考量語意及引用分析之研究主題趨勢分析方法

王京盛
國立成功大學資訊管理研究所
chuchu1128@gmail.com

王惠嘉
國立成功大學資訊管理研究所副教授
hcwang@mail.ncku.edu.tw

摘要

隨著資料的數位化，文件逐漸轉變為便於流通的電子化形式。然而，由於電子化刊物資料量的快速增加，使得研究人員雖能輕易的收集資料，但卻無法從中擷取重要資訊。為了能夠更有效率地進行研究，便利用主題偵測與追蹤技術整理出研究資料集中的代表主題及趨勢的追蹤，但以往的主題偵測技術，僅考慮單一資料類別(如:期刊)，且並未針對研究趨勢進行分析。故本研究利用多樣資料類別(研討會與期刊)，並利用研討會與期刊的先後影響關係來發現較多訊息，並在特徵選取時考慮語意及引用次數，來增加特徵選取的效率，提供研究人員熱門的研究主題及趨勢的分析，使研究人員在找尋研究方向時，能有更快速的參考依據。

關鍵詞：主題偵測與追蹤、趨勢分析、特徵選取、分群

壹、緒論

在現今資料通訊網路的蓬勃發展下，資料的流通及保存的方式從傳統的紙本資料轉變為電子化資料。電子化資料擁有快速傳輸及存放便利的優點，但也因為此特性，在網路流通的資訊量正快速地增加。而目前研究人員獲取資訊的來源，主要是從各種電子資料庫進行檢索，如何從為數眾多的電子化文件中，應用文件探勘的相關技術找尋特定領域的資料便顯得更為重要。同時，由於各研究領域的廣度及深度都在不斷地上升，若研究人員都必須先投入相當長的時間閱讀文件，只為了找到適合且符合時勢的研究主題，可能會是目前各研究進展的瓶頸所在。

以往在文件的檢索上，大多是使用網站所提供的關鍵字檢索系統(Lin, Shih, Chen,Ho, Ko, & Huang, 1998)。然而，即便經過了關鍵字的過濾，可能仍包含過多不適合的文件。Tu & Seng(2009)認為找研究主題的新趨勢是目前相當重要的議題，對於研究人員更是相當重要的。為了能更有效地便利研究人員從大量電子資料中取得所需資訊，文件主題偵測及趨勢的追蹤便成為可能解決此問題的一種方法(吳偉銘，民 97;林宜瑩，民 99)。主題偵測主要能夠以主題來表現不同的群集，而這些群集是將為數眾多的文件，依照其內容的相似程度來進行組織。

過去學術研究的主題偵測及趨勢追蹤，主要是透過分析期刊論文的資訊，找出各時期的主題，進而進行趨勢的分析。例如林宜瑩（民 99）提出的主題偵測與追蹤方法，便是以著名期刊的論文文件資料集當作資料來源。雖然其結果具有參考價值，然而，由於一篇期刊論文從投搞到刊出之間，其審查編修時間通常會超過一年 (Tu & Seng, 2009)，所以單純以期刊論文作為資料集所找出的趨勢，可能已經不是目前該領域最新的主題，相較之下，每年各大研討會所徵求的主題常能夠代表各領域目前最熱門的研究主題，若能有效地將此類資訊進行分析，便得以獲得比以往更具時效性的熱門主題趨勢，也可利用研討會以及期刊研究趨勢時間上可能的關係，來進行趨勢走向的分析。此外，同義詞的使用、拼字變化、縮寫以及或多或少與作者意思表達相關的用字問題，都會阻礙這方面的分析(Li, Xia, Zong, & Huang, 2009)，林宜瑩（民 99）也忽略了同義字及近義字，使其在主題偵測與追蹤時，無法將同意義的字詞進行關聯，若能將同義字及近義字納入考量，便可以減少因上述原因所引發的問題。除此之外，研究論文被引用的次數可以拿來追蹤研究論文本身的影響力(Chiu & Ho, 2007; Li & Ho, 2008)，故可利用被引用的次數來衡量研究論文的重要性。而針對研究所產生的結果，並未對研究熱門程度趨勢的轉折做討論及分析，在結果的呈現上，僅提供研究人員作為參考。

綜合以上所述，本研究目的便是將各大研討會的資料集與電子化期刊論文資料庫進行彙整，接著依研究領域分門別類，針對各個領域的資料集，利用主題偵測的技術將文件進行分群，並歸納出近年來各研究領域的過去與現在的研究趨勢及研究熱度，進而去找尋適合投入的研究主題。由於本研究加入考量了語意、引用次數的因子，期望能夠改善主題偵測與追蹤之於趨勢分析的可靠度，並利用研討會及期刊不同的資料來源產生的結果，做進一步趨勢走向的分析。

貳、文獻探討

本章節將針對本研究所使用的技術進行文獻的回顧及探討，用以了解目前此領域的研究發展。以下主要針對主題偵測與追蹤、資料檢索、特徵選取、文件分群、學術論文特性等，進行詳細的介紹。

一、主題偵測與追蹤(Topic Detection and Tracking, TDT)

