新聞事件自動偵測與追蹤及事件摘要系統實作與評估

An Experiment and Evaluation of a News Events Detection and Tracking System with Multi-document Summarization

黃純敏　　戴尚孝　　郭家良

國立雲林科技大學資管系　國立雲林科技大學資管系　國立雲林科技大學資管系

huangcm@yuntech.edu.tw　gmi013@yuntech.edu.tw　gmi127@yuntech.edu.tw

摘要

一般電子新聞分類多以人工方式依新聞敘述大致分類（如政治、社會、體育…），本研究改變傳統分類模式，希以更直覺的分類方式，將各新聞來源的新聞文件依「事件」群聚，讓讀者能清楚的了解正在發生或已經發生的事件，並提供自動持續追蹤事件發展的功能，以協助讀者快速、完整且通盤了解事件全貌。尤其利用多文件自動摘要技術，本系統可隨事件發展不斷調整事件摘要內容。研究結果顯示本系統所整理之事件能快速及有效幫助讀者了解新聞事件的完整來龍去脈；新聞事件群集經評定各集合間的新聞有高度相關性，標題之給定也頗具代表性。至於多文件自動摘要，則認為不僅能有效縮短閱讀時間、協助讀者了解事件，對於摘要可讀性及字數的適當性都有極為正面的評價。

關鍵詞：事件偵測、事件追蹤、自動摘要、多文件摘要。

Abstract

In this study, we proposed an integrated system for Internet users to browse the news from multiple news sites. To provide a further intuitive way to search the news, we use the “event” concept as a news grouping method. That means an event with various statements would be put into the same cluster and displayed in the same category for comparison. Through our system, users can acquire what is exactly happening or has happened by reading the event reports. The system also provides a tracking function to automatically detect the follow-up reports as the events evolve over time. Moreover, this system provides a brief description of each event by using multi-document summarization technique to have the task done.The experimental result reveals that our system provides promising effects on event detecting and tracking. The automatic news abstraction is also gained high remarks especially on the context readability and the default abstract length.

Keywords: Event detection、Event tracking, Multi-document summarization、Automatic abstraction

壹、緒論

網路的便捷與全民資訊意識的提昇，使得電子新聞已然成為民眾掌握瞬息萬變的時事與獲取最新訊息的最佳媒介之一。不過也由於網路的便捷，造成網路上的文件以驚人的數量增加，並將資訊過載（information overloading）問題帶給所有網際網路的使用者。目前各大入口網站及新聞網站都提供線上新聞閱讀服務，並對新聞文件依性質約略分類（如政治、社會、體育…），讀者需依其分類架構閱讀新聞。由於新聞報導有別於一般文件，相同事件有多人同時撰述之特點，因此新聞記者可能因立場差異、切入角度不同、或個人專業素養差距，所報導之事件與實情或有所出入。欲客觀掌握事件實況，需瀏覽數個網站，針對特定主題比較報導內容，否則難以窺其全貌。此外，新聞事件著重新聞性及時效性，對於熱門事件多有一窩蜂爭相報導、獨家專訪及後續發展報導特性。一般讀者如需回顧某一事件過往資訊，礙於個人時間或未諳查詢功能，多半僅以現成報導對照記憶以勾勒事件梗概。是以，若能提供主動偵測及追蹤事件機制，藉以有效發覺匯入不同來源之相關事件新聞與後續報導，對於提供讀者了解事件來龍去脈發展應有不小的助益。此外，鑒於同一或相關事件之新聞報導，常常動輒數十篇甚至近百篇，若在偵

測與追蹤機制加入摘要功能，應可大幅減少讀者閱讀時間。至於讀者之接受度與對處理後新聞事件理解成效，則有賴實證數據支持。

基於上述理由，本研究提出一個可幫助讀者以更快速、更有效率的方式，瀏覽其感興趣的新聞文件發展的機制。本機制不惟可主動偵測事件發生並可將相同事件的新聞群聚，尤其可將後續報導的文件自動歸類到適當的事件群集，以達到事件追蹤目的。特別的是，本系統結合多文件摘要技術，提供事件群集一段簡短及具代表性之文字敘述，該事件之簡要敘述並可隨事件發展不斷調整內容。希冀本系統所整理之事件能快速及有效幫助讀者了解新聞事件的完整來龍去脈。

