語意關聯主題模型於資訊檢索之研究
Semantic Associative Topic Models for Information Retrieval

吳孟淞
中央研究院資訊科學研究所
Institute of Information Science
Academia Sinica, Taipei, Taiwan
wums@iis.sinica.edu.tw

王新民
中央研究院資訊科學研究所
Institute of Information Science
Academia Sinica, Taipei, Taiwan
whm@iis.sinica.edu.tw

摘要

主題模型(topic model)被廣泛地應用在各種文件建模以及語音識別、資訊檢索和本文探勘系統中，有效地擷取文件或字詞的語意和統計資料。大多數主題模式，例如機率潛在語意分析(probabilistic latent semantic analysis)和潛在狄利克里分配(latent Dirichlet allocation)，主要都透過一組潛藏的主題機率分布來描述文件與字詞之間的關係，並用以擷取文件的潛在語意資訊。然而，傳統的主題模型受限於詞袋(bag-of-words)的假設，其潛藏主題僅能用來擷取個體詞(individual word)之間的語意資訊。雖然個體詞可傳達主題信息，但有時會缺乏本文準確的語意知識，容易造成文件的誤判，降低檢索的品質。為了改善主題模型的缺點，本論文提出一種新穎的語意關聯主題模型(semantic associative topic models)，考慮多元字詞(multi-words)之間的語意關聯資訊，基於關聯式探勘(association mining)法擷取出多元字詞之間的相互關聯資訊，並透過線性模型結合的方式，有效地改善傳統的機率潛在語意分析模型。我們以華爾街日報和美聯社新聞文件集進行實驗評估。實驗結果顯示新方法相較於傳統主題模型具有較優的文件模組化，並在文件檢索的效率上亦有良好的改善。

關鍵詞：主題模型、機率潛在語意分析、關聯探勘法、語言模型、資訊檢索。

1. 前言

科技的日益發達和網際網路的普及加速了數位圖書館的實現。如何以適當的形式表示文件中的資訊，以利資訊檢索的應用，是資訊檢索系統必須面對的問題。將文件以詞袋(bag-of-words，BoW)表示是文件檢索常用的模式之一[5][14][20]，此法不考慮語法和詞序列的資訊，而是利用統計方法，以字詞出現的頻率(term frequency)和反逆文件頻率(inverse document frequency)做為文件特徵，來建構文件模式，此方法亦稱為向量空間模型[20]。在處理文件資料時，向量空間模型是文件表達的常用簡便方法，然而，此方法卻有一些缺失，即無法區別文中字詞間的關聯性以及同義詞(synonym)和多義詞(polysemy)的問題[12]。另外，此方法的空間維度表示相當於字典個數的大小，這意謂有許多的參數必須被估計，容易導致效能的降低。另一種檢索模式是由Ponte 和Croft學者[18]提出，透過自然語言的*N*-連(*N*-gram)模型化，達到統計式文件檢索的目標，但其方法缺乏長距離資訊[8]以及語意資訊[12]。

近幾年來，在機器學習(machine learning)領域中，有許多方法被引用到資訊檢索和語音辨識的研究上，提供不同的觀點來探討語言和本文模型以及訓練文集。在文獻中[5][12][14]，主題模型被提出用來發掘完整文件或歷史詞序列中所隱含的語意資訊或是語句結構資訊等，針對文字中隱含的主題進行擷取，以改善BoW模型的缺點。這些主題資訊有助於深入理解使用者的查詢需求，進而達到更精確的檢索結果。潛在語意分析(latent semantic analysis, LSA)[12]探討隱藏在字詞背後的某種關係，這種關係並非以詞典中的定義為基礎，而是參考字詞的使用環境，其基本概念是以低維度的共同語意因子呈現原始文件和字詞之間的關連，透過奇異值(singular value decomposition, SVD)分解將文件映射到一個低維度的語意空間，以向量空間為分析模型，利用基底來呈現本文資料集(text corpus)中不同字詞和文件之間的關係，同時也解決在高維度的情況下參數量過多和訓練文集不足的問題，並假設每一奇異值及其對應的奇異向量(singular vector)代表其潛在主題或概念，且每一文件可由右奇異矩陣

轉置的行向量表示。文獻中[3][4][16]已證明潛在語意分析在資訊檢索和語音辨識領域上市有價值的分析工具。不同於潛在語意分析，機率潛在語意分析(probabilistic latent semantic analysis, PLSA)[14]以及潛在狄利克里分配(latent Dirichlet allocation, LDA)[5]為最具代表的機率式文件模型。PLSA模型作法是擷取與文件關聯的意向模型(aspect model)[15]，使用機率密度函數作為已觀察到的文件和字詞之間潛在語意的呈現方式，並利用最大相似度法則(maximization likelihood, ML)，結合期望值最大化(expectation maximization, EM)[13]演算法推估隱含的模型參數。然而，PLSA模型有幾項缺點[5]，首先，沒有直接的方法將機率分配給先前未出現(unseen)的文件；其次，參數數量會隨著文件數量線性擴增。而LDA模型[5]為一個較完整的文件生成模型，與PLSA模型主要的不同點在於將每一篇文件的機率視為潛在主題中隨機字詞機率的混合模型，進而求得該篇文件出現的機率值。然而，其近似推論演算法並不容易實現。在文獻上，LDA和PLSA模型已被廣泛地使用在許多領域，包括資訊檢索[5][14][23][25]、語音辨識和語言模型調適[1][6][9][10]等。