主題偵測與追蹤起源於 1996 年，DARPA((Defense Advanced Research Projects Agency)在找尋能夠不用人力參與便能從新聞串流偵測出主題的技術。而在 1998 年，Allan 等人建立了第一個主題偵測與追蹤的系統(Allan,Carbonell, Doddington, Yamron, Yang, Umass, &Umass, 1998)，主題偵測與追蹤的相關研究隨後在 1998 及 1999 年皆有不少的進展。

為了改善主題偵測與追蹤的效率，便有了一些更進一步的研究，例如 Walls, Jin, Sista, & Schwartz(1999)建立了一個依照主題將新聞及網頁群組化的非監督式的主題偵測與追蹤系統，該研究使用了 k-means 演算法來進行分群，也使用了向量空間模型及文章相似度矩陣(Salton, Wong, & Yang, 1975)的方法來比較事件與事件的關係。

雖然過去已經有不少的研究在進行主題偵測與追蹤方法的發展及改善，但通常都應用在像新聞文件或是電子郵件等等非常具時間敏感度的資料集，而並沒有被廣泛的應用在時間區隔較長的學術研究文件上。吳偉銘（民 97）其所使用的資料集不再是新聞文件，利用主題偵測與追蹤方法來進行電子期刊論文的主題偵測，並考慮了時間因子來進行時間演進時，電子刊物的內容變動或消長趨勢。但其研究指出，若能夠考慮名詞片語的分析，將能使結果更具參考性。

林宜瑩（民 99）則利用了主題偵測與追蹤的方法，來進行文獻主題的追蹤。由各期刊中所發表的期刊論文來當作資料集，其目的是利用主題偵測的技術，將期刊論文進行主題性的分群，並由系統歸納出近年來期刊的整體研究趨勢，以節省過去研究人員為了找到有價值的資訊所耗費的人力及時間。其考慮了名詞片語的分析，但卻忽略了語意的考量，對於同義字或近義字等並無進行任何處理。

二、資料檢索

資料檢索(Information Retrieval, IR)是指從大量的資料中取出和使用者需求相關的文集(Cordon, 2003)。常見的資料檢索模型有布林模型(Boolean Model)、機率模型(Probabilistic Model)以及向量空間模型(Vector Space Model, VSM)(Cordon, 2003)。其中，布林模型僅考慮字詞是否出現，而不將字詞本身對文件的重要性納入考量；機率模型則忽略了字詞共同出現的涵義。因此，目前以向量空間模型較被廣為使用(Özgür & Güngör, 2010)。在向量空間模型中，計算相似度的方法較為常見的有 Cosine Measure、Dice's Measure、Jaccord's Measure 以及 Overlap Measure。由於 Cosine Measure 在向量空間模型中，計算相似度的表現及成效是最好的(Wan, 2007)，Cosine Measure 的運算複雜度與成本與其他方法相比皆為中等，故本研究在相似度計算上將會採用 Cosine Measure 作為主要方法。

三、特徵選取(Feature Selection)

一份文件的特徵，顧名思義便是文件中較具代表性的一個字詞、片語或是句子。若一個字詞在文件中所出現的頻率高且與其主題具有高度關聯性時，則將其稱之為該文件的特徵(Frakes & Baeza-Yates, 1992)。若將所有可能的特徵聚集成一個集合，則該集合稱之為特徵空間(Feature Space)或特徵集合(Feature Set)，其數量可能從數十個到數以千計不等。特徵空間的高維度是在做文件分群或分類困難的主要原因之一(Joachims, 1998)。所以特徵選取的主要目標便是從大量的特徵中挑選出部份較適合的特徵，使辨識率及分類正確率能達到最佳化，並且透過特徵選取的步驟，可進一步降低特徵空間的維度(Li et al., 2009)。本研究在特徵選取方法的選擇上，考量到計算的複雜程度，且 DF(Document Frequency)若將 TF(Term Frequency)也納入考量之後，其成效會大幅提升(Xu et al., 2008)，故本研究將利用文件的被引用次數修正 TF-IDF 來進行特徵的選取，同時考量 TF 與 DF 的重要性。

四、文件分群

文件分群的目的是要將文件集合切割成數個特定的群集，而群集之內的文件相似度要高，群與群之間的相似度要低(Mahdavi, Chehreghani, Abolhassani, & Forsati, 2008)。在進行文件主題偵測時，文件分群(Literature Clustering)是一個非常強力的資料探勘技術(Luo, Li, & Chung, 2009)；文件分群法是一個被廣泛用來找文件集合中的主題的方法(Anaya-Sánchez, Pons-Porrata, & Berlanga-Llavori, 2010)。由於文件集合最後可能會產生的群數以及各群集的共同主題是未知的，故文件分群法是屬於非監督式的機器學習(Unsupervised Machine Learning)。

分群演算法大約可依其特性分成以下四種：密度式分群(Density-Based Clustering)、網格式分群(Grid-Based Clustering)、階層式分群(Hierarchical Clustering)以及分割式分群(Partition Clustering)。以下將介紹本研究使用的分割式分群。