貳、文獻探討

2.1 字詞處理技術

資訊檢索領域中，對於文件內容處理，需要使用字詞處理技術來分析文件，藉以篩選出能代表該文件的特徵（feature）或關鍵字詞。由於中英文分屬不同語系，中文字不同於英文有明顯分隔符號（空白符號），因而有斷詞處理及詞彙判斷的問題。一般而言中文斷詞的方法可歸納為以下三種[3]：

1. 詞庫斷詞法

利用已經建置好的詞庫，比對文件內文字資料，以擷取對應詞彙。使用本方法，所依據的詞庫必須有相當的權威性。然而由於詞庫更新不易，多半需要搭配人工選錄新詞及專業名詞以維護詞庫品質。

2. 統計式斷詞法

需有大量文件做為字詞處理基礎之語料庫，以字（Gram）在語料庫中出現的次數達到訂定的門檻值，便認定可能為有意義的詞彙。依照選取相鄰字數的長短，可區分為2-Gram、3-Gram 至 N-Gram。本法優點為不受到詞庫固定詞彙的限制。缺點是易呈現語料庫相依(corpus dependent)的特性，所據以斷出之詞彙代表性與可用性亦值得質疑。

3. 混合斷詞法

此法先進行詞庫式斷詞法，再輔以統計式斷詞法萃取新詞。由於兼顧詞的品質與新詞之納入，此法已漸漸成為研究者採行的方式。

惟經過斷詞的程序，仍不足以產出代表文件的關鍵詞。要篩選出具代表性的字詞，需計算該字詞在文件中的權重。字詞權重的給定是藉由計算該字詞在單一文件的重要性（local weight）及在整個文件集的重要性（global weight）而來。目前最常使用的字詞權重計算方式為 TFIDF（Term Frequency Inverse Document Frequency，詞頻反轉文件頻率）。

2.2 新聞事件偵測與追蹤

美國國防部高等研究計劃局曾主導「主題偵測與追蹤」(Topic Detection and Tracking，TDT)計畫，該計劃的研究主題為從新聞廣播的串流中偵測及追蹤新的事件，而「新聞事件偵測與追蹤」為其中之一項目。所謂的事件（event）定義為：「在一些特定的時間及地點所發生的事情。」[6]。例如「在某年某月某日在某地發生車禍」可被視為一個事件，但單獨討論「車禍」這種較廣泛的議題則不算是一個事件。

CMU(Carnegie Mellon University) 與 Umass (University of Massachusetts)都曾進行類似研究，由於評論者對於 CMU 與 Umass 的研究各有正面評價[7][9][10]，本研究參考 CMU 所提出的方法，但考量中文特性與事件追蹤效度，而進行部份技術改良。

2.2.1 事件偵測

所謂事件偵測（detection）可定義為：「發現包含在連續新聞串流中有關新的或先前未發現的事件」[6]，是一種非監督式的學習工作。此外，又可分為「回顧偵測（retrospective detection）」及「線上偵測（on-line detection）」兩種[6]。本研究採用線上偵測的方法。線上偵測是指從一連串接踵而來的即時新聞中標定新事件開始，其後依抵達時間先後輸入新聞文件，再針對進入系統的新進文件判斷是否為新事件，而給予 YES/NO 的輸出決定。

2.2.2 事件追蹤

事件追蹤的目的在於將後續報導文件歸類至先前的事件中[9]。是一種監督式的學習工作，也可說是文件分類的一種應用。CMU 是使用 kNN 分類法 (k-Nearest Neighbor Classification)，並針對 TDT 評估的需要（每個事件都要能獨立的追蹤，而事件中不含其他事件的分類知識），將一般 M-way 的 kNN 法加以修改，成為 2-way kNN 法[9][10]。

2.3 文件摘要技術

2.3.1 文件摘要的定義

摘要是指能正確表達文件內容的一段簡短文字，摘要的目的是產生一個言簡意賅的文件描述，它比文件標題更具相關性，但又短的讓人一眼就明瞭[1]。摘要可以幫助使用者決定是否一篇文件是其所感興趣的，不但能節省使用者的時間，並能提高閱讀原文的理解力。