傳統的主題模型受限於詞袋的假設，使得潛在主題僅能用來擷取個體詞之間的語意資訊，但在某些檢索情況下，一群個體詞並不足以準確地代表文本的語義信息[7]。例如，使用者想查詢與電腦病毒"Friday 13th Virus"相關的文件，若文件模型將"Friday"、"13th"或"Virus"個別獨立的字詞視為文件特徵的表達方式，那麼一些不匹配的文件，如影片"勝利之光(Friday Night Lights)"、"驚爆十三天(Thirteen Days)"以及新型流感"H1N1 病毒(H1N1 Virus)"將被假定具有相關性而被檢索出來。基於上述原因，本研究考慮以主題模型下字詞之間的關聯性替代傳統以個體詞的資訊來表達文件。一般來說，上下文(context)中任何一組字詞，若其字詞與字詞之間具有強大的關聯性，則可稱為關聯字詞(associated word)[8][11]。例如{德國、章魚、預測}這樣一個有意義的個體詞的組合足以用來描述世界杯足球比賽的訊息，但這些字詞在文件中不一定需要出現在相鄰的位置上。Chien[8]嘗試利用文句結構的特性，使用文句前面所提供的資訊來建立文句中前後文的關聯法則，將其用於*N*-連模型以擷取長距離資訊。利用關聯字詞來表達文件資訊雖然會增加文件模型的複雜性，但可以避免主題模型在詞袋假設下所造成的模糊性[7]。

許多學者將字詞序列(word order)和字詞共同出現(word co-occurrence)的資訊嵌入至主題模型，以改善傳統主題模型的不足[7][17][22]。其中，雙連主題模型(bigram topic model, BTM)[17][22]結合了傳統*N*-連語言和主題模型的優點，可利用*N*-連語言模型擷取短距離的字詞連接資訊，解決主題模型下字詞序列被忽略的問題，亦可獲取隱含的語意資訊或是語句結構資訊等，以補足*N*-連語言模型的不足。在自然語言中，常存在許多高關聯性的詞組，比方說“章魚”、“預測”等經常出現於同一句子之中，但由於它們在句子中並不一定相連，所以BTM模型並沒有辦法擷取到這些字詞之間的相關資訊。文獻[7]提出以詞對袋(bag-of-word pairs, BoWP)為基礎的LSA模型架構，利用詞對(word pairs)的資訊來表示文件。由於模型所產生的詞對數會過於龐大，因此利用字詞與字詞之間在LSA空間的距離來作為選擇詞對的依據。這些詞對不受限於的字詞的順序和位置，亦即在同一文件中任何兩個字詞均可被選擇作為一個詞對。藉此，BoWP可以保存被BoW方法所忽略的更詳細的語意信息。另外，在語音辨識上，亦有學者[3]直接將PLSA模型與傳統*N*-連語言模型以線性插補法(linear interpolation)作結合，提供在給定歷史詞序列的條件下，每一個候選詞發生的機率。雖然上述的模型可以補償原有主題模型的缺點，但是這些模型僅考慮相鄰的字詞或是詞對表示型態關係，並未考慮多元詞組間的關聯性。

為了改善主題模型的缺點，本研究嘗試提出一個新穎的語意關聯主題模型(semantic associative topic models, SATM)，主要是以PLSA模型為基礎，將關聯探勘技術[2]應用在主題模型上，以關聯字詞視為文件的特徵來建構本文模型，並透過線性模型結合的方式，將兩種模型結合成為一個聯合機率模型，使得能夠完整反應語言模型的特性。在選取具有較高關聯性的詞組方面，是以pointwise Kullback-Leibler (KL) divergence作為評估，此為關鍵字組(keyphrase)檢測技術中，用來計算語言模型轉換架構的方法[21]。本研究主要貢獻在於以有效地探勘文件和字詞間的關聯性，改善傳統利用詞袋為基礎的主題模型無法有效分辨本文中相關概念及準確的語意知識，而造成檢索結果不正確的缺點。我們將其應用在文件模組化和文件檢索的領域

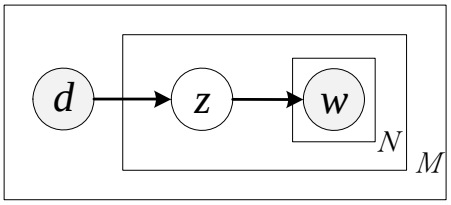

圖 1、PLSA 模型之圖形模型表示法

上，期望能改善正確率。

本論文接下來的章節組織如下。第二章探討文獻中各種相關的研究方法。第三章說明本研究所提出的方法，並比較幾種主要模型的差異。第四章分析實驗結果，以證明新方法的效益。最後，第五章提供結論以及未來的研究方向。