（一）分割式分群

分割式分群是利用反覆重新分配群心的位置，直到分群結果達到所需的群數為止。其中最常見的分割式分群演算法是 k-means，先選擇 k 筆資料當成起始群心，接著將每一筆資料分配到與其距離最近的群心的群集，進行反覆運算，然而 k-means 的分群結果不但容易受到初始群心與群數 k 的影響，同時也容易因資料集內的雜訊值與離群值而導致分群的品質不佳(Xu & Wunsch, 2005)。另一種分割式分群法為 k-medoids，其與 k-means 是非常相似的演算法，在群心的選擇上，k-medoids 以群集中最接近中心的資料當做群心，與 k-means 以群集中各資料的平均值當做中心點是不同的，因此與 k-means 演算法稍有不同，但分群方式上仍然差別不大。

分割式分群法的運算成本較其他方法低，因此分割式分群法是目前最廣泛，也被視為最適合用於大量文件的分群法(Mahdavi et al., 2008; Steinbach, Karypis, & Kumar, 2000)。由於林宜瑩(民，99)在進行文獻主題追蹤時，在分群方法上使用 k-medoids 得到的結果較使用 k-means 來得好，故本研究將透過修正及改良 k-medoids，作為分群方法。

（二）分群效度評估

在分群效度的評估上，可使用分群效度指標(Davies-Bouldin Index, DBI)(Davies & Bouldin, 1979)，來進行分群結果的評量。DBI 是利用計算同一群集內聚合程度以及不同

群集之間分散的情形來評估分群的效度。故本研究考慮利用分群效度指標作為分群的中止條件。

五、學術論文

學術論文通常被分為兩類：研討會論文及期刊論文，而除了各有不同的特性外，研討會論文與期刊論文之間也有值得探討的關係，研討會論文及期刊論文是有關聯的，很多學者一開始將其研究主題發佈至研討會，接收來自其他學者的意見之後，接著將其研討會論文進一步編纂，再發佈至期刊上(Tu & Seng, 2009)。

若要進一步探討研討會論文及期刊論文之間的關係，可將其之間的關係列為以下四種可能，如表 1：

表 1 研討會論文與期刊論文關係

跟隨角色 / 領導角色	研討會論文(P)	期刊論文(A)
研討會論文(P)	P → P(關係 1)	P → A(關係 2)
期刊論文(A)	A → P(關係 3)	A → A(關係 4)

研究結果發現，關係 2 具有顯著效應(Tu & Seng, 2009)。所以可以得到以下的推論，根據研討會論文及期刊論文的特性，研究人員可能會在研討會上發佈其研究初期的版本，之後經過研討會的審查、討論及收集各方的建議，之後再進一步將更完整的研究發佈在期刊上。

七、小結

本研究將以特徵選取及文件分群技術，進行主題偵測與追蹤並利用研討會論文及期刊論文兩種資料集進行趨勢的分析。文件的特徵選取的部份，將以”利用語意及引用次數修正的 TF-IDF 作為字詞重要性的計算方法。在分群技術上，以 k-medoids 分割式分群法為基礎，進一步依研究需求來改良並以分群效度指標 DBI 作為中止條件。而文件的表示上，本研究將採向量空間模型進行表示，並以 Cosine Similarity Measure 來進行文件相似度計算。

參、研究方法

本研究將提出一套考慮結合研討會論文與期刊論文當作資料集的研究主題趨勢偵測與追蹤法，透過此方法給予研究人員特定領域整體及個別研究主題的趨勢分析，使其能瞭解各研究主題的概況。以下章節將對本研究架構及方法做詳細的介紹。

一、研究架構

本研究架構如圖 2 所示，分為資料收集及處理模組、主題偵測模組以及趨勢分析模組：

（一）資料收集及處理模組(Data Collection Module)

主要任務為依照發表年份收集研討會論文及期刊論文，並截取文章內容所需的資料，並進行資料的前處理。

（二）主題偵測模組(Topic Detection Module)

