領域)的時間序列，並計算其趨勢。各文件依據發表年代以時間區間加以群集。計算各時間區間內文件數目多寡所得出的年代篇數序列（此序列在本文中亦稱爲趨勢），可用以瞭解該類主題在時間軸上的變化。

以上的計算並未區分不同來源對該主題演變的貢獻度差異。爲了考慮來源貢獻度的差異性，有學者使用特徵趨勢(Eigen-trends)來計算。(Chi , Tseng, & Tatemura, 2006)其特點是利用奇異值分解(Decomposition)從上述簡單趨勢拆解(Break down)得到的矩陣，以產生新的時間序列。(Lathauwer, Moor, & Vandewalle, 2000)具體而言，若所有的文件來自於 m 個來源(所謂來源，可能爲期刊或國家) ，在每個時間間隔中各來源所貢獻的文件數可以用下面的矩陣 *D* 表示：

$$D=\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{bmatrix}$$

藉由奇異值分解，D矩陣可以分解爲以下三個連乘的矩陣$D=USV^T$，如下所示：

$$\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1m} \\ u_{21} & u_{22} & \cdots & u_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ u_{m1} & u_{m2} & \cdots & u_{mm} \end{bmatrix} \times \begin{bmatrix} s_{11} & 0 & \cdots & 0 \\ 0 & s_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_{mn} \end{bmatrix} \times \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nn} \end{bmatrix}$$

多種特徵趨勢可從以上的拆解獲得。其中最重要（包含主要且最多的趨勢資訊）是第一個特徵趨勢，其計算方式是將*S*矩陣中的第一個特徵值s_{11}乘以V^T矩陣中的第一行而得。每個資料來源的權威度，亦即每個來源對於整體趨勢的貢獻程度，亦可從以上的拆解得知。*U*矩陣中的第一行即爲第一個特徵趨勢對應的各個來源權威性。相較於特徵趨勢及特徵權威性，簡單趨勢及簡單權威性可由矩陣*D*求得，其即爲*D*的行和以及列和。以上的敘述，可由下面的計算式清楚表示：

簡單趨勢：$[d_1, d_2, \ldots, d_n]$, 當 $d_j = \sum_{i=1}^{m} d_{ij}$ 對於每一個時間區間 j

簡單權威性：$[a_1, a_2, \ldots, a_m]$, 其中 $a_i = \sum_{j=1}^{n} d_{ij}$ 對於每一個來源 i

(第一個) 特徵趨勢 $[s_{11}v_{11}, s_{11}v_{21}, ..., s_{11}v_{n1}]$

(第一個) 特徵權威性：$[u_{11}, u_{21}, \ldots, u_{m1}]$

Chi 等人展示過，特徵趨勢對高權威來源的（論文篇數）變動比較敏感（其特徵趨勢變化較大），而對低權威來源的變動比較不敏感（即便該來源的篇數變動大，其特徵趨勢變化仍較小）。他們的展示例子顯示，一個來源（期刊或國家）要成爲權威來源（權威期刊或高影響國家），必須對某一主題有長期而持續的刊載或貢獻，才能成爲權威來源，進而可以支配特徵趨勢。只在某個時候有大量的文獻出現，這樣的來源，對特徵趨勢影響不大。

二、趨勢指標

根據上述的時間序列(簡單趨勢或者是特徵趨勢)，可用各類趨勢預測指標來呈現該主題的未來傾向。本文將比較過去研究中常用的趨勢指標，並對比出何者有較佳的預測效力。

第一個指標是平均增加百分比 (Average percentage of increase，以 api 表示)

$$api = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{d_{i+1} - d_i}{d_i}$$

該指標被用於日本的前瞻研究中心(Science and Technology Foresight Center, STFC, 2004)，而當 n=2 時，則是 Noyons 等人常用的指標。(Noyons & van Raan, 1998)

第二個指標，是時間序列的線性迴歸線的斜率 (註 1)，以 slp 表示：

$$slp = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

其中 $x_i = i - \frac{1}{n}\sum_{i=1}^{n} i$ 且 $y_i = d_i - \frac{1}{n}\sum_{i=1}^{n} d_i$.

這個指標在統計學的預測方法中，是最常見的指標(Mendenhall & Sincich, 2003)，但是，因各主題時間序列裡的篇數差異度可能很大，使得這個指標的變異數也大，增添人爲觀測預測時的難度。例如，某個主題 A 整體篇數多，slp=20，而另一個主題 B 整體篇數少，使得其 slp=5，但實際上主題 B 可能才是真正的漸現主題，而 A 只是一個已經成熟但還沒老化的主題。

爲了尋求上述問題(趨勢指標變異數太大)的解決，第三個指標是將時間序列的值轉爲標準化的 z值之後，亦即每個d_i值，都轉成$z_i=(d_i-avg)/stderr$，再計算其迴歸線斜率，以slp_z表示。其中avg爲原始時間序列內篇數的平均值，stderr則爲其標準差。

第四個指標爲第一個指標與第二個指標的整合：將原時間序列中的篇數轉換增加百分比(Percentage of increase)之後，，再計算此新序列的線性迴歸線，以其斜率作爲趨勢指標，以slp_{pi}表示。因此新的時間序列的長度會較原序列爲短。該指標的意涵在於若要得到正的slp_{pi}趨勢值，則篇數成長的幅度需要隨時間而越來越大。因此理論上來講，這個指標最可以捕捉到漸增趨勢。

