以文件倉儲概念實現動態群聚與多重文件摘要之研究—以中文電子新聞為例[1]

魏玲玉
高雄第一科技大學資訊管理系

曾守正
高雄第一科技大學資訊管理系

摘要

由於電子文件的數量成爆炸性成長，如何有效率地將文件歸納，以方便日後快速瀏覽與查詢，已經是知識管理領域中刻不容緩的課題。傳統上仰賴反轉索引檔(Inverted Index File) 為基礎的全文檢索技術，往往搜尋出相當龐大且雜亂的文件資料，所以還需經過進一步的篩選，才能找到真正有用的文件。這樣的應用模式已經無法滿足使用者快速瀏覽與查詢的需求。在本論文中，我們應用文件倉儲的概念將文件予以結構化儲存，配合多維度查詢的機制，找出具有相關性的文件以進行多重文件摘要與動態群聚之研究。整體概念透過實作 DNCSS 系統 (Dynamic News Clustering and Summarization System) 來驗證其效果，我們應用資料倉儲處理數值資料的概念來處理文件資料，建立文件倉儲將文件所包含的結構化資訊應用在文件儲存、搜尋與整合上，並提供多維度查詢。更運用動態群聚的概念，幫助使用者組織對文件倉儲作查詢所回傳之查詢結果。最後以多文件摘要系統對每一個文件群聚結果產生一份多文件摘要，方便使用者瀏覽文件集合的精要內容，以更有效率的方式取得有用的資訊。我們以台灣地區各大網路新聞文件為實例來驗證本系統之效果，經人工評估後獲得相當正面之評價，顯示本研究確實能提供使用者快速且有效地獲取符合需求的文件資訊。

關鍵字：資訊檢索、文件倉儲、多文件摘要、文件群聚

---

[1] This research was partially supported by National Science Council, TAIWAN, ROC, under Contract No. NSC 94-2416-H-327-009.

A Study on Multi-Document Summarization Based on Document Warehousing and Dynamic Clustering －Using Internet News as Examples

Ling-Yu Wei
Department of Information Management, National Kaohsiung First University of Science and Technology
Frank S.C. Tseng
Department of Information Management, National Kaohsiung First University of Science and Technology

Abstract

As electronic documents proliferate drastically, for contemporary knowledge management, it is indispensable to provide a mechanism for integrating and sorting huge volume of documents for quick browsing and efficient query processing. Traditionally, full-text searching systems were usually based on inverted-index, which is usually huge in volume and unsorted. That makes users suffer from easily determining the information embedded in the collection. Therefore, for document searching over the Internet, such systems are no longer satisfactory for user's need. In this paper, we propose a general framework for document clustering and multi-document summarization based on the concept of document warehousing. Based on our framework, we have implemented a prototype system, named DNCSS (Dynamic News Clustering and Summarization System) to be the test bed of our approach. The system adopts the concept of document warehousing, which models text-oriented documents into multi-dimensional viewpoints. The constructed document warehouse can be regarded as the main repository for our system and it flexibly organizes document structure information for user's searching and querying. Moreover, the retrieved documents from the document warehouse will be further clustered by some clustering techniques to provide a more organized structure. Finally, our system generates a multi-document summary for each cluster to support users finding distilled information more efficiently. We have collected the most famous on-line news in TAIWAN from the Internet as the testing examples to verify the effectiveness of our system. The evaluation result shows that our approach positively alleviates users from reading large amount of related news and elaborating the necessary conclusion effectively.

Keywords: Information Retrieval, Document Warehouse, Multi-Document Summarization, Document Clustering.

壹、引言

網際網路的興起，讓過去手寫紙抄的書寫習慣，已經被數位檔案所取代。而文書信件等資訊的傳送，現在也大量透過網際網路來進行。人們接收與傳送資訊的管道比以往變得更多元，內容也更豐富。各種資訊與文件隨著數位化資訊時代的來臨，不斷累積增長。資訊科技的變革也對資訊的儲存、呈現、處理與交換的方式產生很大的影響，徹底改變了人類資訊傳遞與知識累積的過程。

透過網際網路，可以找到大量的文件資料，但龐大的數量也往往讓人無所適從。在知識管理的目標中，除了收集與整理文件外，更需要能進一步瞭解、詮釋及組織這些文件，使之成為具有價值的知識。傳統透過反轉索引檔 (Inverted Index File) 的檢索方法，由於所搜尋出來的文件數量龐大，往往還需經過進一步的篩選，才能找到真正想要的文件，因此已經無法符合當代使用者快速組織文件的需求了。

一、研究動機與目的

運用資料倉儲 (Data Warehouse) 管理大量的資料，已是行之有年的成熟技術。然而，資料倉儲主要是針對數值形態資料做處理，而文字導向 (Text-Oriented) 的文件則在儲存結構、資料來源、目的與所使用的檢索技術與數值資料有所差異，因此我們需要將概念擴充以形成所謂的「文件倉儲」(Document Warehouse)。一份文件的內容，其實包含了許多階層概念的資訊 (Hierarchical Information)，如：時間的計算方式含有階層的關係：年可細分為季；季可細分為月；月可細分為日；地點表示法也有相同的概念：國家可細分為省份、省可細分為縣 (市)；縣 (市) 又可細分為鄉 (鎮)。本研究應用文件倉儲的概念，建立完整的文件倉儲系統讓這些結構化資訊可以被應用在文件搜尋與整合上。

此外，文件群聚 (Document Clustering) 的技術已被證明是一種能夠加強資訊檢索效能的方法 (Salton & McGill 1983)。Van Rijsbergen (1979) 對於群聚的基本假設是能將相關與不相關的資料分別歸屬於不同的群集。近年來，群聚的技術也已經成功地被用在文件檢索的查詢上 (Carey *et al.* 2000; Hearst & Pedersen 1996; Leuske 2001; Wu *et al.* 2001; Zamir & Etzioni 1998)。此種針對查詢結果做群聚的方式稱為 Post-Retrieval Clustering，其目的是組織檢索系統所檢索出來的結果，並讓這些結果易於瀏覽。但在某些情況下，群集所給予的資訊 (如：關鍵字列表)，並不足以清楚表達文件群集的內容，此一現象雖然讓群聚結果得以幫助使用者發現具有高度相關的文件群集，但卻沒有辦法提供更進一步簡化的摘要資訊供使用者快速瀏覽。

本研究嘗試結合文件倉儲、文件群聚與多重文件摘要 (Multi-Document Summarization) 的概念，有效組織與處理文件資料，並以新聞文件為實驗範例，實現了 DNCSS 系統 (Dynamic News Clustering and Summarization System)，讓使用者在面對大量文件時，能夠依個人不同的需求選擇適當的維度切入，再運用動態群聚的概念，動態群聚查詢結果，使得相關的文件歸屬於相同的群聚，然後針對各群聚產生一份多

重文件摘要，方便使用者瀏覽文件集的精要內容，或選擇符合其需求之群集，以便更有效率地取得所需資訊。系統完成後，我們藉由人工評估的結果，獲得了相當正面之評價，顯示本研究確實能提供使用者快速且有效地獲取符合需求的文件資訊。

二、研究範圍與限制

我們的研究對象以中文新聞文件為主，選用中文新聞文件有資料取得方便且較一般文件主題更明確的優點，然而因為資料來源本身的限制，也會有以下的研究限制：

1. 僅處理中文：本研究之對象為中文新聞文件，且所使用的語言為中文。而新聞文件或多或少都會夾雜英文字詞，由於本系統僅針對中文做處理，因此在英文字詞的部分暫不處理之。
2. 詞彙使用差異：由於寫作習慣的不同，在詞彙的使用上也會有所差異，本研究分成以下兩種情況來說明：
   (1) 譯名的差異：新聞中部分由英文或其他語言翻譯而來的字彙，會有譯名不同的差異，例如：人名「Jordan」可能會翻譯「喬登」或「佐頓」；地名「San Francisco」會翻譯成「舊金山」或「三藩市」。本研究將採同義字典的方式，蒐集同義詞彙處理大部分翻譯名詞的問題。
   (2) 慣用字的差異：在中文裡有一些字是可以相互替代的，如：「台灣」、「臺灣」；「撞球檯」、「撞球枱」。其中「台」和「臺」，「檯」和「枱」是可以相互替換的。因此，為避免在查詢時造成誤差，本研究只取其中一種字詞用法，將其他的用字統一替換成一樣的用法。

貳、文獻探討

我們希望結合文件倉儲的概念與資訊檢索技術，協助使用者將文件做有系統的整合，以便更容易找到所需的文件。相關研究分述如下：

一、文件倉儲簡介

根據 Survey.com (http://www.survey.com) 的分析結果顯示：企業所需要的商業智慧大約只有 20% 是由存放在傳統關聯式資料庫中的結構化資料所推導出來的，其餘 80% 左右的商業智慧必須要到各式各樣的商業文件中去找尋。然而，目前企業界對於這些文件的管理上也僅止於文件本體的管理，對於文件的內容仍然是以人為閱讀的方式來吸收，效率不彰且可能流於以偏蓋全。再者，資料倉儲所針對的是數值形態的資料，而文字導向 (Text-oriented) 的文件則在儲存結構、資料來源、目的與所使用的檢索技術上與數值資料有所差異。一份文件的內容其實包含了許多階層概念的資訊 (Hierarchical Information) ，我們可以應用資料倉儲處理數值形態資料的概念處理文件資料，透過「線上分析處理」(OLAP, On-Line Analytical Processing) 的技術，讓這些結構化資訊可以被應用在文件搜尋與整合上。