2.3.2 文件摘要的分類

根據文件摘要所要達到的目的，可以分為下列四種[8]：

1. 指示性摘要（Indicative Abstract）：
   提示使用者文件的存在，並提供足夠資訊，使其決定是否需要閱讀原始文件。

2. 資訊性摘要（Informative Abstract）：
   提供豐富的內容資訊，有時甚至可以用來取代原始文件。

3. 評論性摘要（Critical Abstract）：
   以摘要的型式對原文做評論。此種摘要目前電腦技術尚無法處理。

4. 摘錄（Extract）：
   直接由原文字句中，選取提供事實資料的文句、段落等，視情形而定可能是指示性或資訊性的性質。

若從文件摘要所依據的原始文件數量，文件摘要又可分為單文件摘要（single document summarization）及多文件摘要（multi-document summarization）。單文件摘要把單篇文件內容精簡化與重點化，留下真正能代表文件內涵的資料；多文件摘要則是將多篇探討類似主題的文件結合在一起，刪減及過濾在多篇文章所重複出現的資訊。

2.3.3 中文多文件摘要

目前有關中文多文件摘要的研究仍然很稀少[2][5]。本研究採用過去研究者的發展技術[2]，然針對新聞文件需線上即時處理之特性做若干修改。進行流程如下：

1. 從特定新聞網站定時讀取中文新聞文件。
2. 使用斷詞程式比對文件的中文詞，並標註每一個詞的詞性。
3. 依據關鍵詞相似度進行文件自動分類演算。
4. 進行文件自動摘要處理。

參、系統架構

本研究的系統架構如圖 1 所示，我們嘗試結合並改良先前研究的事件偵測及追蹤技術，並提供每個事件中的新聞文件綜合摘要，以網頁的方式即時呈現，以供讀者快速、完整的了解即時新聞資訊。本研究系統架構可分成三個部分：「網路新聞收集器」、「事件偵測與追蹤系統」、與「多文件摘要系統」。以下針對各研究步驟說明如下。

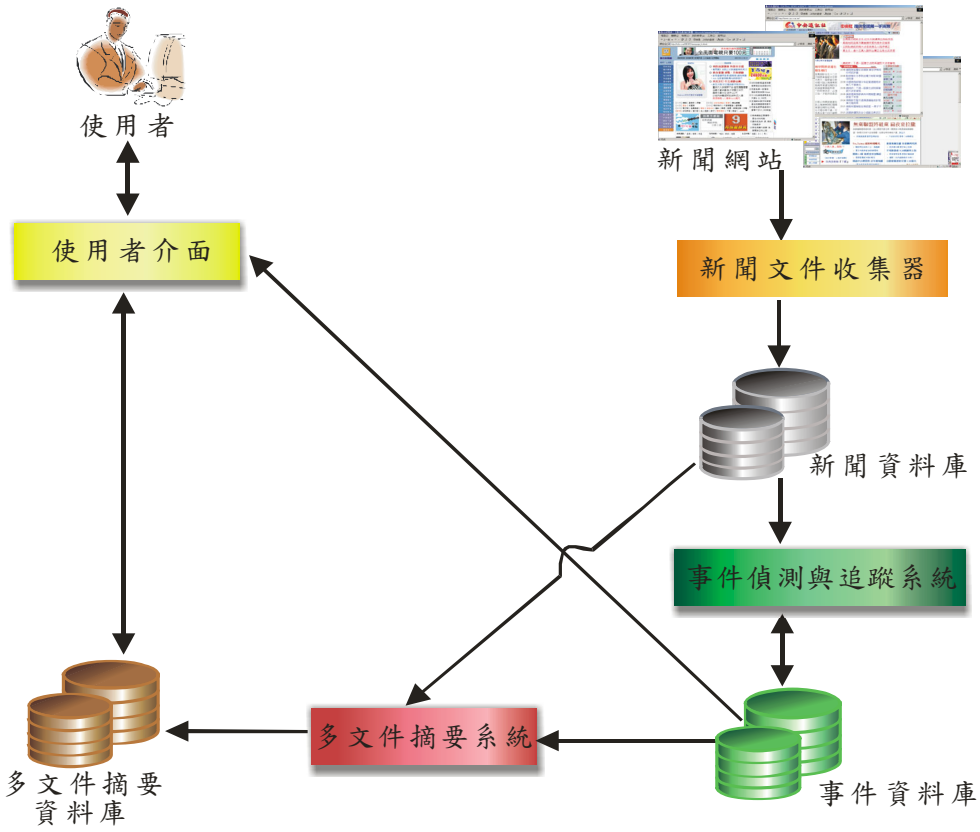

圖 1 本研究系統架構圖

3.1 網路新聞收集器

為獲取即時且多元的電子新聞，做為本研究樣本，本研究以聯合新聞網、中央社新聞及東森新聞報做為中文新聞文件的語料來源。為蒐集所需之新聞文件，我們利用網路新聞收集器至網路入口網站，下載上述三家新聞網站的新聞資料。由於資料量甚大，本研究僅擷取一個月份的電子新聞做為訓練及評估的語料庫，惟俟系統訓練調整完成後，已可設定每間隔適當之時間，自動至上述新聞網站偵測收集新的新聞文件。