第五個指標爲slp_c，是將本來的序列，經過以國家爲來源進行拆解(Break down)所得到的(第一個)特徵趨勢的線性迴歸斜率值。

第六個指標是slp_j，其爲原本序列以期刊爲來源進行拆解所得到的(第一個)特徵趨勢的線性迴歸斜率值。

三、趨勢指標的評估

爲了幫專家找出具有增加趨勢的主題，將自動辨識出來的知識結構或主題，以上述各項趨勢指標計算，並由高至低排序。若這些趨勢指標是一個好的預測指標，則排序後的結果將有助於快速發現哪些主題具有上升的趨勢（因爲排序在前面，人工檢視時，容易很快就發現）。爲了要鑑別哪些趨勢指標具有較佳的預測能力，本研究採用 trec_eval (註 2)這項工具。這個工具廣泛地運用於資訊檢索的成效評估中，例如 TREC(註 3)、NTCIR(註 4)以及 CLEF(註 5)等皆爲其應用的範例。從此工具可以得到兩個值，其一是 NAP(Non-interpolated Average Precision rate)以及 Pre@R(Precision rate at Recall position)。NAP 的定義爲：

$$NAP = \frac{1}{R}\sum_{i=1}^{R}\frac{i}{Rank_i}$$

其中R是相關的項目數（真正的漸現趨勢或真正的熱門主題），而$Rank_i$是第i個相關項目排序之名次。Pre@R則是描述當使用者看到第R個排序時，所累計到的精確率，也就是r/R，其中的r是前R個項目中，相關項目的個數。

爲了更清楚了解這兩個值的涵義，表 1 以範例呈現此兩個值應用在三項排序結果的評估情形。假設 A-E 以及 V-Z 爲被排序的 10 個項目，其中 A-E 爲所謂的相關項目(即我們有興趣的項目，也就是真正的熱門主題)，V-Z 則否。如表 1 所示，在 S1 的排序中，五個相關項目都排在最前面五名，在 S2 的排序中，五個相關項目都排在後五名，而在 S3 的排序中，五個相關項目在所有項目的排序間均勻交叉分布。在這三個排序中，S1 爲最佳的排序、S3 次之，而 S2 最差。這兩個評估值計算此三序列的結果如下所示：

NAP(S1) = (1/1+2/2+3/3+4/4+5/5)/5=1.0,
Pre@R(S1) = 5/5=1.0

NAP(S2) = (1/6+2/7+3/8+4/9+5/10)/5=0.3547,
Pre@R(S2) = 0/5=0.0

NAP(S3)=(1/1+2/3+3/5+4/7+5/9)/5=0.6787,
Pre@R(S3) = 3/5=0.6

表 1

以 NAP 及 Pre@R 來評估三項排序之範例

名　次	S1	S2	S3
1	**A**	V	**A**
2	**B**	W	V
3	**C**	X	**B**
4	**D**	Y	W
5	**E**	Z	**C**
6	V	**A**	X
7	W	**B**	**D**
8	X	**C**	Y
9	Y	**D**	**E**
10	Z	**E**	Z
NAP	1.00	0.35	0.68
Pre@R	1.00	0.00	0.60

從計算的結果可以發現，NAP 以及 Pre@R 計算出的值雖然不同，但都可完美的反應其優先順序。值得注意的是，計算這兩個指標的先決條件，是要先知道哪些項目是相關的。Pre@R 是以真正熱門趨勢的項目個數作爲門檻，由上至下排序至該門檻，再計算門檻內有多少個熱門議題。NAP 則將排序中最後一個熱門議題出現的位置設爲門檻，計算加權後的密度值。整體來說，Pre@R 是較爲粗糙的判斷指標，但亦易於解釋。相對地，NAP 較爲複雜，但較可精確地呈現排序上的些微差異。

肆、測試資料

本研究的實證資料集包含兩個部分。其一是來自於安全農業領域、另一個則是關於資訊檢索的領域。實證的步驟如下所示：

一、安全農業：文件蒐集及主題萃取

由國研院科技政策研究與資訊中心的專家定義出六個農業安全相關領域，包括了食品安全、農作物保護、家畜、漁業、林業以及環境。每個領域由一串關鍵字定義，並至 Web of Science 檢索相關的 SCI 以及 SSCI 文章。從 1996 年至 2005 年十年間，共包括了 72,500 篇文章。在這些文章中，我們下載了其書目資料，包括以下幾個欄位: AU(作者)、AB(摘要)、TI(論文標題)、SO(期刊名)、SC(領域別)、DE(關鍵詞)、ID(描述詞)、C1(主要作者住址將之轉換爲作者國別代碼)、CR(參考文獻)、NR(參考文獻數)、TC(被引用次數)、PY(發表年)、UT(論文識別碼)。表 2 陳列出每一年的文章發表數，可以清楚地發現一個上升增加的趨勢。

表 2

1996 年至 2005 年具有安全農業相關之論文數

年	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	Total
篇數	5,448	6,056	6,363	6,211	6,773	7,475	8,028	7,700	9,178	9,268	72,500

爲了辨識出此資料集裡包含的研究主題，本研究嘗試了多種方法。第一個方式是利用書目對(Biliographic coupling)進行主題歸類(Clustering)。初步的分析顯示，此份資料中共含有多達 2,765,938 筆參考文獻，形成了 11,248,898 個書目對。此數量超過系統負荷而無法得到任何結果。在移除了共引文獻小於 5 篇的書目對後，剩下 145,471 個書目對，而形成了 20,795 個集群(Clusters)。集群中包含的文章數目呈現了偏態的分布，亦即大部分的集群(93.85%)包含的文章小於 4 篇，只有 16 個集群裡包含的文章超過了 10 篇，而即便是最大的集群也僅包含 16 篇文章。因此之故，此結果並未用於後續的評估當中。