文件倉儲 (Document Warehousing) 是近幾年才興起的概念，因此目前的研究仍處於起步階段。整體目標是希望提供文件資料的線上分析處理，正如同資料倉儲 (Data Warehousing) 希望提供數值資料的線上分析處理一般。此概念的落實有賴 Sullivan

(2001) 提出，隨後由 McCabe 等人 (2000ab) 提出一個透過多維度模式與 MDX 查詢語法，以達成布林檢索與相似度排序等檢索機制的文件倉儲雛形。Bleyberg & Ganesh (2000) 與 Bleyberg & Paranjape (2001) 則是將文件倉儲的重點放在語意分類 (Atomic Semantic Categories) 上，希望能透過語意上的階層架構達到文件分類的目的，其研究使用自然語言處理技術與語法規則，配合決策樹 (Decision Tree) 的使用，建構完整的語意分類，但對於使用者的查詢與文件的呈現並無具體成果。近來我們也曾就文件倉儲的多維度查詢語言進行完整的擴充設計 (Tseng 2005)，並提出整體的系統架構設計 (Tseng & Chou 2006)，以及相關的索引設計技術 (Tseng & Lin 2006)。在本研究中，我們希望能進一步從實務面著手，將文件倉儲的概念應用在文件群聚與自動化摘要的過程中。以下我們對文件 (Document)、維度 (Dimension) 與文件方塊 (Document Cube) 等做一番正規的定義。

定義一： 所謂的「文件」(Document) 是指一份具有邏輯意義的本文 $T = \{k_1, k_2, \ldots, k_n\}$，其本身的邏輯意義是由 T 本身所內含的數個關鍵字 $k_1, k_2, \ldots, k_n$ 所描述。

將許多文件組成文件方塊後，可以透過多維度的文件呈現 (Multi-Dimensional Document View) 方式瀏覽與查詢文件。而要建構文件方塊之前必須要先建立各個維度 (Dimension)，我們定義如下：

定義二： 「維度」(Dimension) D 是一個具有 m ($m \geq 1$) 個階層 (Level) 的樹狀結構，用來表示一群關鍵字彼此間的階層關係 (Hierarchical Relationships)。結構中的節點稱為「成員」(Member)，其根節點 (Root)代表著整個維度下所有節點的整體概念，我們以"(*All D*)"或 $D(1) = \{(All\ D)\}$ 來表示之；同理 $D(2)$ 則代表階層 2 的整體概念，也就是說 $D(2)$ 為階層 2 上的所有節點所組成的集合，依此類推。另外，我們以 $D(0)$ 表示該維度中所有節點的集合，即 $D(0) = \cup_{1 \leq i \leq h} D(i)$，其中 h 為該維度的樹狀結構高度。

對於任一個維度 D 而言，我們可以進行兩個基本運算，稱為「向下鑽研」(Drill-Down) 與「向上提昇」(Roll-Up)，其中由任一個內部節點向下展開所有子節點稱為「向下鑽研」(Drill-Down)。相反地，由任一節點往父節點收縮則稱為「向上提昇」(Roll-Up)。

範例一：圖 1 為「地區」維度的示意圖，它代表著臺灣地區的各大城市階層關係。假定此維度稱為 D，則 $D(1) = \{(\text{All 地區})\}$，$D(2) = \{\text{北, 中, 南, 東}\}$，依此類推。在此維度中，由「南」這個節點向下鑽研，可以得到「台南」、「高雄」、「屏東」三個子節點；由「高雄」這個節點向上提升，則可以得到「南」這個父節點。

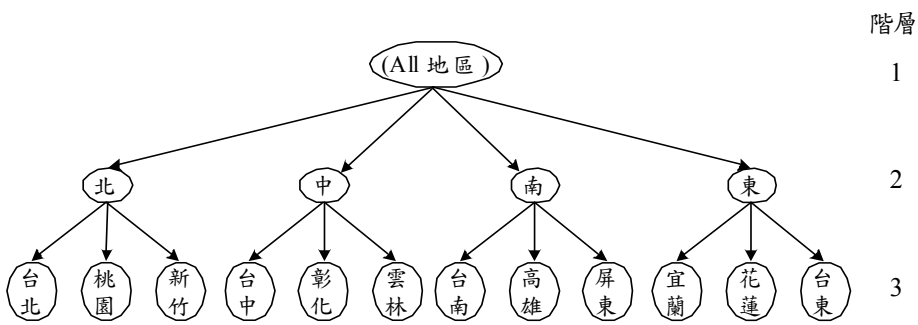

圖 1：「地區」維度示意圖

透過維度的概念，我們可以對文件進行掃描以得到「文件索引」(Document Index)，定義如下：

定義三：對於具有唯一識別編號為 id_T 的文件 T 而言，我們將其定義於 n 個維度 $(D_1, D_2, \ldots, D_i, \ldots, D_n)$ 上之「文件索引」(Document Index) 表示為 $x = (id_T, K_T)$，其中 $K_T = (K_1 K_2,\ldots, K_i,\ldots, K_n)$ 是由 n 個關鍵字集 K_i 所構成的值組 (Tuple)，使得所有 K_i 中的元素 k_{ij} 都會在 T 中出現，而且 $k_{ij} \in D_i(0)$，$1 \le i \le n$。 □

關於定義三，要注意的是：如果某個維度 D_i 中的關鍵字並未出現在文件 T 中時，則上述文件索引 $x = (id_T, K_T)$ 的 K_T 中之 K_i 會是空集合。對於任一個文件索引的 $K_T = (K_1 K_2,\ldots, K_i,\ldots, K_n)$ 而言，如果所有的 $|K_i|$ 都等於 1，$1 \le i \le n$，則我們稱此文件索引為「基底文件索引」(Base Document Index)，而 $K_T = (K_1, K_2,\ldots, K_i,\ldots, K_n) = (\{k_1\}, \{k_2\}, \ldots, \{k_i\}, \ldots, \{k_n\})$ 便可以簡化為 $K_T = (k_1, k_2, \ldots, k_i, \ldots, k_n)$；若其中有一個 $|K_i| > 1$，而且其它的 $|K_j|$ 都等於 1，$1 \le j \ne i \le n$，則我們稱此文件索引為「複合文件索引」(Composite Document Index)；如果有某個 $|K_i| = 0$，則我們稱此文件索引為「退化性文件索引」(Degenerate Document Index)。一個具有 $|K_i| = 0$ 的退化性文件索引，在我們的研究中，將會以該相對維度的最上層元素集 $D_i(1)$ 來取代 K_i，以便將空集合推廣成整體維度的概念，避免該文章在後續透過該維度搜尋時被忽略。

接下來，我們定義文件方塊的基本組成結構—「方格」(Cell) 如下：

定義四：對於一個定義於 n 個維度 $(D_1, D_2, \ldots, D_n)$ 之上的「方格」(Cell) 表示為 $c = (t_c, X_c)$，其中 $t_c = (c_1, c_2,\ldots, c_i,\ldots, c_n)$，$c_i \in D_i(0)$，$1 \le i \le n$，而且 $X_c = \{x_1, x_2,\ldots, x_j,\ldots, x_k\}$ 是一個文件索引所構成的集合，裡面包含了 k 個文件索引 $x_j = (id_{Tj}, (K_1, K_2,\ldots, K_i,\ldots, K_n))$，可用來索引到文件 T_j 的唯一識別編號 id_{Tj}。而所有可以透過方格 c 索引到的文件唯一識別編號所成的集合，我們以 $ID(c)$ 來表示之。另外，對於任一方格 $c = (t_c, X_c)$ 來說，其中 $t_c = (c_1, c_2,\ldots, c_i,\ldots, c_n)$，我們稱 c 是一個「基本方格」(Base Cell)，若且唯若任何 c_i 都對應到其所屬維度上的葉節點；反之，則稱 c 是一個「非基本方格」(Non-Base Cell)。

以下我們定義「文件方塊」(Document Cube) 如下，其示意圖如圖 2 所示：

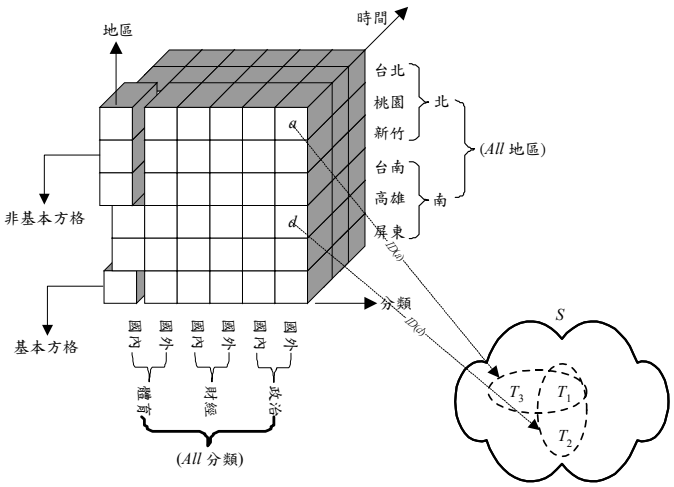

圖 2：文件方塊示意圖

定義五：一個定義於 n 個維度 $(D_1, D_2, \ldots, D_n)$ 之上的「文件方塊」(Document Cube) 表示為 $DC = (S, (D_1, D_2, \ldots, D_n))$，其中 S 是一個文件集合，而 DC 是一個由所有方格 $c_i = (t_{c_i}, X_{c_i})$ 所組成的方塊結構，其中 $t_{c_i} \in \bigtimes_{1 \le j \le n} D_j(0)$ 而且 $ID(c_i) \subseteq S$。 □