3.2 事件偵測與追蹤系統

當新的新聞文件被收集到資料庫後，隨即送至事件偵測與追蹤系統中做進一步的處理。

3.2.1 斷詞與斷句

進行事件偵測與追蹤前首先需進行斷句。斷句是依照句號、問號、驚嘆號來做為句子分隔之依據。首先以程式去除所有超文件標籤，擷取網頁文件內文敘述，並取其標題及文章內容，隨即進行分句作業。句子係依照句號、驚嘆號和問號為句子分隔之依據。在字詞處理方面，以中研院八萬詞庫中的動詞與名詞進行字詞比對擷取作業，關鍵詞選取原則以二至九字詞為限，且以長詞為優先選取對象。為提高字詞處理效率，先刪除一般性字詞，例如：關於、然後等，共 176 個。考慮到在新聞文件中，人名、地名、機關名稱的重要性，以及新詞的辨識，本研究搭配經驗法則、教育部所提供的新詞、以及網路上的資料補強詞庫中有關新詞、人名、地名之不足。

3.2.2 字詞權重計算

本研究以 TFIDF 計算字詞權重。計算公式為：

$$W_{ij} = tf_{ij} \times \log(\frac{N}{df_i}) \qquad (1)$$

其中 W_{ij} 表示字詞 i 在文件 j 的權重，tf_{ij} 為字詞 i 在文件 j 的詞頻，df_i 表示字詞 i 的文件頻率。研究並針對新聞標題所出現的關鍵詞給予較高的權重，並以(TF*3)作為強化該詞重要性的加權計算。由於網路新聞內容長度差距不大，是以並未進行字句長度正規化調整。

3.2.3 事件偵測

事件偵測主要利用 CMU 所提出的方法，亦即利用 single-pass clustering 的概念，並輔以時間區間（time window）來進行來事件偵測。首先計算新進新聞文件與現有事件群集的相似度，若新進新聞文件為系統資料庫的第一篇文件，則將該文件視為一個事件，否則新進文件需與現有事件群集進行相似度比較。本研究相似度計算採用 cosine 相似度公式：

$$sim(x,c) = \frac{\sum_{j=1}^{M} w_{jx} * w_{jc}}{\sqrt{\left(\sum_{j=1}^{M} w_{jx}^2\right) * \left(\sum_{j=1}^{M} w_{jc}^2\right)}} \qquad (2)$$

其中 sim(x,c)代表新進新聞文件 x 對某事件群集的相似度，w_{jx} 為字詞 j 在新進新聞文件 x 的權重，w_{jc} 為字詞 j 在群集 c 的權重，M 為文件集中字詞的總數。

每個事件以向量值表示，並計算出該事件群集內所有新聞文件向量質心（centroid）—即事件群集的平均權重，作為衡量新進文件與各群集相似度的計算標的。此外，考量事件的重要性將隨時間的遞增而衰減，因此特別加上時間區間的計算。如新進文件在時間區間內經計算後，相似值愈小者，則我們認定它是新事件的信心度值 score(x)則愈大。公式為：

$$score(x) = 1 - \max_{c_i \in window} \{(1 - \frac{k}{m}) \times sim(\vec{x}, \vec{c}_i)\} \qquad (3)$$

其中 x 代表新進的文件，c_i 為時間區間中的第 i 個群集，而 i=1,2,…,n，m 為時間區間中所含的文件數，k 為群集 c_i 中最新的一篇文件收錄時間至新進文件 x 到達的時間之間所增加的文件數目。若算出來的 score 大於設定的門檻值，則標定新進的文件為「new」，代表該新進新聞文件為新的事件；反之則標定為「old」，並交由事件追蹤來決定該新聞應歸屬至哪個新聞事件群集。