2. 相關文獻探討

本章先介紹機率潛在語意分析和詞對袋模型的基本概念，再簡單說明關聯式探勘法如何擷取字詞之間的相關性。

2.1 機率潛在語意分析

機率潛在語意分析(PLSA)[14]的概念是從潛在語意分析(LSA)延伸而來，其相異點在於LSA是將高維度的字詞向量與文件向量投影到低維度的潛在語意空間，表示字詞與文件的關係，而PLSA是以aspect model作為主要架構[15]，以機率方式針對字詞和文件共同事件，透過隱藏的主題，建立一生成模型(generative model)，如圖 1 所示。假設本文資料集是由文件-字詞對 (w,d) 所組成，文件以 $\mathbf{d}\in\{d_1,\ldots,d_N\}$ 表示，其個數為 N；字詞以 $\mathbf{w}\in\{w_1,\ldots,w_M\}$ 表示，意味字典是由 M 個字詞所形成之集合。假設每一字詞由給定的文件的潛在主題 $\mathbf{z}\in\{z_1,\ldots,z_K\}$ 產生，可以將文件-字詞對 (w,d) 共同出現(co-occurrence)的聯合機率表示成：

$$\begin{aligned} P(d_j,w_i) &= \sum_{k=1}^{K} P(z)P(w_i \mid z_k)P(d_j \mid z_k) \\ &= P(d_j)\sum_{k=1}^{K} P(w_i \mid z_k)P(z_k \mid d_j) \end{aligned} \quad (1)$$

其中 z_k 代表一個潛在主題，具有某種語意結構成分；$p(w_i \mid z_k)$ 是給定潛在主題 z_k 的情況下，字詞 w_i 出現的機率；$p(z_k \mid d_j)$ 是文件產生潛在主題 z_k 的機率。$p(w_i \mid z_k)$ 和 $p(z_k \mid d_j)$ 可以利用最大相似度估測法則，以期望值最大化(EM)演算法[13]推估出來。EM演算法是在機率模型中尋找參數最大相似度估計的演算法，其中機率模型依賴於無法觀測的隱藏變數(latent variable)。經過兩個步驟交替進行計算，第一階段是計算期望值(E)，利用對隱藏變數的現有估計值，計算其最大相似度估計值；第二步是最大化(M)，最大化在 E 階段所求得的最大相似度估計值來估算參數的值。M階段找到的參數估計值被用於下一個 E 階段計算，這個過程不斷交替進行。針對PLSA模型，對數相似度可以表示成：

$$L=\sum_{j=1}^{M}\sum_{i=1}^{N} n(w_i,d_j)\log\sum_{k=1}^{K} p(w_i \mid z_k)p(z_k \mid d_j), \quad (2)$$

其中 $n(w,d)$ 表示字詞在文件中的數量。在 E-step 中，利用目前估計的參數來計算潛在變數的事後機率，其式子如下：

$$p(z_k \mid w_i,d_j)=\frac{p(w_i \mid z_k)p(z_k \mid d_j)}{\sum_{l=1}^{K} p(w_i \mid z_l)p(z_l \mid d_j)}. \quad (3)$$

在 M-step 中，利用潛在變數在 E-step 中的估測，使得觀察的聯合對數相似度的期望最大化，得到參數的更新式如下：

$$p(w_i \mid z_k)=\frac{\sum_{j=1}^{M} n(w_i,d_j)p(z_k \mid d_j,w_i)}{\sum_{l=1}^{N}\sum_{j=1}^{M} n(w_l,d_j)p(z_k \mid d_j,w_l)}, \quad (4)$$

$$p(z_k \mid d_j)=\frac{\sum_{i=1}^{N} n(w_i,d_j)p(z_k \mid w_i,d_j)}{n(d_j)}. \quad (5)$$

然而，PLSA模型存在一些問題[5]。首先，假設在給定某一個潛在主題的前提下，文件與詞的關係是獨立的。其次，隨著所收集到的訓練集中文件數的增加，PLSA模型所需的參數也會呈線性增加，有可能會讓模型參數過度符合(overfitting)訓練資料。且對於描述未見過的(unseen)文件中的詞，PLSA沒有具備健全的預測能力。因此，當用於估測一篇新文件或查詢文句之主題模型時，會受到原始訓練資料的限制。Blei等人[5]提出潛藏狄利克里分配，改善PLSA模型參數量增長和只對在訓練文集中出現的文件估測模型參數的問題，而簡和吳[9]提出累進式學習(incremental learning)演算法，將統計學上的近似貝氏估測(Quasi-Bayes estimate)法則應用於求取PLSA模型最佳參數，同時使用累進觀測到的調整文集不斷的將PLSA模型調整到最新環境。