在本研究中，我們將運用文件倉儲的概念，透過維度先將文件進行初步分類，以利後續的群聚處理，以及多重文件摘要作業。

二、中文文件處理

本研究對象以中文文件為主。由於中文字之間並不像西方語文一樣使用明顯的空白做區隔，所以必須要先經過斷詞 (Word Segmentation) 的程序把中文句子切割分成合理的字詞單元，才能夠做後續的處理 (Yeh & Lee 1991; Chen & Kiu 1992; Nie *et al.* 1996)。

一般特徵詞的擷取技巧主要有下列四種方法：

1. 詞庫比對法 (Dictionary-Based Method)：透過詞庫的建立，以擷取出文件中出現在詞庫的字詞，如：Carey *et al.* (2001)、陳永德 (1997)。
2. 文法剖析法 (Syntax-Based Method)：主要是以自然語言處理技術剖析文法規則，以取出名詞片語，例如：Trigg & Weiser (1987)。
3. 統計分析法 (Statistical-Based Mthtod)：透過大量文件或語料庫 (Corpus) 的訓練，取得斷詞參數的統計資訊，如：Sproat & Shih (1990)、曾元顯 (1997)。
4. 混合式斷詞法 (Hybrid Method)：混合上述的方法而成，如：Meyrowitz & van Dam (1982) 等。

另外，由於中文斷詞具歧義問題，陳永德 (1997) 的研究列出了中文歧義的類型有：句子結構歧義、詞彙歧義、詞類歧義、及詞間歧義等四種。為了解決歧義問題，一般我們會採用三大斷詞規則：長詞優先、前詞優先、詞間頻率對比，來降低歧義性所造成的斷詞錯誤問題。

在經過斷句斷詞取得每篇文件的所有詞彙後，必須透過衡量該詞彙在文件中的權重以計算其在文件中的重要性。詞彙權重可藉由同時考量該字詞在特定文件中的權重 (Local Weight)，以及整體文件集中的權重 (Global Weight) 來計算 (Popescu 2001)，並經正規化後得到詞彙權重，其表示法如公式 1，其中 d_{ij} 表示詞彙 i 在文件 j 中的權重；L_{ij} 表示詞彙 i 在文件 j 中的 Local Weight；G_i 表示詞彙 i 的 Global Weight；N_j 則表示文件 j 的正規化因子 (Normalization Factor)。

$$d_{ij} = L_{ij} \; G_i \; N_j \qquad \text{(公式 1)}$$

表 1 則列出 Popescu (2001) 所整理的 Local Weight 及 Global Weight 計算方式。其中 f_{ij} 為詞彙 i 在文件 j 出現的次數，我們稱為「詞頻」(Term Frequency)；a_j 為文件 j 中所有詞彙詞頻的平均數；x_j 為文件 j 中出現次數最多的詞彙的數目；N 為文件集中之文件總數；n_i 為詞彙 i 出現的文件數，我們稱為「文件頻率」(Document Frequency)；F_i 為詞彙 i 在整個文件集的出現總次數；而 l_j 則為文件 j 中個別詞彙的數目。另外，依照先前學者的研究經驗顯示 slope 通常是取 0.2 為最適當。

表 1：常見的權重計算計算公式 (資料來源：Popescu (2001))。

Local weight 公式	公式名稱
$1 \quad if\ f_{ij} > 0$ $0 \quad if\ f_{ij} = 0$	Binary
F_{ij}	Within-document frequency (Term Frequency)
$1+\log f_{ij} \quad if\ f_{ij} > 0$ $0 \quad if\ f_{ij} = 0$	Log
$(1+\log f_{ij})/(1+\log a_j) \quad if\ f_{ij} > 0$ $0 \quad if\ f_{ij} = 0$	Normalized Log
$0.5 = 0.5(f/x) \quad if\ f_{ij} > 0$ $0 \quad if\ f_{ij} = 0$	Augmented normalized term frequency

Global weight 公式	公式名稱
$\log(N/n_i)$	Inverse document frequency
$\log[(N-n_i)/n_i]$	Probabilistic inverse
$1-\sum_{j=1}^{N}\frac{\frac{f_{ij}}{F_i}\log\frac{f_{ij}}{F_i}}{\log N}$	Entropy
F_i/n_i	Global Frequency IDF
1	No global weight

Normalization factor 公式	公式名稱
$\frac{1}{l_j}$	Document length
$\frac{1}{\sqrt{\sum_{i=0}^{L_j}(G_iL_{ij})^2}}$	Cosine normalization
$\frac{1}{(1-slope)\,pivot+slope\,l_j}$	Pivoted unique normalization

三、向量空間模型

「向量空間模型」(Vector Space Model, VSM) 是由 Salton *et al.* (1975) 所提出來最廣為使用的資訊檢索模型，它是一種由關鍵詞與文件所組成的向量矩陣空間，我們說明如下。

(一) 向量表示法

向量空間模型的概念是以文件中所含的詞彙 (Term) 所組成的向量空間來表示文件集。為了能將向量空間模型以數值化的型態表示，該模型將字詞權重拿來表示向量中的元素，因此文件可以寫成 $D = (w_1, w_2, \ldots, w_i, \ldots, w_n)$，其中 w_i 為該詞彙在文件 D 中的權重。

利用「詞彙-文件矩陣」，可以很容易表達出詞彙與文件的觀念。圖 3 為一個有 k 份文件集合與 i 個詞彙的「詞彙-文件矩陣」。其中 W_{ik} 為字詞 i 在文件 k 的權重。在向量空間中「詞彙-文件矩陣」的表示方法。圖 4 為在向量空間中「詞彙-文件矩陣」的表示方法。

$$\begin{array}{c} \quad Term1\ Term2\ \cdots\ \cdots\ Termi \\ \begin{matrix} Doc1 \\ Doc2 \\ \vdots \\ \vdots \\ Dock \end{matrix}\begin{pmatrix} W_{11} & W_{12} & \ldots & \ldots & W_{1i} \\ W_{21} & W_{22} & \ldots & \ldots & W_{2i} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ W_{ki} & W_{k2} & \ldots & \ldots & W_{ki} \end{pmatrix} \end{array}$$

圖 3：詞彙-文件矩陣

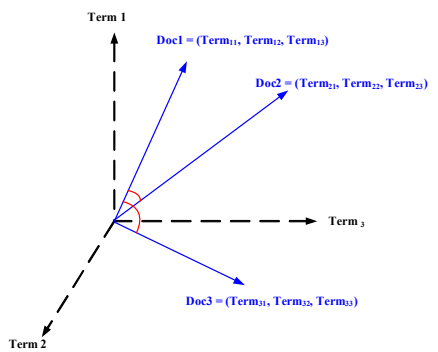

圖 4：向量空間模型

(二) 相似度計算

當文件以向量空間模型量化表示後，我們就可以利用表 2 中所列的相似度計算公式 (資料來源：Salton (1988)) 來算出文件或文句間的相似程度，以進行群聚或分類的處理。

表 2：常見相似性計算公式 (資料來源：Salton (1988))

Similarity Measure sim(X, Y)	Evaluation for Binary Term Vectors	Evaluation for Binary Term Vectors
Inter product	$\lvert X \cap Y \rvert$	$\sum_{i=1}^{t} x_i \cdot y_i$
Dices coefficient	$2\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$	$\frac{2\sum_{i=1}^{t} x_i \cdot y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2}$
Cosine coefficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert^{1/2} + \lvert Y \rvert^{1/2}}$	$\frac{\sum_{i=1}^{t} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{t} x_i^2 \cdot \sum_{i=1}^{t} y_i^2}}$
Jaccard coefficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert \cup \lvert Y \rvert}$	$\frac{\sum_{i=1}^{t} x_i \cdot y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2 - \sum_{i=1}^{t} x_i \cdot y_i}$

$X = (x_1, x_2, \cdots, x_t)$
$\lvert X \rvert$ = number of terms in X
$\lvert X \cap Y \rvert$ = number of terms appearing jointly in X and Y

四、文件群聚技術介紹

群聚技術 (Clustering)，屬於非監督式學習 (Non-Supervised Learning)，不預先設定分類所代表的意義，而把資料以群聚技術將性質內容相近的資料聚集成群集後，再分析定義各群聚的意義，因此與監督式學習 (Supervised Learning) 的分類 (Classification) 技術並不相同。群聚的方法有很多，其中以「切割式群聚法」與「階層式群聚法」為最常見的兩種演算法 (Jain & Dubes 1988)。

切割式群聚法需要由使用者事先決定要將資料分為多少個群集，也就是要將資料分成指定的 k 個群集，常用的方法有 k-means (MacQueen 1967; Jain *et al.* 1999; Leuske 2001)，以及 k-medoid (Kaufman & Rousseeuw 1990) 演算法等。

在階層式群聚法的研究中，可分為：「分裂式階層群聚法」(Divisive Hierarchical Clustering)，以及「凝聚式階層群聚法」(Agglomerative Hierarchical Clustering) 兩種 (Jain & Dubes 1988; Jain *et al.* 1999)。分裂式階層群聚法其進行的方向是 Top-Down：先將所有物件分在同一群，然後一步一步將目前最不相似的群分開。凝聚式階層群聚

法則是 Buttom-Up：先將所有物件當作個別一群，再一步一步將群間相似度最高的群合併。在凝聚式階層群聚法中，根據群和群之間相似度量測的定義不同，有以下四種作法 (Jain & Dubes 1988)：單一連結法 (Single-Linkage Method)、完整連結法 (Complete-Linkage Method)、平均連結法 (Average-Linkage Method) 及沃德法 (Ward's Method)。