第二個方式是共用字分析(Co-word analysis)，其分析的來源基礎是標題及摘要的自由詞彙。利用曾元顯的演算法(Tseng, 2002)，找出其中的關鍵字及共現字(在同一個句子裡常常共同出現的詞彙) 。若各文件包含一定程度以上的共用字，則這些文件被聚集成一集群。結果產生了 423 個以上的集群(集群數依據設定的門檻而有所差異) 。雖然各集群中文章的篇數分布偏態程度較前述方法爲低，但是以這些詞彙形成的集群主題品質並不一致。集群主題描述不夠精確可能使得專家在進行趨勢判讀時效率降低。因此，雖然此方法可以涵蓋文件集裡較充分的資訊，本研究仍決定暫時保留此結果，而不做進一步的評估。

最後一個方法乃是採用以控制詞彙爲基礎的共用字分析。詞彙的來源爲書目資料中的 SC 以及 DE 欄位。初步統計平均每篇文章有 1.8 個 SC 詞彙，而只有 2.8%的 SC 詞彙出現在標題或摘要中。就 DE 詞彙來說，平均每篇文章有 5.52 個 DE 詞彙，而有 46.35%會出現在標題或摘要中。整體而言，有 179 個 SC 詞彙出現在 10 篇以上（含）的文件中，而有 3,632 個 DE 詞彙出現在 10 篇以上（含）的文件中。簡而言之，控制詞彙的數量是在可分析的範圍中，而且可以呈現出與自由詞彙不同的分析面向。

在控制詞彙的共用字分析中，紀錄了每一對 SC 詞彙或 DE 詞彙共同出現在同一篇文章的次數。此成對出現的次數再以個別詞彙出現的次數加以標準化。依此計算得來的相似度應用在完全連結(Complete link)的歸類演算方法中。此方法可從 179 個 SC 詞彙獲得 80 個集群，而從 3,632 個 DE 詞彙中，可獲得 1,617 個集群。

值得一提的是，SC 是屬於較廣義的詞彙，呈現的是一個領域中的主題類別。將之加以群集得到的結果類似主題間的聯集或交集。相形之下，DE 較接近自由詞彙，描述的是較精確的概念，其集群的品質受到了涵蓋範圍的影響。舉例來說，圖 1 顯示了兩個擁有相似概念的集群。其一是利用自由詞彙形成的集群(方法二)，另一則是從 DE 詞彙得來的群集。吾等可發現前者涵蓋較多的詞彙因而主題內容上似乎較爲完整。相較之下，DE 詞彙因詞彙少，導致雜訊也少，因而包含較有效的主題範圍。

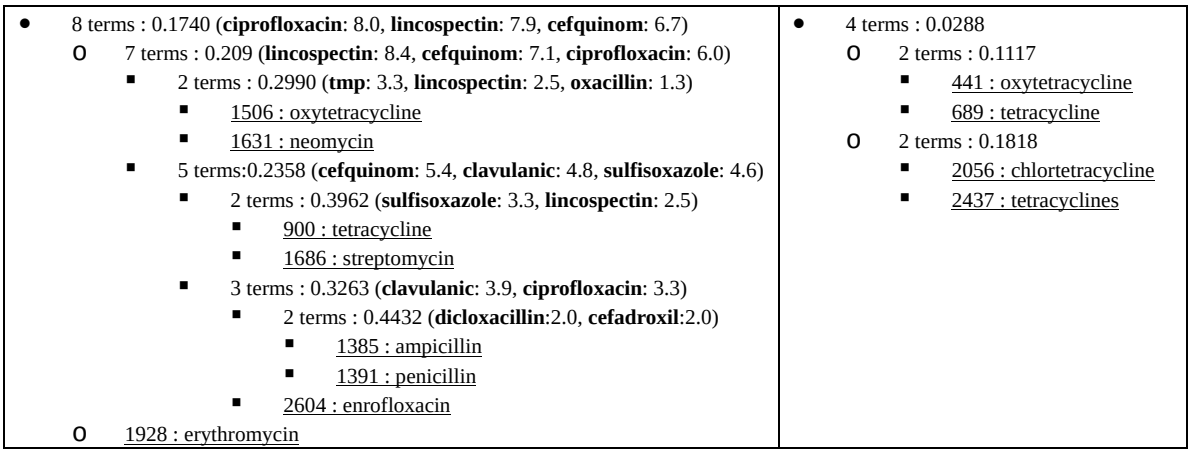


圖 1　共用字分析之結果範例。左邊的結果來自於標題及摘要之自由詞彙，括號中表示同一共用字；右邊則為 DE 欄位所得之共用字分析結果。

二、安全農業：專家之判別與趨勢標示

爲了減少上述集群數至一個可以操作的量，本研究隨機抽取了 50%的 SC 集群以及 10%的 DE 集群，並移除了小於 30 個文件數量的集群。剩下來的集群交由專家判別趨勢，給予適當的符號顯示，以作爲標竿比較的對象。趨勢判斷專家包含了六位教授、兩位研究者以及一位科技政策與研究中心專案管理人員，這些人士都是對安全農業未來發展有一定程度瞭解並投注關心的人。