2.2 詞對袋模型

先前提到傳統的主題模型主要是將文件視為詞袋的表示法。在這樣的前提假設下，容易造成其所訓練出的潛在主題內的個體詞無法準確代表文本的

$$\mathbf{A}=\begin{bmatrix} t(w_1,w_1,d_1) & \cdots & t(w_1,w_1,d_j) & \cdots & t(w_1,w_1,d_M) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ t(w_i,w_n,d_1) & \cdots & t(w_i,w_n,d_j) & \cdots & t(w_i,w_n,d_M) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ t(w_N,w_N,d_1) & \cdots & t(w_N,w_N,d_j) & \cdots & t(w_N,w_N,d_M) \end{bmatrix}$$

圖 2、詞對袋模型之矩陣表示法

語義信息。詞對袋模型[7]的概念主要是從詞袋模型延伸而來，建構在潛藏語意分析模型的架構下，找出有用的詞對做為文件特徵。圖 2 顯示所有可能詞對和文件所構成的一個矩陣，以列代表詞對，以行代表文件。矩陣中 $t(w_i,w_n,d_j)$ 表示詞對 $\{w_i,w_n\}$ 出現在文件中的正規化詞頻，可表示成：

$$t(w_i,w_n,d_j)=(1-\varepsilon(w_i,w_n))\cdot\frac{n(w_i,w_n,d_j)}{\sum_{w_x,w_y\in d_j} n(w_x,w_y,d_j)} \quad (6)$$

其中，$n(w_i,w_n,d_j)$ 為詞對 $\{w_x,w_y\}$ 在文件 d_j 中出現的次數，$\varepsilon(w_i,w_n)$ 表示詞對之正規化熵(normalized entropy)。由於所產生的詞對種類數會太大，必須經由一些機制來刪減詞對數。詞對的選擇是基於詞與詞之間在LSA 空間的距離，可用兩個向量的夾角值計算而得[7]。給定一篇文件，詞對的選擇條件如下：

$$WP(w_i,w_n)=\Big\{w_i,w_n \Big| idf(w_i)>\alpha, idf(w_n)>\alpha, \frac{\mathbf{u}(w_i)\mathbf{S}^2\mathbf{u}(w_n)}{\|\mathbf{u}(w_i)\mathbf{S}\|\cdot\|\mathbf{u}(w_n)\mathbf{S}\|}>\beta\Big\} \quad (7)$$

其中，α 和 β 為經驗臨界值，*idf*表示反逆文件頻率(inverse document frequency)，代表全域權重，即某字詞在整個本文資料集中出現在越多的地方，則代表愈不重要。給定詞對-文件矩陣後，如同LSA[12]的運算，利用奇異值分解將文件投射到一個低維度的語意空間，利用基底來呈現本文資料集中不同詞對和文件之間的關係，如下所示：

$$\mathbf{A}=\mathbf{USV}^T \quad (8)$$

其中 **S** 為奇異值之對角矩陣，亦可視為對應的潛在語意空間，其值為 $s_1\ge s_2\ge\cdots\ge s_k\ge 0$ ； **U** 和 **V** 分別表示左奇異向量和右奇異向量。

2.3 關聯式探勘法

資料關聯性之研究為資料探勘重要的問題之一。它的目的是要從資料庫中，發現項目(item)間的關聯，在此，我們將其運用於本文資料集。在本文中字詞是語言中承載語意訊息的最小單位，當我們進行以語言為基礎的研究或運用時(例如文件檢索和分類)，必須先匯集整理出本文資料集中所可能使用到的不同字詞才能進行各種後續處理。在文件中，若發現某些字詞的出現會引發其他字詞的出現，這樣的字詞關係，即可以用關聯規則的型式加以表達，例如：世足賽 → 預測 → 章魚。在關聯規則探勘[2]中有兩個重要的評定標準，分別為信賴度(confidence)和支持度(support)，信賴度是一種量測關聯法則強弱的標準，而支持度則是表示統計上出現的頻率，當探勘演算法找出的規則滿足使用者訂定的最小信賴度與支持度的門檻，擷取出來之關聯法則才算成立。傳統的Apriori 關聯探勘演算法[2]有兩個階段，第一階段是先找出滿足最小支持度的項目集合(即大項目集合large itemset)，第二階段就根據前階段所找出之大項目集合，計算出所有符合最小信賴度的關聯規則。其中第一個步驟決定了整個作業的效能，它佔了作業的大部分時間，所以在探討關聯規則的挖掘時，均將焦點放在如何有效率的找出大項目集合。然而，一般的使用者都無法事先知道該如何選擇合適的支持度門檻，如果選擇了一個不合適的支持度門檻，往往造成最後產生的關聯規則沒有用。為了避免上述問題，本研究依據資訊理論的技術來選擇合適的關聯字詞。