另外，Salton & McGill (1983) 也提出計算成本比較小的啟發式群聚法 (Heuristic Clustering Method) 方法，稱為 Single-Pass Clustering。啟發式的群聚法的效果雖然不及完全連結法與單一連結法來得好，但其優點是快速且簡單，計算的方式也較為直覺，較適合使用在即時 (Real-Time) 或線上 (On-Line) 處理的群聚系統。此種群聚方法所需的計算成本較小，時間複雜度僅 O($n \log n$)，比單一連結法的 O(n^2) 及完全連結法的 O(n^3) 要好。其他相關的群聚方法，請參考 Jain *et al.* (1999) 以及 Sebastiani (2002)。

五、自動文件摘要

摘要就是以最精簡的文字表達原始文件中所要傳達的訊息，也可以說是原始文件的精簡版本 (Edmundson & Wyllys 1961; Salton 1988; Hearst & Pedersen 1996; Mani *et al.* 1998; Guo & Stylios 2003)。Hovy & Lin (1998) 提出了 Summarization = Topic Identification + Concept Fusion + Generation 的概念，也就是先經過主題的辨認將文件內文中所描述最重要的主題擷取出來，然後透過詮釋並融合具有相同涵義的主題以取出更精簡的意義，最後再重組 (Reformulate) 這些語句以產生有意義的摘要。

自動文件摘要的方法可分成兩種：Extraction 和 Abstraction (Edmundson & Wyllys 1961; Salton 1988; Guo & Stylios 2003)。Extraction 是從原始文件中抽取出許多小片段的文字，並將之組合成一段短文。Abstraction 則是將原文經過釋意 (Paraphrased) 後，使用精簡的字彙去表示原始的文件 (Guo & Stylios 2003)，具有更高的挑戰性 (Hahn & Mani 2000)。而本研究之摘要方法將以 Extraction 為主，我們的主要架構依據 Edmundson & Wyllys (1961) 所提出之 Extraction 流程做進一步改良 (該架構也曾被 Guo & Stylios (2003) 修正採用)，以配合文件倉儲架構與適合中文文件處理。

此外，根據原始文件數量的多寡，自動文件摘要又可分為單一文件摘要與多重文件摘要 (Hearst & Pedersen 1996)。單一文件摘要是把單篇文件的內容精簡化與重點化，注重的是能否有效地刪減不必要的資訊，並留下真正能代表文件內涵的內容；而多重文件摘要則是把多篇探討類似主題或事件的文件融合在一起，除了刪減無用的資料外，尚需有效率地過濾重複在多篇文章中所出現的資訊。本研究將以多重文件摘要為研究對象。

參、DNCSS 系統架構

本研究之系統架構圖如圖 5 所示。主要分成網路新聞收集器、文件分析子系統、文件倉儲、動態文件主題群聚、與多重文件摘要子系統五大部分：

1. 網路新聞收集器 (News Collector)：以自動化方式，收集特定新聞網站之特定類別新聞文件。

2. 文件分析子系統 (Text Process Subsystem)：透過分析系統將所蒐集到的新聞文件做前置處理 (Pre-Processing)，取出文件中之句子、關鍵詞與 Metadata 等相關資訊。
3. 文件倉儲 (Document Warehouse)：將經過前置處理後之文件與相關資訊以多維度的文件倉儲型式存放，協助使用者整理與歸納龐大的文件資料，以利後續的瀏覽與查詢處理。
4. 動態文件主題群聚 (Dynamic Document Clustering by Topic)：針對使用者於文件倉儲之查詢結果做主題群聚，將描述相關事實之新聞文件歸納於同一群集中，方便使用者依文件群集作瀏覽，並將群聚後之結果提供下面的多文件摘要子系統做進一步處理。。
5. 多重文件摘要子系統 (Multi-Document Summarization Subsystem)：對每一個文件群集提供一份多重文件摘要。

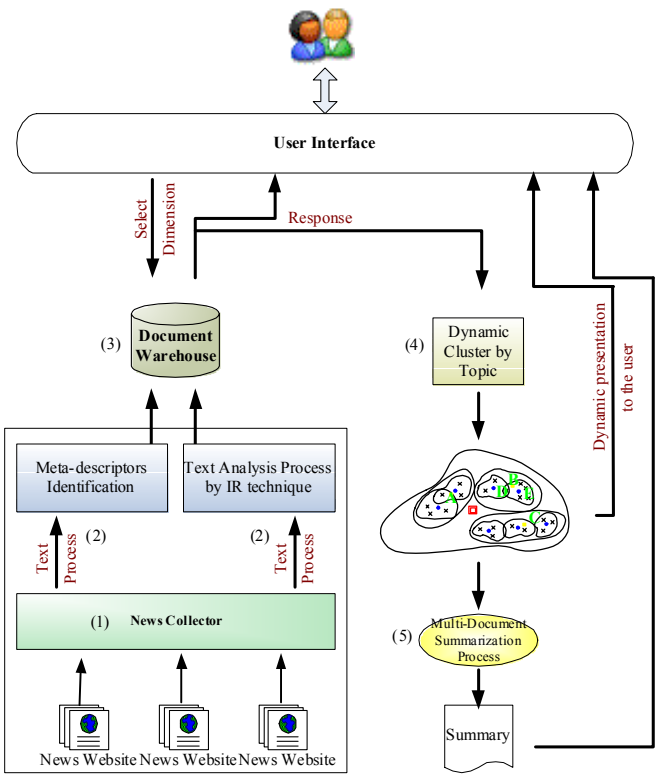

圖 5：DNCSS 系統架構圖

一、網路新聞收集

本研究以自動化的方式批次收集下列六家電子新聞網站之「政治」與「社會」兩類電子新聞文件為資料來源，資訊統計表如表 3 所示。

表 3：新聞文件語料資訊統計表

分類 / 媒體	政治	社會	合計
聯合新聞網	78	65	143
中時電子報	711	450	1,161
中央通訊社	1,289	2,291	3,580
中廣新聞網	1,157	707	1,864
TVBS 新聞	860	1,025	1,885
民視	660	788	1,448
合計	4,775	5,326	10,081

二、文件分析子系統

DNCSS 之多重文件摘要子系統以句子 (Sentence) 為基本單位，故根據教育部國語會所編定之「重訂標點符號手冊」中 14 種標點符號，取置於句末之句號、問號及驚嘆號三種標點符號為斷句之依據，斷句所得的結果我們稱為「句元」(Sentence Unit, SU)。

在斷詞部分，我們採用「中央研究院資訊科學研究所」之《中文詞知識庫資料》為主，以詞庫式斷詞法進行斷詞作業，並輔以「教育部國語推行委員會」所公佈的《新詞語料彙編》，以及本研究所蒐集的詞彙資料，增加詞庫之完整性。另建置贅字表 (Stop Word List) 將位置詞、時間名詞、定詞、分類述詞等納入 (如：上、下、左、右、過去、從前、當初、現在、以上、以下、之於、之後，等)，以增進關鍵詞篩選的準確率。

關鍵詞的篩選方式則以一般最常用的 $tf \times idf$ (Term Frequency × Inverse Document Frequency) 權重計算方式為主 (Popescu 2001)。公式如下：

$$w_{ij} = tf_{ij} \times \log \frac{N}{df_i} \qquad \text{(公式 2)}$$

其中 w_{ij} 為字詞 i 在文件 j 中的權重，tf_{ij} 為字詞 i 在文件 j 的詞頻 (Term Frequency)，df_i 代表整個文件集中有出現字詞 i 的文件數目，N 為整個文件集的數目，取 $\log(N/df_i)$ 後稱為 Inverse Document Frequency (Trigg & Weiser 1987; Popescu 2001)。整個公式的精神在於：字詞在文件中的重要性是與其出現的在文件中的次數成正比，但與其出現的文件數成反比。

另外，由於一般新聞文件的字數差異並不會太大，本研究為增進字詞處理的速度，故省略正規化的動作。

三、動態文件主題群聚

在動態文件群聚方面，本研究採用 Salton & McGill (1983) 的「啟發式群聚法」(Heuristic Clustering Method) 為本研究之群聚方法，主要的考量在於群聚之文件來源是根據使用者查詢的結果而得來的，只要使用者所下達的查詢條件不同，文件的數量與內容皆會有所變動，所以文件的來源會根據使用者的查詢結果動態產生。若採用其他的群聚方法，時間複雜度太高，在效率上的表現不佳。同時，為了計算文件間之相似度，我們採用 cosine coefficient 公式來計算，公式如下：

$$C(X,Y)=\frac{f_{X\cap Y}}{\sqrt{f_X}\times\sqrt{f_Y}} \qquad \text{(公式 3)}$$

其中 f_X 為 X 文件中關鍵詞的數目，f_Y 為 Y 文件中關鍵詞的數目，而 $f_{X\cap Y}$ 為 X 文件與 Y 文件中重複之關鍵詞數目。若計算出來的相似度大於門檻值，則代表這兩篇文件可能是描述相同的事實，系統便會把這兩篇文件歸為一群，然後重新計算該群集之質心。若算出來的相似度小於於門檻值，則自成一群。重複此一步驟直至所有的文件都歸屬於某一群集為止。

四、文件倉儲

在本研究中，我們藉由定義「時間」、「地區」、「新聞媒體」、「新聞分類」等維度，將新聞文件先行透過文件索引組織成文件方塊，以利系統後續能依據這些維度與使用者所給定的關鍵字進行快速的群聚處理。由於四個維度在使用者介面的展現上很難同時展現於螢幕上，故我們採用下拉選單的方式，由使用者自行選擇維度中的成員，然後取出各維度成員所交叉出來的文件方格 (Cell) 顯示在螢幕上。

五、文件摘要子系統

根據群聚後的結果，我們透過文件摘要子系統，將文件群集內所描述事實之摘要資訊擷取出來。如果一個文件群集內只有一份文件時，我們則針對此份文件產生一份單文件摘要。若是群集內有兩份以上的文件，我們便針對整個文件群集內的所有文件產生一篇多重文件摘要。其主要步驟如圖 6 所示，並說明如下：