在進行趨勢判斷時，專家們依據下列原則，對集群主題做出趨勢類別的標示。最優先的原則，是根據集群的類別描述詞，推測其主題範圍後判斷出其未來發展趨勢。若專家們無法單就類別描述詞判別趨勢，則可參考各集群的時間序列資料作出判斷。若此資料仍讓專家無法判斷，則可進一步參考各集群內所包含的文章資料。若以上的方法皆無法讓專家判斷出趨勢，則該集群便標示爲「無法判讀(Inconclusive)」，以「?」標註。專家們將集群類別以五種趨勢型態標示：急速上升、上升、波動、下降以及急速下降，分別是以「++」、「+」、「=」、「-」、「--」標示。

表 3 顯示以 DE 關鍵字形成的集群範例，其中 Trend 欄位標示的是由專家判斷出來的趨勢型態，而其第一欄標示的是集群的辨識碼、第二欄則是包含在集群以及子集群中的詞彙數、第三欄是同一個集群內詞彙間的最小相似度，其餘的欄位內容如表頭所示。

表 4 顯示了專家判斷趨勢的統計結果，其中 SC 集群有 43 個爲無法判讀，佔了所有集群的 27.74%，無法判讀的比例在 DE 形成的集群中則有 22.89%。因爲整體的趨勢是向上的，各集群中並沒有急速下降的部分，即便是列於下降的類型亦極爲少見。

表 3

專家回應趨勢種類標示範例

cid	nt	Sim	Trend	DE Terms	df	96	97	98	99	00	01	02	03	04	05
9	6	0.025	=	osmoregulation; chloride cell; metamorphosis; thyroid hormone; flounder; flatfish	147	3	9	17	14	16	26	17	12	14	19
9	4	0.066	=	osmoregulation; chloride cell; metamorphosis; thyroid hormone	106	2	4	16	11	13	16	11	10	9	14
9	2	0.178	-	osmoregulation; chloride cell	74	1	2	13	8	10	14	7	4	7	8
9	2	0.192	+	metamorphosis; thyroid hormone	36	1	2	3	3	4	3	4	6	2	8
9	2	0.120	?	flounder; flatfish	46	1	5	1	4	5	11	6	2	5	6
28	5	0.032	++	food allergy; anaphylaxis; IgE; gelatin; allergen	102	4	6	4	5	10	15	15	14	15	14
28	2	0.138	++	Food allergy; anaphylaxis	67	3	5	4	4	5	12	10	8	9	7
28	3	0.164	++	IgE; gelatin; allergen	46	1	2	1	2	6	5	6	8	7	8
28	2	0.196	+	IgE; gelatin	34	1	2	1	1	6	4	3	8	4	4

表 4

專家對各主題進行趨勢判讀之統計結果

Field	(sub-)clusters	++	+	=	-	--	?
SC	155	18	57	37	0	0	43
DE	249	20	61	97	14	0	57

三、資訊檢索：文件蒐集、群聚及熱門主題

第二部份的資料，使用的是 Smeaton 等(2003)過去研究所蒐集的資料。其資料範圍爲歷屆 ACM SIGIR 研討會文章的標題、作者名與摘要，共 25 屆合計 853 篇文章。他們利用一商業套裝軟體 Glustan Graphics 將這些文章歸類成 29 個互不重疊的集群。之後以人工檢視每一個集群並給予各個集群對應的主題描述詞，以顯示各集群所包含論文的內涵。爲了建立集群間的結構關係，並探討各主題歷屆的分佈擴散情形，Smeaton 等將每集群中的文章篇數按年代分佈陳列，如表 5 所示。表 5 的列表示的是各集群的主題，這些主題，大致上以其第一次出現的年代，以及其包含文章的篇數，來加以排序。ID 欄則是本研究依照排序順序給予之編號，以方便後續討論。

表 5

Smeaton 等研究者建立 SIGIR 會議之文章主題的群集及排序

集群 \ 年	ID	71	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	總和
Databases, NL Interfaces	29	8	4	1	6		5	10	1	3	5	2	5	2	4	1		3	1	1	2	2						66
General	28	5	2	9	2	9	5	7	10	10	6	10	6	2	5	8	6	2	2	4	3	1		4	2	5	1	126
Models	27	1			2	1	1		4	1	2	1	2	1	2		2	2	2	2	3	1						30
Question answering	26	1			1	1	1						1					1		1			1		4	4	1	17
Syntactic phrases & SDR	25	1					1		1		2	1	6	3	3	2	3	2	1	1	2	1	1	3	1	1	1	37
Conceptual IR, KB IR	24	1			4	4	1	3	3	4	3	5	7	5	1	6	3	5	3	2	3	4	1	3	2	1	1	75
Compression	23	1								1	2	2	1	1	1	3	1	1		1			2			1		18
Clustering	22		2		1	1		2		3	3	2					1	2		1	1	2	1		1		3	26
Relevance feedback	21		1	1	1		2			1	1		1			1	2	4	3		1	2	1	1	1	1		25
Inverted files & Implementations	20		1				1			1			2	1	3	1			2	1				1		1	3	18
Term weighting	19			1	3	2	1	2	1	1	5	3	3				1		2	1	1	1	1			1	1	31
Message understanding & TDT	18			1	1						1						3		2			3	4	2	4	5	5	31
Filtering	17			1						1			1			1			1		4	1	1	1	1	2	3	18
Hypertext IR, Multiple evidence	16											1	3	1	1	2	1	2	2	2	1	4	3	1	5	2	2	33
Image retrieval	15				1				1			1			1	1				2	1	1						9
Probabilistic & Language models	14				1	1	1						3	1		3	4	2	2	3	2	1	3	1		3	3	34
Boolean & extended Boolean	13							1		2	1				1			1	1			1		1	1			10
Japanese & Chinese IR	12									1					1			2		3	2	3	1	1				14
DBMS & IR	11				1		1		1											1	1							5
Users & Search	10					2		3	3	2	2	4		3	2	2	3	1	3	3	1		1	2	1			38
Visualisation	9										1		1	1		1			1		2	1	1	2		1		12
Signature files	8								1		1		1	2	2		1	1										9
Distributed IR	7						1		2	1		2		1		1				3	1	1	3	4	2	1	1	24
Evaluation	6																	3	4	4	2	1	7		2	3	8	34
Topic distillation & Linkage retrieval	5																						1		3	3	2	9
Latent semantic indexing	4												1				1		1					2	1			6
Text categorisation	3													1					3	3	3	1	3	1	3	3	2	23
Document summarisation	2																			2				2	2	3	3	12
Cross lingual	1																				1	3	3	1	1	3	4	16