根據模組一處理後的資料集，進行特徵選取，同時以不同時間區間當作單位，利用所選特徵進行分群，決定各時間區間的熱門主題，利用所產生出來的結果進行主題趨勢的分析

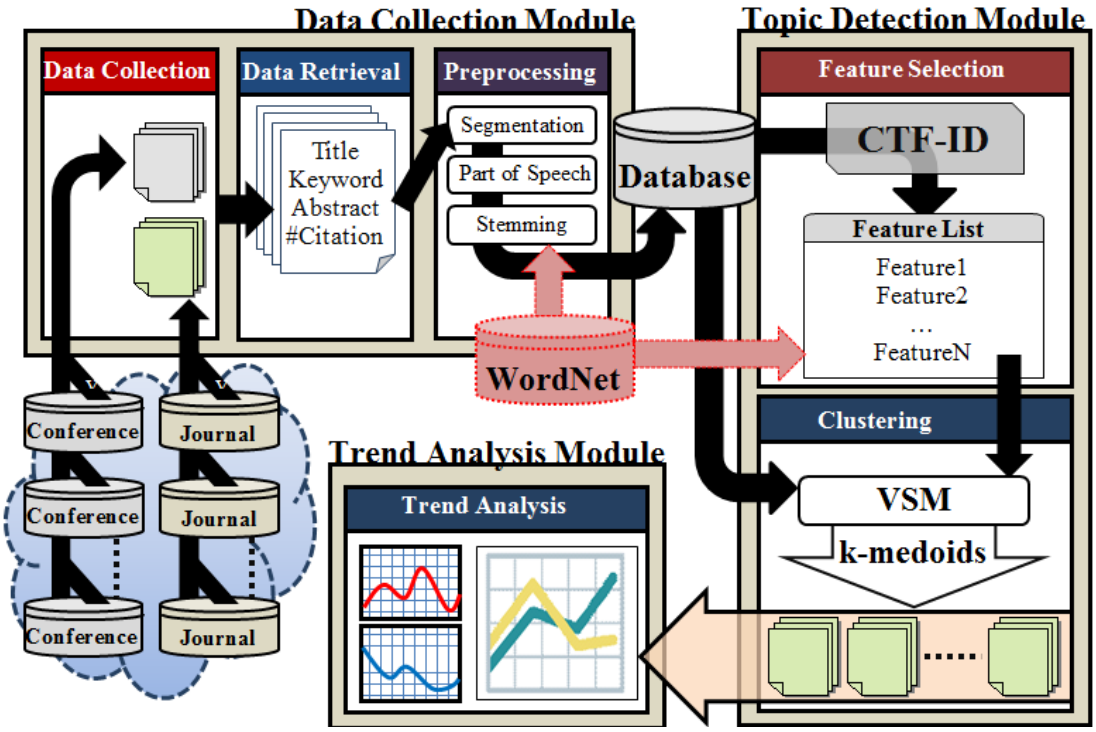

圖 1 研究架構圖

（三）趨勢分析模組(Trend Analysis Module)

利用產生出來的各主趨勢圖，進行整體、個別的研究主題趨勢分析，並利用研討會與期刊的不同特性，進行趨勢走向的分析。

二、資料收集及處理模組

在進行主題偵測前，資料的蒐集及前處理是必要的工作，流程如下所示：

（一）資料蒐集

首先，必須確定研討會C_i及期刊論文J_i的電子資料來源，接著從電子資料庫中進行論文的收集，收集特定年份的所有論文文章，匯集作為實驗的文件集，$P_{ij} \in C_i$，C_i為第i個研討會，而P_{ij}則是第i個研討會內的第j篇文章之集合。在文章資料的收集上，標題通常都包含的作者最想要表達給讀者的資訊(Xie, Zhang, & Ho, 2008)，Zhang and Wang(2010)則透過分析文章摘要，以找出有關作者所強調的重點，關鍵字分析可以提供學者所感興趣的研究趨勢資料(Li, Wang, & Ho, 2011)，而研究論文被引用次數可以拿來追蹤研究論文本身的影響力(Chiu & Ho, 2007; Li & Ho, 2008)，故本研究將針對標題、摘要、關鍵字及引用次數四個部分來進行資料擷取，因此可將一篇文章表示為$P_{ij} = \{T_{ij}, Abs_{ij}, K_{ij}, C_{ij}\}$，其中$T_{ij}$為標題集合，$Abs_{ij}$為摘要集合，$K_{ij}$為關鍵字集合，$C_{ij}$

則為被引用次數。期刊論文也以同樣的方式進行收集，可表示成$A_{ij} \in J_{ij}$，而A_{ij}則是第i本期刊內的第j篇文章之集合，$A_{ij} = \{T_{ij}, Abs_{ij}, K_{ij}, C_{ij}\}$。

當資料收集完成之後，可接著進行資料的前處理(Preprocessing)，可分為斷句(Segmentation)、詞性標記(Part of Speech)、字根還原(Stemming)。

（二）斷句

斷句主要是針對句子的分段及切割，由於文章摘要通常一個以上的句子組成，在進行詞性標記以前，必須先將摘要切割成個別獨立的句子，以S_{ijk}表示，$S_{ijk} \in Abs_{ij}$，S_{ijk}代表在第 i 個研討會或期刊內第 j 篇文章的摘要Abs_{ij}中的第 k 句。

（三）詞性標記

詞性標記則是將前一步驟切割好的句子S_{ijk}，以及文章標題T_{ij}進行文法分析，並對每一個字加以標記詞性。而在進行完詞性標記以後，由於林宜瑩（民 99）證實名詞片語(Noun Phrase, NP)的表現比單一字詞在進行特徵選取時成效較佳，故本研究僅考慮名詞片語，而 Zheng, Kang, and Kim(2009)將名詞片語分為「名詞+名詞」及「形容詞+名詞」兩大類，故只需保留名詞及形容詞

將摘要句S_{ijk}及標題句T_{ij}，利用 POS 相關的解析器(Parser)－Stanford Parser，如圖 3-3，進行詞性分析，然後從中保留屬於名詞片語的字詞，可以得到以下SN_{ijk}與TN_{ij}的名詞集合：

$$SN_{ijk} = \{SW_{ijk1}, SW_{ijk2}, \dots, SW_{ijkn} | where\ SW_{ijkl} \in NP\} \quad (7)$$

$$TN_{ij} = \{TW_{ij1}, TW_{ij2}, \dots, TW_{ijm} | where\ TW_{ijl} \in NP\} \quad (8)$$