1. 內容分析 (Content Reconstruction)：此一步驟是針對文件的內容作分析處理，也就是對文件做斷句斷詞的處理取出關鍵詞。
2. 群聚同義句元 (Clustering Sentences)：利用關鍵詞彙將各文件中描述相同事實的句子群聚起來形成句元群聚。而本研究計算句子間相似度是採用 cosine coefficient 方式來計算。
3. 篩選群聚 (Clustering-Filtering)：將句元經過相似度計算形成句元群聚後，我們根據摘要的目的來選擇適當的句元群聚並從中選擇句元輸出。我們相信一個句元群聚內所涵蓋的句元數越多，則代表這個群聚內的句元是常常被提及且所陳述的事實有顯著的重要性。因此，我們會選擇從涵蓋句元數最多的句元群聚開始輸出句元。
4. 句元選擇 (Sentence Selection)：由於句元群聚中有多個代表相同意義之句元，

因此我們必須從句元群聚中選擇一個具代表性的句元輸出成為摘要。多文件摘要的意涵在於去除重複的資訊，本研究選擇句元的方式，採權重值最高的句子優先輸出。也就是將各句中所含的關鍵詞權重相加，再挑出權重相加結果最高的句子。最後，在決定句元輸出的順序與位置方面，我們採用位置法決定句子輸出的順序，參考該句子在原始文件的相對的位置來決定。

5. 長度控制 (Size Control)：本研究之文件摘要系統所產生之摘要長度有 200、300、與 400 字供使用者選擇。若使用者欲產生 300 字之文件摘要，則會在顧及文句的完整性的前提下以 300 字為原則，完整收錄最後一句的內容。所以摘要的長度並不一定會剛剛好是 300 字，而是至少有 300 字的摘要內容。這是為了保持摘要的可讀性，避免句子被中斷而造成語意上的不完整。

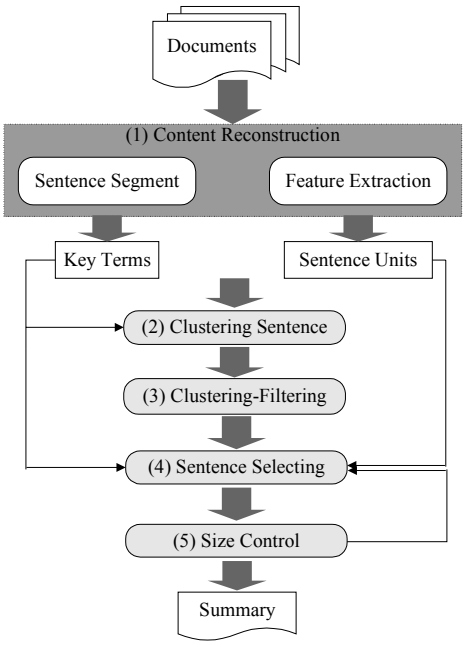

圖 6：本研究多文件摘要流程圖

肆、系統實驗成果

一、使用者介面

本研究所開發之 DNCSS 系統，資料來源分別取自聯合新聞網、中時電子報、中央通訊社、中廣新聞網、TVBS 新聞、與民視，共六家新聞媒體之政治與社會類別之 10081 篇新聞文件作為語料來源。系統共分三個頁面，「資料呈現」、「文件預覽」、與「文件動態群集與多文件摘要」。使用者先在「資料呈現」頁面下達查詢條件後，於下方呈現查詢結果，如圖 7 所示。

這種對文件倉儲的查詢做法，基本上類似於對資料倉儲的查詢。也就是說，對於一個已經建置完成的文件方塊 $DC = (S, (D_1, D_2, \ldots, D_n))$ 來說，使用者可以提供一個基

本方格或非基本方格 $c = (t_c, X_c)$ 來對系統做查詢，其中 $t_c = (c_1, c_2, \ldots, c_i, \ldots, c_n)$，$c_i \in D_i(0)$，$1 \le i \le n$，則文件倉儲系統會精確地回傳由 c 所索引的文件子集 $S_c \subseteq S$，同時也可以快速地依據某些維度進行向上提升 (Roll-Up) 或向下鑽研 (Drill-Down) 的搜尋。

由於在三或四個維度以上，在使用者介面的設計上很難同時展現於螢幕上，故我們採用圖 7 下拉選單的方式，由使用者自行選擇維度中的成員，構成上述的基本方格或非基本方格 c，然後取出各維度成員所交叉出來的文件子集 S_c 顯示在螢幕上。

「文件預覽」的頁面主要是幫助使用者瀏覽查詢結果，使用者選擇欲瀏覽的文件後，新聞文件的內容將呈現於頁面的下方，如圖 8 所示。「文件動態群集與多文件摘要」的頁面，主要分成動態群集和多文件摘要兩部分。動態群集會將使用者的查詢結果，按照使用者所訂定的歸群相似度門檻值，將新聞文件做主題群聚。經群聚的過程得到文件群聚後，使用者可以瀏覽文件群集所含的文件，以及其文件內容，並可依群集產生各個群集的摘要，如圖 9 所示。

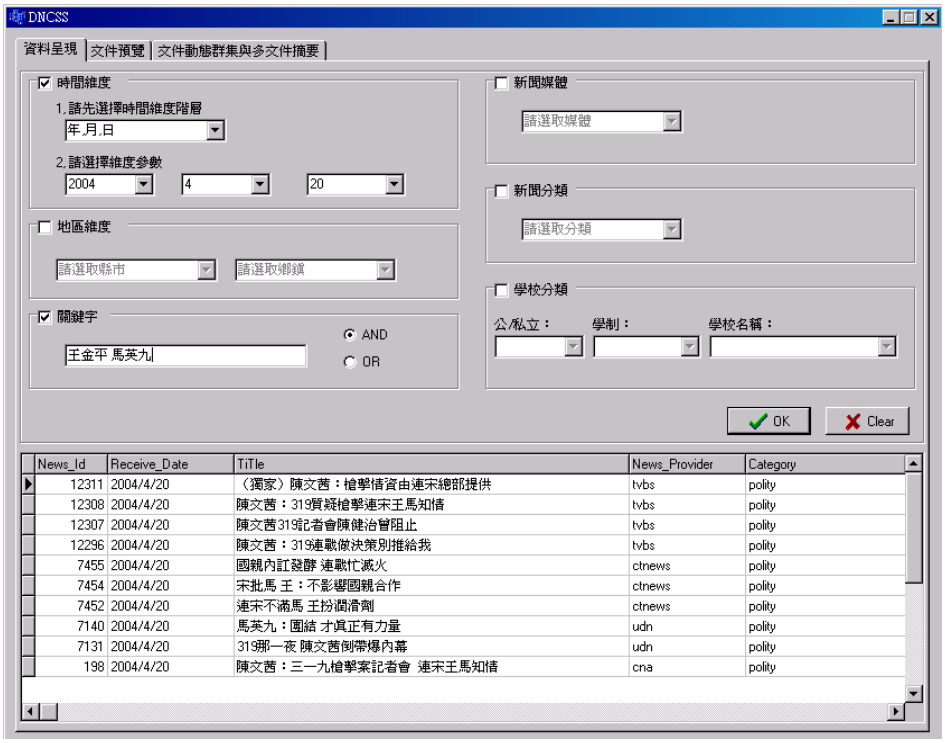

圖 7：「資料呈現」頁面圖

圖 8：「文件預覽」頁面

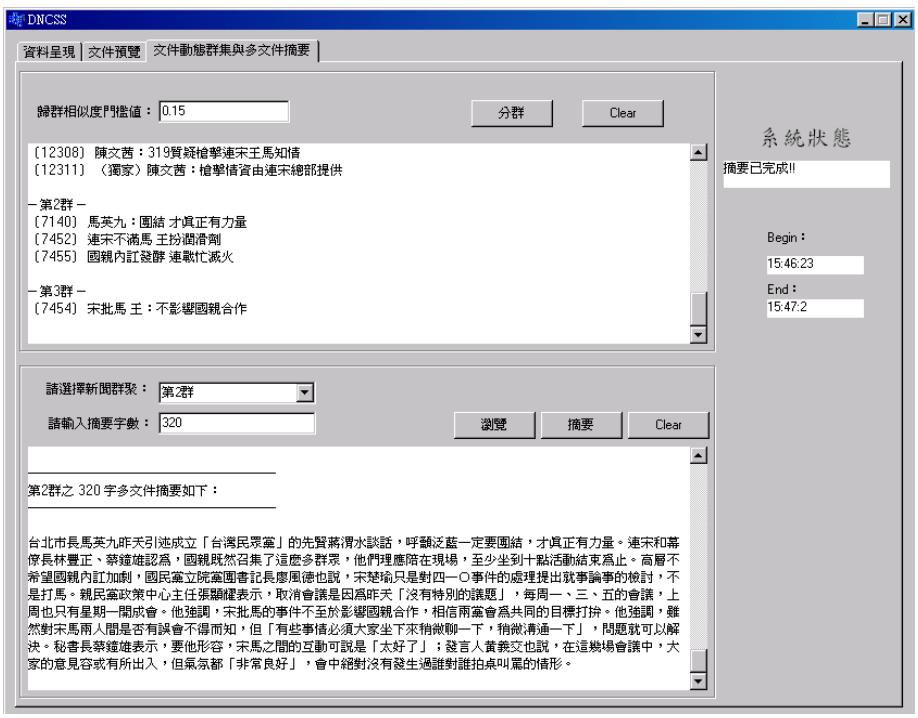

圖 9：「文件動態群集與多文件摘要」頁面

二、實驗結果

以下查詢條件為：時間維度 2004 年 3 月、新聞分類為「政治」類、關鍵字為「陳水扁」與「李昌鈺」並選擇 AND 作為運算元。所得到符合查詢條件之結果共 29 篇新聞文件。經對查詢結果群聚後，文件群集內容如表 4 所列。