在 Smeaton 等人的研究中，除了嘗試去追蹤資訊檢索領域的主題演進，也推論及預測後續可能的熱門主題(Hot topics)。根據他們的預期，出現在 2003 年 ACM SIGIR 研討會的理想主題爲 "**Evaluation** of a **Language Model** Implementation of a **Topic-Based, Cross-Lingual Question-Answering** and **Summarisation System**"。本研究將和此標題相關的主題拆解出來，並列於表

6 中。表 6 共分爲三群，原因在於吾等無法確認在上述理想標題中"Topic based"包含的範圍。因此在 J5 中去除此詞，並在 J8 以及 J10 中逐漸加入與"Topic based"相關的主題。

爲了瞭解在 2003 年 SIGIR 研討會實際發表的文章內容，我們分析了該年研討會論文集的目錄。表 7 列示了各場次主題，並加上一些更能符合其內容的補充詞，第二欄列出了該主題的論文篇數，第三以及第四欄則列出了對應表 5 的集群 ID。此對應是基於表5的集群標題以及表7的場次主題的相似度而來。但爲了容納不同判斷的標準，我們做出兩種對應：第三欄中的對應，其判斷比較寬鬆，而第四欄的對應判斷比較嚴謹。

表 6

以 Smeaton 等學者預測之理想論文標題為基礎，產生之三群 SIGIR 2003 年的熱門議題

J5	J8	J10
26: Question answering 14: Probabilistic & Language models 6: Evaluation 2: Document summarisation 1: Cross lingual	26: Question answering 14: Probabilistic & Language models 6: Evaluation 2: Document summarisation 1: Cross lingual **22: Clustering** **17: Filtering** **3: Text categorisation**	26: Question answering 14: Probabilistic & Language models 6 :Evaluation 2: Document summarisation 1: Cross lingual 22: Clustering 17: Filtering 3: Text categorisation **18: Message understanding & TDT** **5: Topic distillation & Linkage retrieval**

表 7

2003 年 SIGIR 會議中的 14 個議程標題（主題）

主　　題	df	S13	S10
Retrieval Models (Language Models, Evaluation)	3	14	14
Question Answering (Evaluation)	3	26	26
Web (Hyperlink, Classification)	3	5	
Human Interaction	6	10	
Text Categorization	6	3	3
Multimedia Information Retrieval	3	15	15

（續下表）

（接上表）

Structured Documents (XML)	2		
Text Representation (Term Modelling)	2	19	
IR Theory (LSA)	3	4	4
Filtering and Retrieval Models (LSA)	3	17	17
Clustering	3	22	22
Distributed Information Retrieval (Source Selection, Topic Segmentation)	3	7	7
Novelty and Topic Change (Text Segmentation)	3	18	18
Cross-Lingual Information	3	1	1

伍、研究結果

一、安全農業

本研究將急速上升及上升的情形獨立出來討論，其趨勢指標的表現如表 8 所示。根據NAP以及Pre@R來判斷，slp以及slp_z兩個指標表現最佳，其次是兩個特徵趨勢指標，最差的則爲增加百分比類型的指標。

結果顯示線性迴歸的斜率是最佳的趨勢指標，這也驗證了其在趨勢分析上被廣泛應用的現象。特徵趨勢指標表現並不如預期突出，原因在於各來源的貢獻度沒有很大的差異性，因而無法突顯出特徵趨勢指標的優異性。針對資料進行驗證，本研究發現簡單權威度向量與特徵權威度向量幾乎一致；簡單趨勢的斜率亦與特徵趨勢的斜率一致。在這種情況下，藉由奇異值分解法將趨勢矩陣拆解並無法得出較佳的結果。而增加百分比類型的指標在時間區隔爲一年時表現最差，此結果令人感到意外，因爲此指標十分直觀且易於解釋。除此之外，標準化的迴歸斜率指標slp_z在偵測急速發展的趨勢表現並不佳，這也顯示了趨勢偵測的複雜性。