3. 語意關聯主題模型

本研究將傳統主題模型中詞袋假設的作法延伸至關聯詞袋。文件被表示成無順序性的關聯字詞特徵集合。本研究主要在探索當主題模型包含更複雜的關聯性詞組時，是否會比簡單的個體詞群組擁有更強建的語意訊息。

3.1 關聯字詞選取

對建構關聯字詞並將語意訊息合併至主題模型而言，為了使得所找出的關聯字詞能夠反應語言模型的特性，我們修改原始的關聯探勘法[2]，從訓練資料中擷取出現頻率較高的關聯字詞。假設有一本文資料集 $D=\{d_1,\cdots,d_N\}$ ，每篇文章均由字詞集合 $W=\{w_1,\cdots,w_M\}$ 所構成。每個字詞出現的頻率可經由訓練文集統計而得，其單一字詞之詞頻定義為 $L_1=\{w_i\}$ ，而 L_{aw} 代表在文件中關聯字詞組 W_{aw}^i 的最大集合，可表示為 $L_{aw}=\{W_{aw}^i\}=\{W_{aw-1}^i;w_i\}$ ，而 aw 值代表最大的關聯詞組數($aw\ge 1$)。在模型中，對於所有可能出現在本文中的關聯字詞組應該皆被選取。然而，這樣的詞組數目可能過大。因此，所

選取的詞對必須經過刪減。為了找出 L_{aw+1}，關聯探勘法執行兩個步驟，高頻項目集合併(joining)和刪減(pruning)候選項目集，反覆地執行上述步驟，直到無法發掘出高頻關聯字詞組為止。其過程如下：
(1)高頻項目集合併步驟: 將 L_{aw} 中具前 *aw-1* 個相同項目之字詞兩兩組合，得到長度 *aw+1* 之候選項目集(candidate sets) C_{aw} 。
(2)刪減候選項目集: 若長度為 *aw* 之候選項目集其長度為 *aw-1* 之子集合不屬於 L_{aw-1}，則刪除此候選項目集(亦即高頻項目集之子集合必為高頻項目集)。掃瞄整個資料庫，根據刪減後之 C_{aw+1} 進行比對，找出 L_{aw+1} 。

在傳統關聯探勘法[2]裡，支持度門檻的選擇會影響到最後產生關聯詞組的結果。因此，我們採用資訊理論的技術來作為選取的依據。觸發序對語言模型[19]是根據平均相互資訊(average mutual information, AMI)來選擇觸發序對，它只能用來評估兩個詞 w_i 和 w_n 之間的關聯性大小。我們採語言模型轉換架構中關鍵字組(keyphrase)檢測技術 pointwise KL divergence (PKL)[21] 作為選取關聯詞組的評估方式，PKL是用來衡量二個語言模型之間差異程度的方法，其數學式表示如下：

$$\mathrm{PKL}(W^i_{aw-1};w_i) = p(W^i_{aw-1},w_i)\log\frac{p(W^i_{aw-1},w_i)}{p(W^i_{aw-1})p(w_i)} \quad (9)$$

其中 $p(W^i_{aw-1},w_i)$ 代表 W^i_{aw-1} 、 w_i 出現在同一視窗大小(window size)的機率，在此我們將視窗大小訂為文句的長度。透過 PKL 評估標準，可以找出具有強健性語意關聯之字詞組。當 PKL 值越小，表示 W^i_{aw-1} 和 w_i 兩者之間越有相關性，亦即說明此關聯詞組在主題模型中更能代表文件的特徵。

3.2 模型

和 PLSA 模型作法相似，本研究嘗試透過一組共享的潛藏主題分布，估算訓練資料中多元字詞間的語意關連性，稱之為語意關聯主題模型(semantic associative topic model, SATM)。不同於 PLSA 模型，SATM 直接對訓練資料中多元字詞組 $\{W^i_{aw}\}=\{W^i_{aw-1};w_i\}$ 的聯合機率 $p(W^i_{aw-1},w_i)$ 透過一組潛在主題分布所建構的語意空間作機率分解，其條件機率式可表示為：

$$\begin{aligned} P(W^i_{aw-1};w_i \mid d_j) &= P(W^i_{aw} \mid d_j) \\ &= \sum_{k=1}^{K} P(W^i_{aw} \mid z_k) P(z_k \mid d_j) \end{aligned} \quad (10)$$