SW_{ijkl}及TW_{ijl}分別表示為摘要句S_{ijk}及標題句T_{ij}裡第l個字，n為摘要句的字數，m為標題句的字數

（四）字根還原

資料前處理的最後步驟為字根還原，本研究利用機器可讀式字典(Machine Readable Dictionary, MRD)來輔助處理字根還原，採用的字典為 WordNet，將SN_{ijk}、TN_{ij}及K_{ij}還原成字根，避免產生單複數造成的誤判，K_{ij}處理後可表示成$KN_{ij} = \{KW_{ij1}, KW_{ij2}, \dots, KW_{ijo}\}$。

三、主題偵測模組

完成資料收集及前處理之後，每篇研討會及期刊論文皆會產生摘要、標題、關鍵字的字詞集合，分別為SN_{ijk}、TN_{ij}及KN_{ij}，可將每篇文章P_{ij}及A_{ij}表示如下：

$$AN_{ij} = \{SN_{ijk}, TN_{ij}, KN_{ij}\} \quad (9)$$

AN_{ij}表示為第i本研討會或期刊內，第j篇文章的處理過後的字詞集合。接著可利用這些字詞集合進行主題偵測，詳細流程如圖 3。

（一）特徵選取

首先進行特徵選取。在特徵選取的方法上，林宜瑩（民 99）在做期刊資料集的特徵

選取時，使用的方法是 TF-IDF，其考慮到各研討會或期刊文件數並不相同，故在計算時進行正規化，將計算出來的 TF-IDF 值再除以各研討會或期刊總文件數。而本研究提出 Citation-based TF-IDF(CTF-IDF)額外考慮了引用次數，故將其修正為：

$$tf-idf \cdot \frac{1}{N_i} \cdot \frac{(N_c + C_{ij})}{N_c} \cdot w$$
$$= \frac{n_{AN_{ijl},ij}}{n_{ij}} \cdot \log_2 \frac{N}{N_{ij}} \cdot \frac{1}{N_i} \cdot \frac{(\max\ C_{ij} + C_{ij})}{\max\ C_{ij}} \cdot w \quad (10)$$

N_i為第i個研討會或是第i本期刊中的文件數，N_c為所有文件之中最高的被引用次數，α為用來調整引用次數影響程度的係數，C_{ij}為該文件的被引用次數，$n_{AN_{ijl},ij}$為AN_{ij}中字詞AN_{ijl}出現的次數，n_{ij}代表文件AN_{ij}中所有字詞出現次數總合，N為文件集合中的文件總數，N_{ij}則為文件集中包含字詞AN_{ijl}的文件數而w則為該字詞分屬於標題、摘要或是關鍵字所給予不同的權重，分別為w_t、w_a、w_k。