表 8

不同趨勢指標的表現 Avg 列為 SC 列及 DE 列之平均

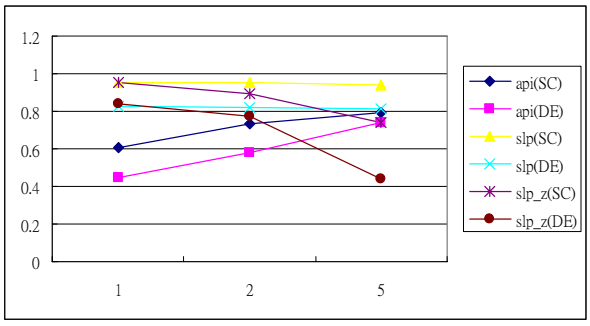
指　標	判 讀 欄 位	+ 或 ++		++	
		NAP	Pre@R	NAP	Pre@R
api	SC	0.6093	0.6533	0.1733	0.1111
	DE	0.4454	0.5062	0.1587	0.1500
	Avg	0.5273	0.5798	0.1660	0.1306
slp	SC	0.9521	0.8800	0.4552	0.3333
	DE	0.8254	0.7407	0.4293	0.5500
	Avg	**0.8887**	**0.8104**	**0.4423**	**0.4417**
slp_z	SC	0.9524	0.8933	0.2992	0.2778
	DE	0.8424	0.7654	0.5689	0.5500
	Avg	**0.8974**	**0.8294**	0.4340	0.4139

（續下表）

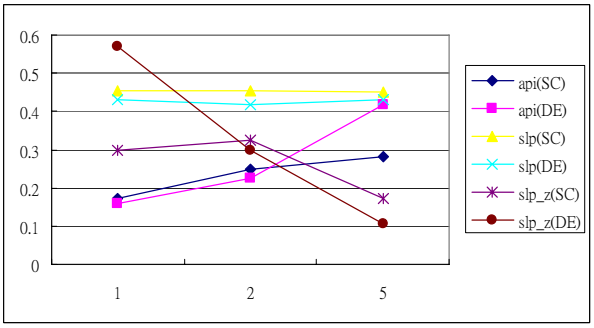
（接上表）

slp_{pi}	SC	0.7250	0.7733	0.2051	0.1667
	DE	0.3807	0.3704	0.0867	0.0000
	Avg	0.5528	0.5719	0.1459	0.0833
slp_c	SC	0.9295	0.8533	0.4457	0.2222
	DE	0.7846	0.6914	0.4574	0.4500
	Avg	0.8570	0.7723	0.4516	0.3361
slp_j	SC	0.9214	0.8267	0.4722	0.4444
	DE	0.8041	0.7284	0.4084	0.3500
	Avg	0.8627	0.7775	0.4403	0.3972

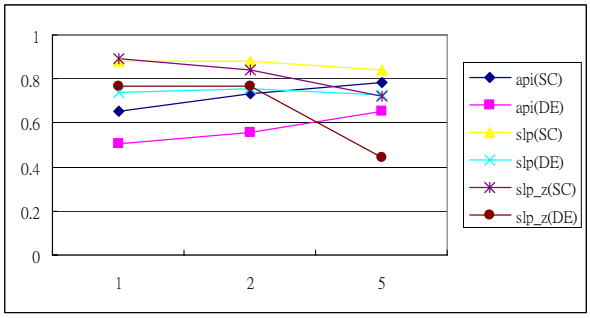
過去研究用以建立時間序列的時間間隔有許多種方式，本研究亦分析不同指標api, slp以及slp_z在不同的時間間隔下觀察趨勢的表現情形，其結果如圖 2 所示。時間間隔分別爲 1、2 以及 5 年。5 年的時間區隔意謂著將 10 年區分成兩個階段。令人意外的是，即便是只區分成兩個階段，slp的表現水準仍與區分成較多時段時類似。而api指標則隨著時間區間的加大而有較佳的表現結果。相反地，slp_z則隨著時間間隔的拉大而表現急速下降。這是因爲當僅存在兩個時間間隔時，該指標僅有三個值＋2、0 以及-2(註 6)。差異性過小的情況造成其難以有效地追蹤趨勢。



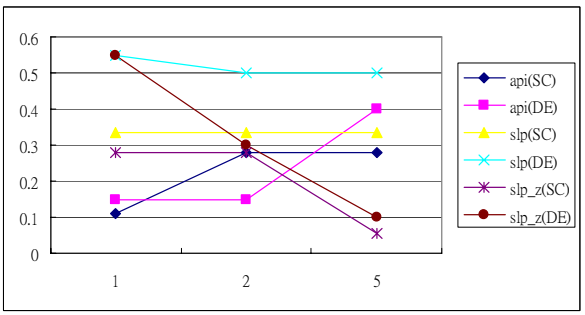
(a) NAP 對於 + or ++.



(b) NAP 對於 ++.



(c) Pre@R 對於 + or ++.



(d) Pre@R 對於 ++.

圖 2　不同時間區間 1、2 以及 5 年之預測效果

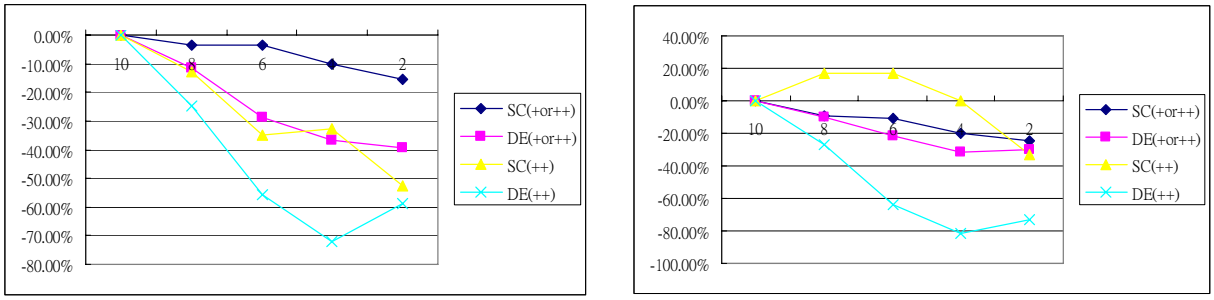


圖 3　僅使用 n 年資料以 slp 進行預測，有效性下降比例，n=10、8、6、4 及 2，左圖為以 NAP 衡量，右圖為 Pre@R.