為了使關聯字詞模組更能反應語言模型的特性，我們將所提出之 SATM 模型與 PLSA 模型結合為一個

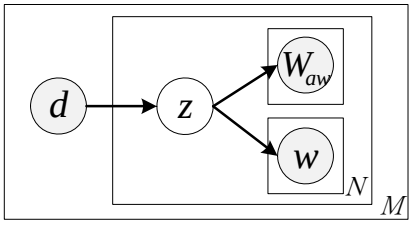

圖 3、混合式主題模型之圖形模型表示法

聯合機率模型，可以視為一混合式語意主題模型(hybrid semantic topic model, HSTM)。根據式(1)和(10)，兩個模組的分解共享相同的文件-特定信息(document-specific)混合部份 $p(z_k \mid d_j)$，其圖形模型如圖 3 所示。如此一來獲得的資訊比傳統的 PLSA 模型多了更強健的語意關聯資訊，亦保留原有 PLSA 模型的資訊。我們使用可調整權重參數 $\lambda \in (0,1)$ 結合兩種模型，結合之後，對數相似度可表示成：

$$\begin{aligned} L = \sum_{d_j} \Bigg[& \lambda \sum_{i=1}^{N} \log \sum_{k=1}^{K} p_{\mathrm{PLSA}}(w_i \mid z_k) p(z_k \mid d_j) \\ & + (1-\lambda) \sum_{W^i_{aw}} \log \sum_{k=1}^{K} p_{\mathrm{SATM}}(W^i_{aw} \mid z_k) p(z_k \mid d_j) \Bigg] \end{aligned} \quad (11)$$

當 $\lambda = 0$ 時，對數相似度函數只考慮關聯語意信息，而 $\lambda = 1$ 時，則退化為傳統的 PLSA 模型。透過期望值最大化演算法最大化訓練文集資料相似度可求得模型參數 $p(w_i \mid z_k)$ 、 $p(W^i_{aw} \mid z_k)$ 和 $p(z_k \mid d_j)$ 。
在 E-step 中，利用目前估計的參數來計算潛在變數的事後機率，其式子如下：

$$p(z_k \mid w_i, d_j) = \frac{p(w_i \mid z_k) p(z_k \mid d_j)}{\sum_{l=1}^{K} p(w_i \mid z_l) p(z_l \mid d_j)} \quad (12)$$

$$p(z_k \mid W^i_{aw}, d_j) = \frac{p(W^i_{aw} \mid z_k) p(z_k \mid d_j)}{\sum_{l=1}^{K} p(W^i_{aw} \mid z_l) p(z_l \mid d_j)} \quad (13)$$

在 M-step 中，利用潛在變數在 E-step 中的估測，使得觀察的聯合對數相似度的期望最大化，得到參數的更新式如下：

$$p(w_i \mid z_k) = \frac{\sum_{j=1}^{M} n(w_i, d_j) p(z_k \mid d_j, w_i)}{\sum_{l=1}^{N} \sum_{j=1}^{M} n(w_l, d_j) p(z_k \mid d_j, w_l)} \quad (14)$$

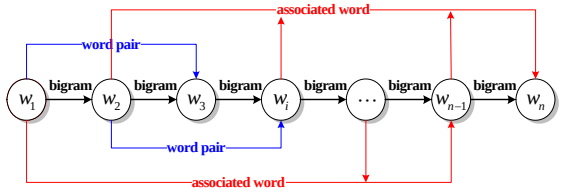

圖 4、不同字詞關聯型態之比較

$$p(W_{aw}^{i} \mid z_k) = \frac{\sum_{j=1}^{M} n(W_{aw}^{i}, d_j)\, p(z_k \mid W_{aw}^{i}, d_j)}{\sum_{W_{aw}^{i'}} \sum_{j=1}^{M} n(W_{aw}^{i'}, d_j)\, p(z_k \mid W_{aw}^{i'}, d_j)} \quad (15)$$

在考慮文件 d_j 是和字詞 w_i 以及關聯詞組 W_{aw}^{i} 共生之下， $p(z_k \mid d_j)$ 的估測隨著混合比例表示為

$$p(z_k \mid d_j) \propto \lambda \sum_{i=1}^{N} p(z_k \mid w_i, d_j) + (1-\lambda) \sum_{W_{aw}^{i}} p(z_k \mid W_{aw}^{i}, d_j) \quad (16)$$

3.3 主題模型中不同字詞關聯型態之比較

語意關聯主題模型與詞對袋模型最大不同在於我們透過關聯探勘法可以獲得多元詞組之間的關聯性。圖 4 為不同型態之字詞間關聯性的示意圖，圖中箭頭表示字詞與字詞之間的關係。如同文獻[8]中所述，由圖 4 可以看出雙連模型中字詞的關係是由先前緊鄰的詞來估測目前所出現的詞之條件機率，即文句間的關聯性是循序的。詞對袋模型跳脫此限制，只要是同一段文句中所出現的詞都可以有相互間的關聯性存在，雖然改善雙連模型的不足，但僅限於擷取字詞與字詞之間的關係。本研究所提之方法找出的字詞則可獲得多元詞組之間的相互關係，可以說是詞對的延伸研究。

4. 實驗與討論

4.1 實驗文集和設定說明

本論文的實驗使用文件檢索會議(Text REtrieval Conference, TREC)所收集的測試文集，分別為 Associated Press newswire (AP) 1988 年份，包含 79,919 篇文件，和Wall Street Journal (WSJ) 1987 年份，包含 46,488 篇文件[24]。所有文件都先經過stop word過濾和stemming前處理。我們分別對此兩文集