3.2.4 事件追蹤

事件追蹤的處理是以 CMU 原有方法加以改良的 2-way kNN 分類法。我們的方法為分別計算新聞文件與現有各事件群集之相關分數（relevance score）。 Positive document(正向文件)代表目標事件群集中所包含的新聞文件；negative document(負向文件)則代表目標事件群集以外的其他群集中所包含的新聞文件，公式如下：

$$rel\_score(\vec{x}, kp, kn, D) = \frac{1}{|U_{kp}|}\sum_{\vec{y}\in U_{kp}}\cos(\vec{x},\vec{y}) - \frac{1}{|V_{kn}|}\sum_{\vec{z}\in V_{kn}}\cos(\vec{x},\vec{z}) \quad (4)$$

其中中$\vec{x}$為新進新聞文件之向量，$\vec{y}(\vec{z})$為positive(negative) document的向量，D為整個新聞文件集，k_p為positive document中對於新進新聞文件x的k個最近鄰，k_n為negative document中的對於新進新聞文件x的k個最近鄰，U_{kp}為kp的集合，V_{kn}為kn的集合。計算出來的相關分數若大於所設定的門檻值，則將該新進文件與該群集的關係標定為「YES」，表示該新進文件與該群集相關，反之標定為「NO」。採用此公式的原因是因為2-way kNN計算相關分數時，能同時對於positive document與negative document都取k個最近鄰做比較，以解決先前研究指出所設定之k值太小時，可能取不到positive document的可能性問題[9][10]。

3.3 多文件摘要系統

當一個事件群集有兩份以上文件，本系統即產出一篇多文件摘要，以協助使用者藉由閱讀多文件摘要，快速了解該事件不同報導內容的綜合簡要，以減省分途閱讀全文之時間耗費。誠如前述，本研究在多文件摘要子系統實作是參考Chen and Huang[5]所提出的方法，並做修正而得，主要步驟分為「斷句與斷詞」、「群聚語句」、「形成多文件摘要」等三個部分。

3.3.1 斷詞與斷句

斷詞與斷句的處理，直接使用之前事件偵測的字詞前置處理結果。

3.3.2 群聚語句

群聚語句之做法為針對各文件中描述同一事實（fact）的句子進行群聚，再從各語句群集各取一句代表句以組成摘要，如此即可避免輸出重複描述的語句。句子群聚採用計算句子間相似度的方式進行，而本研究計算句子間相似度也是採用cosine相似度方式來計算。

3.3.3 形成多文件摘要

形成語句群集後，再從中選出代表性句子輸出成摘要。首先決定由哪些詞句群集輸出語句，選取原則是基於假設同一事實(語句群集)，如愈多語句提及，則應表示該事實愈重要，因此以選取涵蓋語句數較多者為摘要候選考量對象。是以語句群集所涵蓋之語句數，依次由大到小輸出摘要語句。

為避免輸出重複句子，一個候選輸出語句群集僅取權重值最高之句子。最後句子輸出的順序則參考該句子在原始文件的位址相對來決定。公式如下：

P=句子在原始文件的位置/原始文件的總句數 (5)

計算所有輸出句子的P值後，P值小的句子會先輸出，如遇P值相同，則依文件號順序，最後形成一篇多文件的摘要。

有鑒於網路新聞文件多半簡短，故本研究以選取2~3句，約175字左右為[預設摘要]；另增加一般論文規定的300字數為「事件內容摘要」。為顧及摘要文意之完整性，所擷取之最後一句雖已超過字數長度限制，仍以完整收錄為準。

肆、系統實作與評估

4.1 系統實作

完成上述處理步驟後，所有新聞文件已被歸類到適當事件，並且產出適當之摘要，表1為部分實作結果。

表1 事件偵測與追蹤結果

事件編號	1847
事件新聞來源	亞太經合會議貿易部長會議揭幕（cna／財經 - 2003/6/2） 林義夫感謝泰菲巴紐三友邦關切台灣SARS疫情（cna／國際 - 2003/6/1） 參加APEC部長會議 林義夫抵泰（udn／國內要聞 - 2003/5/31） APEC貿易部長會議明日正式開會三天（cna／財經 - 2003/5/31） 林義夫率團抵坤敬市參加APEC貿易部長會議（cna／財經 - 2003/5/31） 經長啟程赴泰APEC貿易部長會議預期中共不會有打壓（bcc／兩岸 - 2003/5/31） 林義夫率團赴泰參加APEC貿易部長會議（cna／財經 - 2003/5/31）
預設摘要	二十一個經濟體部長下午開始討論，主辦的泰國官員指出，主要議題有反恐怖主義和安定貿易SECURETRADE，APEC透明化標準、APEC對今年九月間在墨西哥舉行第五屆世界貿易組織的貢獻，在世貿組織杜哈開發議程DOHADEVELOPMENTAGENDA下發展經濟的信心建構，APEC與商業社會