此外，本研究亦評估各指標的預測能力。圖 3 顯示了僅用前 n 年資料來進行預測的情況。舉例來說，以 NAP 衡量 slp 的正確性，若僅採取前八年的資料，預測急速發展的主題的成效將會減少 25%。亦即，若我們在第八年的時間點來預測第十年的趨勢，其績效表現爲利用整整十年時的績效的 75%。

二、資訊檢索

在資訊檢索的部份，本研究比較上述趨勢指標將主題排序的結果以及Smeaton的排序結果（如表 5）。各指標及Smeaton的預測成效如表 9 的前六欄所示。其中，slp_{pi}表現得出乎意外地好，這可能是因爲此份資料相關的主題呈現持續熱門的趨勢。slp_{pi}得出的前十大趨勢主題如表 10 所示，與表 6 相較，在J10 部分 10 個主題中有高達 9 個可與slp_{pi}預測出的結果相對應。預測績效次之的是斜率指標 slp。Smeaton的排序則僅在判斷情況J5 時，能與此兩個指標相提並論，在其他的判斷情況下時，均不若上述兩個最佳指標……。

另外，以實際出現於 2003 年 SIGIR 研討會的主題爲準，比較不同指標排序所得效果則如表 9 的最後四欄所示。雖與先前結果相比有稍許的不同，但 slp 仍然較其他的指標來得有效(註 7)。Smeaton 的排序表現則不若預期。

表 9

以 Smeaton 資料為基礎，各趨勢指標預測 SIGIR 2003 實際章節的主題的效果

Judgment	J5		J8		J10		S13		S10	
Index	NAP	Pre@R	NAP	Pre@R	NAP	Pre@R	NAP	Pre@R	NAP	Pre@R
api	0.2527	0.4000	0.4824	0.3750	0.5724	0.5000	0.5432	0.5714	0.3691	0.4286
slp	**0.5852**	**0.4000**	**0.6584**	**0.7500**	**0.8234**	**0.8000**	0.5404	0.5714	**0.3994**	**0.5000**
slp_z	0.3958	0.4000	0.5597	0.6250	0.7340	0.8000	0.5582	**0.6429**	**0.4023**	**0.5714**
slp_{pi}	**0.8083**	**0.6000**	**0.7698**	**0.7500**	**0.9625**	**0.9000**	**0.5727**	**0.5714**	0.3808	0.5000
Smeaton	**0.5956**	**0.4000**	0.6253	0.5000	0.6712	0.5000	**0.5910**	0.5000	0.3854	0.3571

表 10

以表 6 資料為基礎，slp_{pi}預測之 10 大主題

Rank	Topic	slp_{pi}	Rank	Topic	slp_{pi}
1	Evaluation	0.0462	6	Cross lingual	0.0173
2	Question answering	0.0287	7	Filtering	0.0156
3	Topic distillation & Linkage retrieval	0.0235	8	Probabilistic & Language models	0.0151
4	Document summarisation	0.0213	9	Text categorisation	0.0138
5	Message understanding & TDT	0.0213	10	Compression	0.0113

三、研究結果意涵及結論

回顧科學計量學的過去文獻，吾等發現過去研究曾用不同的時間區間及指標來創造時間序列，以觀察時間趨勢。但是，關於時間區隔及趨勢指標這些參數的有效性，鮮少用量化且比較性的角度加以呈現。本研究根據各趨勢指標計算出的結果，將各主題的趨勢以降冪方式排序，並以資訊檢索的相關工具衡量其排序成果。我們比較了多種呈現趨勢的方式（增加百分比 vs.迴歸）、趨勢公式（簡單趨勢 vs.特徵趨勢）及時間區隔（多種時間區隔及預測長度），並且試驗於不同的領域（農業安全以及資訊檢索），不同的文件規模（72,500 篇論文 vs.853 篇論文），以及不一樣的主題來源（SC vs.DE 欄位）。

本研究的結果可做爲趨勢分析時，選擇合適指標的參考。舉例來說，若欲以 api 指標來進行趨勢預測，必須注意該指標在大的時間區間以及時間序列很短的時候才適用。當所有的文件按時間區分爲兩期時，slp 以及 api 是合宜的指標。因此，Noyons 等(1998a, 1998b, 1999)常使用 api 來分析論文發表量的兩期變化，有其合理性。不論在哪種情形下，slp 在評估中皆有良好的表現，是推薦使用的指標。

以上結果的有效度，是基於專家趨勢判斷的有效程度。特別值得注意的是，專家進行趨勢判斷的過程中，可接觸到趨勢資料作爲參考，此過程並不會扭曲效度。如表 5 顯示的 SIGIR 的例子，即便專家們可以看到各年趨勢，他們仍會將排名較後面的主題(例如"Question Answering"排名爲第 26 名）視爲熱門主題，反之亦然（如"Distributed IR"排名 7 但專家仍不視爲一熱門主題）。此外，儘管有不同的趨勢判斷標準，如表 9 所示，表現良好的指標一般說來皆會表現良好。這個現象在牽涉人爲判斷的情況往往會出現，如在 Tseng & Teahan (2004)的研究中可見。因此，不可避免的人爲誤差是可以被忽略的。