表 4：文件群聚結果範例

-- 第 1 群 --	
27458	美法醫魏區：陳總統傷口應是槍傷造成
34539	美鑑識專家：陳總統傷口已接近癒合無法判定受槍傷時間
34563	美 3 專家檢視槍擊案證物
34624	李昌鈺石台平 受邀返台鑑識
36593	美鑑識專家：陳總部腹部的確受到槍傷
36615	美法醫魏區：陳總統傷口應是槍傷造成
36682	三名美國刑事鑑識專家來台，強調不介入政治問題
36818	盧仁發：高檢署成立 319 鑑識小組，協助專案小組辦案
39466	美 3 鑑定專家確定陳總統受到槍傷
39477	李昌鈺：辦案不分黨派與政治無關
-- 第 2 群 --	
27511	邱義仁：調查槍擊真相不需總統頒緊急命令
27583	陳總統四項聲明籲民眾平和理性表達訴求
27589	陳總統：團結國家團結全民是自己責任義務
34523	藍營：3 訴求無回應 410再上街頭
34552	盧仁發：邀國外專家 屬體制內
34614	【回應 327 集會】扁：盼和平落幕 平安回家
36712	扁周一要與連宋見面同意司法全面驗票
36734	陳水扁四點聲明回應 327 集會
-- 第 3 群 --	
27550	陳總統駁斥 319 槍擊案是假的謠言
34605	扁回應調查槍擊案 若連宋可複製相同槍傷 願辭總統
36706	槍擊若假陳水扁願辭職
39562	回應 327 總統駁斥槍擊案自導自演
-- 第 4 群 --	
34562	侯友宜怒答：不管誰，辦到底！
36666	侯友宜：不惜生命偵辦槍擊案
-- 第 5 群 --	
36576	丁守中：扁只會拖延 四一〇再度上街頭
-- 第 6 群 --	
36607	李昌鈺人未到證物未看 民進黨團就質疑不公
-- 第 7 群 --	
39536	商業週刊：台朝野續對立影響經濟

選擇以第三群文件群集做多文件摘要，其新聞內容如表 5，摘要後之 400 字與 500 字之結果如表 6 所示。

表 5：第三群文件群集新聞內文

文件編號 ＝27550 標題 ＝ 【陳總統駁斥 319 槍擊案是假的謠言】 媒體： cna 分類： polity 日期： 2004/3/27
對在野陣營質疑三一九槍擊案是假的，陳水扁總統今晚嚴正駁斥，這是對個人人格非常大侮辱，如可自導自演，請連宋聘請全球最好神槍手，站在靜止吉普車上，若能打成這樣，阿扁馬上辭掉總統職務，你要重新選舉，就重新選舉，若不敢，請你閉嘴，請所有說出可以自導自演者都閉嘴。在野陣營近來不斷散播陳水扁總統、副總統呂秀蓮三一九槍擊事件是假的謠言，對此，陳總統晚間在總統府召開記者會，親自予以嚴正駁斥。他說，有關槍擊事件，真的沒辦法接受竟然有人懷疑兩人的挨槍是假的，甚至有人說是自導自演，難道一定要自己和呂秀蓮死掉、肚破腸流，才是真的嗎？─還要叫阿扁和呂秀蓮對天發誓我們沒有說謊，做為國家的領導者，真的不宜對天來發誓，但絕不是不敢發誓。─他強調，有一點可以這樣講，如果這件事可以自導自演，─請連主席、宋主席雇到全世界最好的神槍手，錢我來負責就好，請他們兩位站在吉普車，不必行進中的吉普車，靜止中的吉普車就好，請那位神槍手同樣打兩槍，如果剛好像阿扁和呂秀蓮的傷勢，那你們所謂的自導自演，才可能取信於人。─陳總統表示，如果靜止狀況可以打成這樣的傷勢，─阿扁馬上接受，我辭掉總統的職務，五二Ｏ阿扁不會就職，你要重新選舉，就讓你重新選舉，如果你不敢做，請你閉嘴，請所有還說出可以自導自演的人全都閉嘴。─陳總總指出，這是對個人人格非常大的侮辱，包括那麼多的醫護人員、專業人士所參與的醫療作業，都可以被這樣污衊、全盤否定，那麼這個社會還有什麼信任、和諧可言？此外，陳總統重申，在二十三日和五院院長茶敘時就表示，歡迎、感謝大家來推薦，包括反對陣營也可以推薦適當、最好的鑑識人員、刑事專家、彈道專家來參與協助這次的辦案。他說，自己非常希望能夠知道答案，為什麼有人要開槍？為什麼要暗殺正、副總統？到底是什麼動機？希望趕快知道答案，─難道我不心急？難道我的家人不心急？─他說，為了選舉目的，可以理解，但否定這麼多專業的東西，污衊相關的當事人，包括南台灣最好的奇美醫院，院長詹啟賢院長從頭到尾都在旁邊，看得非常清楚，能夠作假嗎？他曾是國民黨執政時代的衛生署長，他的人格、道德、為人所敬重，─難道他會袒護阿扁嗎？─因此，陳總統重申，非常感謝最高檢察署檢察總長盧仁發表達要成立獨立專案鑑識小組，希望在野黨推薦國際知名刑事鑑識專家李昌鈺和石台平醫師能趕快來台灣，協助相關鑑識工作，更歡迎大家能幫忙推薦全世界最有公信力、專業的機構或個人來參與調查。陳總統強調，政府對這個案子一切公開、透明，沒什麼好掩飾，希望讓這個案子早日查個水落石出，─難道我希望案子沈寂，揹負自導自演的惡名，有什麼好處？─，他也說─連主席、宋主席，如果挨槍的是你們，你們會講這種話嗎？─930327
文件編號 ＝34605 標題 ＝ 【扁回應調查槍擊案 若連宋可複製相同槍傷 願辭總統】 媒體： ctnews 分類： polity 日期： 2004/3/28
陳水扁總統昨晚回應連宋要求「調查槍擊」表示，如果連宋可以在靜止狀態下被神槍手打成他的傷口，他願意辭去總統職務，五二○不會就職，如果連宋做不到就請閉嘴。他並強調，奇美醫院院長詹啟賢是國民黨時代衛生署長，不可能幫他說謊，他歡迎在野推薦的李昌鈺、石台平參加鑑識，一切坦然盼案情早日水落石出。有關槍擊事件，陳總統激動地表示，竟然有人懷疑，他和呂秀蓮的挨槍是假的，甚至說是自導自演，難道要他們死才是真的？要我們肚破腸流才是真的嗎？有人要我們對天發誓沒有說謊，但做為國家領導者不宜對天發誓，而不是我們不敢發誓。陳水扁反問，難道他會希望這個案子就這樣沈寂嗎？他揹這樣「自導自演」的惡名，有什麼好處？他反問連宋，如果今天挨槍的是他們，還會講這種話嗎？陳水扁強調，如果這種事可以自導自演，他建議連宋兩位主席，是不是能雇全世界最好的神槍手，錢他來負責，請連宋站在吉普車上，不必行進中，靜止狀態就好，也挨兩槍，打出相同的傷勢，他馬上接受，辭掉總統的職務，五二○不會就職，要重新選舉重新選舉！但若做不到、不敢做，請連宋閉嘴，請所有說出槍擊事件可以自導自演的人，「通通閉嘴。」陳水扁表示，槍擊造假的說法，是對他人格非常大的羞辱，那麼多的醫務人員所參與的醫療，可以這樣誣衊，全盤否定，這個社會還有什麼信任與和諧可言？陳水扁重申，他歡迎感謝大家推薦適當的彈道、刑事、傷口專家來協助辦案，他也很希望知道答案，為什麼有人要開槍？為什麼要暗殺正副總統？到底是什麼動機，他也希望知道。陳水扁也感謝最高檢察總長盧仁發已成立一個獨立的專案小組，由兩位在野陣營所推薦的李昌鈺博士、石台平法醫，希望兩位專家能趕快回台協助相關鑑識工作，有公信力的機構或相關個人都能參與調查和鑑識，一切透明，沒什麼好掩飾的，一切坦然，盼案情早日水落石出。
文件編號 ＝36706 標題 ＝ 【槍擊若假陳水扁願辭職】 媒體： bcc 分類： polity 日期： 2004/3/28
319 槍擊案至今仍然真相未明，對於外界質疑這起槍擊案是陳、呂自導自演，甚至有人質疑槍擊案是假的。陳水扁總統表示，他無法接受有人質疑槍擊案作假，而陳水扁也說，如果連宋也願意站上吉普車讓神槍手開兩槍，同時造成和扁、呂一樣的傷勢，證明 319 槍擊案可能自導自演的話，他願意辭去總統職務。陳水扁星期六晚間召開記者會，回應國親陣營所提出的質疑。陳水扁表示，他無法接受有人質疑 319 槍擊案是假的，或是他自導自演。陳水扁說，如果連宋能夠登上吉普車作試驗，他願意付錢請來神槍手，朝連、宋身上開兩槍，如果能證明槍擊案可以自導自演，他願意馬上辭職；陳水扁也表示，他是在 319 槍擊案中挨槍的人，不只希望儘快知道兇手開槍的動機，也不想背負著自導自演槍擊案的惡名，因此，除了希望李昌鈺、石台平等專業的鑑識人員儘快來台協助調查之外，也希望 319 槍擊案的調查能夠公開透明，讓案情早日水落石出。
文件編號 ＝39562 標題 ＝ 【回應 327 總統駁斥槍擊案自導自演 】 媒體： tvbs 分類： polity 日期： 2004/3/28
327 泛藍大會師，要求查明槍擊案真相，陳總統昨晚召開記者會，強力駁斥在野陣營對他自導自演的指控！總統說，他願意接受國親推薦的彈道與鑑識專家進行專業調查，因為他比任何人都想知道槍擊案的真相！總統還說，在野黨罵他自導自演，他建議在野黨自己去模擬槍擊，如果不敢做，就請閉嘴！總統陳水扁：「難道阿扁跟呂副總統一定要死掉才是真的嗎？一定要肚破腸流，這樣才真的嗎？」神情肅穆、語氣嚴厲，面對在野黨不斷質疑，槍擊案是總統自導自演，陳總統強力駁斥。總統陳水扁：「不過有一點我可以這樣講，如果這種事情可以自導自演，請他們 2 位同樣地站在吉普車上，不必行進中，靜止狀態就好，請那位神槍手同樣的打 2 槍，剛好也是像阿扁與呂副總統這樣的一個傷勢，那你們所謂的自導自演，才可能取信於人，靜止狀態可以打成這樣，阿扁馬上接受，我辭掉總統的職務，520 阿扁不會就職，你要重新選舉就讓你去重新選舉，如果你做不到，如果你不敢做請你閉嘴，請所有還說出所謂可以自導自演的人，請大家閉嘴，我覺得這是對個人人格非常大的一個羞辱。」陳總統說，他比任何人都急於想知道槍擊案的真相。總統陳水扁：「總統副總統是挨槍的人，我們難道不心急嗎？我的家人難道不急嗎？阿扁非常感謝昨天最高檢察署盧總長擔任召集人，正式成立一個獨立的專案鑑識小組，特別邀請由在野陣營所推薦的 2 位鑑識專家，李昌鈺博士跟石台平醫師，我們希望 2 位由在野陣營所推薦的鑑識專家能夠趕快來台灣，來協助相關的鑑識工作，我們非常歡迎、也非常感謝，我們希望李博士、石醫師也能夠來幫忙推薦全世界最有名的最有專業的、有公信力的機構或者相關的個人，來參與這個案子的調查跟鑑識。」選在 327 泛藍大會師的這一天，陳總統以嚴厲口吻，回應在野陣營對槍擊案的各項質疑，希望循開放透明的專業途徑，還原槍擊案真相。