	的交互影響等。林義夫等與巴布亞紐幾內亞副總理兼貿易工業部長馬拉帶領的代表團會談時，台方敦促巴方早日在台灣設置貿易代表團，以增進雙方經貿往來，尤其是漁業合作。
事件內容摘要	二十一個經濟體部長下午開始討論，主辦的泰國官員指出，主要議題有反恐怖主義和安定貿易 SECURETRADE，APEC 透明化標準、APEC 對今年九月間在墨西哥舉行第五屆世界貿易組織的貢獻，在世貿組織杜哈開發議程 DOHADEVELOPMENTAGENDA 下發展經濟的信心建構，APEC 與商業社會的交互影響等。林義夫等與巴布亞紐幾內亞副總理兼貿易工業部長馬拉帶領的代表團會談時，台方敦促巴方早日在台灣設置貿易代表團，以增進雙方經貿往來，尤其是漁業合作。參加這項會議的貿易部長包括中華民國經濟部長林義夫、美國貿易代表左雷克 ROBERTB. ZOELLICK、新加坡貿易工業部長楊榮文、馬來西亞國際貿易兼工業部長拉菲達女士 DATO'SERIRAFIDAAZIZ、南韓外交兼貿易部長黃斗淵、越南貿易部長董汀端 TRUONGDINHTUYEN、日本通產省副大臣高市早苗女士 MS.SANAETAKICHI、中共商業部副部長安民、俄羅斯經濟開發兼貿易部次長皮司克比 ROALDF.PISKOPPEL 等。

使用者可透過使用者界面一「新聞事件瀏覽器」，以事件為單位，瀏覽本研究中所收集的新聞網站新聞，本系統最多可瀏覽 7 天內曾經發生或是有後續發展的事件。為方便使用者瀏覽，我們將收集來的新聞，根據其在原新聞網站中的分類，簡要劃分為三大類「國內要聞」、「兩岸國際」及「財經股市」如表 2 所示。事件歸類則依據事件群集所包含之新聞文件類別在三類中何者所佔比例較大，而定其歸屬。對於每一事件群集標題的給定是以該事件群集中，TFIDF 加總值最高之新聞文件標題做為事件的標題，並於標題下方列出該事件摘要(如圖 2)。協助使用者了解該事件的內容描述，也節省使用者尋找感興趣的事件。

表 2 事件分類規則表

事件類別	國內要聞	兩岸國際	財經股市
原新聞文件類別	政治、社會、生活、健康、國內要聞	兩岸、國際、兩岸國際	股市理財、產業財經、財經

整個網站架構可以分成三個部分：

1. 首頁

於新聞事件瀏覽器首頁中(如圖 2)，系統預設會顯示當日發生最重要的數個事件。當中「頭條大代誌」配置的是當日發生或有更新的事件中，所包含新聞文件 TFIDF 加總值最高事件群集。而「焦點新聞事件」部分則放置三大類新聞事件中，除卻置於「頭條大代誌」的事件群集外，各分類 TFIDF 加總最高的事件。並提供「預設摘要」供使用者閱讀。

圖 2 新聞事件瀏覽器首頁

2. 分類事件清單頁面

使用者點選進入各分類網頁後(如圖 3)，可以看到各分類新聞事件清單，系統預設顯示該分類於當日發生或有後續報導的新聞事件，且依照各事件包含新聞之 TFIDF 加總值由大到小依序顯示。此外，分類事件清單頁面如同首頁一樣，亦提供使用者「預設摘要」。