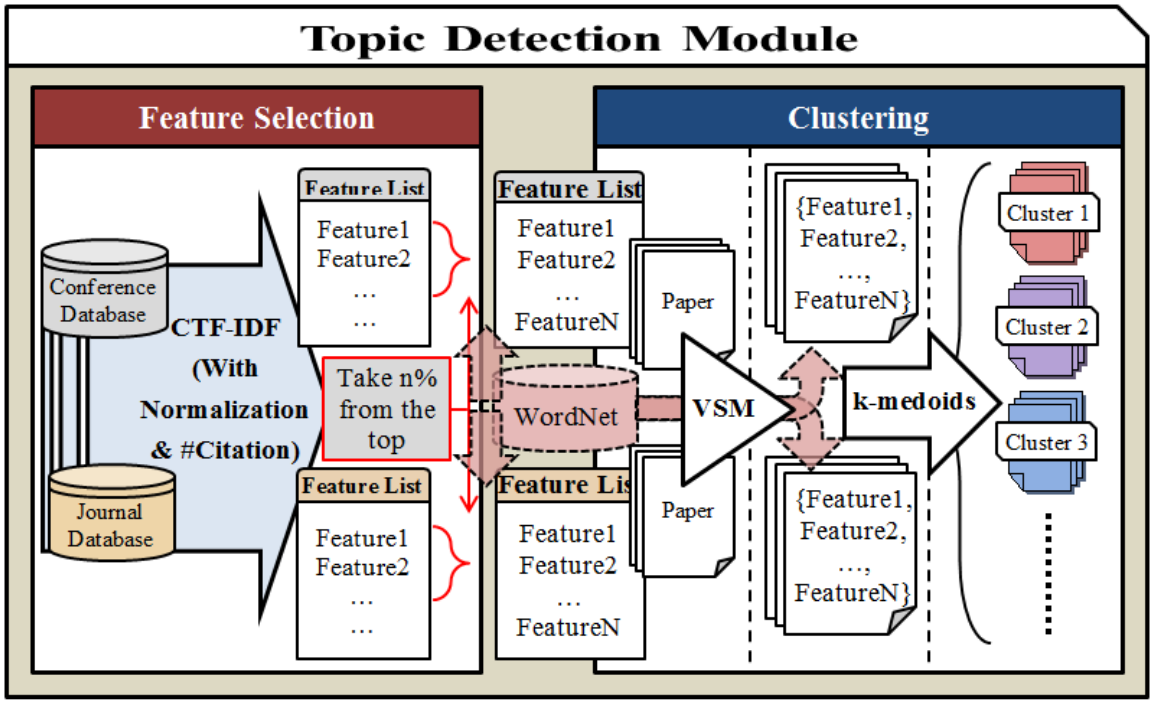

圖 3 主題偵測流程圖

將所有字詞進行特徵值的計算，並進行特徵值排序，為了考量同義不同字的同義字問題，本研究利用 WordNet 進行同義特徵的合併，並以排序較前的特徵字詞作為代表。接著再從特徵列表中選取前 n%的字詞，作為關鍵特徵字詞。接著利用所選取的特徵字詞，以特徵向量來表示$\{Feature_1, Feature_2, \ldots, Feature_f\}$，將所有的文件轉為向量空間模型表示法，如下所示：

$$A_{ij} = \{Feature_1, Feature_2, \ldots, Feature_f\} \quad (11)$$

向量中的各個值為各特徵在文件A_{ij}的特徵值，透過 TF 進行計算，但在表示文章時，

同樣加入同義字及近義字作為考量，式子如下：

$$Feature\ value = \frac{n_{F,ij} + n_{syn\_F,ij} + \beta n_{hyp\_F,ij}}{n_{ij}} \quad (12)$$

其中$n_{F,ij}$為特徵F在文件A_{ij}出現的次數，$n_{syn_F,ij}$為與特徵F為同義字的字詞出現次數，$n_{hyp_F,ij}$則為與特徵F為近義字的字詞出現次數，同義字與近義字則是透過連結WordNet 來進行判定，近義字僅考慮與特徵屬一層關係的上下義字，β則是給予近義字的權重係數。

（二）分群

將所有文件利用特徵向量表示完成之後，接著可進行分群的動作，以將相似內容的文件群聚，進而找出該群的主題。在進行分群以前，必須先決定時間區間，若把時間區分為多個區間D_y，時間單位可為任何長度，本研究由於考慮研討會及期刊發表大多為一年一次，故以年作為單位。

而在進行分群時，必須要進行文件之間相似度的計算，也就是文件在向量空間模型中的距離，本研究選用 Cosine Similarity Measure。若要計算兩文件d_1與d_2之間的相似度，可由以下式子計算：

$$Similarity\ Score = \frac{\sum_F (w_{F,d_1} w_{F,d_2})}{\sqrt{\sum_F w_{F,d_1}^2 \sum_F w_{F,d_2}^2}} \quad (13)$$

F為經過特徵選取出來的特徵字詞，$w_{F,d}$為特徵F在文章d中的特徵值。

本研究修正並使用 k-medoids 進行分群，詳細流程如下：

1. 在起始區間D_1時，先將所有的文件選出 k 筆作為分群的初始群心。
2. 將除了群心外的每份文件與所有群心進行相似度的計算，找出與該文件距離最近的群心，並將該文件歸屬於該群心。
3. 若計算出來的距離小於門檻值λ，則以該文件為群心產生新的主題群。
4. 當所有文件都被分別分屬到距離最近，相似度最高的群集中後，再以各群中所有的文件的特徵值計算新的群心值。
5. 以最靠近新群心值的文件作為新群心，重複 2～4 直到滿足中止條件。

*. 若為第一年則會重覆步驟 1~5 數次，來找到最佳初始群心，再進行步驟 2~4，以消除不同初始群心對於分群結果的可能偏差狀況。

本研究利用分群效度指標 DBI 作為中止條件，若 DBI 值在反覆的過程仍在下降則繼續進行新群心的分群，反之若趨近收斂則中止。完成以上分群後，在起始區間D_1可以決定出 k 個以上的主題群，接著可進行時間區間D_2的分群計算，但以前一時間區間的最後的群心特徵值當作起始群心來計算，進行步驟 2~5。並持續進行分群直到完成所有時間區間的文件分群。

（三）主題偵測

得到所有時間區間的分群結果後，為了能夠了解各群的內容，必須針對每個群集都

進行群集主題的偵測，以找出足以代表該群集的關鍵字詞。在進行分群結果的主題表示時，Decker & Scholz(2007)利用群心文件的文件標題作為該群主題，Anaya-Sánchez et al.(2010)列出該群最常出現的五個字詞作為主題的描述，Chen & Chundi(2011)則是收集了該群所有文件的關鍵字，計算出現次數後列出五個出現次數最多的關鍵字，作為該群集的高度相關主題。

本研究考慮利用在特徵選取步驟所產生的特徵集合來進行主題的表示，並結合相關字詞來作為主題的描述，步驟如下：

1. 將該群集所有文件的特徵值進行加總，並找出特徵值最高的特徵，以該特徵作為該群集的主題。
2. 除了主題外，另取三個次高的特徵作為主題的相關描述。

四、趨勢分析模組

完成所有文件分群及主題的表示後，可得到每個時間區間內的群集結果及主題，可利用圖表來進行趨勢的分析。

首先分別將研討會論文及期刊論文的分群結果以折線圖表示，橫軸為時間，縱軸為主題的熱門程度，以該群集論文數所佔總論文數的比例作為表示，可得到研討會以及期刊整體的研究主題趨勢變化，從結果可以觀察並比較不同主題之間的變化。接著可再針對特定主題，產生單一主題的趨勢變化圖，以便做單一主題的追蹤。