■ 為 400 字摘要所節錄之句子。

■ 為 500 字摘要多於 400 字摘要所節錄之句子

表 6：第三群文件群集之多文件摘要結果

第 3 群之 400 字多文件摘要如下：
對在野陣營質疑三一九槍擊案是假的，陳水扁總統今晚嚴正駁斥，這是對個人人格非常大侮辱，如可自導自演，請連宋聘請全球最好神槍手，站在靜止吉普車上，若能打成這樣，阿扁馬上辭掉總統職務，你要重新選舉，就重新選舉，若不敢，請你閉嘴，請所有說出可以自導自演者都閉嘴。─因此，陳總統重申，非常感謝最高檢察署檢察總長盧仁發表達要成立獨立專案鑑識小組，希望在野黨推薦國際知名刑事鑑識專家李昌鈺和石台平醫師能趕快來台灣，協助相關鑑識工作，更歡迎大家能幫忙推薦全世界最有公信力、專業的機構或個人來參與調查。陳水扁說，如果連宋能夠登上吉普車作試驗，他願意付錢請來神槍手，朝連、宋身上開兩槍，如果能證明槍擊案可以自導自演，他願意馬上辭職；陳水扁也表示，他是在 319 槍擊案中挨槍的人，不只希望儘快知道兇手開槍的動機，也不想背負著自導自演槍擊案的惡名，因此，除了希望李昌鈺、石台平等專業的鑑識人員儘快來台協助調查之外，也希望 319 槍擊案的調查能夠公開透明，讓案情早日水落石出。
第 3 群之 500 字多文件摘要如下：
對在野陣營質疑三一九槍擊案是假的，陳水扁總統今晚嚴正駁斥，這是對個人人格非常大侮辱，如可自導自演，請連宋聘請全球最好神槍手，站在靜止吉普車上，若能打成這樣，阿扁馬上辭掉總統職務，你要重新選舉，就重新選舉，若不敢，請你閉嘴，請所有說出可以自導自演者都閉嘴。希望趕快知道答案，一難道我不心急？─因此，陳總統重申，非常感謝最高檢察署檢察總長盧仁發表達要成立獨立專案鑑識小組，希望在野黨推薦國際知名刑事鑑識專家李昌鈺和石台平醫師能趕快來台灣，協助相關鑑識工作，更歡迎大家能幫忙推薦全世界最有公信力、專業的機構或個人來參與調查。陳水扁星期六晚間召開記者會，回應國親陣營所提出的質疑。陳水扁說，如果連宋能夠登上吉普車作試驗，他願意付錢請來神槍手，朝連、宋身上開兩槍，如果能證明槍擊案可以自導自演，他願意馬上辭職；陳水扁也表示，他是在 319 槍擊案中挨槍的人，不只希望儘快知道兇手開槍的動機，也不想背負著自導自演槍擊案的惡名，因此，除了希望李昌鈺、石台平等專業的鑑識人員儘快來台協助調查之外，也希望 319 槍擊案的調查能夠公開透明，讓案情早日水落石出。神情肅穆、語氣嚴厲，面對在野黨不斷質疑，槍擊案是總統自導自演，陳總統強力駁斥。

雖然群聚後的結果看起來彼此都非常相關，但其中還是有許多不相似的主題可以加以區分。因為這些新聞雖然都繞著 319 槍擊案打轉，但其中有些報導的重點是有關旅美專家李昌鈺博士查案的相關報導 (第 1 群)、有的以立法委員在立院質疑調查真相為主的新聞 (第 2 群)、有的是在講連宋的看法 (第 3 群)、有的則是闡述警方辦案的決心 (如第 4 群)。因此，可以看出本系統的確可以有效協助與提昇後續的多重文件摘要作業。

三、系統評估與結論

因 DNCSS 系統是隨著使用者之查詢條件不同，動態針對不同的查詢結果產生文件群集與文件摘要，因此本研究先隨機挑選 10 組查詢條件，透過 DNCSS 系統在相同系統參數下產生共 67 個文件群集。在此 67 個文件群集中，我們挑選包含三篇新聞文件以上之文件群集為評估對象，共計挑選 31 個文件群集。再針對此 31 個文件群集，以本系統之多文件摘要模組對每一個文件群集產生一份多文件摘要。

為提高本研究評估作業之可行性，評估作業以四位研究所以上學歷，具資訊科技背景與資訊檢索專長，並有閱讀電子新聞文件三年以上經驗之受測者進行之。當受測者仔細觀看完每一個文件群集後，必須填答一份評估量表，以五個尺度去衡量受測者對本系統在文件群聚與多文件摘要之效果。我們的評估項目探討以下三個問題，以評估本研究之系統效果，驗證本研究以文件倉儲進行文件群聚與多文件摘要之適用性。

問題一：新聞文件群集所包含之新聞報導內容相關程度?

問題二：多文件摘要是否能有效幫助了解該新聞文件群集所描述之主題?

問題三：多文件摘要之語句是否具連貫性?

表 7：受測者填答統計圖表

新聞文件群集所包含之新聞報導內容相關程度？		
相關程度	次數	百分比
1	1	0.806452
2	8	6.451613
3	27	21.77419
4	60	48.3871
5	28	22.58065
總和	124	100
多文件摘要能否幫助了解新聞群集所描述之主題？		
幫助程度	次數	百分比
1	0	0
2	9	7.258065
3	35	28.22581
4	65	52.41935
5	15	12.09677
總和	124	100
多文件摘要之語句是否具連貫性？		
連貫程度	次數	百分比
1	2	1.612903
2	8	6.451613
3	47	37.90323
4	50	40.32258
5	17	13.70968
總和	124	100

在經過四位受測者對本系統進行評估後，均有相當不錯的評估結果，評估結果如表 7。受測者在「新聞群集內所包含之新聞報導相關程度」評估項目上，認為「相關」與「非常相關」佔七成以上。而在「多文件摘要是否能有效幫助了解該新聞群集之主題？」上，受測者認為「有幫助」與「非常有幫助」亦有六成以上之認同。「多文件摘要之語句是否具連貫性？」的問題，受測者對連貫性程度「連貫」與「非常連貫」的認同在五成左右。然而，受測者對摘要連貫性大多認為在「普通」與「連貫」的範圍內，此一數據反映出受測者對於摘要連貫性的認知差異甚大，因此在摘要的連貫性上可待後續研究者作更進一步的努力。

伍、結論與未來研究方向

本論文提出透過文件倉儲並結合文件分類、多重文件摘要等技術，幫助使用者透過自動化處理的機制將文件資訊做系統化呈現。本研究之 DNCSS 系統在經人工評估之後，也有不錯的滿意度。我們的貢獻在於能以自動化的方式剖析文件內容，將這些文件相關資訊存放進文件倉儲中，並透過文件倉儲的運作提供使用者以維度作查詢。然後以文件群聚的技術幫助使用者自動將相關的新聞議題群聚在同一個群集內，並輔以多文件摘要技術對每一個群集提供一份文件摘要資訊。總括來說，具有以下之優點：

1. 讓使用者在除了依關鍵字查詢之外，可以按照其需求從時間、地點、媒體、分類等不同的查詢組合去找到所需的新聞資料。
2. 對查詢後結果作群聚，幫助使用者將新聞文件做歸納。