圖 3 分類事件清單頁面

3. 事件內容頁面

使用者點選進入事件群集後，即可看到事

件所包含的新聞文件(如圖 4)。新聞文件依收錄日期排序，愈新收錄的新聞排在前面。當使用者欲閱讀新聞文件的內容，點選新聞標題後隨即顯示該新聞內容。

圖 4 事件內容頁面

4.2 系統評估

為驗證本研究研發之新聞事件自動偵測與追蹤與多文件摘要系統之適用性，我們亦設計供使用者瀏覽及評估事件的系統。本研究以雲林科技大學資管所之研究生為受測對象，先由研究者向受測者說明系統操作方式、版面編排與問卷題目後，便由受測者上網測驗本研究之「新聞事件瀏覽器」。在評估過程中，受測者可以任意瀏覽新文事件，並於三個新聞事件分類中，各取三個有興趣的事件做評估。整個評估過程共有 25 位有效受測者參與。本研究線上問卷評估項目分為三個部分，第一部分為「使用新聞網站需求與意見分析」、第二部分為「新聞事件偵測與追蹤效益評估」、第三部分為「多文件摘要效益評估」。各項評估分析如次：

4.2.1 使用新聞網站需求與意見分析

此部分主要是調查受測者對於使用目前電子新聞網站的經驗與感受，以及受測者對於閱讀相關新聞和持續關注新聞發展的需求程度，共有四個題目。

分析結果如圖 5 所示，絕大部分受測者對於其感興趣的新聞報導，在時間許可下，願意花時間尋找及閱讀相關新聞及持續關注新聞發展。然而約有半數的受測者對於現今的新聞網站在須尋找及搜尋相關新聞感到不便。

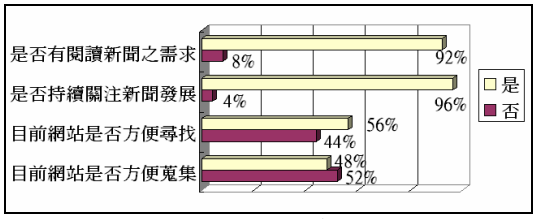

圖 5 使用新聞網站需求與意見分析

4.2.2 新聞事件偵測與追蹤效益評估

新聞事件偵測與追蹤效益評估項目共有四個題目，原題目以五等距方式要求受測者勾選，為方便統計，本研究將答題結果歸併為三等距。資料結果顯示有七成六的受測者肯定新聞事件群集的相關程度，，即認為本系統群聚之新聞文件所描述之事件性質相關;六成受測者認為新聞事件群集標題具有代表性(如圖 6)。約有七成的受測者對於本系統新聞事件群集所蒐集之新聞報導的完整性，以及事件群集協助了解報導的來龍去脈表示很有幫助(如圖 7)。

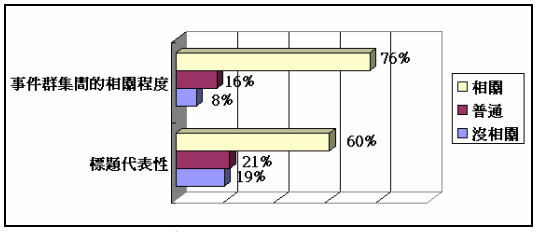

圖 6 新聞事件群集及標題代表性評估

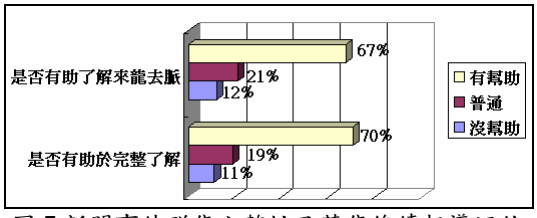

圖 7 新聞事件群集完整性及蒐集後續報導評估

4.2.3 多文件摘要效益評估資料分析

多文件摘要效益評估項目有三，即：摘要有助於了解新聞事件的程度、可讀性以及字數的適當性，並分別針對「預設摘要」及「事件內容摘要」進行評估，共有六個題目，分析結果如圖 8、9、10 所示。