表 1、不同模型在字詞關係層級為 2 之語言複雜度實驗結果比較

	APLM	BoWP	SATM
Perplexity	248.6	232.49	217.8

以文件模組化和文件檢索驗證本研究所提方法之正確性與可行性。比較對象包括*N*-連語言模型[18]、機率潛在語意分析模型(PLSA)[14]及詞對袋模型(BoWP)[7]。實驗中潛在變數*k*設為 32，關聯詞對的選擇視窗長度設為文句長度。實驗分為兩個部分。第一是以語言複雜度(perplexity, pp)評估文件模組化的效果。語言複雜度是由資訊理論發展而來，可視為語言模型的平均分支度[10]，語言複雜度越低，表示所訓練的語言模型所遇到的分支越少，代表模型對於分支的描述較為集中。第二是評估各個模型應用在文件檢索上的效能，以精確-召回曲線(precision-recall curve)和平均精確率(mean average precision, mAP)作為評估的準則[20]。平均精確率是反應系統在全部相關文件上檢索效能的單值指標，相關文件排越前面，系統平均精確率越高。

4.2 不同模型在文件模組化之評估

文件模組化實驗以WSJ為實驗資料文集。本實驗將文件分為兩個部份，百分之九十的資料量作為基礎模型的訓練資料集，剩餘部份做測試資料集。基礎實驗結果得到單連語言模型(unigram)和PLSA模型的語言複雜度分別為 970.48 和 795.94。我們將字詞關係層級設定為 2，比較語意關聯主題模型(SATM)、關聯語言模型(APLM)[8]和詞對袋(BoWP)[7]的效能，其結果如表 1 所示。從表中可以看出SATM明顯優於BoWP和APLM，語言複雜度分別由 232.49 和 248.6 降至 217.8。另外，我們探討SATM模型與PLSA模型結合之後(即為混合式語意主題模型，HSTM)對語言複雜度的影響，其模型插補法調適的權重係數 λ 設定為 0.5。圖 5 表示混合式語意主題模型HSTM和關聯語言模型APLM在關聯字詞層級 2 到 6 之語言複雜度的比較。由圖 5 可以發現隨著最大層級的增加，其語言複雜度有明顯的下降，由此可以得知利用以PKL為基礎的關聯探勘對於主題模型有一些改善。由實驗結果亦可看出混合式語意主題模型HSTM的語言複雜度比關聯語

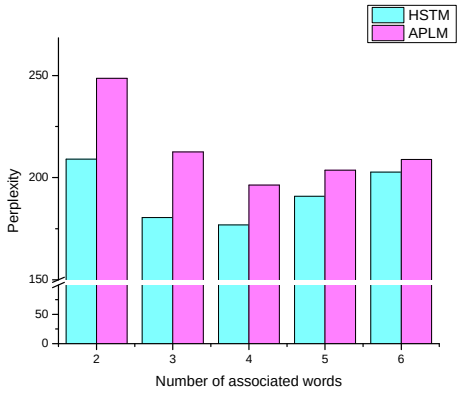

圖 5、混合式語意主題模型與關聯語言模型在不同層級所得之語言複雜度實驗結果

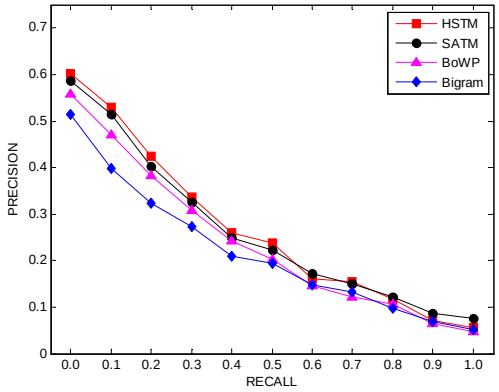

圖 6、精確-召回曲線在不同方法比較之實驗結果

言模型APLM更低。在最大層級為 4 時，混合式語意主題模型HSTM得到的最低語言複雜度為176.92，優於關聯語言模型APLM之語言複雜度196.37。從以上的結果可看出主題模型相較於語言模型可以得到較佳的文件模組化。

4.3 不同模型在檢索效能之評估

本實驗比較雙連語言模型 Bigram LM、詞對袋模型 BoWP 和語意關聯模型 SATM 及混合式語意主題模型 HSTM 的文件檢索效能。圖 6 表示在關聯層級為 2 時，不同模型之精確-召回曲線。由這四組實驗數據可以發現以主題為基礎的文件模型，皆比雙連語言模型有更好的效能。其原因在於主題模型相較於雙連語言模型可以獲得較多的語意訊息，且雙連語言模型侷限於有限距離，以致於流失長距離資訊。另外，我們也比較不同模型之平均精確率。根據實驗數據，混合式主題模型 HSTM 的平均精確率為 0.2384，其結果優於雙連模型(0.2082)、詞對袋模型(0.2296)和 SATM 模型(0.2328)。由於 HSTM 相較於 SATM 多考慮個體詞對模型的影響，因此更能反應語言模型的特性。