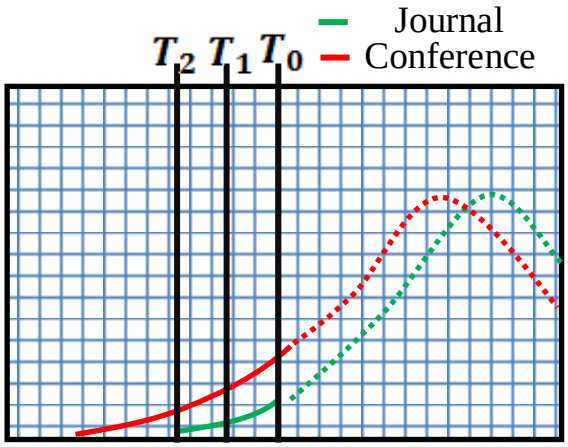

圖 4 趨勢變化圖

根據 Tu & Seng(2009)研究指出，研討會論文的主題與期刊論文主題具有時間上的先後關係，研討會論文的主題會對接下來發佈的期刊論文的主題產生影響，從研討會發佈開始，大約對之後一至兩年的期刊論文有顯著影響。本研究接著便利用此特性，並結合本研究所產生的研討會及期刊主題趨勢，進行主題趨勢的成長或衰退分析。

由上述概念，本研究利用各主題研討會當下的情況來預測期刊下一時間區段的主題走向，如圖 4，若以T表示主題的熱門程度，選定一期刊主題，T_0為同時間區間研討會該主題的熱門程度，T_1為前一時間區間的熱門程度，T_2則為前二時間區間的熱門程度。利用熱門程度的變化，可以整理出如表 2。

表 2 趨勢變化表

	趨勢變化 1	趨勢變化 2
成長期	若$T_0 - T_1 \geq 0$、$T_1 - T_2 \geq 0$且$T_0 - T_1 > T_1 - T_2$，表示該主題在研討會的熱門程度正在上升，對於投入研究該主題是良好的先鋒拓荒時期。	若$T_0 - T_1 \geq 0$、$T_1 - T_2 \geq 0$且$T_0 - T_1 < T_1 - T_2$，表示該主題在研討會的熱門程度正接近顛峰，成長趨於平緩，投入研究正處熱門時機。

	趨勢變化 3	趨勢變化 4
停滯期	若$T_0 - T_1 < 0$、$T_1 - T_2 \geq 0$，表示該主題該時段在研討會的熱門程度剛過顛峰，正逐漸下降。	若$T_0 - T_1 \geq 0$、$T_1 - T_2 < 0$，表示該主題該時段在研討會的熱門程度經過一陣低潮，但有回升的現象。
	趨勢變化 5	趨勢變化 6
衰退期	若$T_0 - T_1 < 0$、$T_1 - T_2 < 0$且$T_1 - T_0 > T_2 - T_1$，表示該主題在研討會的熱門程度正在下降，投入該主題研究可能稍嫌過晚。	若$T_0 - T_1 < 0$、$T_1 - T_2 < 0$且$T_1 - T_0 < T_2 - T_1$，表示該主題在研討會的熱門程度正在失去關注。

產生各主題的趨勢圖後，便可利用以上六個規則進行各主題趨勢走向的分析，整理出處於各個階段的研究主題，提供研究人員作為投入研究時，研究主題選擇的參考。

肆、結論與討論

本研究提出了考量語意及被引用次數進行特徵選取的主題偵測與追蹤方法，並進行進一步的各主題趨勢分析，利用 WordNet 加入同義字及近義字的考量，提升主題偵測的準確性。

將以所收集的研討會及期刊論文進行主題偵測與追蹤系統的實作，系統頁面將包含所有主題列表，以及熱門主題的整理。研究人員可以透過此頁面找到瀏覽主題的論文文件，同時可以了解研究趨勢的走向，若尚未有特定主題的需求，也可以透過熱門主題的列表來了解當下較熱門的主題。

本研究尚有諸多可改進的地方，可進一步在後續的研究上進行修正，以下將提出本研究對於後續研究發展及方向的建議：

1. 語意考量的進階：每一個字詞在 WordNet 當中可能包含不只一個的意義(Sense)，由於本研究在考量同義字及近義字時，僅考慮其中一個意義作為判斷，如果能夠進一步將所有意義加入考慮，語意比對可能會更加完善。再者，本研究也只考慮一層關係近義字，若能利用 WordNet 本身提供的階層關係，給予不同層級的上下義字不同的權重，將能增加考量語意的完善度。
2. 專有名詞及縮寫的考量：由於很多研究中，除了專有名詞外，也會將特定的字詞進行縮寫，造成進行詞性標記時，無法將其判定為本研究所收集的字詞，而忽略了這些字詞。如果能透過外部資源或是其他方式來進行專有名詞及縮寫的判讀，便能將這類的資訊也加以利用，增加字詞判讀的能力。