3. 每一個文件群集均提供一份多文件摘要，使用者不需一篇篇的去看內容便能瞭解此一文件集合的精要資訊。

雖然，本研究所採用的方法僅是將傳統的文件處理方法架在文件倉儲結構上來實驗，並未提出新的方法。但是，我們認為能夠得到不錯的實驗結果應該具有雙重的意義：

1. 傳統的文件處理方法有許多是以西方語文的處理角度所開發出來的，而本論文則以中文文件來加以驗證之。後者在斷字與斷詞的處理上，由於中文文字間沒有空白區隔而讓難度提高許多，而我們的實驗結果顯示這些方法用在中文文件上一樣有不錯的效果。
2. 利用文件倉儲的概念先將文件以多維度的結構來組織，對整體工作也扮演著相當大的助力。由於使用文件倉儲可以很有效率並精確地以多維度角度索引到所需的文件子集，如圖 2 所示。如果不使用文件倉儲的話，則可能在相同效率的情況下，所找到的文件子集是相關性較低的 (Recall 與 Precision 較低)。因此我們可以說：若不使用文件倉儲，則在找尋所需文件時可能會產生效率不彰，或只能以某單一角度來為之，效果較差應該是顯而易見的。我們不敢說能得到不錯的多重文件摘要結果完全是文件倉儲結構的功勞，但某種程度上對整體測試效果有所幫助應該是可以受到肯定的。

關於未來的研究方向，我們認為可以朝向以下幾點進行：

1. 文件倉儲的部分，未來可結合本體論 (Ontology) 的概念以擴展文件的維度與概念階層，相信能提高文件倉儲在資訊呈現上的效果，讓使用者能從更上一層的語意概念做搜尋。
2. 關聯法則 (Association Rule) 是資料探勘領域中頗受重視的一種方法，目的是找出項目間彼此的關聯性。加入關聯法則的應用，可做為文件間相關程度比對與文件倉儲查詢句擴展之參考。
3. 目前限於人力僅針對新聞文件進行測試，未來將擴充至各類型的文件資料。
4. 中文字詞處理的部分，我們將在斷詞技術上進一步加強，以解決中文斷詞歧義問題，並加強關鍵詞擷取能力。另外，如何辨識新詞與未知詞也是未來努力的重點。
5. 事件偵測與追蹤 (Event Detection and Traction) 是目前國外非常熱門之研究議題，若能在本研究中加入事件偵測與追蹤的技術，在回傳查詢結果給使用者時附加新聞事件發展的資訊，相信能提供給使用者更多的選擇。
6. 由於文件中各種資訊的處理均需要大量的比對和計算，受限於現今硬體設備的運算速度，因此不得不在運算速度與系統效能間作取捨。若能在硬體設備上有所改進，加快運算速度，則採時間複雜度較高但運算結果較佳之演算法，所得之結果必定能更有突破。

陸、致謝

本研究承蒙國科會計劃的部分經費贊助，計劃編號為 NSC 94-2416-H-327-009。

柒、參考文獻

1. Bleyberg, M.Z. and Ganesh, K., "Dynamic multi-dimensional models for text warehouses," *IEEE International Conference on Systems, Man, and Cybernetics*, Vol. 3, Oct. 2000, pp. 2045- 2050.
2. Bleyberg, M.Z. and Paranjape, P.S., "A content delivery strategy for text warehouses," *IEEE International Conference on Systems, Man, and Cybernetics*, Vol. 4, Oct. 2001, pp. 2322–2325.
3. Carey, M., Kriwaczek, F., and Ruger, S., "A visualization interface for document searching and browsing," *Proc. of Workshop on the New Paradigms in Information Visualization and Manipulation* (*NPIVM'2000*), *Washington, D.C.*, 2000, pp. 24-28.
4. Chen, K.J. and Kiu, S.H., "Word identification for mandarin chinese sentences," The *Fifth International Conference on Computational Linguistics*, 1992, pp. 101-107.
5. Edmundson, H.P and Wyllys, R.E., "Automatic Abstracting and indexing—survey and recommendations," *Communications of the ACM*, Vol. 4, No. 5, May 1961, pp. 226-234.
6. Guo, Y. and G. Stylios, "A new multi-document summarization system," *Proceedings of 2003 Workshop on Text Summarization* **(**with the 2003 *Human Language Technology Conference)* May 31-June 1, Edmonton, Canada, 2003, pp. 102-109.
7. Hahn, U. and Mani, I. "The challenges of automatic summarization," *IEEE Computer* Vol. 33, No. 11, November 2000, pp. 29-36.
8. Hearst, M.A. and Pedersen, J.O., "Reexamining the cluster hypothesis: Scatter/gather on retrieval results," *Proceedings of the 19th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval (Zurich, Switzerland)*, New York, 1996, pp.76–84.
9. Hovy, E.H. and Lin, C.-Y., "Automated text summarization and the SUMMARIST system," TIPSTER Text Program Phase III final report, October 1998.
10. Jain, A.K. and Dubes, R., *Algorithms for Clustering Data*, Prentice-Hall, Englewood Cliffs, NJ, 1988.
11. Jain, A.K., Murty, M.N., and Flynn, P.J., "Data Clustering: A Survey," *ACM Computing Surveys*, Vol. 31, No. 3, 1999, pp. 264-323.
12. Kaufman, L., and Rousseeuw, P.J., *Finding Groups in Data: an Introduction to Cluster Analysis*, John Wiley & Sons, 1990.
13. Lee, J., Grossman, D., Frieder, O., McCabe, M.C., "Integrating Structured Data and Text: A Multi-dimensional Approach", *Proc. International Conference on Information Technology: Coding and Computing*, (*ITCC 2000*), March 2000, pp. 264-271.
14. Leuske, A., "Evaluating document clustering for interactive information retrieval," *Proceedings of 10th International Conference on Information and Knowledge Management (CIKM'01)*. 2001, pp.33–40.
15. MacQueen, J.B, "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, Vol. 1, 1967, pp.281-297.

16. Mani, I., House, D., Klein, G., Hirschman, L., Obrst, L., Firmin, T., Chrzanowski, M., and Sundheim, B., "The TIPSPER SUMMAC Text Summarization Evaluation", *Automatic Text Summarization Conference*, 1998, pp. 77-85.
17. McCabe, M.C., Lee, J., Chowdhury, A., Grossman, D., Frieder, O., "On the Design and Evaluation of a Multi-dimensional Approach to Information Retrieval", *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 2000, pp. 363-365.
18. Meyrowitz, N. and van Dam, A., "Interactive Editing Systems," *ACM Computing Surveys*, Vol. 14, No. 3, Sep. 1982, pp. 321-415.
19. Nie, J. Y., Brisebois, M. and Ren, X., "On Chinese text retrieval," *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, August 1996, pp. 225-233.
20. Popescu, A. R., "Implementation of term weighting in a simple IR system," Kursprojekt, June 2001. http://www.cs.helsinki.fi/u/popescu/docs/ir.pdf
21. Salton, G., and McGill, M. J., *Introduction to Modern Information Retrieval,* McGraw-Hill, 1983, pp. 338-341.
22. Salton, G., *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley Publishing Company, 1988.
23. Salton, G., Wong, A., and Yang, C.S., "A vector space model for automatic indexing," *Communications of the ACM*, Vol. 18, No. 11, 1975, pp. 613-620.
24. Sebastiani, F. "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, Vol. 34, No. 1, March 2002, pp. 1-47.
25. Sproat, R. and Shih, C., "A Statistical Method for Finding Word Boundaries in Chinese Text," *Computer Processing of Chinese and Oriental Languages*, Vol. 4, No. 4, March 1990, pp. 336-351.
26. Sullivan, D., *Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing and Sales,* John Wiley & Son, Inc., 2001.
27. Trigg, R.H. and Weiser, M., "Text Net: A Network Based Approach to Text Handling," *ACM Transactions on Office Information Systems*, Vol. 4, No. 1, Jan. 1987, pp. 1-23.
28. Tseng, F.S.C. and Chou, A.Y.H., "The Concept of Document Warehousing for Content Management of Enterprise Business Intelligence," *Decision Support Systems*, Accepted and forthcoming, 2006.
29. Tseng, F.S.C. and Lin, W.-P., "D-Tree: A Multi-Dimensional Indexing Structure for Constructing Document Warehouses," *Journal of Information Science and Engineering*, Accepted and forthcoming, 2006.
30. Tseng, F.S.C., "Design of a Multi-Dimensional Query Expression for Document Warehouses," *Information Sciences*, Vol. 174, No. 1-2, July 2005, pp. 55-79.
31. Van Rijsbergen, C. J., *Information Retrieval*, 2nd Ed., Buttersworth, London, 1979.
32. Wu, M., Fuller, M., and Wilkinson, R., "Using clustering and classification approaches in interactive retrieval," *Information Processing and Management*, Vol. 37, Issue 3, 2001, pp.459–484.

33. Yeh, C.L. and Lee, H.J., "Rule-Based Word Identification for Mandarin Chinese Sentences-A Unification Approach," *Computer Processing of Chinese and Oriental Languages*, Vol. 5, No. 2, March 1991, pp. 97-118.
34. Zamir, O. and Etzioni, O., "Web document clustering: a feasibility demonstration," *Proceedings of the 21st Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia), 1998, pp.46–54.
35. 陳永德，中文斷詞中長詞優先、詞頻對比及前詞優先規則之使用，國立臺灣大學心理學研究所碩士論文，1997 年。
36. 曾元顯，"關鍵詞自動擷取技術之探討"，中國圖書館學會會訊，第 5 卷，第 3 期 (第 106 期)，1997 年 9 月，頁 26-29。