5. 結論

本論文對於傳統的主題模型做了概要性的介紹，包含模型的建立、評估以及優缺點的討論。我們探討關聯字詞對主題模型的影響，利用關聯探勘法來建立本文中多元字詞之間的關聯性，並將語意訊息合併至主題模型中。為了使其不失語言特性，我們並將傳統的機率潛在語意模型結合本論文所提之方法，實驗結果顯示對於文件模組化可以有效地降低語言複雜度，對於文件檢索也有一定幅度的改善。未來，我們擬將語意關聯主題模型運用於中文檢索、雙語檢索(bilingual retrieval)和口述文件檢索(spoken document retrieval)等方面。另外，對於混合權重值 λ 的設定擬以最大期望法找出較佳的權重值以代替現有經驗值的設定。再者，由於在發掘關聯字詞的過程中，必須多次讀取文本訓練資料，造成效率的降低，未來擬尋求一快速的方法，減少掃描本文資料的次數。

參考文獻

[1] Y. Akita and T. Kawahara, “Language model adaptation based on PLSA of topics and speakers”, in *Proc. of International Conference on Spoken Language Processing*, pp. 1045-1048, 2004.

[2] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules”, in *Proc. of the International Conference on Very Large Data Bases*, pp. 48-499, 1994.

[3] J. R. Bellegarda, “Exploiting latent semantic information in statistical language modeling”, *Proceeding of the IEEE*, vol. 88, no. 8, pp. 1279-1296, 2000.

[4] M. W. Berry, S. T. Dumais and G. W. O’Brien, “Using linear algebra for intelligent information retrieval”, *SIAM: Review*, vol. 37, no. 4, pp. 573-595, 1995.

[5] D. M. Blei, A. Y. Ng and M. I. Jordan, “Latent

Dirichlet allocation", *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[6] H. L. Chang, Y. C. Pan and L. S. Lee, "Latent semantic retrieval of spoken documents over position specific posterior lattices", *IEEE Workshop on Spoken Language Technology*, pp.285-288, 2008.

[7] L. Chen, KK Chin and K. Knill, "Improved language modelling using bag of word pairs", in *Proc. of INTERSPEECH*, pp. 2671-2674, 2009.

[8] J. T. Chien, "Association pattern language modeling", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1719-1728, 2006.

[9] J. T. Chien and M. S. Wu, "Adaptive Bayesian latent semantic analysis", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 198-207, 2008.

[10] J.-T. Chien, M.-S. Wu and H.-J. Peng, "Latent semantic language modeling and smoothing", *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 9, no. 2, pp. 29-44, August 2004.

[11] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography", *Computational Linguistics*, vol. 16, pp. 22-29, 1990.

[12] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis", *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391-407, 1990.

[13] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1-38, 1977.

[14] T. Hofmann, "Probabilistic latent semantic analysis", in *Proc. of Uncertainty in Artificial Intelligence*, pp. 289-296, 1999.

[15] T. Hofmann, "Unsupervised learning from dyadic data", *Advances in Neural Information Processing Systems*, vol. 11. MIT Press, 1999.

[16] T. G. Kolda and D. P. O'Leary, "A semi-discrete matrix decomposition for latent semantic indexing in information retrieval", *ACM Transactions on Information Systems*, vol. 16, no. 4, pp. 322-346, 1998.

[17] J. Nie, R. Li, D. Luo and X. Wu, "Refine bigram PLSA model by assigning latent topics unevenly", in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 401-406, 2007.

[18] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval", in *Proc. of ACM Conference on Research and Development in Information Retrieval*, pp. 275-281, 1998.

[19] R. Rosenfield, "A Maximum Entropy approach to adaptive statistical language modeling", *Computer Speech and Language*, Vol. 10, pp.187-228, 1996.

[20] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, New York: McGraw-Hill, 1983.

[21] T. Tomokiyo and M. Hurst, "A language model approach to keyphrase extraction", in *Proc. of the ACL Workshop on Multiword Expressions: Analysis, Acquisition & Treatment*, pp. 33-40, 2003.

[22] H. M. Wallach, "Topic modeling: beyond bag-of-words", in *Proc. of the 23rd International Conference on Machine Learning*, pp.977-984, 2006.

[23] X. Wei and W. B. Croft, "LDA- based document models for ad-hoc retrieval", in *Proc. of ACM Conference on Research and Development in Information Retrieval*, pp. 178-185, 2006.

[24] M. S. Wu and J. T. Chien, "Minimum rank error training for language model", in *European Conference on Speech Communication and Technology*, pp. 614-617, 2007.

[25] G.-R. Xue, W. Dai, Q. Yang and Y. Yu, "Topic-bridged PLSA for cross-domain text classification", in *Proc. of ACM Conference on Research and Development in Information Retrieval*, pp. 627-634, 2008.