參考文獻

1. 吳偉銘（民 97）。基於語意及時間因素之主題偵測法。國立成功大學資訊管理研究所

碩士論文，未出版，台南市。
2. 林宜瑩（民 99）。利用時間因子與名詞片語之文獻主題追蹤法。國立成功大學資訊管理研究所碩士論文，未出版，台南市。
3. Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y., Umass, J., . . . Umass, M. (1998). *Topic Detection and Tracking Pilot Study Final Report.* Paper presented at the In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.
4. Anaya-Sánchez, H., Pons-Porrata, A., & Berlanga-Llavori, R. (2010). A document clustering algorithm for discovering and describing topics. *Pattern Recognition Letters, 31*(6), 502-510.
5. Chen, W., & Chundi, P. (2011). Extracting hot spots of topics from time-stamped documents. [Article]. *Data & Knowledge Engineering, 70*(7), 642-660.
6. Chiu, W. T., & Ho, Y. S. (2007). Bibliometric analysis of tsunami research. [Article]. *Scientometrics, 73*(1), 3-17.
7. Cordon, O. (2003). A review on the application of evolutionary computation to information retrieval. *International Journal of Approximate Reasoning, 34*(2-3), 241-264.
8. Davies, D., & Bouldin, D. (1979). A Cluster Separation Measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, PAMI-1*(2), 224-227.
9. Frakes, W. B., & Baeza-Yates, R. (1992). *Information Retrieval: Data, Structures and Algorithms*: Pretice Hall.
10. Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features (pp. 137-142): Springer Verlag.
11. Li, S., Xia, R., Zong, C., & Huang, C.-R. (2009). *A framework of feature selection methods for text categorization*. Paper presented at the Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2, Suntec, Singapore.
12. Lin, S.-H., Shih, C.-S., Chen, M. C., Ho, J.-M., Ko, M.-T., & Huang, Y.-M. (1998). *Extracting classification knowledge of Internet documents with mining term associations: a semantic approach*. Paper presented at the Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia.
13. Luo, C., Li, Y., & Chung, S. M. (2009). Text document clustering based on neighbors. *Data & Knowledge Engineering, 68*(11), 1271-1288.
14. Mahdavi, M., Chehreghani, M. H., Abolhassani, H., & Forsati, R. (2008). Novel meta-heuristic algorithms for clustering web documents. [Article]. *Applied Mathematics and Computation, 201*(1-2), 441-451.
15. Ö zgür, L., & Güngör, T. (2010). Text classification with the support of pruned dependency patterns. *Pattern Recognition Letters, 31*(12), 1598-1607.

16. Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM, 18*(11), 613-620.
17. Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques.
18. Tu, Y.-N., & Seng, J.-L. (2009). Research intelligence involving information retrieval – An example of conferences and journals. *Expert Systems with Applications, 36*(10), 12151-12166.
19. Walls, F., Jin, H., Sista, S., & Schwartz, R. (1999). Topic Detection in broadcast news *In Proceedings of the DARPA Broadcast News Workshop* (pp. 193-198): Morgan Kaufmann Publishers, Inc.
20. Wan, X. (2007). A novel document similarity measure based on earth mover's distance. *Information Sciences, 177*(18), 3718-3730.
21. Xie, S. D., Zhang, J., & Ho, Y. S. (2008). Assessment of world aerosol research trends by bibliometric analysis. [Article]. *Scientometrics, 77*(1), 113-130.
22. Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. [Review]. *Ieee Transactions on Neural Networks, 16*(3), 645-678.
23. Xu, Y., Wang, B., Li, J., & Jing, H. (2008). *An extended document frequency metric for feature selection in text categorization*. Paper presented at the Proceedings of the 4th Asia information retrieval conference on Information retrieval technology, Harbin, China.
24. Zhang, X., & Wang, T. (2010). Topic Tracking with Dynamic Topic Model and Topic-based Weighting Method. *Journal of Software, 5*(5), 482-489.
25. Zheng, H.-T., Kang, B.-Y., & Kim, H.-G. (2009). Exploiting noun phrases and semantic relationships for text document clustering. *Information Sciences, 179*(13), 2249-2262.

A Research Trend Analyzing Method Based on Semantics and Citation Count

Ching Sheng Wang
National Cheng Kung University
chuchu1128@gmail.com

Hei-Chia Wang
National Cheng Kung University
hcwang@mail.ncku.edu.tw

Abstract

With the digitization of knowledge, all kind of documents are gradually transformed into electronic form in order to transfer easily. However, due to the rapid increase of the amount of data, researchers cannot extract important information even though they can collect research data easier. In order to find research materials efficiently, people use topic detection and tracking technology to generalize topics of research papers and trends of research topics. Nevertheless, the methods in the past only have one date set, and no one focused on analyzing research trends. Therefore, this paper takes a advantages of the relations of papers between conferences and journals to do topic tracking. Besides, we also take semantics and citation count into consideration on feature selection to increase the efficiency of clustering. We can help researchers to reduce the wasting of time working on selecting research field by providing hot topics and trend analysis of each topic to them.

Keywords: Topic Detection and Tracking, Trend Analyzing, Clustering, Feature Selection