本研究的目標，在探索最佳的方法，以便從大量的文獻資料中，萃取主題、監控科研發展趨勢。本研究結果的重要性，在於檢驗並瞭解了各種情況下各指標的適用性。一個合適的趨勢預測指標，意謂著利用該指標排序時，將會有效的減輕監測大量文獻資料所需花費的時間和精力。

儘管本研究已實驗多種情況，然更多仍然值得繼續再做。本研究以資訊檢索的成效評估方法爲基礎，爲驗證各類趨勢指標的成效，提供了一個客觀且可重複操作的程序。例如，是否高品質的期刊真得需要給更高的權重，可作爲後續研究的議題之一，雖然本研究的安全農業並未顯示出該現象。

在本研究中，各主題的趨勢發展僅以各主題包

含的文件數量來表現。在後續的研究中，可加入內在或外在的資訊結構，以更精準的反應趨勢發展。舉例來說，可加入複雜的演化模型或各主題間的引用資訊等。運用這些模式去預測未來趨勢，可能會使效果更爲提升。

（收稿日期：2008 年 9 月 2 日）

致謝

本研究由國科會計畫 NSC 96-2221-E-003-017-及 NSC 96-2524-S-003-001-部分贊助；本篇論文改寫自：Yuen-Hsien Tseng, Yu-I Lin, Yi-Yang Lee, Wen-Chi Hung, and Chun-Hsiang Lee, “A Comparison of Methods for Detecting Hot Topics”, 被 Scientometrics 於 2008 年 4 月 28 日接受。

註釋

註 1：我們用 Perl 程式語言中的 Statistics:Regression 模組進行迴歸的計算。

註 2：A similar tool trec_eval.prl rewritten in Perl provided by the host of the NTCIR Workshop was actually used

註 3：Text REtrieval Conference, Retrieved August 28, 2008, from http://trec.nist.gov/

註 4：NTCIR (NII Test Collection for IR Systems) Project, Retrieved August 28, 2008, from http://research.nii.ac.jp/ntcir/ CLEF

註 5：The Cross-Language Evaluation Forum, Retrieved August 28, 2008, from http://www.clef-campaign.org/

註 6：設其數列爲 (x1, x2)。 轉化爲 Z 數列 ((x1-avg)/stderr, (x2-avg)/stderr)=((x1-x2)/2 / |(x1-x2)/2|、 (-x1+x2)/2 / |(x1-x2)/2|)。因而該數列僅會有 3 種值: (-1, 1)、 (1, -1)、(0, 0)，造成僅有 3 種可能的斜率值：+2、-2 以及 0。

註 7：因相關的主題數增加（如 Pre@R，表 7 中 R=14，較表 6 中 R=5、8 或 10），因而衡量出來的值較小，而排序仍相同）。

參考書目

Buckley, C. trec_eval IR evaluation package. Retrieved August 28, 2008, from http://trec.nist.gov/trec_eval/

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359-377.

Chi, Y., Tseng, B. L., & Tatemura, J. (2006). Eigen-trend: trend analysis in the blogosphere based on singular value decompositions. In P. S. Yu, V. J. Tsotras, E. A. Fox, & B. Liu (Eds.), T*he Proceedings of the 15th ACM International Conference on Information and Knowledge Management* (pp. 68-77). New York: ACM.

Lathauwer, L. D., Moor, B. D., & Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications, 21*(4), 1253-1278.

Mendenhall, W., & Sincich, T. L. (2003). *A Second course in statistics: regression analysis.* (6th ed.). Upper Soddle River, N.J. : Prentice-Hall.

Noyons, E. C. M., Moed, H. F., & van Raan, A. F. J. (1999). Intergrating research performance analysis and science mapping.

Scientometrics, 46(3), 591-604.

Noyons, E. C. M., & van Raan, A. F. J. (1998a). *Mapping scientometrics, informetrics, and bibliometrics.* Retrieved November 23, 2006 from http://www.cwts.nl/ed/sib/home.html

Noyons, E. C. M., & van Raan, A. F. J. (1998b). Monitoring science developments from dynamic perspective: Self-organized structuring to map neural network research. *Journal of the American Society for Information Science and Technology, 49*(1), 68-81.

Smeaton, A. F., Keogh, G., Gurrin, C., McDonald, K., & Sodfing, T. (2003). Analysis of papers from twenty-five years of SIGIR conferences: What have we been doing for the last quarter of a century ? *ACM SIGIR Forum, 37*(1), 49-53.

Science and Technology Foresight Center. (2004). The 8th Science and technology foresight survey: Study on rapidly-developing research areas *(Interim Repor).* Japan: National Institute of Science & Technology Policy.

Tseng, Y.-H. (2002). Automatic thesaurus generation for Chinese documents. *Journal of the American Society for Information Science and Technology, 53*(13), 1130-1138.

Tseng, Y.-H., Lin, C.-J., Chen, H.-H., & Lin, Y.-I. (2006). Toward generic title generation for clustered documents. In H. Toung, M.-K. Leong, M.-Y. Ken, & J. Donghong (Eds.), *Information retrieval technology: Third Asia Information Retrieval Symposium*(pp.145-157). Singapore: Springer.

Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Information Processing and Management, 43*(5), 1216-1247.

Tseng, Y.-H., & Teahan, W. J. (2004). Verifying a Chinese collection for text categorization. In K. Jârrelin, J. Allen, P. Bruza, & M. Sanderson (Eds.), *The 27th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp.556-557). New York: ACM.