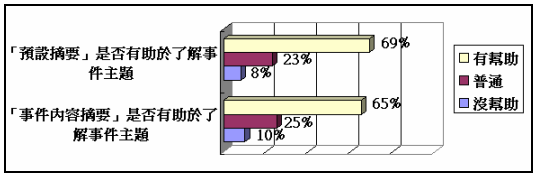

圖 8 摘要是否有助了解事件程度之分析圖表

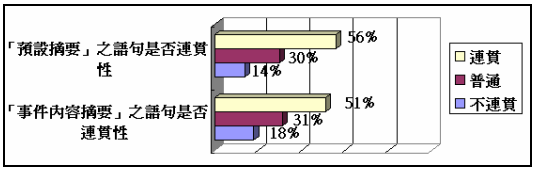

圖 9 摘要語句是否連貫性之分析圖表

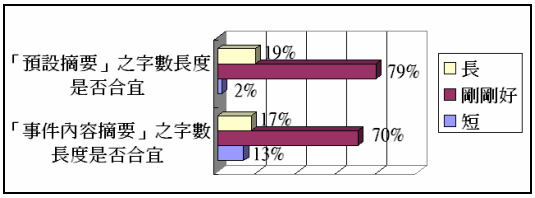

圖 10 摘要字數長度是否合宜之分析圖表

由於本系統並無法保證每篇新聞文件都與新聞事件群集的內容主題有關，因此提高摘要內容代表性與可讀性，成為本研究後半最重要之挑戰。惟經受測者評估結果發現，對於摘要之各項評估，在語句連貫性已有半數以上的受測者持肯定的看法；近七成的受測者肯定預設摘要有助於了解新聞事件主題；近八成的受測者認為預設摘要的字數長度是適合的。上述資料顯示本系統的多文件摘要技術已達到一定的水平。

伍、結論與未來研究方向

本研究利用事件偵測與追蹤以及多文件摘要技術所建構之系統，能自動偵測新聞事件的發生及持續追蹤其發展，且利用摘要技術產出多文件摘要。研究結果顯示本系統確實可有效幫助使用者了解新聞事件報導的來龍去脈，對於摘要的字數長度多持正面肯定的看法，摘要語句連貫性也有半數以上受測者的支持。建議後續研究可著眼於：

1. 關鍵詞選錄問題: 由於時事變化快速，所包括之人、地、時、物等詞彙數量，難以計數。本研究以詞庫式斷詞法為主要斷詞方法，雖加入教育部公佈新詞，仍有掛一漏萬之虞，也因此無法取出適當的文件特徵。
2. 文件特徵權重值計算的加強，為加強新聞聚類，未來研究可針對某些特徵值進行加權計算，或可提高文件辨類率。當文件中真正重要的特徵能被擷取出來後，在聚類及摘要計算結果的品質必然會有所提升。
3. 新聞事件群集標題的改進，本研究新聞事件標題的給定是以新聞事件群集中 TFIDF 加總值最高的一篇為準，然而依據單篇新聞標題可能無法代替事件主題。
4. 事件摘要品質的提升，本研究所產生的摘要仍僅限於重要文句萃取與組合，未來研究如語意技術可突破，可朝智慧型改寫摘要發展。

陸、參考文獻

[1] 黃純敏、吳郁瑩, “網路文件自動摘要”, 台灣區網際網路研討會 TANET 99, 國立中山大學承辦, 1999.

[2] 翁鴻加, ”多文件摘要一些新技術及評估模型之建立”, 國立台灣大學資訊工程研究所碩士論文, 2001.

[3] 顧皓光,”網路文件自動分類”,國立台灣大學資訊管理研究所碩士論文,1996.

[4] H. H. Chen, and S. J. Huang, ”A summarization system for Chinese News from multiple sources.” *Proceedings of 4th International Workshop on Information Retrieval with Asia Language*, pp. 1-7,1999.

[5] H. H. Chen and C. J. Lin, “A Multilingual News Summarizer,” *Proceedings of 18th International Conference on Computational Linguistics*, pp. 159-165, 2000.

[6] J. Allan, J. Carbonell, G. Doddington, J. Yamron and Y. Yang, ”Topic detection and tracking pilot study: Final report,” Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[7] J. Allan, R. Papka and V. Lavrenko, ”On-line New Event Detection and Tracking,” *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 37-45, 1998.

[8] J. E. Rush, R. Salvador, and A. Zamora, "Automatic abstracting and indexing.II. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria", *Journal of the American Society for Information Science*, Vol.22, No.3, pp.260-274, 1971.

[9] Y. Yang, J. G. Carbonell, R. Brown, T. Pierce, B. T. Archibald and X. Liu, “Learning approaches for detecting and tracking news events,” *IEEE Intelligent System*,Vol.14,No.3, pp. 32-43, 1999.

[10] Y. Yang, T. Ault and T. Pierce, ”Improving text categorization methods for event tracking,” *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 65-72, 2000.